CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Nuclear Sciences and Physical Engineering

Department of Physics

Research project

Analysis of empirical time series

Martin Prokš

Supervisor: Ing. Petr Jizba, Ph.D.

Prague, 2015

Acknowledgement

I would like to thank my supervisor, Ing. Petr Jizba, Ph.D., for his support and bottomless patience. I am also grateful to Dr. Hamid Shefaat and Dr. Tatjana Carle for providing intraday data and challenging discussion full of invaluable advices.

Contents

List of Figures

Chapter 1

Information theory

Information theory was founded by Shanon 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 needed information to select element from first set and then from the other.

Shanon 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 expected. Hence we can say that entropy is an expected value of measure of surprise.

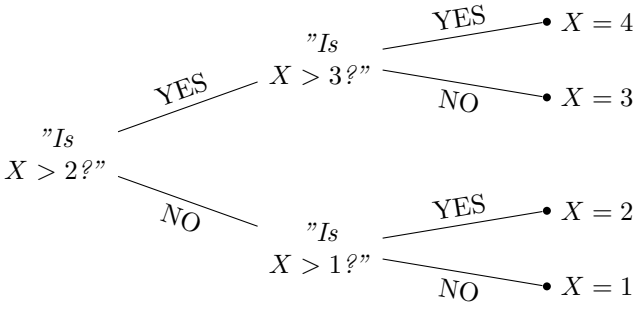
1.1.1 Coding theory and Huffman code

According to Shanon, the entropy is the averaged number of bits needed to optimally encode random variable X with its probability distribution $p(x)$. It means that entropy is averaged number of questions which can be answered only by yes or no and that bring us from absolute randomness to complete knowledge of random variable X. Let us proceed the following example to 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 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 $E[Q] = 2$ and this corresponds 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 smartly started asking from the most probable value):

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

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 value 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 encode in binary code so that finding the least averaged number of questions is equivalent to finding the shortest averaged binary code.

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 all code. Such a code is already uniquely decipherable. The existence of such a code is guaranteed by Kraft's inequality.

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

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

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

Theorem 1.2. *(Shanon's noiseless coding theorem 1948)*
Let the lengths of codes $\{l_1, \ldots, l_N\}$ 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 averaged length.

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

For fixed source i.e. 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 averaged length is

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

and from Shanon theorem it follows that L^* is minimum and thus lengths l_i^* 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 since the property of Ceiling function $x \leq \lceil x \rceil \leq x+1$ and this choice of code lengths is called Shanon-Fano code. The averaged length then satisfies well known inequality.

$$H_D(X) \leq L < H_D(X) + 1 \tag{1.3}$$

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 averaged length per symbol. We see that coding longer sequences allows arbitrary approach to theoretical boundary.

Let us 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. So that after rearrangement of the 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.

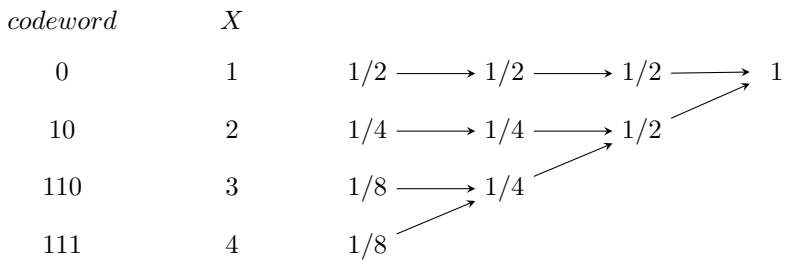
From these properties we can simply construct the optimal code. In what follow assume $D = 2$ i.e. binary code. The construction is:

1. Assign to 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. Let illustrate the procedure by example. Recall the random variable X from the beginning of this subsection and let 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 Shanon there is no code with averaged length less than $H(X)$. Let see what excepted 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 end let us note the connection between code and questions which bring us to complete knowledge about some the 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 exchanging ones by zeros the codes exactly match.

Hence, we can ask whether asking 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 question may be yielded 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 set A?" we must know what the first digit in the code is and accordingly we choose the set A.

1.1.2 Entropy interpretation in statistical physics

Let us show brief evolution of the word "entropy" as it emerges in two fields of physics namely thermodynamics and statistical physics. At the end we will intimate connection with information definition of entropy.

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

$$\mathrm{d}S = \frac{\mathrm{d}Q}{T}$$

The definition was motivated by the fact that heat received by the system during any reversible process depends on the path in state space i.e. dQ is not a total differential of some state function. Luckily, for reversible processes there always is 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. 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 stays that for reversible (quasi-static) adiabatic process the entropy is conserved. However, for non-static (irreversible) processes the entropy increase and difference $\mathrm{d}S > 0$ can be regarded as a measure of irreversibility or, in other words, the loss of information that is necessary to back trace the process.

For isolated system entropy increases for non-static processes. Let have system at equilibrium state and by sudden change of external parameters shift it to another non-equilibrium state (1). 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 microstates of the system satisfying given constraint, such as specific energy level. The Boltzmann equation expresses this interpretation.

$$S = k \ln \Gamma$$

In other words, every physical system is incomplete 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 is amount of information about the system that is needed for description of microscopic structure.

There were none clues that entropy defined by Shanon 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.

1.1.3 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 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 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.

Next, the 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*

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

Someone 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*

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

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

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). And 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 emphasize thus the system is completely unpredictable. Any no uniform distribution may be understood as additional information and therefore leads to decrease of entropy (or uncertainty).

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

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

$$h(p) \text{ is nonnegative for } \forall p \in (0,1\rangle \tag{1.4}$$

$$\begin{gathered} h(p) \text{ is additive for independent events i.e.} \\ h(pq) = h(p) + h(q),\ p,q \in (0,1\rangle \end{gathered} \tag{1.5}$$

Someone may ask whether there is another function satisfying these two conditions (axioms). The answer gives the following theorem [3].

Theorem 1.5. *The only function satisfying conditions 1.4 and 1.5 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 1.4 and 1.5 also normalization:

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

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

1.1.4 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 individuals ones states 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.

1.1.5 Conditional entropy

Let X and Y are 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:*

$$H(X|Y) = \sum_{y \in \mathcal{Y}} p(y) H(X|Y = y) \tag{1.6}$$

After inserting the definition of $H(X|Y = y)$ into expression 1.6 we get:

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

that is just an expected value of $-\log p(X|Y)$.

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 change 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 gained

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

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

$$H(X, Y) = H(X) + H(Y|X) \tag{1.7}$$

Theorem 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*

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

Here we use notation $H(X_1|X_0, \ldots, X_1) = H(X_1)$.

1.1.6 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 does not hold.

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:*

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

It is worth mentioning that relative entropy is only 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 are used as a different measures of distinction between probability distributions.

In coding theory the relative entropy represents averaged number of unnecessarily bits used in encoding of random variable X if we use bad distribution $q(x)$ instead of underlaying probability distribution $p(x)$.

Relative entropy is used for definition of 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 it is depicted at the Figure 1.1.

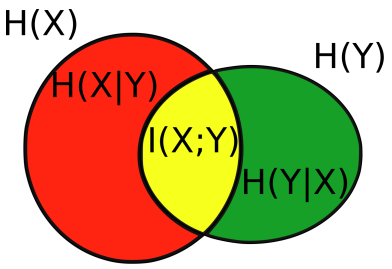


Figure 1.1: Relations between entropies

Thus mutual information is helpful measure of dependence of random variables (or time series as we will see later). Actually, 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.

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

$$H(X) = I(X;Y) + H(X|Y) \tag{1.8}$$

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 diagram 1.1.

The next expression is clear form 1.8 (since $H(X|X) = 0$) and intuitively reasonable as well.

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

From equation 1.8 we express the mutual information and by conditioning both sides we get so-called **conditional mutual** information:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{1.9}$$

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 as a redundancy in variables X and Y given Z.

There is relation similar to chain rule for entropy:

$$I(X_1,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_{i-1},\ldots,X_1) \tag{1.10}$$

For more variable we cannot depict the situation in a Venn diagram because it becomes indecipherable. But it is still possible to imagine that this relation just says: "Common information about n random variables in Y is union of individual information about X_i in Y."

The predictability property of mutual information may be regarded as a redundancy and we can define m dimensional mutual information (redundancy):

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

which represents number of saved bits when group of m events are encoded with one codeword instead of encoding events separately.

1.1.7 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 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 constant (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 1.8 that:

$$H(X) \geq H(X|Y) \tag{1.11}$$

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 $H(X)$.

1.1.8 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 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 holds the following theorem.

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 next step given the past never decreases.

1.1.9 Differential entropy

We shortly mention generalization of Shanon 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 example, 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.

Entropy of Normal distribution Let us compute entropy of Normal distribution. We just interchange the sum with integral and use 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 calculation 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. Continuous entropy cannot be defined for δ distribution because $\log \delta$ is not well defined.

Remark to Normal distribution We should mention that Normal distribution is the less bias 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 system. Similarly for n random variables and given covariance matrix entropy is maximized by n dimensional Gaussian distribution.

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

$$H(X) = \log\left(e\lambda\right)$$

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

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

and then we simply sum over all boxes. The entropy diverges with finer partitioning ($\epsilon \mapsto 0$), see [10] 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 we are not able to distinguish all small details this infinity does not have to scary us.

Chapter 2

Rényi entropy

In this chapter we will generalize Shanon information measure according to Rényi, see [4]. 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 to generalize them.

2.1 Information quantities

2.1.1 Entropy

To motivate Rényi entropy we should have a look at the way how Shanon 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 Shanon's formula.

From above generalization of Hartley information measure we can see that Shanon 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 Shanon and considered Kolmogorov–Nagumo generalized mean

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

where f is continuous and strictly monotone (i.e. invertible). It represents the most general mean compatible with Kolmogorov axiomatics. 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 [5]. The linear function leads to already known Shanon entropy and exponential function gives Renyi entropy:

$$H_\alpha(\mathcal{P}) = \frac{1}{1-\alpha} \log_2 \sum_{i=1}^{N} p_i^\alpha \tag{2.1}$$

Though, the left hand of 2.1 is defined for all $\alpha \neq 1$ we should disable non-positive values since for $\alpha < 0$ 2.1 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 $H_0(\mathcal{P}) = \log_2 N$. As long as the limit α tends to 1 is well defined and equals to Shanon's entropy (it can be readily clarified by L'Hopital rule) we may conclude that 2.1 defines suitable measure of information for $\alpha > 0$.

It is easily seen that H_α is non-negative and equal to 0 if and only if the probability distribution is degenerate, consider $p_i^\alpha \leq p_i$ 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 Shanon entropy.

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

Simple derivation of 2.1 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 get that Rényi entropy is decreasing function of α. This means that Shanon 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. Let have 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 $\log_2 1/p - \log_2 1/q$ 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 $\log_2 \frac{q}{p}$ or average increases of uncertainty $\log_2 \frac{p}{q}$ and result multiply by (-1).

In Shanon'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 [4]. The first method

suggests undesirable properties of information gain and hence the other method is used.

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*

$$D_\alpha(q||p) = \frac{1}{\alpha - 1} \log_2 \left(\sum_{k=1}^{n} \frac{q_k^\alpha}{p_k^{\alpha-1}} \right) \tag{2.2}$$

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 relate 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 fist chapter only linear averaging is replaced by generalized mean. Let have two random variables X and Y then 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 Shanon conditional entropy is easily broaden to Rényi conditional entropy so we have

$$0 \leq H_\alpha(X|Y) \leq H_\alpha(X) \tag{2.3}$$

with the same conditions to equality as in Shanon'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 [4].

We remark another definition of conditional information that is based on the additive property of Shanon entropy for dependent variables, equation 1.7. We can postulate this equation also for Rényi entropy and define conditional entropy as

$$\tilde{H}_\alpha(X|Y) = H_\alpha(X,Y) - H_\alpha(Y) = \frac{1}{1-\alpha}\log_2\left(\frac{\sum_{k=1}^{n} q_k^\alpha \left(\sum_{h=1}^{m} p_{h|k}^\alpha\right)}{\sum_{k=1}^{n} q_k^\alpha}\right), \quad (2.4)$$

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

Escort distribution

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

$$\rho_{qk} = \frac{q_k^q}{\sum_{k=1}^{n} q_k^q}$$

This new probability distribution has interesting property that it emphasizes probable events and suppresses rare ones for $q > 1$. The greater q is, the bigger 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. 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 2.4 to

$$\tilde{H}_\alpha(X|Y) = \frac{1}{1-\alpha}\log_2\left(\sum_{k=1}^{n} \rho_{\alpha k} 2^{(1-\alpha)H_\alpha(X|Y=y_k)}\right), \quad (2.5)$$

which means that to fulfill condition [odkaz v prvni kapitole] we have to average with respect to escort distribution instead of original distribution.

It can be shown, see [5], 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 opposite implication does not generally hold. It is necessary to be valid $\tilde{H}_\alpha(X|Y) = H_\alpha(X)$ for all $\alpha > 1$ or $0 < \alpha < 1$ to imply independence of X and Y, see [6].

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 Shanon mutual information. The ambiguity is caused by the fact that restriction to one property of Shanon mutual information violates some other one. Hence, in application we should pick up such a definition that best fits our requirements.

The Shanon 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 2.2 and define generalized mutual information in the same way. Unfortunately, this definition would violate desirable property of mutual information, namely

$$I_\alpha(X;Y) \leq H_\alpha(X), \tag{2.6}$$

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 Shanon mutual information

$$I(X;Y) = H(X) - H(X|Y) \tag{2.7}$$

This would give us generalized mutual information in the form

$$I_\alpha(X;Y) = \frac{1}{1-\alpha} \log_2 \left(\frac{\sum_{h=1}^{m} p_h^\alpha}{\sum_{h=1}^{m} \sum_{k=1}^{n} \frac{r_{hk}^\alpha}{q_k^{\alpha-1}}} \right) \tag{2.8}$$

However, Rényi preferred in his paper [4] another way of defining mutual information. He noticed that Shanon 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 2.2 and generalized mean instead of linear averaging results in

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

which satisfies 2.6 with the same conditions for equality as in Shanon's case. The drawback is that neither 2.9 nor 2.8 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 2.9 and 2.8 are different. First of them represents decrease of uncertainty while the second one is average information gain on X from observing Y.

In next chapter we will use different definition of mutual information that is based on property of Shanonon mutual information.

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

By inserting Rényi entropies we arrive to the formula

$$I_\alpha(X;Y) = \frac{1}{1-\alpha}\log_2\frac{\sum\limits_{h,k}(p_h q_k)^\alpha}{\sum\limits_{k,h} r_{hk}^\alpha} \tag{2.10}$$

This quantity is symmetric and might have been obtained also by using equation 2.7 and the second definition of conditional entropy, equation 2.4.

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, 2.10 can be negative and according to Rényi it is inappropriate to have negative mutual information. However, it was examined in [6] that mutual information defined in 2.10 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 the next chapter.

2.2 Operational definition

Renyi entropy is information measure as well as Shanon 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 Shanon 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 term needed to send one bit. Nevertheless, in some situations it is reasonable to use various cost function, for example, in storing data when exponential cost function may be used for "pricing" of allocated free space. Thus, we are not interested in the shortest code but the cheapest one.

Campbell dealt with the problem of exponential weighing in his paper [7]. He suggested 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 proposed to minimize logarithm of C

$$L(t) = \frac{1}{t} \log_D \left(\sum_{i=1}^{N} p_i D^{tl_i} \right) \tag{2.11}$$

to get elegant connection with generalized mean, i.e. 2.11 corresponds to Kolmogorov-Nagumo generalized mean with $\varphi(x) = D^{tx}$.

The following theorem is the analogy of well-known Shanon 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

$$H_\alpha \leq L(t), \tag{2.12}$$

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 2.12 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 Shanon 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 of finite buffer used for transmitting when 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 to defining different information measures it is convenient to postulate some basic requirements that suitable information measure should fulfill. Fadeev proposed 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 [8], that these axioms holds if and only if Shanon 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 Shanon's and Rényi's 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* -
$$\begin{aligned} &H(\{p_1q_1, \ldots, p_nq_1, p_1q_2, \ldots, p_nq_2, \ldots, p_1q_m, \ldots, p_nq_m\}) = \\ &H(\{p_1, \ldots, p_n\}) + H(\{q_1, \ldots, q_m\}) \end{aligned} \tag{2.13}$$
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 Shanon 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 Shanon's and Rényi's entropy that define also an conditional entropy in the form 2.5 which will be used in the next chapter.

1. Let X be a discrete random variable with probability distribution $\{p_i\}$, then $H(X)$ is 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 $\{p_i = 1/n,\ i = 1, \dots, n\}$ 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 $\{p_1, \dots, p_n\}$ its distribution, and if X' has probability distribution $\{p_1, \dots, p_n, 0\}$ 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 [9] in order to include Rényi's entropy. It can be shown that the only possible functions in axiom 3 are either linear or exponential function which corresponds to Shanon's and Rényi's entropy, see [5].

Chapter 3

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 [10] 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 [11] and [12]. 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 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.

3.1 Shannonian transfer entropy

We introduce transfer entropy in similar way as was done in [11] with slight modification in the form of time dependency.

Let have discrete stochastic process $\mathbf{X} = \{X_t\}$ (time series) then 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 $X_t, \ldots, X_{t-m+1}$.

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. For prediction the difference

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

is very important because it represents conditional entropy (see chain rule 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 stated in chapter 1 we have inequality.

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

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

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

$$RE_{\mathbf{X}}(t,m) = 1 - \frac{h_{\mathbf{X}}(t,m)}{H(X_{t+1})} \tag{3.1}$$

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 next step. Similarly, increase of $RE_{\mathbf{X}}(t,m)$ with respect to m until some 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

$$T^{(m,l)}_{\mathbf{Y}\mapsto\mathbf{X}}(t) = h_{\mathbf{X}}(t,m) - h_{\mathbf{XY}}(t,m,l), \tag{3.2}$$

where the conditional entropy for two processes is

$$h_{\mathbf{XY}}(t,m,l) = H_{\mathbf{XY}}(t+1,m+1,l) - H_{\mathbf{XY}}(t,m,l) \tag{3.3}$$

and

$$H_{\mathbf{XY}}(t,m,l) = -\sum p(x_t^{(m)}, y_t^{(l)}) \log_2 p(x_t^{(m)}, y_t^{(l)}) \tag{3.4}$$

where $x_t^{(m)}$ and $y_t^{(l)}$ substitutes history in X and Y respective, i.e. $x_t^{(m)} = x_t, \ldots, x_{t-m+1}$ and similarly $y_t^{(l)} = y_t, \ldots, y_{t-l+1}$.

From equation 3.2 we see that transfer entropy is always nonnegative since any extra knowledge about random variable never increase uncertainty, see equation 1.11, and transfer entropy vanish if and only if the next step in $\mathbf{X}$ process is independent on history up to $t-l+1$ of $\mathbf{Y}$.

After inserting 3.3 to 3.2 and with help of 3.4 we get following explicit formula for transfer entropy that can be used for numerical evaluation.

$$T^{(m,l)}_{\mathbf{Y}\mapsto\mathbf{X}}(t) = \sum p(x_{t+1}, x_t^{(m)}, y_t^{(l)}) \log_2 \frac{p(x_{t+1}|x_t^{(m)}, y_t^{(l)})}{p(x_{t+1}|x_t^{(m)})}, \tag{3.5}$$

where sum is taken over all possible outcomes of $(x_{t+1}, x_t^{(m)}, y_t^{(l)})$.

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

$$\begin{aligned} T^{(m,l)}_{\mathbf{Y}\mapsto\mathbf{X}}(t) = {} & \text{Uncertainty about next step in } \mathbf{X} \text{ knowing history in } \mathbf{X} \\ - {} & \text{Uncertainty about next step in } \mathbf{X} \text{ knowing history in } \mathbf{X} \text{ and } \mathbf{Y} \end{aligned} \tag{3.6}$$

By using 1.10 and generalization of 1.9 for more random variables we get transfer entropy in form of flow of information from process $\mathbf{Y}$ to $\mathbf{X}$.

$$\begin{aligned} T^{(m,l)}_{\mathbf{Y}\mapsto\mathbf{X}} &= I(X_{t+1}; X_t^{(m)}, Y_t^{(l)}) - I(X_{t+1}; X_t^{(m)}) \\ &= I(X_{t+1}; Y_t^{(l)} | X_t^{(m)}) \end{aligned} \tag{3.7}$$

The causality or directionality of transfer entropy is provided by non-symmetry property of conditional information. 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.

Other advantage of transfer entropy that arises from using information approach is that it takes into account all kinds of dependence unlike cross-correlation function which considers only linear correlation. Figures 3.1 and 3.2 demonstrate this benefit of information approach for lagged mutual information and auto-correlation function of London stock exchange index. We see that auto-correlation drops immediately to zero while lagged mutual information shows correlation even for one hour time lag.

We see that transfer entropy depends on two parameters (m and l). These parameters should correspond to order of Markov process, i.e. time series (discrete stochastic process) should be Markov process with specific order. 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 actual value of transfer entropy but that

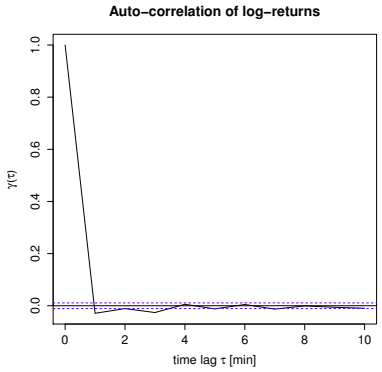


Figure 3.1: ACF of log-returns $r(t)$

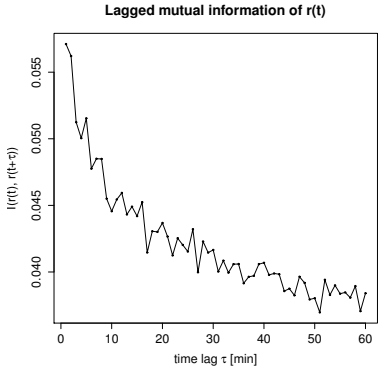


Figure 3.2: Lagged mutual information

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 3.7. Thus generally speaking, low m overestimates transfer entropy while low l, i.e. short history in source series, underestimates information flow. To avoid spurious information from target series it is common to set $l = 1$ and m as large as possible.

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

$$REA(m, l, t) = \frac{T^{(m,l)}_{\mathbf{Y}\mapsto\mathbf{X}}(t)}{h_{\mathbf{X}}(t, m)} \tag{3.8}$$

This quantity tell us how many percent of information about 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, arise in practical application since we have to somehow obtain the probability distribution in equation 3.5. This is done by observing one long realization of process and computing the relative frequencies, hence, the processes in consideration should be also ergodic.

3.2 Rényian transfer entropy

Generalization of Shannonian transfer entropy for Renyi's entropy may be done according to information representation of transfer entropy, i.e., equation 3.7, see [6].

$$\begin{aligned} T^{(m,l)}_{q;\mathbf{Y}\mapsto\mathbf{X}}(t) &= H_q(X_{t+1}|X_t^{(m)}) - H_q(X_{t+1}|Y_t^{(l)}, X_t^{(m)}) \\ &= I_q(X_{t+1}; X_t^{(m)}, Y_t^{(l)}) - I_q(X_{t+1}; X_t^{(m)}) \end{aligned} \tag{3.9}$$

Using definition of conditional entropy in equation 2.4 and mutual information in form 2.10 we can rewrite aforementioned equation to

$$\begin{aligned} T^{(m,l)}_{q;\mathbf{Y}\mapsto\mathbf{X}}(t) &= \frac{1}{1-q}\log_2 \frac{\sum \rho_q(x_t^{(m)}) p^q(x_{t+1}|x_t^{(m)})}{\sum \rho_q(x_t^{(m)}, y_t^{(l)}) p^q(x_{t+1}|x_t^{(m)}, y_t^{(l)})} \\ &= \frac{1}{1-q}\log_2 \frac{\sum \rho_q(x_t^{(m)}) p^q(y_t^{(l)}|x_t^{(m)})}{\sum \rho_q(x_{t+1}, x_t^{(m)}) p^q(y_t^{(l)}|x_{t+1}, x_t^{(m)})} \end{aligned} \tag{3.10}$$

As we know from preceding chapter, 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 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 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 2.10.

3.3 Simulated data

To test properties of transfer entropy estimator we simulated simple linear coupling:

$$X(t) = r(t) + \epsilon Y(t-1), \tag{3.11}$$

$$Y(t) = s(t), \tag{3.12}$$

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

3.3.1 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 date 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., $x_t^{(m)} = (x_t, \ldots, x_{t-m+1})$ and due to stationarity we get $x_t^{(m)} = (x_0, \ldots, x_{-m+1})$ 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 is available, see [11], hence we can write y_0 instead of $y^{(l)}$. Then transfer entropy 3.5 may be written as

$$T_{\mathbf{Y}\mapsto\mathbf{X}}^{(m,1)} = \sum_{x_{m+1}} \sum_{x^{(m)}} \sum_{y_0} p(x_{m+1}, x^{(m)}, y_0) \log_2 \frac{p(x_{m+1}, x^{(m)}, y_0) p(x^{(m)})}{p(x^{(m)}, y_0) p(x_{m+1}, x^{(m)})} \tag{3.13}$$

From equation 3.11 and 3.12 we see that successive values of X process are mutually independent and identical 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 3.13

$$p(x_{m+1}, x^{(m)}, y_0) = p(x_{m+1}, y_0) p(x^{(m)}) \tag{3.14}$$

$$p(x^{(m)}, y_0) = p(y_0) p(x^{(m)}) \tag{3.15}$$

$$p(x_{m+1}, x^{(m)}) = p(x_{m+1}) p(x^{(m)}) \tag{3.16}$$

After inserting equations 3.14, 3.15 and 3.16 into 3.13 and simplification of the fraction we arrive at

$$T_{\mathbf{Y}\mapsto\mathbf{X}}^{(m,1)} = \sum_{x_{m+1}} \sum_{x^{(m)}} \sum_{y_0} p(x_{m+1}, y_0) p(x^{(m)}) \log_2 \frac{p(x_{m+1}, y_0)}{p(y_0) p(x_{m+1})} \tag{3.17}$$

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

$$T_{\mathbf{Y}\mapsto\mathbf{X}}^{(m,1)} = \sum_{x,y=1}^{S} p(x, y) \log_2 \frac{p(x, y)}{p(y) p(x)} \tag{3.18}$$

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 according to standard deviation.

We will need joint probability density function to calculate probabilities occurring in equation 3.18. The density may be written in the form

$$\rho(x,y) = \frac{1}{2\pi}\exp\left\{-\frac{(x-\epsilon y)^2 - y^2}{2}\right\} \tag{3.19}$$

which follows from definitional equations 3.11 and 3.12 (recall that x and y represents $X(t+1)$ and $Y(t)$ respectively). Then, necessary probabilities are obtained by integrating over appropriate limits depending on chosen partition.

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

$$2\log_2 S + \sum_{x,y=1}^{S} p(x,y)\log_2 p(x,y) \tag{3.20}$$

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

$$\sum_{x,y=1}^{3} p(x,y)\log_2 p(x,y) - \sum_{x=1}^{3} p(x)\log_2 p(x) - \sum_{y=1}^{3} p(y)\log_2 p(y) \tag{3.21}$$

Numerical evaluation of equations 3.20 and 3.21 gives theoretical values 0.010 and 0.009 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 3.3 along with straight lines denoting precise theoretical value 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 see appendix A.

From figure 3.3 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 [11] the same example was 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*

$$T^{Eff}_{Y\to X} = T_{Y\to X} - T_{Y_{sh}\to X} \tag{3.22}$$

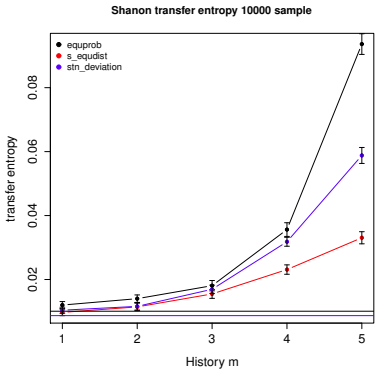


Figure 3.3: Transfer entropy

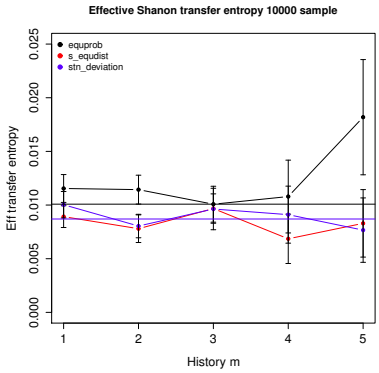


Figure 3.4: Effective transfer entropy

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 3.1: Standard errors

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 3.22 instead of 3.5.

Estimator of Effective entropy is depicted in figure 3.4. Though it is clear that Effective entropy is much closer to theoretical values for both partitions than transfer entropy estimator (notice different scale on y axis) it still considerably fluctuates for different values of m even for relatively large sample length used $N = 10000$, see also [11] where comparison between transfer entropy and Effective transfer entropy was done for length 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. 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 3.4 can

be seen that this partition is rather stable with respect to m and therefore should be used in later application. 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 3.1 for small $m \in \{1, 2\}$.

3.3.2 Rényian flow

The same linear coupling was analyzed even with help of Rényian transfer entropy. Here, instead of struggling with explicit formula 3.10 that is suitable for unknown systems, we profit from symbolic representation of transfer entropy, equation 3.9. Using the same argumentation about independence as used in Shanonian case we get

$$\begin{aligned} T^{(m,l)}_{q;\mathbf{Y}\mapsto\mathbf{X}}(t) &= H_q(X_{t+1}) + H_q(Y_t) - H_q(X_{t+1}, Y_t) \\ &= \frac{1}{1-q}\left(\log_2 \sum_y p^q(y) + \log_2 \sum_x p^q(x) - \log_2 \sum_{x,y} p^q(x,y)\right) \end{aligned} \tag{3.23}$$

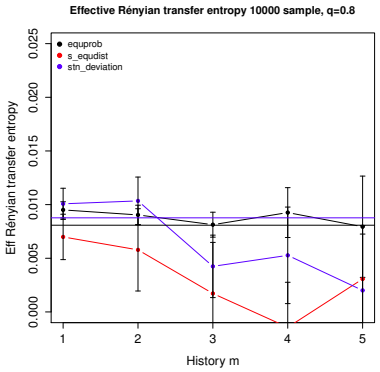


Figure 3.5: Rényian effective transfer entropy, $q = 0.8$

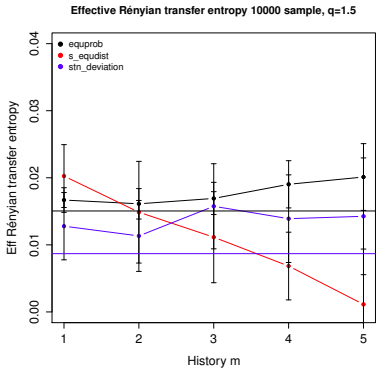


Figure 3.6: Rényian effective transfer entropy, $q = 1.5$

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 3.5, 3.6.

In order to compare Shanonnian 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. Actually, time series are usually already discretized in time so we can say that every time series analysis uses 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.

3.4 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 appendix with figure depicting means and variances calculated in individual blocks along with its error bars. 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 preserve non-stationarity. Since our basic analysis assumes stationarity of time series we had to select only part of data where all indexes in particular continent have at least approximately the

same mean and variance within their blocks. Mean does not violate stationarity assumption so much but variance is more diverse in particular blocks, see appendix B.

In the case of Europe we analyzed three indexes, 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. We choose data points from around 13th to 15th block which shows the lowest variance diversity. This data corresponds to period from $23rd$ of August to $7th$ of September. Only two weeks may seem rather short but acquisition at high frequency assures sufficient amount of data $\simeq 5000$.

In America the biggest stock indexes was 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 therefore 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 indexes 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 has left which is reason why only 11 blocks were used to test stationarity. See in appendix that it is impossible to find stationary part in all five indexes, hence, we opted only Shenzhen Stock Exchange Composite Index, Korea Stock Exchange KOSPI 200 Index and HSI - The Hang Seng Index Hong Kong from 13th of August to 6th of September. This period gave us over 3000 data points.

3.4.1 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 remarkable and for this reason we decided to use only transfer entropy and parameters $m = l = 1$ as had been done in [12] where similar amount of date had been analyzed.

This approach leads to slightly overestimating 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 appears. 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 [11].

3.4.2 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 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 are particular stock exchanges only slightly coupled (note different scales in heat maps).

3.4.3 Time dependent information flow

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

We can see from figure 3.8 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 other stronger correlation around 8th week. At the end of examined period we see restoration of previous lower connected state.

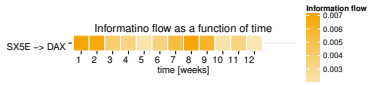


Figure 3.7: Heat map of Shanonnian transfer entropy as a function of time

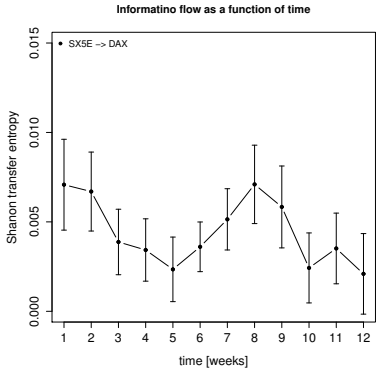


Figure 3.8: Shanonnian transfer entropy as a function of time

Appendix A

Estimators, errors and Bootstrap

A.1 Estimators

From data we want to infer some feature A of 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 following properties.

- *consistency* - $\lim_{N \mapsto \infty} \hat{A}_N = A$
- *unbiasedness* - $E[\hat{A}_N] = A$
- *effectiveness* - Every other estimator of A fulfilling conditions above must have higher variance.

For example biased estimator of variance is

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

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 common practice in estimation of parameters. However, this estimator is biased and the unbiased one is

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

where $\bar{x}$ is sample mean.

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 of the estimate and is used for error bars in plotting.

Let have a look at the simples case, i.e. population mean. Its variance, provided the date is 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 properly estimate standard error of $\bar{x}$ we need to find suitable estimator $\hat{\sigma}$ of population standard deviation σ. Such an estimator is, see [13]:

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

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 $\hat{\sigma}_{\bar{x}} = s/\sqrt{N}$. We use this relation for testing stationarity of time series in the last chapter.

The standard error of variance estimator is:

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

we used the fact that $\frac{(N-1)s^2}{\sigma^2}$ has chi-squared distribution with $N-1$ degrees of freedom and such a distribution has variance $2(N-1)$. Thus estimated error of sample variance, which we also use in the last chapter for basic weak stationary justification, is:

$$\hat{\sigma}_{s^2} = s^2 \sqrt{\frac{2}{N-1}}$$

A.2 Bootstrap

These two examples are specific because there is analytical derived standard error of the estimate. In practice we need to estimate other, much more difficult, parameters also called statistics θ, i.e. function of hidden probability distribution $\theta = t(F)$. For this purpose, we suggest some estimate $\hat{\theta}$ of θ and then we need to know its standard error. In many cases we are not able to derive exact formula and it is where bootstrap come handy. Bootstrap has been using since 1979 when computers power became capable of processing huge amount of data in reasonable time, see [14].

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

After that, we calculate estimator for this new sample $\hat{\theta}^*(\mathbf{x}^*)$ and 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 of the sample of estimators. The length m of bootstrap sample is usually taken in range $25 - 200$, see [14].

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 identical distributed restriction may be satisfied for stationary time series but the independence is general problem in most time series. The simplest solution is differencing time series and hope that new time series of differences is already independent, as it is case, e.g. for random walk. The differencing of time series is based on some a priori known structure or model of the system. Hence, provided we have faithful model of date we may bootstrap only the extracted random already independent noise or residuals and then reconstruct resampled time series. However, in many cases no suitable model exist and we are left with nonparametric bootstrap.

We analyze financial data, particularly stock indexes values which are, more precisely its logarithm, according to old and austere theory motivated by Bachelier regarded as a random walk. Therefore, someone would expect there is no problem, 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 suggest uncorrelated date but as we have already seen auto-correlation cannot detect nonlinear correlation and more precise analysis with mutual information points out nonnegligible 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, these blocks are less correlated and the original structure of time series remains unchanged. 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.

Appendix B

Figures

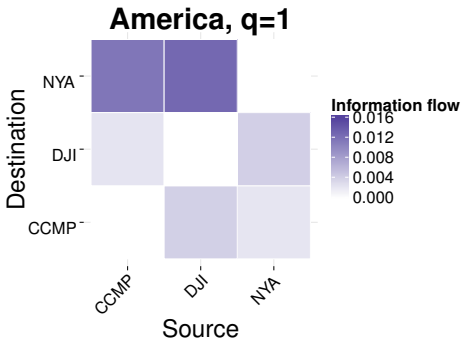


Figure B.1: America, heat map of Shanonian transfer entropy

	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

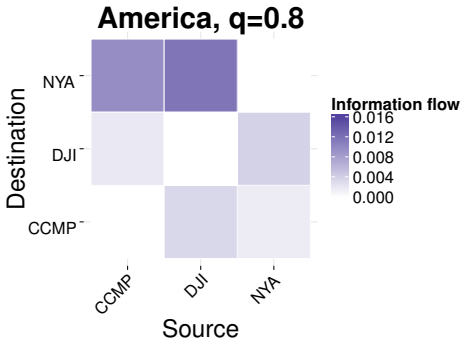


Figure B.2: America, heat map of Rényian transfer entropy q=0.8

	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

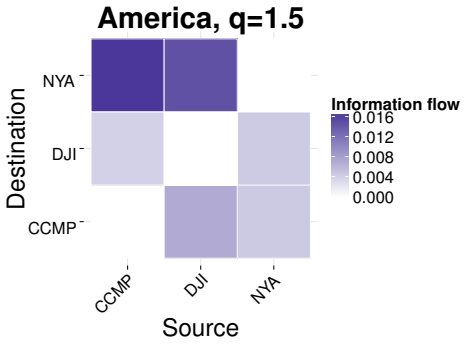


Figure B.3: America, heat map of Rényian transfer entropy q=1.5

	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

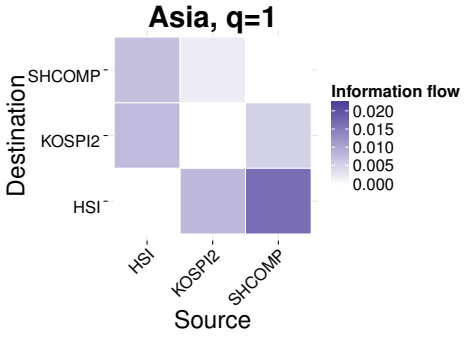


Figure B.4: Asia, heat map of Shanonian transfer entropy

	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

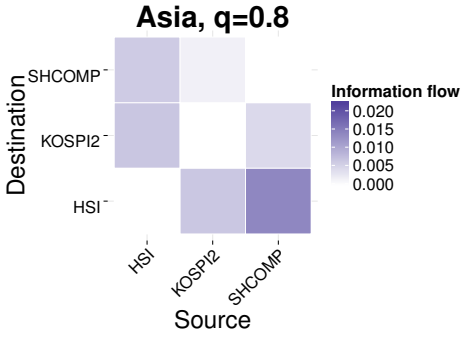


Figure B.5: Asia, heat map of Rényian transfer entropy q=0.8

	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

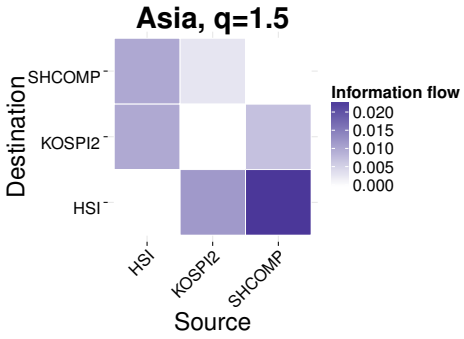


Figure B.6: Asia, heat map of Rényian transfer entropy q=1.5

	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

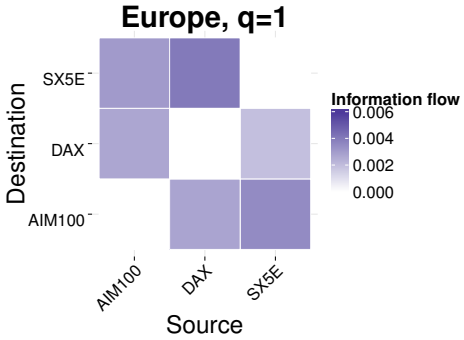


Figure B.7: Europe, heat map of Shanonian transfer entropy

	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

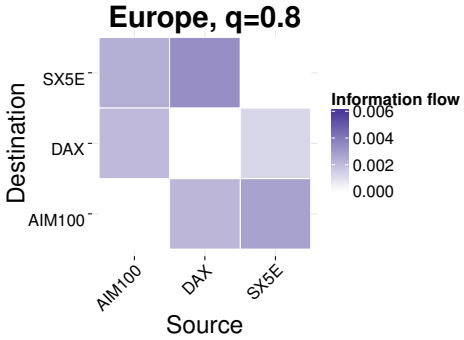


Figure B.8: Europe, heat map of Rényian transfer entropy q=0.8

	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

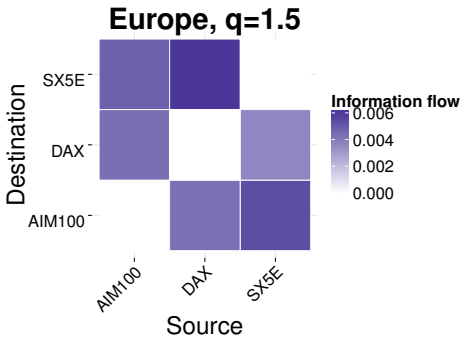
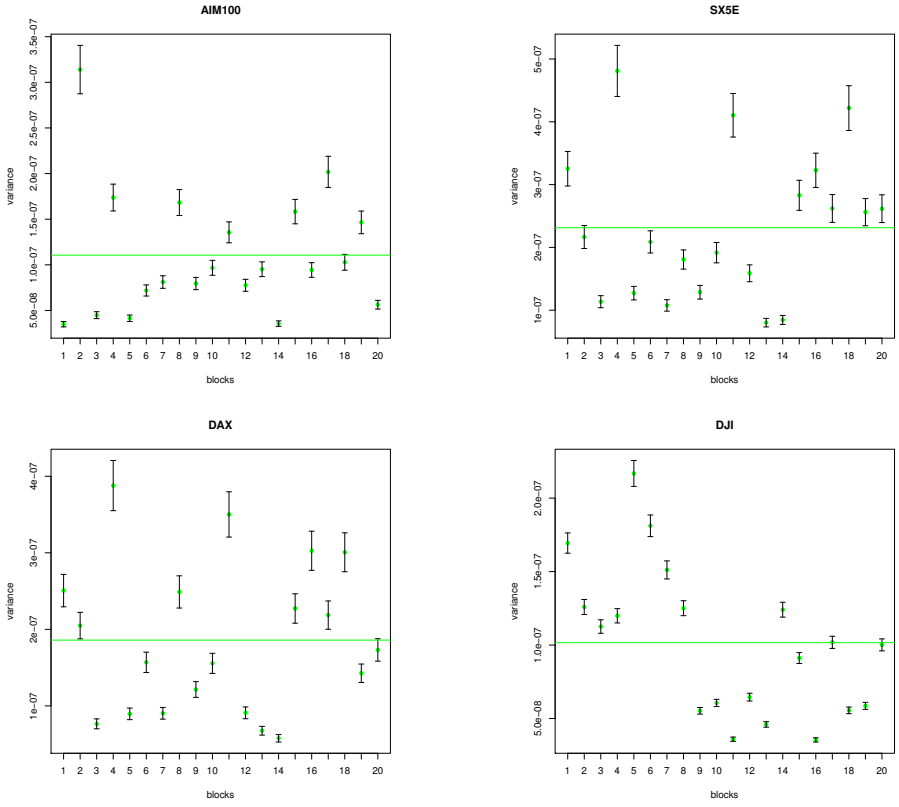
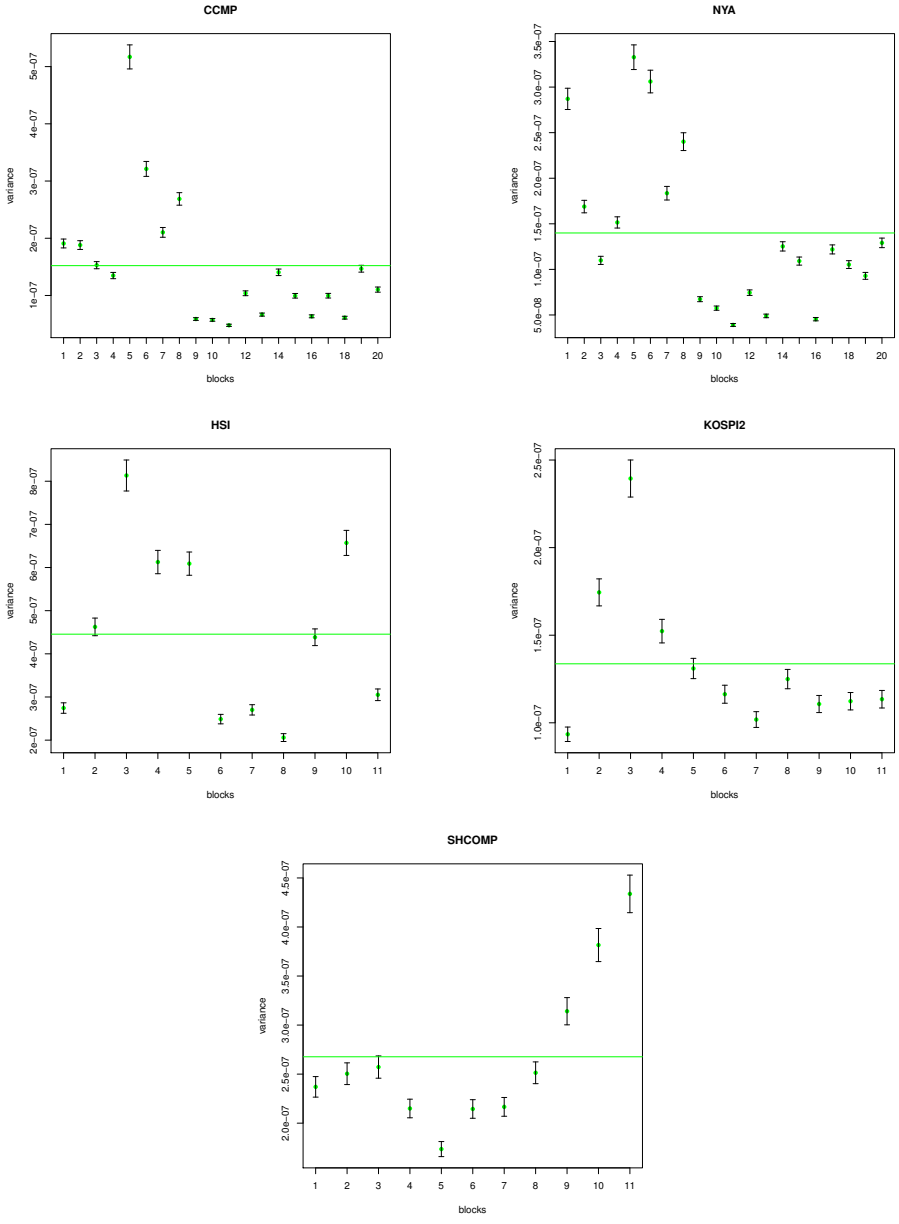


Figure B.9: Europe, heat map of Rényian transfer entropy q=1.5

	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




CCMP
variance
blocks
NYA
variance
blocks
HSI
variance
blocks
KOSPI2
variance
blocks
SHCOMP
variance
blocks

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

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

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

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

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

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

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

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

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

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

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

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

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

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