

Tampere International Center for Signal Processing. TICSP series # 56

Boris Ryabko, Jaakko Astola & Mikhail Malyutov

Compression-Based Methods of Prediction and Statistical Analysis of Time Series: Theory and Applications

Tampere International Center for Signal Processing
Tampere 2010

ISBN 978-952-15-2444-8
ISSN 1456-2774

TICSP Series # 56

COMPRESSION–BASED
METHODS OF PREDICTION
AND STATISTICAL ANALYSIS OF
TIME SERIES:
THEORY AND APPLICATIONS

Boris Ryabko, Jaakko Astola, Mikhail Malyutov

TICSP Series

Tampere International Center for Signal Processing
Tampere University of Technology
P.O. Box 553
FI-33101 Tampere
Finland

ISBN 978-952-15-2444-8
ISSN 1456-2774

Tampereen Yliopistopaino Oy
2010

Preface

Initially, in nineteen sixties, universal codes for lossless data compression were developed for information storage and transmission. Those codes can efficiently compress sequences generated by stationary and ergodic sources with unknown statistics. In twenty years, it was realized that universal codes can be used for solving many important problems of prediction and statistical analysis of time series. This book describes recent results in this area.

The first part of this book is mainly devoted to general description of statistical methods which are based on universal codes. This part is written by B. Ryabko and J. Astola.

The second part describes a sketch of theory and many applications of a simplified homogeneity test between literary texts based on universal compressors. In particular, this test is used to attributing authorship, if training texts written by competitive candidates are available. This part is written by M. Malyutov.

Part 1

Statistical Methods Based on Universal Codes

Abstract

We show how universal codes can be used for solving some of the most important statistical problems for time series. By definition, a universal code (or a universal lossless data compressor) can compress any sequence generated by a stationary and ergodic source asymptotically to the Shannon entropy, which, in turn, is the best achievable ratio for lossless data compressors.

First we show how universal codes can be used for solving some problems of time series analysis and then apply these methods to several real problems.

We consider finite-alphabet and real-valued time series and the following problems: estimation of the limiting probabilities for finite-alphabet time series and estimation of the density for real-valued time series, the on-line prediction, regression, classification (or problems with side information) for both types of the time series and the following problems of hypothesis testing: goodness-of-fit testing, or identity testing, and testing of serial independence. It is important to note that all problems are considered in the framework of classical mathematical statistics and, on the other hand, everyday methods of data compression (or archivers) can be used as a tool for the estimation and testing.

It turns out, that quite often the suggested methods and tests are more powerful than known ones when they are applied in practice.

The applications are intended to show the practical efficiency of the obtained methods and concern the prediction of currency rates and testing the style homogeneity between literary texts.

1 Introduction

Since C. Shannon published the paper “A mathematical theory of communication” [50], the ideas and results of Information Theory have played an important role in cryptography [27, 51], mathematical statistics [3, 8, 16, 26], and many other fields [6, 7], which are far from telecommunications. Universal coding, which is a part of Information Theory, also has been efficiently applied in many fields since its discovery [22, 13]. Thus, application of results of universal coding, initiated in 1988 [36], created a new approach to prediction [1, 20, 28, 30]. Maybe the most unexpected application of data compression ideas arises in experiments that show that some ant species are capable of compressing messages and are capable of adding and subtracting small numbers [45, 46].

In this chapter we describe a new approach to estimation, prediction and hypothesis testing for time series, which was suggested recently [36, 40, 44, 39]. This approach is based on ideas of universal coding (or universal data compression). We would like to emphasize that everyday methods of data compression (or archivers) can be directly used as a tool for estimation and hypothesis testing. It is important to note that the modern archivers (like *zip*, *arj*, *rar*, etc.) are based on deep theoretical results of the source coding theory [10, 21, 25, 33, 49] and have shown their high efficiency in practice because archivers can find many kinds of latent regularities and use them for compression.

It is worth noting that this approach was applied to the problem of randomness testing [44]. This problem is quite important for practice; in particular, the National Institute of Standards and Technology of USA (NIST) has suggested “A statistical test suite for random and pseudorandom number generators for cryptographic applications” [34], which consists of 16 tests. It has turned out that tests which are based on universal codes are more powerful than the tests suggested by NIST [44].

All proofs are given in Appendix, but some intuitive indication are given in the body of the paper.

2 Definitions and Statements of the Problems

2.1 Estimation and Prediction for I.I.D. Sources

First we consider a source with unknown statistics which generates sequences $x_1x_2\cdots$ of letters from some set (or alphabet) A . It will be con-

venient now to describe briefly the prediction problem. Let the source generate a message $x_1 \dots x_{t-1}x_t$, $x_i \in A$ for all i , and the following letter x_{t+1} needs to be predicted. This problem can be traced back to Laplace [11, 31] who considered the problem of estimation of the probability that the sun will rise tomorrow, given that it has risen every day since Creation. In our notation the alphabet A contains two letters 0 ("the sun rises") and 1 ("the sun does not rise"), t is the number of days since Creation, $x_1 \dots x_{t-1}x_t = 00 \dots 0$.

Laplace suggested the following predictor:

$$L_0(a|x_1 \dots x_t) = (\nu_{x_1 \dots x_t}(a) + 1)/(t + |A|), \quad (1)$$

where $\nu_{x_1 \dots x_t}(a)$ is denotes the count of letter a occurring in the word $x_1 \dots x_{t-1}x_t$. It is important to note that the predicted probabilities cannot be equal to zero even through a certain letter did not occur in the word $x_1 \dots x_{t-1}x_t$.

Example. Let $A = \{0, 1\}$, $x_1 \dots x_5 = 01010$, then the Laplace prediction is as follows: $L_0(x_6 = 0|x_1 \dots x_5 = 01010) = (3 + 1)/(5 + 2) = 4/7$, $L_0(x_6 = 1|x_1 \dots x_5 = 01010) = (2 + 1)/(5 + 2) = 3/7$. In other words, $3/7$ and $4/7$ are estimations of the unknown probabilities $P(x_{t+1} = 0|x_1 \dots x_t = 01010)$ and $P(x_{t+1} = 1|x_1 \dots x_t = 01010)$. (In what follows we will use the shorter notation: $P(0|01010)$ and $P(1|01010)$).

We can see that Laplace considered prediction as a set of estimations of unknown (conditional) probabilities. This approach to the problem of prediction was developed in 1988 [36] and now is often called on-line prediction or universal prediction [1, 20, 28, 30]. As we mentioned above, it seems natural to consider conditional probabilities to be the best prediction, because they contain all information about the future behavior of the stochastic process. Moreover, this approach is deeply connected with game-theoretical interpretation of prediction [18, 38] and, in fact, all obtained results can be easily transferred from one model to the other.

Any predictor γ defines a measure (or an estimation of probability) by the following equation

$$\gamma(x_1 \dots x_t) = \prod_{i=1}^t \gamma(x_i|x_1 \dots x_{i-1}). \quad (2)$$

And, vice versa, any measure γ (or estimation of probability) defines a predictor:

$$\gamma(x_i|x_1 \dots x_{i-1}) = \gamma(x_1 \dots x_{i-1}x_i)/\gamma(x_1 \dots x_{i-1}). \quad (3)$$

Example. Let us apply the Laplace predictor for estimation of probabilities of the sequences 01010 and 010101. From (2) we obtain $L_0(01010) = \frac{1}{2} \frac{1}{3} \frac{2}{4} \frac{2}{5} \frac{3}{6} = \frac{1}{60}$, $L_0(010101) = \frac{1}{60} \frac{3}{7} = \frac{1}{140}$. Vice versa, if for some measure (or a probability estimation) χ we have $\chi(01010) = \frac{1}{60}$ and $\chi(010101) = \frac{1}{140}$, then we obtain from (3) the following prediction, or the estimation of the conditional probability, $\chi(1|01010) = \frac{1/140}{1/60} = \frac{3}{7}$.

Now we concretize the class of stochastic processes which will be considered. Generally speaking, we will deal with so-called stationary and ergodic time series (or sources), whose definition will be given later, but now we consider may be the simplest class of such processes, which are called i.i.d. sources. By definition, they generate independent and identically distributed random variables from some set A . In our case A will be either some alphabet or a real-valued interval.

The next natural question is how to measure the errors of prediction and estimation of probability. Mainly we will measure these errors by the Kullback-Leibler (KL) divergence which is defined by

$$D(P, Q) = \sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)}, \quad (4)$$

where $P(a)$ and $Q(a)$ are probability distributions over an alphabet A (here and below $\log \equiv \log_2$ and $0 \log 0 = 0$). The probability distribution $P(a)$ can be considered as unknown whereas $Q(a)$ is its estimation. It is well-known that for any distributions P and Q the KL divergence is nonnegative and equals 0 if and only if $P(a) = Q(a)$ for all a [14]. So, if the estimation Q is equal to P , the error is 0, otherwise the error is a positive number.

The KL divergence is connected with the so-called variation distance

$$\|P - Q\| = \sum_{a \in A} |P(a) - Q(a)|,$$

via the the following inequality (Pinsker's inequality)

$$\sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)} \geq \frac{\log e}{2} \|P - Q\|^2. \quad (5)$$

Let γ be a predictor, i.e. an estimation of an unknown conditional probability and $x_1 \cdots x_t$ be a sequence of letters created by an unknown source P . The KL divergence between P and the predictor γ is equal to

$$\rho_{\gamma, P}(x_1 \cdots x_t) = \sum_{a \in A} P(a|x_1 \cdots x_t) \log \frac{P(a|x_1 \cdots x_t)}{\gamma(a|x_1 \cdots x_t)}, \quad (6)$$

For fixed t it is a random variable, because x_1, x_2, \dots, x_t are random variables. We define the average error at time t by

$$\begin{aligned} \rho^t(P|\gamma) &= E(\rho_{\gamma,P}(\cdot)) = \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \rho_{\gamma,P}(x_1 \dots x_t) \quad (7) \\ &= \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \sum_{a \in A} P(a|x_1 \dots x_t) \log \frac{P(a|x_1 \dots x_t)}{\gamma(a|x_1 \dots x_t)}. \end{aligned}$$

Analogously, if $\gamma(\cdot)$ is an estimation of a probability distribution we define the errors *per letter* as follows:

$$\bar{\rho}_{\gamma,P}(x_1 \dots x_t) = t^{-1} (\log(P(x_1 \dots x_t)/\gamma(x_1 \dots x_t))) \quad (8)$$

and

$$\bar{\rho}^t(P|\gamma) = t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log(P(x_1 \dots x_t)/\gamma(x_1 \dots x_t)), \quad (9)$$

where, as before, $\gamma(x_1 \dots x_t) = \prod_{i=1}^t \gamma(x_i|x_1 \dots x_{i-1})$. (Here and below we denote by A^t and A^* the set of all words of length t over A and the set of all finite words over A correspondingly: $A^* = \bigcup_{i=1}^{\infty} A^i$.)

Claim 1 ([36]) *For any i.i.d. source P generating letters from an alphabet A and an integer t the average error (7) of the Laplace predictor and the average error of the Laplace estimator are upper bounded as follows:*

$$\rho^t(P|L_0) \leq ((|A| - 1) \log e)/(t + 1), \quad (10)$$

$$\bar{\rho}^t(P|L_0) \leq (|A| - 1) \log t/t + O(1/t), \quad (11)$$

where $e \simeq 2.718$ is the Euler number.

So, we can see that the average error of the Laplace predictor goes to zero for any i.i.d. source P when the length t of the sample $x_1 \dots x_t$ tends to infinity. Such methods are called universal, because the error goes to zero for any source, or process. In this case they are universal for the set of all i.i.d. sources generating letters from the finite alphabet A , but later we consider universal estimators for the set of stationary and ergodic sources. It is worth noting that the first universal code for which the estimation (11) is valid, was suggested independently by Fitingof [13] and Kolmogorov [22] in 1966.

The value

$$\bar{\rho}^t(P|\gamma) = t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log(P(x_1 \dots x_t)/\gamma(x_1 \dots x_t))$$

has one more interpretation connected with data compression. Now we consider the main idea whereas the more formal definitions will be given later. First we recall the definition of the Shannon entropy $h_0(P)$ for an i.i.d. source P

$$h_0(P) = - \sum_{a \in A} P(a) \log P(a). \quad (12)$$

It is easy to see that $t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log(P(x_1 \dots x_t)) = -h_0(P)$ for the i.i.d. source. Hence, we can represent the average error $\bar{\rho}^t(P|\gamma)$ in (9) as

$$\bar{\rho}^t(P|\gamma) = t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log(1/\gamma(x_1 \dots x_t)) - h_0(P).$$

More formal and general consideration of universal codes will be given later, but here we briefly show how estimations and codes are connected. The point is that one can construct a code with codelength

$$\gamma_{code}(a|x_1 \dots x_t) \approx -\log_2 \gamma(a|x_1 \dots x_t)$$

for any letter $a \in A$ (since Shannon's original research, it has been well known that, using block codes with large block length or more modern methods of arithmetic coding [32], the approximation may be as accurate as you like). If one knows the real distribution P , one can base coding on the true distribution P and not on the prediction γ . The difference in performance measured by average code length is given by

$$\begin{aligned} & \sum_{a \in A} P(a|x_1 \dots x_t) (-\log_2 \gamma(a|x_1 \dots x_t)) \\ & - \sum_{a \in A} P(a|x_1 \dots x_t) (-\log_2 P(a|x_1 \dots x_t)) \\ & = \sum_{a \in A} P(a|x_1 \dots x_t) \log_2 \frac{P(a|x_1 \dots x_t)}{\gamma(a|x_1 \dots x_t)}. \end{aligned}$$

Thus this excess is exactly the error defined above (6) . Analogously, if we encode the sequence $x_1 \dots x_t$ based on a predictor γ the redundancy per letter is defined by (8) and (9). So, from mathematical point of view, the estimation of the limiting probabilities and universal coding are identical. But $-\log \gamma(x_1 \dots x_t)$ and $-\log P(x_1 \dots x_t)$ have a very natural interpretation. The first value is a code word length (in bits), if the "code" γ is applied for compressing the word $x_1 \dots x_t$ and the second one is the minimal possible codeword length. The difference is the redundancy of the code and, at the same time, the error of the predictor. It is worth noting that there are many other deep interrelations between the universal coding, prediction and estimation [33, 36].

We can see from the claim and the Pinsker inequality (5) that the variation distance of the Laplace predictor and estimator goes to zero, too. Moreover, it can be easily shown that the error (6) (and the corresponding variation distance) goes to zero with probability 1, when t goes to infinity. (Informally, it means that the error (6) goes to zero for almost all sequences $x_1 \dots x_t$ according to the measure P .) Obviously, such properties are very desirable for any predictor and for larger classes of sources, like Markov and stationary ergodic (they will be briefly defined in the next subsection). However, it is proven [36] that such predictors do not exist for the class of all stationary and ergodic sources (generating letters from a given finite alphabet). More precisely, if, for example, the alphabet has two letters, then for any predictor γ and for any $\delta > 0$ there exists a source P such that with probability 1 $\rho_{\gamma,P}(x_1 \dots x_t) \geq 1/2 - \delta$ infinitely often when $t \rightarrow \infty$. In other words, the error of any predictor may not go to 0, if the predictor is applied to an arbitrary stationary and ergodic source, that is why it is difficult to use (6) and (7) to compare different predictors. On the other hand, it is shown [36] that there exists a predictor R , such that the following Cesaro average $t^{-1} \sum_{i=1}^t \rho_{R,P}(x_1 \dots x_i)$ goes to 0 (with probability 1) for any stationary and ergodic source P , where t goes to infinity. (This predictor will be described in the next subsection.) That is why we will focus our attention on such averages. From the definitions (6), (7) and properties of the logarithm we can see that for any probability distribution γ

$$t^{-1} \sum_{i=1}^t \rho_{\gamma,P}(x_1 \dots x_i) = t^{-1} (\log(P(x_1 \dots x_t)/\gamma(x_1 \dots x_t))),$$

$$t^{-1} \sum_{i=1}^t \rho^i(P|\gamma) = t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log(P(x_1 \dots x_t)/\gamma(x_1 \dots x_t)).$$

Taking into account these equations, we can see from the definitions (8) and (9) that the Chesaro averages of the prediction errors (6) and (7) are equal to the errors of estimation of limiting probabilities (8) and (9). That is why we will use values (8) and (9) as the main measures of the precision throughout the chapter.

A natural problem is to find a predictor and an estimator of the limiting probabilities whose average error (9) is minimal for the set of i.i.d. sources. This problem was considered and solved by Krichevsky [24, 25]. He suggested the following predictor:

$$K_0(a|x_1 \dots x_t) = (\nu_{x_1 \dots x_t}(a) + 1/2)/(t + |A|/2), \quad (13)$$

where, as before, $\nu_{x_1 \dots x_t}(a)$ is the number of occurrences of the letter a in the word $x_1 \dots x_t$. We can see that the Krichevsky predictor is quite close to the Laplace's one (35).

Example. Let $A = \{0, 1\}$, $x_1 \dots x_5 = 01010$. Then $K_0(x_6 = 0|01010) = (3 + 1/2)/(5 + 1) = 7/12$, $K_0(x_6 = 1|01010) = (2 + 1/2)/(5 + 1) = 5/12$ and $K_0(01010) = \frac{1}{2} \frac{1}{4} \frac{1}{2} \frac{3}{8} \frac{1}{2} = \frac{3}{256}$.

The Krichevsky measure K_0 can be represented as follows:

$$K_0(x_1 \dots x_t) = \prod_{i=1}^t \frac{\nu_{x_1 \dots x_{i-1}}(x_i) + 1/2}{i - 1 + |A|/2} = \frac{\prod_{a \in A} (\prod_{j=1}^{\nu_{x_1 \dots x_t}(a)} (j - 1/2))}{\prod_{i=0}^{t-1} (i + |A|/2)}. \quad (14)$$

It is known that

$$(r + 1/2)((r + 1) + 1/2) \dots (s - 1/2) = \frac{\Gamma(s + 1/2)}{\Gamma(r + 1/2)}, \quad (15)$$

where $\Gamma(\cdot)$ is the gamma function [23]. So, (14) can be presented as follows:

$$K_0(x_1 \dots x_t) = \frac{\prod_{a \in A} (\Gamma(\nu_{x_1 \dots x_t}(a) + 1/2) / \Gamma(1/2))}{\Gamma(t + |A|/2) / \Gamma(|A|/2)}. \quad (16)$$

The following claim shows that the error of the Krichevsky estimator is a half of the Laplace's one.

Claim 2 For any i.i.d. source P generating letters from a finite alphabet A the average error (9) of the estimator K_0 is upper bounded as follows:

$$\begin{aligned} \bar{\rho}_t(K_0, P) &\equiv t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log(P(x_1 \dots x_t)/K_0(x_1 \dots x_t)) \equiv \\ t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log(1/K_0(x_1 \dots x_t)) - h_0(p) &\leq ((|A|-1) \log t + C)/(2t), \end{aligned} \tag{17}$$

where C is a constant.

Moreover, in a certain sense this average error is minimal: it is shown by Krichevsky [24] that for any predictor γ there exists such a source P^* that

$$\bar{\rho}_t(\gamma, P^*) \geq ((|A| - 1) \log t + C')/(2t).$$

Hence, the bound $((|A| - 1) \log t + C)/(2t)$ cannot be reduced and the Krichevsky estimator is the best (up to $O(1/t)$) if the error is measured by the KL divergence ρ .

2.2 Consistent Estimations and On-line Predictors for Markov and Stationary Ergodic Processes

Now we briefly describe consistent estimations of unknown probabilities and efficient on-line predictors for general stochastic processes (or sources of information).

First we give a formal definition of stationary ergodic processes. The time shift T on A^∞ is defined as $T(x_1, x_2, x_3, \dots) = (x_2, x_3, \dots)$. A process P is called stationary if it is T -invariant: $P(T^{-1}B) = P(B)$ for every Borel set $B \subset A^\infty$. A stationary process is called ergodic if every T -invariant set has probability 0 or 1: $P(B) = 0$ or 1 whenever $T^{-1}B = B$ [5, 14].

We denote by $M_\infty(A)$ the set of all stationary and ergodic sources and let $M_0(A) \subset M_\infty(A)$ be the set of all i.i.d. processes. We denote by $M_m(A) \subset M_\infty(A)$ the set of Markov sources of order (or with memory, or connectivity) not larger than m , $m \geq 0$. By definition $\mu \in M_m(A)$ if

$$\begin{aligned} \mu(x_{t+1} = a_{i_1} | x_t = a_{i_2}, x_{t-1} = a_{i_3}, \dots, x_{t-m+1} = a_{i_{m+1}}, \dots) & \tag{18} \\ = \mu(x_{t+1} = a_{i_1} | x_t = a_{i_2}, x_{t-1} = a_{i_3}, \dots, x_{t-m+1} = a_{i_{m+1}}) & \end{aligned}$$

for all $t \geq m$ and $a_{i_1}, a_{i_2}, \dots \in A$. Let $M^*(A) = \bigcup_{i=0}^\infty M_i(A)$ be the set of all finite-order sources.

The Laplace and Krichevsky predictors can be extended to general Markov processes. The trick is to view a Markov source $p \in M_m(A)$ as resulting from $|A|^m$ i.i.d. sources. We illustrate this idea by an example [47]. So assume that $A = \{O, I\}$, $m = 2$ and assume that the source $p \in M_2(A)$ has generated the sequence

OOIOIIIOOIIHOIO.

We represent this sequence by the following four subsequences:

*** I ** ** I ** ** ** ,*
*** * O * I ** * I ** * O ,*
*** ** I * * O ** ** * I * ,*
*** ** ** O ** * IO ** .*

These four subsequences contain letters which follow *OO*, *OI*, *IO* and *II*, respectively. By definition, $p \in M_m(A)$ if $p(a|x_t \cdots x_1) = p(a|x_t \cdots x_{t-m+1})$, for all $0 < m \leq t$, all $a \in A$ and all $x_1 \cdots x_t \in A^t$. Therefore, each of the four generated subsequences may be considered to be generated by an i.i.d. source. Further, it is possible to reconstruct the original sequence if we know the four ($= |A|^m$) subsequences and the two ($= m$) first letters of the original sequence.

Any predictor γ for i.i.d. sources can be applied to Markov sources. Indeed, in order to predict, it is enough to store in the memory $|A|^m$ sequences, one corresponding to each word in A^m . Thus, in the example, the letter x_3 which follows *OO* is predicted based on the i.i.d. method γ corresponding to the x_1x_2 - subsequence ($= OO$), then x_4 is predicted based on the i.i.d. method corresponding to x_2x_3 , i.e. to the *OI*- subsequence, and so forth. When this scheme is applied along with either L_0 or K_0 we denote the obtained predictors as L_m and K_m , correspondingly, and define the probabilities for the first m letters as follows: $L_m(x_1) = L_m(x_2) = \dots = L_m(x_m) = 1/|A|$, $K_m(x_1) = K_m(x_2) = \dots = K_m(x_m) = 1/|A|$. For example, having taken into account (16), we can present the Krichevsky predictors for $M_m(A)$ as follows:

$$K_m(x_1 \dots x_t) = \begin{cases} \frac{1}{|A|^t}, & \text{if } t \leq m, \\ \frac{1}{|A|^m}, \prod_{v \in A^m} \frac{\prod_{a \in A} ((\Gamma(\nu_x(va)+1/2) / \Gamma(1/2)))}{(\Gamma(\bar{\nu}_x(v)+|A|/2) / \Gamma(|A|/2))}, & \text{if } t > m, \end{cases} \quad (19)$$

where $\bar{\nu}_x(v) = \sum_{a \in A} \nu_x(va)$, $x = x_1 \dots x_t$. It is worth noting that the representation (14) can be more convenient for carrying out calculations if t is small.

Example. For the word *OOIOIIIOOIIIOIO* considered in the previous example, we obtain

$$K_2(OOIOIIIOOIIIOIO) = 2^{-2} \frac{13}{24} \frac{1113}{2428} \frac{111}{242} \frac{111}{242}.$$

Here groups of multipliers correspond to subsequences *II*, *OIIO*, *IOI*, *OIO*.

In order to estimate the error of the Krichevsky predictor K_m we need a general definition of the Shannon entropy. Let P be a stationary and ergodic source generating letters from a finite alphabet A . The m -order (conditional) Shannon entropy and the limiting Shannon entropy are defined as follows:

$$h_m(P) = \sum_{v \in A^m} P(v) \sum_{a \in A} P(a/v) \log P(a/v), \quad h_\infty(P) = \lim_{m \rightarrow \infty} h_m(P). \quad (20)$$

(If $m = 0$ we obtain the definition (12).) It is also known that for any m

$$h_\infty(P) \leq h_m(P) \quad (21)$$

[5, 14].

Claim 3 *For any stationary and ergodic source P generating letters from a finite alphabet A the average error of the Krichevsky predictor K_m is upper bounded as follows:*

$$-t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log(K_m(x_1 \dots x_t)) - h_m(P) \leq \frac{|A|^m (|A| - 1) \log t + C}{2t}, \quad (22)$$

where C is a constant.

The following so-called empirical Shannon entropy, which is an estimation of the entropy (20), will play a key role in the hypothesis testing. It will be convenient to consider its definition here, because this notation will be used in the proof of the next claims. Let $v = v_1 \dots v_k$ and $x = x_1 x_2 \dots x_t$ be words from A^* . Denote the rate of a word v occurring in the sequence $x = x_1 x_2 \dots x_k$, $x_2 x_3 \dots x_{k+1}$, $x_3 x_4 \dots x_{k+2}$, \dots , $x_{t-k+1} \dots x_t$ as $\nu_x(v)$.

For example, if $x = 000100$ and $v = 00$, then $\nu_x(00) = 3$. For any $0 \leq k < t$ the empirical Shannon entropy of order k is defined as follows:

$$h_k^*(x) = - \sum_{v \in A^k} \frac{\bar{\nu}_x(v)}{(t-k)} \sum_{a \in A} \frac{\nu_x(va)}{\bar{\nu}_x(v)} \log \frac{\nu_x(va)}{\bar{\nu}_x(v)}, \quad (23)$$

where $x = x_1 \dots x_t$, $\bar{\nu}_x(v) = \sum_{a \in A} \nu_x(va)$. In particular, if $k = 0$, we obtain $h_0^*(x) = -t^{-1} \sum_{a \in A} \nu_x(a) \log(\nu_x(a)/t)$.

Let us define the measure R , which, in fact, is a consistent estimator of probabilities for the class of all stationary and ergodic processes with a finite alphabet. First we define a probability distribution $\{\omega = \omega_1, \omega_2, \dots\}$ on integers $\{1, 2, \dots\}$ by

$$\omega_1 = 1 - 1/\log 3, \dots, \omega_i = 1/\log(i+1) - 1/\log(i+2), \dots \quad (24)$$

(In what follows we will use this distribution, but results described below are obviously true for any distribution with nonzero probabilities.) The measure R is defined as follows:

$$R(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_{i+1} K_i(x_1 \dots x_t). \quad (25)$$

It is worth noting that this construction can be applied to the Laplace measure (if we use L_i instead of K_i) and any other family of measures.

Example. Let us calculate $R(00), \dots, R(11)$. From (14) and (25) we obtain:

$$\begin{aligned} K_0(00) &= K_0(11) = \frac{1/2}{1} \frac{3/2}{1+1} = 3/8, \\ K_0(01) &= K_0(10) = \frac{1/2}{1+0} \frac{1/2}{1+1} = 1/8, \\ K_i(00) &= K_i(01) = K_i(10) = K_i(11) = 1/4; \quad , i \geq 1. \end{aligned}$$

Having taken into account the definitions of ω_i (24) and the measure R (25), we can calculate $R(z_1 z_2)$ as follows:

$$\begin{aligned} R(00) &= \omega_1 K_0(00) + \omega_2 K_1(00) + \dots \\ &= (1 - 1/\log 3) 3/8 + (1/\log 3 - 1/\log 4) 1/4 \\ &\quad + (1/\log 4 - 1/\log 5) 1/4 + \dots \\ &= (1 - 1/\log 3) 3/8 + (1/\log 3) 1/4 \approx 0.296. \end{aligned}$$

Analogously, $R(01) = R(10) \approx 0.204$, $R(11) \approx 0.296$.

The main properties of the measure R are connected with the Shannon entropy (20).

Theorem 1 ([36]) *For any stationary and ergodic source P the following equalities are valid:*

$$i) \lim_{t \rightarrow \infty} \frac{1}{t} \log(1/R(x_1 \cdots x_t)) = h_\infty(P)$$

with probability 1,

$$ii) \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{u \in A^t} P(u) \log(1/R(u)) = h_\infty(P).$$

So, if one uses the measure R for data compression in such a way that the codeword length of the sequence $x_1 \cdots x_t$ is (approximately) equal to $\log(1/R(x_1 \cdots x_t))$ bits, he/she obtains the best achievable data compression ratio $h_\infty(P)$ per letter. On the other hand, we know that the redundancy of a universal code and the error of corresponding predictor are equal. Hence, if one uses the measure R for estimation and/or prediction, the error (per letter) will go to zero.

2.3 Hypothesis Testing

Here we briefly describe the main notions of hypothesis testing and the two particular problems considered below. A statistical test is formulated to test a specific null hypothesis (H_0). Associated with this null hypothesis is the alternative hypothesis (H_1) [34]. For example, we will consider the two following problems: goodness-of-fit testing (or identity testing) and testing of serial independence. Both problems are well known in mathematical statistics and there is an extensive literature dealing with their nonparametric testing [2, 8, 9, 12].

The goodness-of-fit testing is described as follows: a hypothesis H_0^{id} is that the source has a particular distribution π and the alternative hypothesis H_1^{id} that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{id} . One particular case, mentioned in Introduction, is when the source alphabet A is $\{0, 1\}$ and the main hypothesis H_0^{id} is that a bit sequence is generated by the Bernoulli i.i.d. source with equal probabilities of 0's and 1's. In all cases, the testing should be based on a sample $x_1 \dots x_t$ generated by the source.

The second problem is as follows: the null hypothesis H_0^{SI} is that the source is Markovian of order not larger than m , ($m \geq 0$), and the alternative hypothesis H_1^{SI} is that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{SI} . In particular, if $m = 0$, this is the problem of testing for independence of time series.

For each applied test, a decision is derived that accepts or rejects the null hypothesis. During the test, a test statistic value is computed on the data (the sequence being tested). This test statistic value is compared to the critical value. If the test statistic value exceeds the critical value, the null hypothesis is rejected. Otherwise, the null hypothesis is accepted. So, statistical hypothesis testing is a conclusion-generation procedure that has two possible outcomes: either accept H_0 or accept H_1 .

Errors of the two following types are possible: The Type I error occurs if H_0 is true but the test accepts H_1 and, vice versa, the Type II error occurs if H_1 is true, but the test accepts H_0 . The probability of Type I error is often called the level of significance of the test. This probability can be set prior to the testing and is denoted α . For a test, α is the probability that the test will say that H_0 is not true when it really is true. Common values of α are about 0.01. The probabilities of Type I and Type II errors are related to each other and to the size n of the tested sequence in such a way that if two of them are specified, the third value is automatically determined. Practitioners usually select a sample size n and a value for the probability of the Type I error - the level of significance [34].

2.4 Codes

We briefly describe the main definitions and properties (without proofs) of lossless codes, or methods of (lossless) data compression. A data compression method (or code) φ is defined as a set of mappings φ_n such that $\varphi_n : A^n \rightarrow \{0,1\}^*$, $n = 1,2,\dots$ and for each pair of different words $x, y \in A^n$ $\varphi_n(x) \neq \varphi_n(y)$. It is also required that each sequence $\varphi_n(u_1)\varphi_n(u_2)\dots\varphi_n(u_r)$, $r \geq 1$, of encoded words from the set A^n , $n \geq 1$, could be uniquely decoded into $u_1u_2\dots u_r$. Such codes are called uniquely decodable. For example, let $A = \{a, b\}$, the code $\psi_1(a) = 0, \psi_1(b) = 00$, obviously, is not uniquely decodable. In what follows we call uniquely decodable codes just "codes". It is well known that if φ is a code then the lengths of the codewords satisfy the following inequality (Kraft's inequality) [14] : $\sum_{u \in A^n} 2^{-|\varphi_n(u)|} \leq 1$. It will be convenient to reformulate this

property as follows:

Claim 4 *Let φ be a code over an alphabet A . Then for any integer n there exists a measure μ_φ on A^n such that*

$$-\log \mu_\varphi(u) \leq |\varphi(u)| \quad (26)$$

for any u from A^n .

(Obviously, this claim is true for the measure $\mu_\varphi(u) = \frac{2^{-|\varphi(u)|}}{\sum_{u \in A^n} 2^{-|\varphi(u)|}}$).

It was mentioned above that, in a certain sense, the opposite claim is true, too. Namely, for any probability measure μ defined on $A^n, n \geq 1$, there exists a code φ_μ such that

$$|\varphi_\mu(u)| = -\log \mu(u). \quad (27)$$

(More precisely, for any $\varepsilon > 0$ one can construct such a code φ_μ^* , that $|\varphi_\mu^*(u)| < -\log \mu(u) + \varepsilon$ for any $u \in A^n$. Such a code can be constructed by applying a so-called arithmetic coding [32].) For example, for the above described measure R we can construct a code R_{code} such that

$$|R_{code}(u)| = -\log R(u). \quad (28)$$

As we mentioned above there exist universal codes. For their description we recall that sequences $x_1 \dots x_t$, generated by a source P , can be "compressed" to the length $-\log P(x_1 \dots x_t)$ bits (see (27)) and, on the other hand, for any source P there is no code ψ for which the average codeword length ($\sum_{u \in A^t} P(u) |\psi(u)|$) is less than $-\sum_{u \in A^t} P(u) \log P(u)$. Universal codes can reach the lower bound $-\log P(x_1 \dots x_t)$ asymptotically for any stationary and ergodic source P in average and with probability 1. The formal definition is as follows: a code U is universal if for any stationary and ergodic source P the following equalities are valid:

$$\lim_{t \rightarrow \infty} |U(x_1 \dots x_t)|/t = h_\infty(P) \quad (29)$$

with probability 1, and

$$\lim_{t \rightarrow \infty} E(|U(x_1 \dots x_t)|)/t = h_\infty(P), \quad (30)$$

where $E(f)$ is the expected value of f , $h_\infty(P)$ is the Shannon entropy of P , see (21). So, informally speaking, a universal code estimates the probability characteristics of a source and uses them for efficient "compression".

In this chapter we mainly consider finite-alphabet and real-valued sources, but sources with countable alphabet also were considered by many authors [4, 17, 19, 41, 42]. In particular, it is shown that, for infinite alphabet, without any condition on the source distribution it is impossible to have universal source code and/or universal predictor, i.e. such a predictor whose average error goes to zero, when the length of a sequence goes to infinity. On the other hand, there are some necessary and sufficient conditions for existence of universal codes and predictors [4, 19, 41].

3 Finite Alphabet Processes

3.1 The Estimation of (Limiting) Probabilities

The following theorem shows how universal codes can be applied for probability estimation.

Theorem 2 *Let U be a universal code and*

$$\mu_U(u) = 2^{-|U(u)|} / \sum_{v \in A^{|u|}} 2^{-|U(v)|}. \quad (31)$$

Then, for any stationary and ergodic source P the following equalities are valid:

$$i) \lim_{t \rightarrow \infty} \frac{1}{t} (-\log P(x_1 \cdots x_t) - (-\log \mu_U(x_1 \cdots x_t))) = 0$$

with probability 1,

$$ii) \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{u \in A^t} P(u) \log(P(u)/\mu_U(u)) = 0.$$

The informal outline of the proof is as follows: $\frac{1}{t}(-\log P(x_1 \cdots x_t))$ and $\frac{1}{t}(-\log \mu_U(x_1 \cdots x_t))$ goes to the Shannon entropy $h_\infty(P)$, that is why the difference is 0.

So, we can see that, in a certain sense, the measure μ_U is a consistent nonparametric estimation of the (unknown) measure P .

Nowadays there are many efficient universal codes (and universal predictors connected with them), which can be applied to estimation. For example, the above described measure R is based on a universal code [35, 36] and can be applied for probability estimation. More precisely, Theorem 2 (and the following theorems) are true for R , if we replace μ_U by R .

It is important to note that the measure R has some additional properties, which can be useful for applications. The following theorem describes these properties (whereas all other theorems are valid for all universal codes and corresponding measures, including the measure R).

Theorem 3 ([35, 36]) *For any Markov process P with memory k*

i) the error of the probability estimator, which is based on the measure R , is upper-bounded as follows:

$$\frac{1}{t} \sum_{u \in A^t} P(u) \log(P(u)/R(u)) \leq \frac{(|A| - 1)|A|^k \log t}{2t} + O\left(\frac{1}{t}\right),$$

ii) the error of R is asymptotically minimal in the following sense: for any measure μ there exists a k -memory Markov process p_μ such that

$$\frac{1}{t} \sum_{u \in A^t} p_\mu(u) \log(p_\mu(u)/\mu(u)) \geq \frac{(|A| - 1)|A|^k \log t}{2t} + O\left(\frac{1}{t}\right),$$

iii) Let Θ be a set of stationary and ergodic processes such that there exists a measure μ_Θ for which the estimation error of the probability goes to 0 uniformly:

$$\lim_{t \rightarrow \infty} \sup_{P \in \Theta} \left(\frac{1}{t} \sum_{u \in A^t} P(u) \log(P(u)/\mu_\Theta(u)) \right) = 0.$$

Then the error of the estimator which is based on the measure R , goes to 0 uniformly too:

$$\lim_{t \rightarrow \infty} \sup_{P \in \Theta} \left(\frac{1}{t} \sum_{u \in A^t} P(u) \log(P(u)/R(u)) \right) = 0.$$

3.2 Prediction

As we mentioned above, any universal code U can be applied for prediction. Namely, the measure μ_U (31) can be used for prediction as the following conditional probability:

$$\mu_U(x_{t+1}|x_1 \dots x_t) = \mu_U(x_1 \dots x_t x_{t+1}) / \mu_U(x_1 \dots x_t). \quad (32)$$

The following theorem shows that such a predictor is quite reasonable. Moreover, it gives a possibility to apply practically used data compressors for prediction of real data (like EUR/USD rate) and obtain quite precise estimation [43] .

Theorem 4 *Let U be a universal code and P be any stationary and ergodic process. Then*

$$i) \lim_{t \rightarrow \infty} \frac{1}{t} E \left\{ \log \frac{P(x_1)}{\mu_U(x_1)} + \log \frac{P(x_2|x_1)}{\mu_U(x_2|x_1)} + \dots + \log \frac{P(x_t|x_1 \dots x_{t-1})}{\mu_U(x_t|x_1 \dots x_{t-1})} \right\} = 0,$$

$$ii) \lim_{t \rightarrow \infty} E \left(\frac{1}{t} \sum_{i=0}^{t-1} (P(x_{i+1}|x_1 \dots x_i) - \mu_U(x_{i+1}|x_1 \dots x_i))^2 \right) = 0,$$

and

$$iii) \lim_{t \rightarrow \infty} E \left(\frac{1}{t} \sum_{i=0}^{t-1} |P(x_{i+1}|x_1 \dots x_i) - \mu_U(x_{i+1}|x_1 \dots x_i)| \right) = 0.$$

An informal outline of the proof is as follows:

$$\frac{1}{t} \left\{ E \left(\log \frac{P(x_1)}{\mu_U(x_1)} \right) + E \left(\log \frac{P(x_2|x_1)}{\mu_U(x_2|x_1)} \right) + \dots + E \left(\log \frac{P(x_t|x_1 \dots x_{t-1})}{\mu_U(x_t|x_1 \dots x_{t-1})} \right) \right\}$$

is equal to $\frac{1}{t} E \left(\log \frac{P(x_1 \dots x_t)}{\mu_U(x_1 \dots x_t)} \right)$. Taking into account Theorem 2, we obtain the first statement of the theorem.

Comment 1. The measure R described above has one additional property if it is used for prediction. Namely, for any Markov process P ($P \in M^*(A)$) the following is true:

$$\lim_{t \rightarrow \infty} \log \frac{P(x_{t+1}|x_1 \dots x_t)}{R(x_{t+1}|x_1 \dots x_t)} = 0$$

with probability 1, where $R(x_{t+1}|x_1 \dots x_t) = R(x_1 \dots x_t x_{t+1}) / R(x_1 \dots x_t)$ [37].

Comment 2. It is known [48] that, in fact, the statements ii) and iii) are equivalent.

3.3 Problems with Side Information

Now we consider the so-called problems with side information, which are described as follows: there is a stationary and ergodic source whose alphabet A is presented as a product $A = X \times Y$. We are given a sequence

$(x_1, y_1), \dots, (x_{t-1}, y_{t-1})$ and side information y_t . The goal is to predict, or estimate, x_t . This problem arises in statistical decision theory, pattern recognition, and machine learning. Obviously, if someone knows the conditional probabilities $P(x_t | (x_1, y_1), \dots, (x_{t-1}, y_{t-1}), y_t)$ for all $x_t \in X$, he has all information about x_t , available before x_t is known. That is why we will look for the best (or, at least, good) estimations for this conditional probabilities. Our solution will be based on results obtained in the previous subsection. More precisely, for any universal code U and the corresponding measure μ_U (31) we define the following estimate for the problem with side information:

$$\mu_U(x_t | (x_1, y_1), \dots, (x_{t-1}, y_{t-1}), y_t) = \frac{\mu_U((x_1, y_1), \dots, (x_{t-1}, y_{t-1}), (x_t, y_t))}{\sum_{x_t \in X} \mu_U((x_1, y_1), \dots, (x_{t-1}, y_{t-1}), (x_t, y_t))}.$$

The following theorem shows that this estimate is quite reasonable.

Theorem 5 *Let U be a universal code and let P be any stationary and ergodic process. Then*

$$\begin{aligned} i) \quad & \lim_{t \rightarrow \infty} \frac{1}{t} \left\{ E \left(\log \frac{P(x_1 | y_1)}{\mu_U(x_1 | y_1)} \right) + E \left(\log \frac{P(x_2 | (x_1, y_1), y_2)}{\mu_U(x_2 | (x_1, y_1), y_2)} \right) + \dots \right. \\ & \left. + E \left(\log \frac{P(x_t | (x_1, y_1), \dots, (x_{t-1}, y_{t-1}), y_t)}{\mu_U(x_t | (x_1, y_1), \dots, (x_{t-1}, y_{t-1}), y_t)} \right) \right\} = 0, \\ ii) \quad & \lim_{t \rightarrow \infty} E \left(\frac{1}{t} \sum_{i=0}^{t-1} (P(x_{i+1} | (x_1, y_1), \dots, (x_i, y_i), y_{i+1})) - \right. \\ & \left. \mu_U(x_{i+1} | (x_1, y_1), \dots, (x_i, y_i), y_{i+1}))^2 \right) = 0, \end{aligned}$$

and

$$\begin{aligned} iii) \quad & \lim_{t \rightarrow \infty} E \left(\frac{1}{t} \sum_{i=0}^{t-1} |P(x_{i+1} | (x_1, y_1), \dots, (x_i, y_i), y_{i+1})) - \right. \\ & \left. \mu_U(x_{i+1} | (x_1, y_1), \dots, (x_i, y_i), y_{i+1})) \right) = 0. \end{aligned}$$

The proof is very close to the proof of the previous theorem.

3.4 The Case of Several Independent Samples

In this part we consider a situation which is important for practical applications, but needs cumbersome notations. Namely, we extend our consideration to the case where the sample is presented as several independent samples $x^1 = x_1^1 \dots x_{t_1}^1$, $x^2 = x_1^2 \dots x_{t_2}^2, \dots$, $x^r = x_1^r \dots x_{t_r}^r$ generated by a source. More precisely, we will suppose that all sequences were independently created by one stationary and ergodic source. (The point is that it is impossible just to combine all samples into one, if the source is not i.i.d.) We denote them by $x^1 \diamond x^2 \diamond \dots \diamond x^r$ and define $\nu_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(v) = \sum_{i=1}^r \nu_{x^i}(v)$. For example, if $x^1 = 0010$, $x^2 = 011$, then $\nu_{x^1 \diamond x^2}(00) = 1$. The definition of K_m and R can be extended to this case:

$$K_m(x^1 \diamond x^2 \diamond \dots \diamond x^r) = \quad (33)$$

$$\left(\prod_{i=1}^r |A|^{-\min\{m, t_i\}} \right) \prod_{v \in A^m} \frac{\prod_{a \in A} ((\Gamma(\nu_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(va) + 1/2) / \Gamma(1/2))}{(\Gamma(\bar{\nu}_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(v) + |A|/2) / \Gamma(|A|/2))},$$

whereas the definition of R is the same (see (25)). (Here, as before, $\bar{\nu}_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(v) = \sum_{a \in A} \nu_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(va)$. Note, that $\bar{\nu}_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(\cdot) = \sum_{i=1}^r t_i$ if $m = 0$.)

The following example is intended to show the difference between the case of many samples and one.

Example. Let there be two independent samples $y = y_1 \dots y_4 = 0101$ and $x = x_1 \dots x_3 = 101$, generated by a stationary and ergodic source with the alphabet $\{0, 1\}$. One wants to estimate the (limiting) probabilities $P(z_1 z_2)$, $z_1, z_2 \in \{0, 1\}$ (here $z_1 z_2 \dots$ can be considered as an independent sequence, generated by the source) and predict $x_4 x_5$ (i.e. estimate conditional probability $P(x_4 x_5 | x_1 \dots x_3 = 101, y_1 \dots y_4 = 0101)$). For solving both problems we will use the measure R (see (25)). First we consider the case where $P(z_1 z_2)$ is to be estimated without knowledge of sequences x and y . Those probabilities were calculated previously and we obtained: $R(00) \approx 0.296$, $R(01) = R(10) \approx 0.204$, $R(11) \approx 0.296$. Let us now estimate the probability $P(z_1 z_2)$ taking into account that there are two independent samples $y = y_1 \dots y_4 = 0101$ and $x = x_1 \dots x_3 = 101$. First of all we note that such estimates are based on the formula for conditional probabilities:

$$R(z|x \diamond y) = R(x \diamond y \diamond z) / R(x \diamond y).$$

Then we estimate the frequencies: $\nu_{0101 \diamond 101}(0) = 3$, $\nu_{0101 \diamond 101}(1) = 4$, $\nu_{0101 \diamond 101}(00) = \nu_{0101 \diamond 101}(11) = 0$, $\nu_{0101 \diamond 101}(01) = 3$, $\nu_{0101 \diamond 101}(10) = 2$,

$\nu_{0101 \diamond 101}(010) = 1$, $\nu_{0101 \diamond 101}(101) = 2$, $\nu_{0101 \diamond 101}(0101) = 1$, whereas frequencies of all other three-letters and four-letters words are 0. Then we calculate :

$$K_0(0101 \diamond 101) = \frac{1}{2} \frac{3}{4} \frac{5}{8} \frac{1}{10} \frac{3}{12} \frac{5}{14} \approx 0.00244,$$

$$K_1(0101 \diamond 101) = (2^{-1})^2 \frac{1}{2} \frac{3}{4} \frac{5}{6} \frac{1}{2} \frac{3}{4} \frac{1}{1}$$

$$\approx 0.0293, \quad K_2(0101 \diamond 101) \approx 0.01172, \quad K_i(0101 \diamond 101) = 2^{-7}, \quad i \geq 3,$$

$$R(0101 \diamond 101) = \omega_1 K_0(0101 \diamond 101) + \omega_2 K_1(0101 \diamond 101) + \dots \approx$$

$$0.369 \cdot 0.00244 + 0.131 \cdot 0.0293 + 0.06932 \cdot 0.01172 + 2^{-7} / \log 5 \approx 0.0089.$$

In order to avoid repetitions, we estimate only one probability $P(z_1 z_2 = 01)$. Carrying out similar calculations, we obtain $R(0101 \diamond 101 \diamond 01) \approx 0.00292$, $R(z_1 z_2 = 01 | y_1 \dots y_4 = 0101, x_1 \dots x_3 = 101) = R(0101 \diamond 101 \diamond 01) / R(0101 \diamond 101) \approx 0.32812$. If we compare this value and the estimation $R(01) \approx 0.204$, which is not based on the knowledge of samples x and y , we can see that the measure R uses additional information quite naturally (indeed, 01 is quite frequent in $y = y_1 \dots y_4 = 0101$ and $x = x_1 \dots x_3 = 101$).

Such generalization can be applied to many universal codes, but, generally speaking, there exist codes U for which $U(x^1 \diamond x^2)$ is not defined and, hence, the measure $\mu_U(x^1 \diamond x^2)$ is not defined. That is why we will describe properties of the universal code R , but not of universal codes in general. For the measure R all asymptotic properties are the same for the cases of one sample and several samples. More precisely, the following statement is true:

Claim 5 *Let x^1, x^2, \dots, x^r be independent sequences generated by a stationary and ergodic source and let t be a total length of these sequences ($t = \sum_{i=1}^r |x^i|$). Then, if $t \rightarrow \infty$, (and r is fixed) the statements of the Theorems 2 - 5 are valid, when applied to $x^1 \diamond x^2 \diamond \dots \diamond x^r$ instead of $x_1 \dots x_t$. (In theorems 2 - 5 μ_U should be changed to R .)*

The proofs are completely analogous to the proofs of the Theorems 2—5.

Now we can extend the definition of the empirical Shannon entropy (23) to the case of several words $x^1 = x_1^1 \dots x_{t_1}^1$, $x^2 = x_1^2 \dots x_{t_2}^2, \dots, x^r =$

$x_1^r \dots x_{t_r}^r$. We define $\nu_{x^1 \diamond x^2 \diamond \dots \diamond x^r}(v) = \sum_{i=1}^r \nu_{x^i}(v)$. For example, if $x^1 = 0010, x^2 = 011$, then $\nu_{x^1 \diamond x^2}(00) = 1$. Analogously to (23),

$$h_k^*(x^1 \diamond x^2 \diamond \dots \diamond x^r) = - \sum_{v \in A^k} \frac{\bar{\nu}_{x^1 \diamond \dots \diamond x^r}(v)}{(t - kr)} \sum_{a \in A} \frac{\nu_{x^1 \diamond \dots \diamond x^r}(va)}{\bar{\nu}_{x^1 \diamond \dots \diamond x^r}(v)} \log \frac{\nu_{x^1 \diamond \dots \diamond x^r}(va)}{\bar{\nu}_{x^1 \diamond \dots \diamond x^r}(v)}, \quad (34)$$

where $\bar{\nu}_{x^1 \diamond \dots \diamond x^r}(v) = \sum_{a \in A} \nu_{x^1 \diamond \dots \diamond x^r}(va)$.

For any sequence of words $x^1 = x_1^1 \dots x_{t_1}^1, x^2 = x_1^2 \dots x_{t_2}^2, \dots, x^r = x_1^r \dots x_{t_r}^r$ from A^* and any measure θ we define $\theta(x^1 \diamond x^2 \diamond \dots \diamond x^r) = \prod_{i=1}^r \theta(x^i)$. The following lemma gives an upper bound for unknown probabilities.

Lemma 1 *Let θ be a measure from $M_m(A), m \geq 0$, and x^1, \dots, x^r be words from A^* , whose lengths are not less than m . Then*

$$\theta(x^1 \diamond \dots \diamond x^r) \leq 2^{-(t-rm)h_m^*(x^1 \diamond \dots \diamond x^t)}, \quad (35)$$

where $\theta(x^1 \diamond \dots \diamond x^r) = \prod_{i=1}^r \theta(x^i)$.

4 Hypothesis Testing

4.1 Goodness-of-Fit or Identity Testing

Now we consider the problem of testing H_0^{id} against H_1^{id} . Let us recall that the hypothesis H_0^{id} is that the source has a particular distribution π and the alternative hypothesis H_1^{id} that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{id} . Let the required level of significance (or the Type I error) be $\alpha, \alpha \in (0, 1)$. We describe a statistical test which can be constructed based on any code φ .

The main idea of the suggested test is quite natural: compress a sample sequence $x_1 \dots x_t$ by a code φ . If the length of the codeword ($|\varphi(x_1 \dots x_t)|$) is significantly less than the value $-\log \pi(x_1 \dots x_t)$, then H_0^{id} should be rejected. The key observation is that the probability of all rejected sequences is quite small for any φ , that is why the Type I error can be made small. The precise description of the test is as follows: *The hypothesis H_0^{id} is accepted if*

$$-\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| \leq -\log \alpha. \quad (36)$$

Otherwise, H_0^{id} is rejected. We denote this test by $T_\varphi^{id}(A, \alpha)$.

Theorem 6 *i) For each distribution $\pi, \alpha \in (0, 1)$ and a code φ , the Type I error of the described test $T_\varphi^{id}(A, \alpha)$ is not larger than α and ii) if, in addition, π is a finite-order stationary and ergodic process over A^∞ (i.e. $\pi \in M^*(A)$) and φ is a universal code, then the Type II error of the test $T_\varphi^{id}(A, \alpha)$ goes to 0, when t tends to infinity.*

4.2 Testing for Serial Independence

Let us recall that the null hypothesis H_0^{SI} is that the source is Markovian of order not larger than m , ($m \geq 0$), and the alternative hypothesis H_1^{SI} is that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{SI} . In particular, if $m = 0$, this is the problem of testing for independence of time series.

Let there be given a sample $x_1 \dots x_t$ generated by an (unknown) source π . The main hypothesis H_0^{SI} is that the source π is Markovian whose order is not greater than m , ($m \geq 0$), and the alternative hypothesis H_1^{SI} is that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{SI} . The described test is as follows.

Let φ be any code. By definition, the hypothesis H_0^{SI} is accepted if

$$(t - m) h_m^*(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| \leq \log(1/\alpha), \quad (37)$$

where $\alpha \in (0, 1)$. Otherwise, H_0^{SI} is rejected. We denote this test by $T_\varphi^{SI}(A, \alpha)$.

Theorem 7 *i) For any code φ the Type I error of the test $T_\varphi^{SI}(A, \alpha)$ is less than or equal to $\alpha, \alpha \in (0, 1)$ and, ii) if, in addition, φ is a universal code, then the Type II error of the test $T_\varphi^{SI}(A, \alpha)$ goes to 0, when t tends to infinity.*

5 Examples of hypothesis testing

In this part we describe results of some experiments and a simulation study carried out to estimate an efficiency of the suggested tests. The obtained results show that the described tests as well as the suggested approach in general can be used in applications.

5.1 Randomness testing

First we consider the problem of randomness testing, which is a particular case of goodness-of-fit testing. Namely, we will consider a null hypothesis

H_0^{rt} that a given bit sequence is generated by Bernoulli source with equal probabilities of 0 and 1 and the alternative hypothesis H_1^{rt} that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{rt} . This problem is important for random number (RNG) and pseudorandom number generators (PRNG) testing and there are many methods for randomness testing suggested in literature. Thus, National Institute of Standards and Technology (NIST, USA) suggested "A statistical test suite for random and pseudorandom number generators for cryptographic applications", see [34].

We investigated linear congruent generators (LCG), which are defined by the following equality

$$X_{n+1} = (A * X_n + C) \text{ mod } M,$$

where X_n is the n -th generated number [22]. Each such generator we will denote by $LCG(M, A, C, X_0)$, where X_0 is the initial value of the generator. Such generators are well studied and many of them are used in practice, see [23].

In our experiments we extract an eight-bit word from each generated X_i using the following algorithm. Firstly, the number $\mu = \lfloor M/256 \rfloor$ was calculated and then each X_i was transformed into an 8-bit word \hat{X}_i as follows:

$$\left. \begin{aligned} \hat{X}_i &= \lfloor X_i/256 \rfloor \text{ if } X_i < 256\mu \\ \hat{X}_i &= \text{empty word if } X_i \geq 256\mu \end{aligned} \right\} \quad (38)$$

Then a sequence was compressed by the archiver *ACE v 1.2b* (see <http://www.winace.com/>).

Experimental data about testing of three linear congruent generators is given in the table 1.

Table 1: Pseudorandom number generators testing.

parameters / length (bits)	400 000	8 000 000
M,A,C, X_0		
$10^8 + 1, 23, 0, 47594118$	390 240	7635936
$2^{31}, 2^{16} + 3, 0, 1$	extended	7797984
$2^{32}, 134775813, 1, 0$	extended	extended

So, we can see from the first line of the table that the 400000–bit sequence generated by the LCG($10^8+1, 23, 0, 47594118$) and transformed according to (38), was compressed to a 390240–bit sequence. (Here 400000 is the length of the sequence after transformation.) If we take the level of significance $\alpha \geq 2^{-9760}$ and apply the test $T_\varphi^{id}(\{0, 1\}, \alpha)$, ($\varphi = ACE$ v 1.2b), the hypothesis H_0^{rt} should be rejected, see Theorem 1 and (36). Analogously, the second line of the table shows that the 8000000–bit sequence generated by LCG($2^{31}, 2^{16} + 3, 0, 1$) cannot be considered random (H_0^{rt} should be rejected if the level of significance α is greater than $2^{-202016}$). On the other hand, the suggested test accepts H_0^{rt} for the sequences generated by the third generator, because the lengths of the “compressed” sequences increased.

The obtained information corresponds to the known data about the considered generators. Thus, it is shown in [23] that the first two generators are bad whereas the third generator was investigated in [29] and is regarded as good. So, we can see that the suggested testing is quite efficient.

In [44] the described method was applied for testing random number and pseudorandom number generators and its efficiency was compared with the mentioned methods from “A statistical test suite for random and pseudorandom number generators for cryptographic applications” [34]. The point is that the tests from [34] are selected basing on comprehensive theoretical and experimental analysis and can be considered as the state-of-the-art in randomness testing. It turned out that the suggested tests, which were based on archivers RAR and ARJ, were more powerful than many methods recommended by NIST in [34]; see [44] for details.

5.2 Simulation study of serial independence testing

A selection of the simulation results concerning independence tests is presented in this part. We generated binary sequences by the first order Markov source with different probabilities (see table 2 below) and applied the test $T_\varphi^{SI}(\{0, 1\}, \alpha)$ to test the hypothesis H_0^{SI} that a given bit sequence is generated by Bernoulli source and the alternative hypothesis H_1^{SI} that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{SI} .

We tried several different archivers and the universal code R described in Appendix 2. It turned out that the power of the code R is larger than the power of the tried archivers, that is why we present results for the

test $T_R^{SI}(\{0, 1\}, \alpha)$, which is based on this code, for $\alpha = 0.01$. The table 2 contains results of calculations.

Table 2: Serial independence testing for Markov source of order 6 ("rej" means rejected, "acc" - accepted. In all cases $p(x_{i+1} = 0|x_i = 1) = 0.5$)

probabilities / length (bits)	2^9	2^{14}	2^{16}	2^{18}	2^{23}
$p(x_{i+1} = 0 x_i = 0) = 0.8$	rej	rej	rej	rej	rej
$p(x_{i+1} = 0 x_i = 0) = 0.6$	acc	rej	rej	rej	rej
$p(x_{i+1} = 0 x_i = 0) = 0.55$	acc	acc	rej	rej	rej
$p(x_{i+1} = 0 x_i = 0) = 0.525$	acc	acc	acc	rej	rej
$p(x_{i+1} = 0 x_i = 0) = 0.505$	acc	acc	acc	acc	rej

We know that the source is Markovian and, hence, the hypothesis H_0^{SI} (that a sequence is generated by Bernoulli source) is not true. The table shows how the value of the Type II error depends on the sample size and the source probabilities.

The similar calculations were carried out for the Markov source of order 6. We applied the test $T_\varphi^{SI}(\{0, 1\}, \alpha)$, $\alpha = 0.01$, for checking the hypothesis H_0^{SI} that a given bit sequence is generated by Markov source of order at most 5 and the alternative hypothesis H_1^{SI} that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{SI} . Again, we know that H_0^{SI} is not true and the table 3 shows how the value of the Type II error depends on the sample size and the source probabilities.

Table 3: Serial independence testing for Markov source of order 6. In all cases $p(x_{i+1} = 0 | (\sum_{j=i-6}^i x_j) \bmod 2 = 1) = 0.5$.

probabilities/length (bits)	2^{14}	2^{18}	2^{20}	2^{23}	2^{28}
$p(x_{i+1} = 0 (\sum_{j=i-6}^i x_j) \bmod 2 = 0) = 0.8$	rej	rej	rej	rej	rej
$p(x_{i+1} = 0 (\sum_{j=i-6}^i x_j) \bmod 2 = 0) = 0.6$	acc	rej	rej	rej	rej
$p(x_{i+1} = 0 (\sum_{j=i-6}^i x_j) \bmod 2 = 0) = 0.55$	acc	acc	rej	rej	rej
$p(x_{i+1} = 0 (\sum_{j=i-6}^i x_j) \bmod 2 = 0) = 0.525$	acc	acc	acc	rej	rej
$p(x_{i+1} = 0 (\sum_{j=i-6}^i x_j) \bmod 2 = 0) = 0.505$	acc	acc	acc	acc	rej

6 Real-Valued Time Series

6.1 Density Estimation and Its Application

Here we address the problem of nonparametric estimation of the density for time series. Let X_t be a time series and the probability distribution of X_t is unknown, but it is known that the time series is stationary and ergodic. We have seen that Shannon-MacMillan-Breiman theorem played a key role in the case of finite-alphabet processes. In this part we will use its generalization to the processes with densities, which was established by Barron [3]. First we describe considered processes with some properties needed for the generalized Shannon-MacMillan-Breiman theorem to hold. In what follows, we restrict our attention to processes that take bounded real valued. However, the main results may be extended to processes taking values in a compact subset of a separable metric space.

Let B denote the Borel subsets of \mathbb{R} , and B^k denote the Borel subsets of \mathbb{R}^k , where \mathbb{R} is the set of real numbers. Let \mathbb{R}^∞ be the set of all infinite sequences $x = x_1, x_2 \dots$ with $x_i \in \mathbb{R}$, and let B^∞ denote the usual product sigma field on \mathbb{R}^∞ , generated by the finite dimensional cylinder sets $\{A_1, \dots, A_k, \mathbb{R}, \mathbb{R}, \dots\}$, where $A_i \in B, i = 1, \dots, k$. Each stochastic process $X_1, X_2, \dots, X_i \in \mathbb{R}$, is defined by a probability distribution on $(\mathbb{R}^\infty, B^\infty)$. Suppose that the joint distribution P_n for (X_1, X_2, \dots, X_n) has a probability density function $p(x_1 x_2 \dots x_n)$ with respect to a sigma-finite measure M_n . Assume that the sequence of dominating measures M_n is Markov of order $m \geq 0$ with a stationary transition measure. A familiar case for M_n is Lebesgue measure. Let $p(x_{n+1}|x_1 \dots x_n)$ denote the conditional density given by the ratio $p(x_1 \dots x_{n+1}) / p(x_1 \dots x_n)$ for $n > 1$. It is known that for stationary and ergodic processes there exists a so-called relative entropy rate \tilde{h} defined by

$$\tilde{h} = \lim_{n \rightarrow \infty} -E(\log p(x_{n+1}|x_1 \dots x_n)), \quad (39)$$

where E denotes expectation with respect to P . We will use the following generalization of the Shannon-MacMillan-Breiman theorem:

Claim 6 ([3]) *If $\{X_n\}$ is a P -stationary ergodic process with density $p(x_1 \dots x_n) = dP_n/dM_n$ and $\tilde{h}_n < \infty$ for some $n \geq m$, the sequence of relative entropy densities $-(1/n) \log p(x_1 \dots x_n)$ convergence almost surely*

to the relative entropy rate, i.e.,

$$\lim_{n \rightarrow \infty} (-1/n) \log p(x_1 \dots x_n) = \tilde{h} \quad (40)$$

with probability 1 (according to P).

Now we return to the estimation problems. Let $\{\Pi_n\}, n \geq 1$, be an increasing sequence of finite partitions of \mathbb{R} that asymptotically generates the Borel sigma-field B and let $x^{[k]}$ denote the element of Π_k that contains the point x . (Informally, $x^{[k]}$ is obtained by quantizing x to k bits of precision.) For integers s and n we define the following approximation of the density

$$p^s(x_1 \dots x_n) = P(x_1^{[s]} \dots x_n^{[s]})/M_n(x_1^{[s]} \dots x_n^{[s]}). \quad (41)$$

We also consider

$$\tilde{h}_s = \lim_{n \rightarrow \infty} -E(\log p^s(x_{n+1}|x_1 \dots x_n)). \quad (42)$$

Applying the claim 2 to the density $p^s(x_1 \dots x_t)$, we obtain that a.s.

$$\lim_{t \rightarrow \infty} -\frac{1}{t} \log p^s(x_1 \dots x_t) = \tilde{h}_s. \quad (43)$$

Let U be a universal code, which is defined for any finite alphabet. In order to describe a density estimate we will use the probability distribution $\omega_i, i = 1, 2, \dots$, see (24) (In what follows we will use this distribution, but results described below are obviously true for any distribution with nonzero probabilities.) Now we can define the density estimate r_U as follows:

$$r_U(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_i \mu_U(x_1^{[i]} \dots x_t^{[i]})/M_t(x_1^{[i]} \dots x_t^{[i]}), \quad (44)$$

where the measure μ_U is defined by (31). (It is assumed here that the code $U(x_1^{[i]} \dots x_t^{[i]})$ is defined for the alphabet, which contains $|\Pi_i|$ letters.)

It turns out that, in a certain sense, the density $r_U(x_1 \dots x_t)$ estimates the unknown density $p(x_1 \dots x_t)$.

Theorem 8 Let X_t be a stationary ergodic process with densities $p(x_1 \dots x_t) = dP_t/dM_t$ such that

$$\lim_{s \rightarrow \infty} \tilde{h}_s = \tilde{h} < \infty, \quad (45)$$

where \tilde{h} and \tilde{h}_s are relative entropy rates, see (68), (42). Then

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} = 0 \quad (46)$$

with probability 1 and

$$\lim_{t \rightarrow \infty} \frac{1}{t} E(\log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)}) = 0. \quad (47)$$

We have seen that the requirement (45) plays an important role in the proof. The natural question is whether there exist processes for which (45) is valid. The answer is positive. For example, let a process possess values in the interval $[-1, 1]$, M_n be Lebesgue measure and the considered process is Markovian with conditional density

$$p(x|y) = \begin{cases} 1/2 + \alpha \operatorname{sign}(y), & \text{if } x < 0, \\ 1/2 - \alpha \operatorname{sign}(y), & \text{if } x \geq 0, \end{cases}$$

where $\alpha \in (0, 1/2)$ is a parameter and

$$\operatorname{sign}(y) = \begin{cases} -1, & \text{if } y < 0, \\ 1, & \text{if } y \geq 0. \end{cases}$$

In words, the density depends on a sign of the previous value. If the value is positive, then the density is more than $1/2$, otherwise it is less than $1/2$. It is easy to see that (45) is true for any $\alpha \in (0, 1)$.

The following two theorems are devoted to the conditional probability $r_U(x|x_1 \dots x_m) = r_U(x_1 \dots x_m x) / r_U(x_1 \dots x_m)$ which, in turn, is connected with the prediction problem. We will see that the conditional density $r_U(x|x_1 \dots x_m)$ is a reasonable estimation of the unknown density $p(x|x_1 \dots x_m)$.

Theorem 9 *Let B_1, B_2, \dots be a sequence of measurable sets. Then the following equalities are true:*

$$i) \lim_{t \rightarrow \infty} E\left(\frac{1}{t} \sum_{m=0}^{t-1} (P(x_{m+1} \in B_{m+1} | x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1} | x_1 \dots x_m))^2\right) = 0, \quad (48)$$

$$ii) E\left(\frac{1}{t} \sum_{m=0}^{t-1} |P(x_{m+1} \in B_{m+1} | x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1} | x_1 \dots x_m)|\right) = 0,$$

where $R_U(x_{m+1} \in B_{m+1} | x_1 \dots x_m) = \int_{B_{m+1}} r_U(x|x_1 \dots x_m) dM_{1/m}$

We have seen that in a certain sense the estimation r_U approximates the unknown density p . The following theorem shows that r_U can be used instead of p for estimation of average values of certain functions.

Theorem 10 Let f be an integrable function, whose absolute value is bounded by a certain constant \bar{M} and all conditions of the theorem 2 are true. Then the following equality is valid:

$$\begin{aligned}
 i) \quad \lim_{t \rightarrow \infty} \frac{1}{t} E \left(\sum_{m=0}^{t-1} \left(\int f(x) p(x|x_1 \dots x_m) dM_m - \int f(x) r_U(x|x_1 \dots x_m) dM_m \right)^2 \right) &= 0, \\
 ii) \quad \lim_{t \rightarrow \infty} \frac{1}{t} E \left(\sum_{m=0}^{t-1} \left| \int f(x) p(x|x_1 \dots x_m) dM_m - \int f(x) r_U(x|x_1 \dots x_m) dM_m \right| \right) &= 0.
 \end{aligned}
 \tag{49}$$

6.2 Example: predicting the exchange rate

The considered example was describe in the paper of Ryabko and Monarev [43]. The problem of predicting the US dollar exchange rate to ruble and euro using daily data on the value of one dollar in rubles and euros was considered. There were numerous experiments with various values of parameters, methods of trend elimination, and archivers used for prediction.

As data compression algorithms, we used the commonly known *Rar*, *arj*, *pkzip*, and *ha* archivers. The suggested method can be used in combination with other approaches and methods used in forecasting. Among such methods, considered below in forecasting currency exchange rates, there are trend elimination and using for the prediction of the next value not all available dynamic series but only its last part, say, the last 1000 or 50 values, which is often referred to as a window, or sliding window. (When using this scheme, it is assumed that statistical characteristics of the process may vary in time, and old data contain no information on new statistical characteristics.)

All experiments were conducted in two stages, which we conventionally call parameter estimation and testing. Data used for parameter estimation were sequences of successive US dollar values, which we denoted by $x_1 x_2 \dots x_n$. In the course of experiment, the value x_{n-99} was predicted from the data $x_1 x_2 \dots x_{n-100}$, the value x_{n-98} from data $x_1 x_2 \dots x_{n-99}$, etc., so that x_n was predicted from the data $x_1 x_2 \dots x_{n-1}$. Then we computed

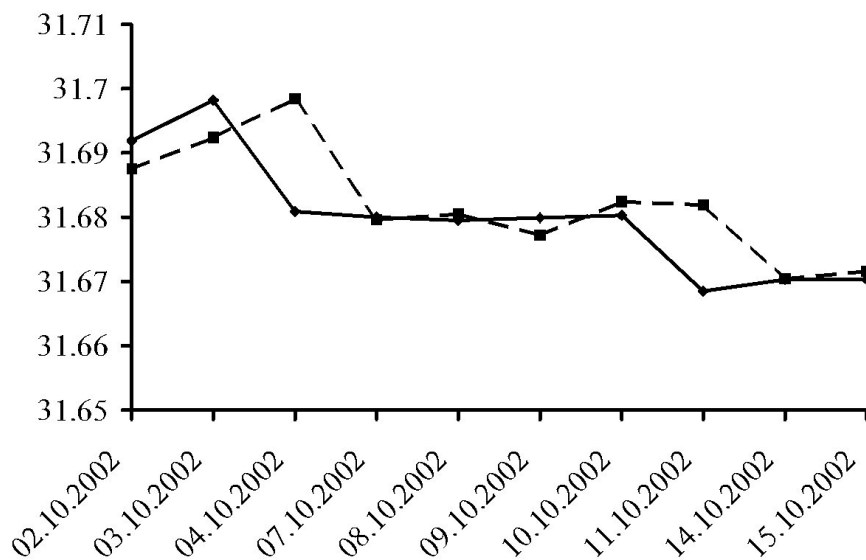
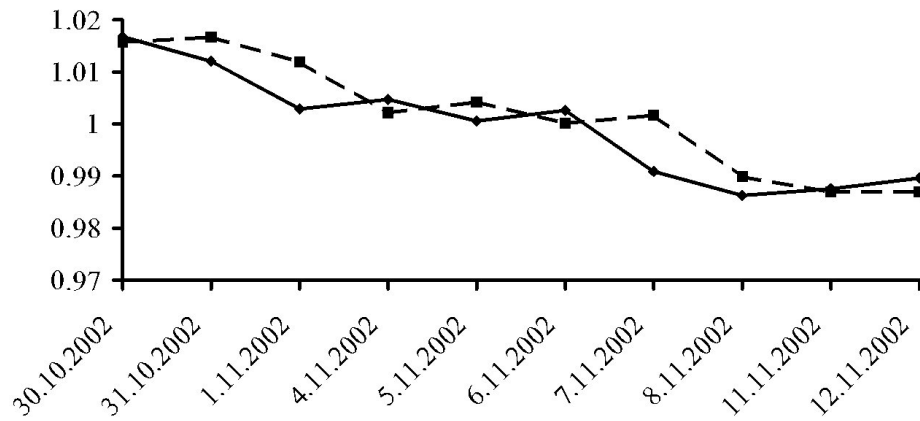
the value

$$\delta = \left(\sum_{i=1}^{100} |x_i - x_i^*| \right) / 100, \quad (50)$$

where x_i^* is the predicted value of x_i . Based on the computations made, we chose a variant with the set of parameters with the minimal value of δ . Here, the estimation stage was completed, and we passed to the testing stage, where the chosen variant was used to predict new (the latest) 100 values, already known but (we emphasize!) not used in the preceding computations. Obtained values of the prediction precision, still evaluated by δ , are given in the table of the US dollar values in rubles and euros. In the table we also indicate the values of the parameters found at the estimation stage. We also used division into intervals of equal length and data preprocessing aimed at trend elimination. Namely, an original sequence x_1, x_2, \dots, x_t was transformed into the sequence of ratios $(x_2/x_1), (x_3/x_2), \dots, (x_t/x_{t-1})$, which was used for the prediction. (Of course, δ in (50) is computed from the absolute values but not relative.)

Table 4

Currency exchange rate	Average precision	Archiver	Number of intervals	"Sliding window" size
US dollar /euro	0,00479 (euro)	Rar	15	50
US dollar /ruble	1,306 (kopeck)	Rar	10	70



We used data on the value of dollar in rubles in the period from January 3, 2001, to August 7, 2002, to find parameters giving the minimal error, and in the period from August 7, 2002, to February 26, 2003, to test the prediction precision. Similarly, we used data on the US dolar/euro rate from March 16, 2001, to July 9, 2002, and from July 9, 2002, to December 2, 2002, respectively. The obtained results are given in the table; Figs. 1 and 2 present the prediction results over 10 days, which gives a general

insight into the prediction precision. In particular, it is seen that the largest prediction errors occur when the rate changes abruptly. It is seen from the table that the average error over 100 days for the US dollar/euro rate is 0.00479 euro, and for the US dollar/ruble rate is 1.306 kopeck, which is close to daily exchange fluctuations. Thus, it is the authors opinion that the presented data demonstrate that methods of data compression (or universal coding) can be a basis for constructing prediction methods of practical interest.

6.3 Hypothesis Testing

In this subsection we consider a case where the source alphabet A is infinite, say, a part of \mathbb{R}^n . Our strategy is to use finite partitions of A and to consider hypotheses corresponding to the partitions. This approach can be directly applied to the goodness-of-fit testing, but it cannot be applied to the serial independence testing. The point is that if someone combines letters (or states) of a Markov chain, the chain order (or memory) can increase. For example, if the alphabet contains three letters, there exists a Markov chain of order one, such that combining two letters into one transforms the chain into a process with infinite memory. That is why in this part we will consider the independence testing for i.i.d. processes only (i.e. processes from $M_0(A)$).

In order to avoid repetitions, we will consider a general scheme, which can be applied to both tests using notations $H_0^{\aleph}, H_1^{\aleph}$ and $T_{\varphi}^{\aleph}(A, \alpha)$, where \aleph is an abbreviation of one of the described tests (i.e. *id* and *SI*).

Let us give some definitions. Let $\Lambda = \lambda_1, \dots, \lambda_s$ be a finite (measurable) partition of A and let $\Lambda(x)$ be an element of the partition Λ which contains $x \in A$. For any process π we define a process π_{Λ} over a new alphabet Λ by the equation

$$\pi_{\Lambda}(\lambda_{i_1} \dots \lambda_{i_k}) = \pi(x_1 \in \lambda_{i_1}, \dots, x_k \in \lambda_{i_k}),$$

where $x_1 \dots x_k \in A^k$.

We will consider an infinite sequence of partitions $\hat{\Lambda} = \Lambda_1, \Lambda_2, \dots$ and say that such a sequence discriminates between a pair of hypotheses $H_0^{\aleph}(A), H_1^{\aleph}(A)$ about processes, if for each process ϱ , for which $H_1^{\aleph}(A)$ is true, there exists a partition Λ_j for which $H_1^{\aleph}(\Lambda_j)$ is true for the process ϱ_{Λ_j} .

Let $H_0^{\aleph}(A), H_1^{\aleph}(A)$ be a pair of hypotheses, $\hat{\Lambda} = \Lambda_1, \Lambda_2, \dots$ be a sequence of partitions, α be from $(0, 1)$ and φ be a code. The scheme for

both tests is as follows:

The hypothesis $H_0^{\aleph}(A)$ is accepted if for all $i = 1, 2, 3, \dots$ the test $T_{\varphi}^{\aleph}(\Lambda_i, (\alpha\omega_i))$ accepts the hypothesis $H_0^{\aleph}(\Lambda_i)$. Otherwise, H_0^{\aleph} is rejected. We denote this test $\mathbf{T}_{\alpha, \varphi}^{\aleph}(\hat{\Lambda})$.

Comment 3. It is important to note that one does not need to check an infinite number of inequalities when applying this test. The point is that the hypothesis $H_0^{\aleph}(A)$ has to be accepted if the left part in (36) or (37) is less than $-\log(\alpha\omega_i)$. Obviously, $-\log(\alpha\omega_i)$ goes to infinity if i increases. That is why there are many cases, where it is enough to check a finite number of hypotheses $H_0^{\aleph}(\Lambda_i)$.

Theorem 11 *i) For each $\alpha \in (0, 1)$, sequence of partitions $\hat{\Lambda}$ and a code φ , the Type I error of the described test $\mathbf{T}_{\alpha, \varphi}^{\aleph}(\hat{\Lambda})$ is not larger than α , and ii) if, in addition, φ is a universal code and $\hat{\Lambda}$ discriminates between $H_0^{\aleph}(A), H_1(A)^{\aleph}$, then the Type II error of the test $\mathbf{T}_{\alpha, \varphi}^{\aleph}(\hat{\Lambda})$ goes to 0, when the sample size tends to infinity.*

7 Conclusion

Time series is a popular model of real stochastic processes which has a lot of applications in industry, economy, meteorology and many other fields. Despite this, there are many practically important problems of statistical analysis of time series which are still open. Among them we can name the problem of estimation of the limiting probabilities and densities, on-line prediction, regression, classification and some problems of hypothesis testing (goodness-of-fit testing and testing of serial independence). This chapter describes a new approach to all the problems mentioned above, which, on the one hand, gives a possibility to solve the problems in the framework of the classical mathematical statistics and, on the other hand, allows to apply methods of real data compression to solve these problems in practise. Such applications to randomness testing [44] and prediction of currency exchange rates [43] showed high efficiency, that is why the suggested methods look very promising for practical applications. Of course, problems like prediction of price of oil, gold, etc. and testing of different random number generators can be used as case studies for students.

8 Appendix

Proof 1 (Claim 1) *We employ the general inequality*

$$D(\mu\|\eta) \leq \log e \left(-1 + \sum_{a \in A} \mu(a)^2 / \eta(a)\right),$$

valid for any distributions μ and η over A (follows from the elementary inequality for natural logarithm $\ln x \leq x - 1$), and find:

$$\begin{aligned} \rho^t(P\|L_0) &= \sum_{x_1 \cdots x_t \in A^t} P(x_1 \cdots x_t) \sum_{a \in A} P(a|x_1 \cdots x_t) \log \frac{P(a|x_1 \cdots x_t)}{\gamma(a|x_1 \cdots x_t)} \\ &= \log e \left(\sum_{x_1 \cdots x_t \in A^t} P(x_1 \cdots x_t) \sum_{a \in A} P(a|x_1 \cdots x_t) \ln \frac{P(a|x_1 \cdots x_t)}{\gamma(a|x_1 \cdots x_t)} \right) \\ &\leq \log e \left(-1 + \sum_{x_1 \cdots x_t \in A^t} P(x_1 \cdots x_t) \sum_{a \in A} \frac{P(a)^2(t + |A|)}{\nu_{x_1 \cdots x_t}(a) + 1}\right) \end{aligned}$$

Applying the well-known Bernoulli formula, we obtain

$$\begin{aligned} \rho^t(P\|L_0) &= \log e \left(-1 + \sum_{a \in A} \sum_{i=0}^t \frac{P(a)^2(t + |A|)}{i + 1} \binom{t}{i} P(a)^i (1 - P(a))^{t-i}\right) \\ &= \log e \left(-1 + \frac{t + |A|}{t + 1} \sum_{a \in A} P(a) \sum_{i=0}^t \binom{t + 1}{i + 1} P(a)^{i+1} (1 - P(a))^{t-i}\right) \\ &\leq \log e \left(-1 + \frac{t + |A|}{t + 1} \sum_{a \in A} P(a) \sum_{j=0}^{t+1} \binom{t + 1}{j} P(a)^j (1 - P(a))^{t+1-j}\right). \end{aligned}$$

Again, using the Bernoulli formula, we finish the proof

$$\rho^t(P\|L_0) = \log e \frac{|A| - 1}{t + 1}.$$

The second statement of the claim follows from the well-known asymptotic equality

$$1 + 1/2 + 1/3 + \dots + 1/t = \ln t + O(1),$$

the obvious presentation

$$\bar{\rho}^t(P\|L_0) = t^{-1}(\rho^0(P\|L_0) + \rho^1(P\|L_0) + \dots + \rho^{t-1}(P\|L_0))$$

and (10).

Proof 2 (Claim 2) *The first equality follows from the definition (9), whereas the second from the definition (12). From (16) we obtain:*

$$\begin{aligned} -\log K_0(x_1 \dots x_t) &= -\log\left(\frac{\Gamma(|A|/2)}{\Gamma(1/2)^{|A|}} \frac{\prod_{a \in A} \Gamma(\nu^t(a) + 1/2)}{\Gamma((t + |A|/2))}\right) \\ &= c_1 + c_2|A| + \log \Gamma(t + |A|/2) - \sum_{a \in A} \Gamma(\nu^t(a) + 1/2), \end{aligned}$$

where c_1, c_2 are constants. Now we use the well known Stirling formula

$$\ln \Gamma(s) = \ln \sqrt{2\pi} + (s - 1/2) \ln s - s + \theta/12,$$

where $\theta \in (0, 1)$ [23]. Using this formula we rewrite the previous equality as

$$-\log K_0(x_1 \dots x_t) = -\sum_{a \in A} \nu^t(a) \log(\nu^t(a)/t) + (|A| - 1) \log t/2 + \bar{c}_1 + \bar{c}_2|A|,$$

where \bar{c}_1, \bar{c}_2 are constants. Hence,

$$\begin{aligned} &\sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) (-\log(K_0(x_1 \dots x_t))) \\ &\leq t \left(\sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \left(-\sum_{a \in A} \nu^t(a) \log(\nu^t(a)/t) + (|A| - 1) \log t/2 + c|A| \right) \right). \end{aligned}$$

Applying the well known Jensen inequality for the concave function $-x \log x$ we obtain the following inequality:

$$\begin{aligned} &\sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) (-\log(K_0(x_1 \dots x_t))) \leq \\ &\quad -t \left(\sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) (\nu^t(a)/t) \right) \\ &\quad \log \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) (\nu^t(a)/t) + (|A| - 1) \log t/2 + c|A|. \end{aligned}$$

The source P is i.i.d., that is why the average frequency

$$\sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \nu^t(a)$$

is equal to $P(a)$ for any $a \in A$ and we obtain from two last formulas the following inequality:

$$\begin{aligned} & \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) (-\log(K_0(x_1 \dots x_t))) \\ & \leq t \left(-\sum_{a \in A} P(a) \log P(a) \right) + (|A| - 1) \log t/2 + c|A| \end{aligned} \quad (51)$$

On the other hand,

$$\begin{aligned} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) (\log P(x_1 \dots x_t)) &= \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \sum_{i=1}^t \log P(x_i) \\ &= t \left(\sum_{a \in A} P(a) \log P(a) \right). \end{aligned} \quad (52)$$

From (51) and (52) we can see that

$$t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log \frac{P(x_1 \dots x_t)}{(K_0(x_1 \dots x_t))} \leq ((|A| - 1) \log t/2 + c)/t.$$

Proof 3 (Claim 3) First we consider the case where $m = 0$. The proof for this case is very close to the proof of the previous claim. Namely, from (16) we obtain:

$$\begin{aligned} -\log K_0(x_1 \dots x_t) &= -\log \left(\frac{\Gamma(|A|/2)}{\Gamma(1/2)^{|A|}} \frac{\prod_{a \in A} \Gamma(\nu^t(a) + 1/2)}{\Gamma((t + |A|/2))} \right) \\ &= c_1 + c_2|A| + \log \Gamma(t + |A|/2) - \sum_{a \in A} \Gamma(\nu^t(a) + 1/2), \end{aligned}$$

where c_1, c_2 are constants. Now we use the well known Stirling formula

$$\ln \Gamma(s) = \ln \sqrt{2\pi} + (s - 1/2) \ln s - s + \theta/12,$$

where $\theta \in (0, 1)$ [23]. Using this formula we rewrite the previous equality as

$$-\log K_0(x_1 \dots x_t) = -\sum_{a \in A} \nu^t(a) \log(\nu^t(a)/t) + (|A| - 1) \log t/2 + \bar{c}_1 + \bar{c}_2|A|,$$

where \bar{c}_1, \bar{c}_2 are constants. Having taken into account the definition of the empirical entropy (23), we obtain

$$-\log K_0(x_1 \dots x_t) \leq th_0^*(x_1 \dots x_t) + (|A| - 1) \log t/2 + c|A|.$$

Hence,

$$\begin{aligned} & \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) (-\log(K_0(x_1 \dots x_t))) \\ & \leq t \left(\sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) h_0^*(x_1 \dots x_t) + (|A| - 1) \log t / 2 + c|A| \right). \end{aligned}$$

Having taken into account the definition of the empirical entropy (23), we apply the well known Jensen inequality for the concave function $-x \log x$ and obtain the following inequality:

$$\begin{aligned} & \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) (-\log(K_0(x_1 \dots x_t))) \leq +c|A| - \\ & t \left(\sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) ((\nu^t(a)/t)) \log \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) (\nu^t(a)/t) + (|A| - 1) \log t / 2 \right). \end{aligned}$$

P is stationary and ergodic, that is why the average frequency

$$\sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \nu^t(a)$$

is equal to $P(a)$ for any $a \in A$ and we obtain from two last formulas the following inequality:

$$\sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) (-\log(K_0(x_1 \dots x_t))) \leq t h_0(P) + (|A| - 1) \log t / 2 + c|A|,$$

where $h_0(P)$ is the first order Shannon entropy, see (12).

We have seen that any source from $M_m(A)$ can be presented as a "sum" of $|A|^m$ i.i.d. sources. From this we can easily see that the error of a predictor for the source from $M_m(A)$ can be upper bounded by the error of i.i.d. source multiplied by $|A|^m$. In particular, we obtain from the last inequality and the definition of the Shannon entropy (20) the upper bound (22).

Proof 4 (Theorem 1) We can see from the definition (25) of R and the Claim 3 that the average error is upper bounded as follows:

$$\begin{aligned} & -t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log(R(x_1 \dots x_t)) - h_k(P) \\ & \leq (|A|^k (|A| - 1) \log t + \log(1/\omega_i) + C) / (2t), \end{aligned}$$

for any $k = 0, 1, 2, \dots$. Taking into account that for any $P \in M_\infty(A)$ $\lim_{k \rightarrow \infty} h_k(P) = h_\infty(P)$, we can see that

$$\left(\lim_{t \rightarrow \infty} t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log(R(x_1 \dots x_t)) - h_\infty(P) \right) = 0.$$

The second statement of the theorem is proven. The first one can be easily derived from the ergodicity of P [5, 14].

Proof 5 (Theorem 2) The proof is based on the Shannon-MacMillan-Breiman theorem which states that for any stationary and ergodic source P

$$\lim_{t \rightarrow \infty} -\log P(x_1 \dots x_t)/t = h_\infty(P)$$

with probability 1 [5, 14]. From this equality and (29) we obtain the statement i). The second statement follows from the definition of the Shannon entropy (21) and (30).

Proof 6 (Theorem 4) i) immediately follows from the second statement of the theorem 2 and properties of \log . The statement ii) can be proven as follows:

$$\begin{aligned} & \lim_{t \rightarrow \infty} E\left(\frac{1}{t} \sum_{i=0}^{t-1} (P(x_{i+1}|x_1 \dots x_i) - \mu_U(x_{i+1}|x_1 \dots x_i))^2\right) = \\ & \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=0}^{t-1} \sum_{x_1 \dots x_i \in A^i} P(x_1 \dots x_i) \left(\sum_{a \in A} |P(a|x_1 \dots x_i) - \mu_U(a|x_1 \dots x_i)|\right)^2 \leq \\ & \lim_{t \rightarrow \infty} \frac{\text{const}}{t} \sum_{i=0}^{t-1} \sum_{x_1 \dots x_i \in A^i} P(x_1 \dots x_i) \sum_{a \in A} P(a|x_1 \dots x_i) \log \frac{P(a|x_1 \dots x_i)}{\mu_U(a|x_1 \dots x_i)} = \\ & \lim_{t \rightarrow \infty} \left(\frac{\text{const}}{t} \sum_{x_1 \dots x_t \in A^t} P(x_1 \dots x_t) \log(P(x_1 \dots x_t)/\mu(x_1 \dots x_t))\right). \end{aligned}$$

Here the first inequality is obvious, the second follows from the Pinsker's inequality (5), the others from properties of expectation and \log . iii) can be derived from ii) and the Jensen inequality for the function x^2 .

Proof 7 (Theorem 5) The following inequality follows from the non-negativity of the KL divergency (see (5)), whereas the equality is obvious.

$$E\left(\log \frac{P(x_1|y_1)}{\mu_U(x_1|y_1)}\right) + E\left(\log \frac{P(x_2|(x_1, y_1), y_2)}{\mu_U(x_2|(x_1, y_1), y_2)}\right) + \dots \leq E\left(\log \frac{P(y_1)}{\mu_U(y_1)}\right)$$

$$\begin{aligned}
& + E\left(\log \frac{P(x_1|y_1)}{\mu_U(x_1|y_1)}\right) + E\left(\log \frac{P(y_2|(x_1, y_1))}{\mu_U(y_2|(x_1, y_1))}\right) + E\left(\log \frac{P(x_2|(x_1, y_1), y_2)}{\mu_U(x_2|(x_1, y_1), y_2)}\right) + \dots \\
& = E\left(\log \frac{P(x_1, y_1)}{\mu_U(x_1, y_1)}\right) + E\left(\log \frac{P((x_2, y_2)|(x_1, y_1))}{\mu_U((x_2, y_2)|(x_1, y_1))}\right) + \dots
\end{aligned}$$

Now we can apply the first statement of the previous theorem to the last sum as follows:

$$\begin{aligned}
& \lim_{t \rightarrow \infty} \frac{1}{t} E\left(\log \frac{P(x_1, y_1)}{\mu_U(x_1, y_1)}\right) + E\left(\log \frac{P((x_2, y_2)|(x_1, y_1))}{\mu_U((x_2, y_2)|(x_1, y_1))}\right) + \dots \\
& E\left(\log \frac{P((x_t, y_t)|(x_1, y_1) \dots (x_{t-1}, y_{t-1}))}{\mu_U((x_t, y_t)|(x_1, y_1) \dots (x_{t-1}, y_{t-1}))}\right) = 0.
\end{aligned}$$

From this equality and the last inequality we obtain the proof of i). The proof of the second statement can be obtained from the similar representation for ii) and the second statement of the theorem 4. iii) can be derived from ii) and the Jensen inequality for the function x^2 .

Proof 8 (Lemma 1) . First we show that for any source $\theta^* \in M_0(A)$ and any words $x^1 = x_1^1 \dots x_{t_1}^1$, ..., $x^r = x_1^r \dots x_{t_r}^r$,

$$\begin{aligned}
\theta^*(x^1 \diamond \dots \diamond x^r) &= \prod_{a \in A} (\theta^*(a))^{\nu_{x^1 \diamond \dots \diamond x^r}(a)} \\
&\leq \prod_{a \in A} (\nu_{x^1 \diamond \dots \diamond x^r}(a)/t)^{\nu_{x^1 \diamond \dots \diamond x^r}(a)}, \tag{53}
\end{aligned}$$

where $t = \sum_{i=1}^r t_i$. Here the equality holds, because $\theta^* \in M_0(A)$. The inequality follows from Claim 1. Indeed, if $p(a) = \nu_{x^1 \diamond \dots \diamond x^r}(a)/t$ and $q(a) = \theta^*(a)$, then

$$\sum_{a \in A} \frac{\nu_{x^1 \diamond \dots \diamond x^r}(a)}{t} \log \frac{(\nu_{x^1 \diamond \dots \diamond x^r}(a)/t)}{\theta^*(a)} \geq 0.$$

From the latter inequality we obtain (53). Taking into account the definition (34) and (53), we can see that the statement of Lemma is true for this particular case.

For any $\theta \in M_m(A)$ and $x = x_1 \dots x_s$, $s > m$, we present $\theta(x_1 \dots x_s)$ as

$$\theta(x_1 \dots x_s) = \theta(x_1 \dots x_m) \prod_{u \in A^m} \prod_{a \in A} \theta(a/u)^{\nu_x(ua)},$$

where $\theta(x_1 \dots x_m)$ is the limiting probability of the word $x_1 \dots x_m$. Hence, $\theta(x_1 \dots x_s) \leq \prod_{u \in A^m} \prod_{a \in A} \theta(a/u)^{\nu_x(ua)}$. Taking into account the inequality (53), we obtain $\prod_{a \in A} \theta(a/u)^{\nu_x(ua)} \leq \prod_{a \in A} (\nu_x(ua)/\bar{\nu}_x(u))^{\nu_x(ua)}$ for any word u . Hence,

$$\begin{aligned} \theta(x_1 \dots x_s) &\leq \prod_{u \in A^m} \prod_{a \in A} \theta(a/u)^{\nu_x(ua)} \\ &\leq \prod_{u \in A^m} \prod_{a \in A} (\nu_x(ua)/\bar{\nu}_x(u))^{\nu_x(ua)}. \end{aligned}$$

If we apply those inequalities to $\theta(x^1 \diamond \dots \diamond x^r)$, we immediately obtain the following inequalities

$$\begin{aligned} \theta(x^1 \diamond \dots \diamond x^r) &\leq \prod_{u \in A^m} \prod_{a \in A} \theta(a/u)^{\nu_{x^1 \diamond \dots \diamond x^r}(ua)} \leq \\ &\prod_{u \in A^m} \prod_{a \in A} (\nu_{x^1 \diamond \dots \diamond x^r}(ua)/\bar{\nu}_{x^1 \diamond \dots \diamond x^r}(u))^{\nu_{x^1 \diamond \dots \diamond x^r}(ua)}. \end{aligned}$$

Now the statement of the Lemma follows from the definition (34).

Proof 9 (Theorem 6) Let C_α be a critical set of the test $T_\varphi^{id}(A, \alpha)$, i.e., by definition, $C_\alpha = \{u : u \in A^t \ \& \ -\log \pi(u) - |\varphi(u)| > -\log \alpha\}$. Let μ_φ be a measure for which the claim 2 is true. We define an auxiliary set $\hat{C}_\alpha = \{u : -\log \pi(u) - (-\log \mu_\varphi(u)) > -\log \alpha\}$. We have $1 \geq \sum_{u \in \hat{C}_\alpha} \mu_\varphi(u) \geq \sum_{u \in \hat{C}_\alpha} \pi(u)/\alpha = (1/\alpha)\pi(\hat{C}_\alpha)$. (Here the second inequality follows from the definition of \hat{C}_α , whereas all others are obvious.) So, we obtain that $\pi(\hat{C}_\alpha) \leq \alpha$. From definitions of C_α, \hat{C}_α and (26) we immediately obtain that $\hat{C}_\alpha \supset C_\alpha$. Thus, $\pi(C_\alpha) \leq \alpha$. By definition, $\pi(C_\alpha)$ is the value of the Type I error. The first statement of the theorem is proven.

Let us prove the second statement of the theorem. Suppose that the hypothesis $H_1^{id}(A)$ is true. That is, the sequence $x_1 \dots x_t$ is generated by some stationary and ergodic source τ and $\tau \neq \pi$. Our strategy is to show that

$$\lim_{t \rightarrow \infty} -\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| = \infty \quad (54)$$

with probability 1 (according to the measure τ). First we represent (54) as

$$\begin{aligned} &-\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| \\ &= t \left(\frac{1}{t} \log \frac{\tau(x_1 \dots x_t)}{\pi(x_1 \dots x_t)} + \frac{1}{t} (-\log \tau(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)|) \right). \end{aligned}$$

From this equality and the property of a universal code (29) we obtain

$$-\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| = t \left(\frac{1}{t} \log \frac{\tau(x_1 \dots x_t)}{\pi(x_1 \dots x_t)} + o(1) \right). \quad (55)$$

From (29) and (21) we can see that

$$\lim_{t \rightarrow \infty} -\log \tau(x_1 \dots x_t)/t \leq h_k(\tau) \quad (56)$$

for any $k \geq 0$ (with probability 1). It is supposed that the process π has a finite memory, i.e. belongs to $M_s(A)$ for some s . Having taken into account the definition of $M_s(A)$ (18), we obtain the following representation:

$$\begin{aligned} -\log \pi(x_1 \dots x_t)/t &= -t^{-1} \sum_{i=1}^t \log \pi(x_i/x_1 \dots x_{i-1}) \\ &= -t^{-1} \left(\sum_{i=1}^k \log \pi(x_i/x_1 \dots x_{i-1}) + \sum_{i=k+1}^t \log \pi(x_i/x_{i-k} \dots x_{i-1}) \right) \end{aligned}$$

for any $k \geq s$. According to the ergodic theorem there exists a limit

$$\lim_{t \rightarrow \infty} t^{-1} \sum_{i=k+1}^t \log \pi(x_i/x_{i-k} \dots x_{i-1}),$$

which is equal to $h_k(\tau)$ [5, 14]. So, from the two last equalities we can see that

$$\lim_{t \rightarrow \infty} (-\log \pi(x_1 \dots x_t))/t = - \sum_{v \in A^k} \tau(v) \sum_{a \in A} \tau(a/v) \log \pi(a/v).$$

Taking into account this equality, (56) and (55), we can see that

$$\begin{aligned} -\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| &\geq \\ &t \left(\sum_{v \in A^k} \tau(v) \sum_{a \in A} \tau(a/v) \log(\tau(a/v)/\pi(a/v)) \right) + o(t) \end{aligned}$$

for any $k \geq s$.

From this inequality and Claim 1 we can obtain that

$$-\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| \geq ct + o(t),$$

where c is a positive constant, $t \rightarrow \infty$. Hence, (54) is true and the theorem is proven.

Proof 10 (Theorem 7) Let us denote the critical set of the test $T_\varphi^{SI}(A, \alpha)$ as C_α , i.e., by definition, $C_\alpha = \{x_1 \dots x_t : (t - m) h_m^*(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)|\} > \log(1/\alpha)\}$. From Claim 2 we can see that there exists such a measure μ_φ that $-\log \mu_\varphi(x_1 \dots x_t) \leq |\varphi(x_1 \dots x_t)|$. We also define

$$\hat{C}_\alpha = \{x_1 \dots x_t : (t - m) h_m^*(x_1 \dots x_t) - (-\log \mu_\varphi(x_1 \dots x_t)) > \log(1/\alpha)\}. \quad (57)$$

Obviously, $\hat{C}_\alpha \supset C_\alpha$. Let θ be any source from $M_m(A)$. The following chain of equalities and inequalities is true:

$$\begin{aligned} 1 &\geq \mu_\varphi(\hat{C}_\alpha) = \sum_{x_1 \dots x_t \in \hat{C}_\alpha} \mu_\varphi(x_1 \dots x_t) \\ &\geq \alpha^{-1} \sum_{x_1 \dots x_t \in \hat{C}_\alpha} 2^{(t-m)h_m^*(x_1 \dots x_t)} \geq \alpha^{-1} \sum_{x_1 \dots x_t \in \hat{C}_\alpha} \theta(x_1 \dots x_t) = \theta(\hat{C}_\alpha). \end{aligned}$$

(Here both equalities and the first inequality are obvious, the second and the third inequalities follow from (57) and the Lemma, correspondingly.) So, we obtain that $\theta(\hat{C}_\alpha) \leq \alpha$ for any source $\theta \in M_m(A)$. Taking into account that $\hat{C}_\alpha \supset C_\alpha$, where C_α is the critical set of the test, we can see that the probability of the Type I error is not greater than α . The first statement of the theorem is proven.

The proof of the second statement will be based on some results of Information Theory. We obtain from (29) that for any stationary and ergodic p

$$\lim_{t \rightarrow \infty} t^{-1} |\varphi(x_1 \dots x_t)| = h_\infty(p) \quad (58)$$

with probability 1. It can be seen from (23) that h_m^* is an estimate for the m -order Shannon entropy (20). Applying the ergodic theorem we obtain $\lim_{t \rightarrow \infty} h_m^*(x_1 \dots x_t) = h_m(p)$ with probability 1 [5, 14]. It is known in Information Theory that $h_m(\varrho) - h_\infty(\varrho) > 0$, if ϱ belongs to $M_\infty(A) \setminus M_m(A)$ [5, 14]. It is supposed that H_1^{SI} is true, i.e. the considered process belongs to $M_\infty(A) \setminus M_m(A)$. So, from (58) and the last equality we obtain that $\lim_{t \rightarrow \infty} ((t - m) h_m^*(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)|) = \infty$. This proves the second statement of the theorem.

Proof 11 (Theorem 8) First we prove that with probability 1 there exists the following limit $\lim_{t \rightarrow \infty} \frac{1}{t} \log(p(x_1 \dots x_t)/r_U(x_1 \dots x_t))$ and this limit is finite and nonnegative. Let $A_n = \{x_1, \dots, x_n : p(x_1, \dots, x_n) \neq 0\}$. Define

$$z_n(x_1 \dots x_n) = r_U(x_1 \dots x_n)/p(x_1 \dots x_n) \quad (59)$$

for $(x_1, \dots, x_n) \in A$ and $z_n = 0$ elsewhere.

Since

$$\begin{aligned}
E_P(z_n | x_1, \dots, x_{n-1}) &= E \left(\frac{r_U(x_1 \dots x_n)}{p(x_1 \dots x_n)} \middle| x_1, \dots, x_{n-1} \right) \\
&= \frac{r_U(x_1 \dots x_{n-1})}{p(x_1 \dots x_{n-1})} E_P \left(\frac{r_U(x_n | x_1 \dots x_{n-1})}{p(x_n | x_1 \dots x_{n-1})} \right) \\
&= z_{n-1} \int_A \frac{r_U(x_n | x_1 \dots x_{n-1}) dP(x_n | x_1 \dots x_{n-1})}{dP(x_n | x_1 \dots x_{n-1}) / dM_n(x_n | x_1 \dots x_{n-1})} \\
&= z_{n-1} \int_A r_U(x_n | x_1 \dots x_{n-1}) dM_n(x_n | x_1 \dots x_{n-1}) \leq z_{n-1}
\end{aligned}$$

the stochastic sequence (z_n, B^n) is, by definition, a non-negative supermartingale with respect to P , with $E(z_n) \leq 1$, [52]. Hence, Doob's submartingale convergence theorem implies that the limit z_n exists and is finite with P -probability 1 (see [52, Theorem 7.4.1]). Since all terms are nonnegative so is the limit. Using the definition (59) with P -probability 1 we have

$$\begin{aligned}
\lim_{n \rightarrow \infty} p(x_1 \dots x_n) / r_U(x_1 \dots x_n) &> 0, \\
\lim_{n \rightarrow \infty} \log(p(x_1 \dots x_n) / r_U(x_1 \dots x_n)) &> -\infty
\end{aligned}$$

and

$$\lim_{n \rightarrow \infty} n^{-1} \log(p(x_1 \dots x_n) / r_U(x_1 \dots x_n)) \geq 0. \quad (60)$$

Now we note that for any integer s the following obvious equality is true: $r_U(x_1 \dots x_t) = \omega_s \mu_U(x_1^{[s]} \dots x_t^{[s]}) / M_t(x_1^{[s]} \dots x_t^{[s]}) (1 + \delta)$ for some $\delta > 0$. From this equality, (31) and (44) we immediately obtain that a.s.

$$\begin{aligned}
\lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} &\leq \lim_{t \rightarrow \infty} \frac{-\log \omega_t}{t} \\
+ \lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{\mu_U(x_1^{[s]} \dots x_t^{[s]}) / M_t(x_1^{[s]} \dots x_t^{[s]})} \\
&\leq \lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|} / M_t(x_1^{[s]} \dots x_t^{[s]})}. \quad (61)
\end{aligned}$$

The right part can be presented as follows:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|} / M_t(x_1^{[s]} \dots x_t^{[s]})}$$

$$\begin{aligned}
&= \lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p^s(x_1 \dots x_t) M_t(x_1^{[s]} \dots x_t^{[s]})}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}} \\
&\quad + \lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{p^s(x_1 \dots x_t)}.
\end{aligned} \tag{62}$$

Having taken into account that U is a universal code, (41) and the theorem 2, we can see that the first term is equal to zero. From (40) and (43) we can see that a.s. the second term is equal to $\tilde{h}_s - \tilde{h}$. This equality is valid for any integer s and, according to (45), the second term equals zero, too, and we obtain that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} \leq 0.$$

Having taken into account (60), we can see that the first statement is proven.

From (61) and (62) we can see that

$$\begin{aligned}
E \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} &\leq E \log \frac{p_t^s(x_1, \dots, x_t) M_t(x_1^{[s]} \dots x_t^{[s]})}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}} \\
&\quad + E \log \frac{p(x_1 \dots x_t)}{p^s(x_1, \dots, x_t)}.
\end{aligned} \tag{63}$$

The first term is the average redundancy of the universal code for a finite-alphabet source, hence, according to the theorem 2, it tends to 0. The second term tends to $\tilde{h}_s - \tilde{h}$ for any s and from (45) we can see that it is equals to zero. The second statement is proven.

Proof 12 (Theorem 9) Obviously,

$$\begin{aligned}
&E\left(\frac{1}{t} \sum_{m=0}^{t-1} (P(x_{m+1} \in B_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m))^2\right) \leq \\
&\frac{1}{t} \sum_{m=0}^{t-1} E(|P(x_{m+1} \in B_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m)| + \\
&\quad |P(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m)|)^2.
\end{aligned} \tag{64}$$

From the Pinsker inequality (5) and convexity of the KL divergence (6) we obtain the following inequalities

$$\begin{aligned}
& \frac{1}{t} \sum_{m=0}^{t-1} E(|P(x_{m+1} \in B_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m)|) \quad (65) \\
& |P(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m)|^2 \leq \\
& \frac{\text{const}}{t} \sum_{m=0}^{t-1} E\left(\log \frac{P(x_{m+1} \in B_{m+1}|x_1 \dots x_m)}{R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m)} + \log \frac{P(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m)}{R_U(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m)}\right) \leq \\
& \frac{\text{const}}{t} \sum_{m=0}^{t-1} \left(\int p(x_1 \dots x_m) \left(\int p(x_{m+1}|x_1 \dots x_m) \log \frac{p(x_{m+1}|x_1 \dots x_m)}{r_U(x_{m+1}|x_1 \dots x_m)} dM\right) dM_m\right).
\end{aligned}$$

Having taken into account that the last term is equal to $\frac{\text{const}}{t} E(\log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)})$, from (64), (65) and (47) we obtain (48). *ii)* can be derived from *i)* and the Jensen inequality for the function x^2 .

Proof 13 (Theorem 10) The last inequality of the following chain follows from the Pinsker's one, whereas all others are obvious.

$$\begin{aligned}
& \left(\int f(x) p(x|x_1 \dots x_m) dM_m - \int f(x) r_U(x|x_1 \dots x_m) dM_m\right)^2 \\
& = \left(\int f(x) (p(x|x_1 \dots x_m) - r_U(x|x_1 \dots x_m)) dM_m\right)^2 \\
& \leq \bar{M}^2 \left(\int (p(x|x_1 \dots x_m) - r_U(x|x_1 \dots x_m)) dM_m\right)^2 \\
& \leq \bar{M}^2 \left(\int |p(x|x_1 \dots x_m) - r_U(x|x_1 \dots x_m)| dM_m\right)^2 \\
& \leq \text{const} \int p(x|x_1 \dots x_m) \log \frac{p(x|x_1 \dots x_m)}{r_U(x|x_1 \dots x_m)} dM_m.
\end{aligned}$$

From these inequalities we obtain:

$$\begin{aligned}
& E\left(\sum_{m=0}^{t-1} \left(\int f(x) p(x|x_1 \dots x_m) dM_m - \int f(x) r_U(x|x_1 \dots x_m) dM_m\right)^2\right) \leq \quad (66)
\end{aligned}$$

$$\sum_{m=0}^{t-1} \text{const } E\left(\int p(x|x_1\dots x_m) \log \frac{p(x|x_1\dots x_m)}{r_U(x|x_1\dots x_m)} dM_{1/m}\right).$$

The last term can be presented as follows:

$$\begin{aligned} \sum_{m=0}^{t-1} E\left(\int p(x|x_1\dots x_m) \log \frac{p(x|x_1\dots x_m)}{r_U(x|x_1\dots x_m)} dM_{1/m}\right) = \\ \sum_{m=0}^{t-1} \int p(x_1\dots x_m) \\ \int p(x|x_1\dots x_m) \log \frac{p(x|x_1\dots x_m)}{r_U(x|x_1\dots x_m)} dM_{1/m} dM_m \\ = \int p(x_1\dots x_t) \log(p(x_1\dots x_t)/r_U(x_1\dots x_t)) dM_t. \end{aligned}$$

From this equality, (66) and Corollary 1 we obtain (49). *ii)* can be derived from (66) and the Jensen inequality for the function x^2 .

Proof 14 (Theorem 11) The following chain proves the first statement of the theorem:

$$\begin{aligned} P\{H_0^{\aleph}(A) \text{ is rejected} | H_0 \text{ is true}\} &= P\left\{\bigcup_{i=1}^{\infty} \{H_0^{\aleph}(\Lambda_i) \text{ is rejected} | H_0 \text{ is true}\}\right\} \\ &\leq \sum_{i=1}^{\infty} P\{H_0^{\aleph}(\Lambda_i) | H_0 \text{ is true}\} \leq \sum_{i=1}^{\infty} (\alpha\omega_i) = \alpha. \end{aligned}$$

(Here both inequalities follow from the description of the test, whereas the last equality follows from (24).)

The second statement also follows from the description of the test. Indeed, let a sample is created by a source ϱ , for which $H_1(A)^{\aleph}$ is true. It is supposed that the sequence of partitions $\hat{\Lambda}$ discriminates between $H_0^{\aleph}(A)$, $H_1^{\aleph}(A)$. By definition, it means that there exists j for which $H_1^{\aleph}(\Lambda_j)$ is true for the process ϱ_{Λ_j} . It immediately follows from Theorem 1 - 4 that the Type II error of the test $T_{\varphi}^{\aleph}(\Lambda_j, \alpha\omega_j)$ goes to 0, when the sample size tends to infinity.

References

- [1] P. Algoet, Universal Schemes for Learning the Best Nonlinear Predictor Given the Infinite Past and Side Information, *IEEE Trans. Inform. Theory*, **45**, 1165-1185, (1999).
- [2] G. J. Babu, A. Boyarsky, Y. P. Chaubey, P. Gora, New statistical method for filtering and entropy estimation of a chaotic map from noisy data, *International Journal of Bifurcation and Chaos*, **14** (11), 3989-3994, (2004).
- [3] A.R. Barron, The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem, *The annals of Probability*, **13** (4), 1292-1303, 1985.
- [4] L.Györfi, I.Páli and E.C. van der Meulen, There is no universal code for infinite alphabet, *IEEE Trans. Inform. Theory*, **40**, 267-271, 1994.
- [5] P. Billingsley, *Ergodic theory and information*. (John Wiley & Sons, 1965).
- [6] R. Cilibrasi and P. M.B. Vitanyi, Clustering by Compression, *IEEE Transactions on Information Theory*, **51** (4), (2005).
- [7] R. Cilibrasi, R. de Wolf and P. M.B. Vitanyi, Algorithmic Clustering of Music, *Computer Music Journal*, **28** (4) 49-67, (2004).
- [8] I. Csiszár and P. Shields, *Notes on information theory and statistics*. (Foundations and Trends in Communications and Information Theory, 2004).
- [9] I. Csiszár and P. Shields, The consistency of the BIC Markov order estimation. *Annals of Statistics*, **6**, 1601-1619, 2000.
- [10] M. Effros, K. Visweswariah, S. R.Kulkarni and S. Verdu, Universal lossless source coding with the Burrows Wheeler transform, *IEEE Trans. Inform. Theory*, **45**, 1315-1321, (1999).
- [11] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol.1. (John Wiley & Sons, New York, 1970).
- [12] L. Finesso, C. Liu, and P. Narayan, The optimal error exponent for Markov order estimation, *IEEE Trans. Inf. Theory*, **42**, (1996).

- [13] B. M. Fitingof, Optimal encoding for unknown and changing statistics of messages, *Problems of Information Transmission*, **2** (2), 3–11, (1966).
- [14] R. G. Gallager, *Information Theory and Reliable Communication*. (John Wiley & Sons, New York, 1968).
- [15] E. N. Gilbert, Codes Based on Inaccurate Source Probabilities, *IEEE Trans. Inform. Theory*, **17**, (1971).
- [16] M. Gutman, Asymptotically optimal classification for multiple tests with empirically observed statistics, *IEEE Trans. Inform. Theory*, **35**(2), 401-408 (1989).
- [17] N.Jevtic, A.Orlitsky and N.P.Santhanam. A lower bound on compression of unknown alphabets. *Theoretical Computer Science*, **332**, 293–311, (2004).
- [18] J. L. Kelly, A new interpretation of information rate, *Bell System Tech. J.*, **35**, 917–926, (1956).
- [19] J. Kieffer. A unified approach to weak universal source coding . *IEEE Trans. Inform. Theory*, **24**, 674–682, 1978.
- [20] J. Kieffer, Prediction and Information Theory, *Preprint*, (available at <ftp://oz.ee.umn.edu/users/kieffer/papers/prediction.pdf/>), 1998.
- [21] J. C. Kieffer and En-Hui Yang, Grammar-based codes: a new class of universal lossless source codes. *IEEE Transactions on Information Theory*, **46** (3), 737–754, (2000).
- [22] A. N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission*, **1**, 3–11, (1965).
- [23] D. E. Knuth *The art of computer programming*. Vol.2. (Addison Wesley, 1981).
- [24] R. Krichevsky, A relation between the plausibility of information about a source and encoding redundancy, *Problems Inform. Transmission*, **4**(3), 48–57, (1968).
- [25] R. Krichevsky, *Universal Compression and Retrieval*, (Kluwer Academic Publishers, 1993).

- [26] S. Kullback, *Information Theory and Statistics*. (Wiley, New York, 1959).
- [27] U. Maurer, Information-Theoretic Cryptography, In: *Advances in Cryptology - CRYPTO '99, Lecture Notes in Computer Science*, Springer-Verlag, vol. 1666, pp. 47–64, (1999).
- [28] D. S. Modha and E. Masry, Memory-universal prediction of stationary random processes. *IEEE Trans. Inform. Theory*, **44**(1), 117–133, (1998).
- [29] O. Moeschlin, E. Grycko, C. Pohl, F. Steinert, *Experimental Stochastics*, Springer-Verlag, Berlin Heidelberg, 1998.
- [30] A. B. Nobel, On optimal sequential prediction, *IEEE Trans. Inform. Theory*, **49**(1), 83–98, (2003).
- [31] A. Orlitsky, N. P. Santhanam, and J. Zhang, Always Good Turing: Asymptotically Optimal Probability Estimation, *Science*, **302**, (2003).
- [32] J. Rissanen, Generalized Kraft inequality and arithmetic coding, *IBM J. Res. Dev.*, **20** (5), 198–203, (1976).
- [33] J. Rissanen, Universal coding, information, prediction, and estimation, *IEEE Trans. Inform. Theory*, **30**(4), 629–636, (1984).
- [34] A. Rukhin and others. *A statistical test suite for random and pseudorandom number generators for cryptographic applications*. (NIST Special Publication 800-22 (with revision dated May,15,2001)). <http://csrc.nist.gov/rng/SP800-22b.pdf>
- [35] B. Ya. Ryabko, Twice-universal coding, *Problems of Information Transmission*, **20**(3), 173–177, (1984).
- [36] B. Ya. Ryabko, Prediction of random sequences and universal coding. *Problems of Inform. Transmission*, **24**(2) 87-96, (1988).
- [37] B. Ya. Ryabko, A fast adaptive coding algorithm, *Problems of Inform. Transmission*, **26**(4), 305–317, (1990).
- [38] B. Ya. Ryabko, The complexity and effectiveness of prediction algorithms, *J. Complexity*, **10**(3), 281–295, (1994).

- [39] B. Ryabko. Compression-Based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series. *IEEE Transactions on Information Theory*, **55**(9), 4309–4315, (2009).
- [40] B. Ryabko, J. Astola and A. Gammerman, Application of Kolmogorov complexity and universal codes to identity testing and non-parametric testing of serial independence for time series, *Theoretical Computer Science*, **359**, 440–448, (2006).
- [41] B. Ryabko, J. Astola and A. Gammerman, Adaptive Coding and Prediction of Sources with Large and Infinite Alphabets, *IEEE Transactions on Information Theory*, **54**(8), (2008).
- [42] B. Ryabko, J. Astola and K. Egiazarian, Fast Codes for Large Alphabets, *Communications in Information and Systems*, **3** (2), 139–152, (2003).
- [43] B. Ryabko and V. Monarev, Experimental Investigation of Forecasting Methods Based on Data Compression Algorithms. *Problems of Information Transmission*, **41**, (1), 65–69, (2005).
- [44] B. Ryabko and V. Monarev, Using Information Theory Approach to Randomness Testing, *Journal of Statistical Planning and Inference*, **133**(1), 95–110, (2005).
- [45] B. Ryabko and Zh. Reznikova, Using Shannon Entropy and Kolmogorov Complexity To Study the Communicative System and Cognitive Capacities in Ants, *Complexity*, **2** (2), 37–42, (1996).
- [46] B. Ryabko and Zh. Reznikova. The Use of Ideas of Information Theory for Studying "Language" and Intelligence in Ants. *Entropy*, **11** (4), 836–853, (2009).
- [47] B. Ryabko and F. Topsoe, On Asymptotically Optimal Methods of Prediction and Adaptive Coding for Markov Sources, *Journal of Complexity*, **18**(1), 224–241, (2002).
- [48] D. Ryabko and M. Hutter, Sequence prediction for non-stationary processes, In proceedings: *Combinatorial and Algorithmic Foundations of Pattern and Association Discovery*, Dagstuhl Seminar, 2006, Germany, <http://www.dagstuhl.de/06201/> see also <http://arxiv.org/pdf/cs.LG/0606077>

- [49] S. A. Savari, A probabilistic approach to some asymptotics in noiseless communication, *IEEE Transactions on Information Theory*, **46**(4), 1246–1262, (2000).
- [50] C. E. Shannon, A mathematical theory of communication, *Bell Sys. Tech. J.* , **27**, 379–423, 623–656, (1948).
- [51] C. E. Shannon, Communication theory of secrecy systems, *Bell Sys. Tech. J.*, **28**, 656–715, (1948).
- [52] A.N. Shiryaev, *Probability*, (second edition), Springer, 1995.

Part 2

Universal Compressors in Testing Style
Homogeneity between Texts: A Review

Abstract

We study a new *context-free* computationally simple stylometry-based attributor: the *sliced conditional compression complexity* (abbreviated as CCC) of literary texts introduced in [25] and inspired by the incomputable Kolmogorov conditional complexity. Other stylometry tools can occasionally almost coincide for different authors. Our CCC-attributor is asymptotically strictly minimal for the true author, if the query text slices are sufficiently large but much less than the training texts, universal compressor is sufficiently good and sampling bias is avoided. This classifier simplifies the homogeneity test in [50] (partly based on compression) **under insignificant difference assumption of unconditional complexities of training and query texts**. This assumption is verified via its asymptotic normality [56] for IID and Markov sources and normal plots for real literary texts. It is *consistent* under large text approximation as a *stationary ergodic sequence* due to the *lower bound for the min-max compression redundancy of piecewise stationary strings* [40] (see also our elementary combinatorial arguments and simulation for IID sources). The CCC is based on the *t-ratio* measuring how many standard deviations are in the mean difference of slices' CCC. This enables evaluation of the corresponding P-value of statistical significance based on slices' CCC *asymptotic normality empirically verified by their normal plots in all cases studied* and expected to be proved soon for simplified statistical models of literary texts.

The *asymptotic CCC study* is complemented by many literary case studies: attributing the Federalist papers agreeing with previous results, significant (beyond any doubt) mean CCC-difference between two translations of Shakespeare sonnets into Russian, between the two parts of M. Sholokhov's early short novel and less so between the two Isaiah books from the Bible, intriguing CCC-relations between certain Elizabethan poems. At the same time, two different S. Brodsky's novels *deliberately written in different styles* and Madison's Federalist papers showed insignificant mean CCC-difference.

Another application of universal compressor-based statistical methodology for screening out sparse active inputs in general systems disturbed by stationary ergodic noise with memory is outlined in [32].

9 Discrimination with Universal Compressors

C. Shannon [53] created a comprehensive theory of information transmission based on Kolmogorov's statistical theory. In particular, **given** a product distribution $P^n(a)$ on n -strings a^n with entries from finite al-

phabet A , the mean length of the Shannon-Fano encoding IID source a^n as binary sequence of minimal integer length not less than $|\log P(a)|$ attains asymptotically the Shannon's entropy lower bound for the length (complexity) of compression. All log are to the base 2.

Kolmogorov in [21] made the next step. Discovering the test for randomness of a long **individual string** which concluded this theory after long development (including the well-known erroneous approach of von Mises), he created an abstract incomputable complexity theory of an *individual string* and a sketch of the **first universal compressor** such that for large strings *belonging to an IID statistical ensemble* their **mean** complexity approximates their entropy.

This groundbreaking idea was used for creating efficient so-called **universal compressors** (UC): Lempel-Ziv (LZ-77, LZ-78) and [48] among others, which adapt to an **unknown stationary ergodic distribution** (SED) of strings attaining asymptotically the Shannon entropy lower bound. \mathbf{P} is the class of SED sources approximated by n -MC's.

Compressor family $\mathbf{L} = \{\mathbf{L}_n : \mathbf{B}^n \rightarrow \mathbf{B}^\infty, n = 1, 2, \dots\}$ is (weakly) **universal**, if for any $P \in \mathbf{P}$ and any $\varepsilon > 0, \mathbf{B} = \{0, 1\}$, it holds:

$$\lim_{n \rightarrow \infty} P(x \in \mathbf{B}^n : |\mathbf{L}_n(\mathbf{x})| + \log \mathbf{P}(\mathbf{x}) \leq n\varepsilon) = 1, \quad (67)$$

where $|L(x)|$ is the length of $L(x)$ and $|L_n(x)| + \log P(x)$ is called *individual redundancy*. Thus for a string generated by a SED, the **UC-compression length is asymptotically its negative loglikelihood** which can be used in **nonparametric** statistical inference, if the *likelihood cannot be evaluated analytically*.

First UC used estimating parameters of approximating n -Markov Chains (n -MC) to adapt for good compression. Both LZ-compressors do not use any statistics of strings at all. Instead, LZ-78 constructs the tree of binary patterns unseen before in the string consecutively, starting from the first digit of the string. Ziv, Wyner and Ziv proved that LZ-78 and LZ-77 are UC implying

$$\lim_{n \rightarrow \infty} P(|L_n(x)|/|x| \rightarrow h) = 1 \quad \text{as} \quad |x| \rightarrow \infty \quad (68)$$

for $P \in \mathbf{P}$, where h is the binary entropy rate (per symbol) proved to be the asymptotic lower bound for compressing a SED source in [54], where SED strings were first singled out as popular models of natural language.

Conversely, an elementary argument shows that (68) implies (67). By nineties, versions of LZ77-78 became everyday tools in computer practice.

Rissanen’s **Minimum Description Length principle (MDL)** published in 1978 and [45], see also [59] for goodness of fit and homogeneity testing, initiated applications of UC to statistical problems for SED sources continued in several recent papers of B. Ryabko with coauthors. Of special interest to us is the **homogeneity test** in [50], where also a consistent estimation of n in approximating SED by n -MC is proposed.

9.1 Homogeneity testing with UC

We always use further binary *txt* encodings of literary texts. Define $|A|$ and $|A_c|$ as the lengths of respectively **binary** string A and its compression A_c .

The *concatenated* string $S = A \cup B$ is the string starting with A and proceeding to string B without stop.

9.1.1 Ryabko-Astola- and U-statistics

If entropy rates of two strings are significantly different, then inhomogeneity proof may be based on their any straightforward consistent estimation. In the opposite case, the Ryabko and Astola homogeneity of two strings test statistic ρ can help which is

$$\rho = h_n^*(S) - |A_c| - |Q_c|, \quad (69)$$

where the empirical Shannon entropy rate h_n^* of the concatenated sample S (based on n -MC approximation) is defined in their formula (6). The local context-free structure (microstyle) of long (several Kbytes) literary texts (LT) can be modeled sufficiently accurately only by binary n -MC with n not less than several dozen. Thus its evaluation for LT is very intensive computationally and unstable for texts of moderate size requiring regularization of small or null estimates for transition probabilities. Therefore, appropriateness of ρ rather than equally computationally intensive Likelihood methods based on n -MC training [47] is questionable. For comparatively short (2Kb) LT, n -Markov approximation may require fitting many hundred times more transition probabilities than the sample size, while for very large LT such as novel affected by long literary form relations (‘architecture’ features such as ‘repeat’ variations), the microstyle describes only a local part of the author’s style as emphasized in [9].

Consider $U(Q, A) = |S_c| - |A_c| - |Q_c|$. Quantity $U(Q, A)$ *mimics the Ryabko and Astola statistic* ρ . In $U(Q, A)$ we *replace their empirical Shan-*

non entropy h^* of the concatenated sample S (based on n -MC approximation) with $|S_c|$ since both are asymptotically equivalent to $h(|Q| + |A|)$ for identical distribution in Q, A with entropy rate h and exceed this quantity for different A, Q .

Test ρ is asymptotically invariant w.r.t. interchanging A, Q and *strictly positive* for *different* laws of A, Q , if $a < |A|/|Q| < 1/a, a > 0$). The last but not the first property would seem to hold also for $U(Q, A)$ in some range of $|A|/|Q|$ due to the straightforward **lower bound for the minimax mean UC-compression redundancy of piecewise-stationary sources** [40], which is logarithmic in $(|Q| + |A|)$.

9.1.2 FAUC

It is proved in [18] that the sliding window size is of order $\log(|Q| + |A|)$ in a version LZ77 of the LZ-compressors is UC. Also, [52] constructed UC with the ‘redundancy price for one jump in piecewise stationary regime’ asymptotically equivalent to the above-mentioned Merhav’s lower bound. Moreover, Ryabko (personal communication) argues that **every UC can be modified in such a way that the FAUC property holds**:

FAUC (Fast Adapting UC) For any given SED P_1, P_2 with equal entropy rates, cross-entropy $D(P_1||P_2) > \varepsilon > 0$, A, Q distributed as P_1 , Q' distributed as P_2 , independent $A, Q, Q', |Q| = |Q'|$ and any $b > 0$, it holds

$$\mathbf{E}[|(A \cup Q')_c| - |(A \cup Q)_c|] = o(|Q|^b).$$

Thus in what follows we can assume that we deal with FAUC. Let us emphasize that the FAUC condition is different from (although related to) small mean redundancy property for SED.

Conjecture. Popular UC used in compressing LT are FAUC.

Claim. The U performance on IID extensive simulations in a large range of $|Q|$ (made recently by NEU PhD student Stefan Savev), was not as predicted above (actually empirical mean of U was negative) due apparently to the additional subtracting of $|Q_c|$. For small $|Q_c|$ this is due to an uncontrollably large ‘adaptation value’ of $|Q_c|$, since ‘entropy’ asymptotics is not yet attained. For large $|Q_c|$, the small increase of U due to inhomogeneity ‘is drowned’ in the large noise of variable $|S_c|$. Averaging different slices of identically distributed moderately large $Q_i, i = 1, \dots$ can make mean U positive, yet this does not seem appropriate in our applications.

9.1.3 CCC- and CC-statistics

Fortunately, another statistic, CCC defined below, overcomes the shortfalls of statistic U .

In our applications $|A|/|Q|$ is large to statistically assess reliability of attribution and upperbounded by an approximate empirical condition $|Q| \geq 2000$ bytes (requiring further study) for appropriateness of SED approximation.

The *Conditional Complexity of Compression* of text B given text A is

$$CCC(Q|A) = |S_c| - |A_c|. \quad (70)$$

$$CCC(B|A) = |S_c| - |A_c|, CCCr(B|A) = CCC(B|A)/|B| \quad (71)$$

The CCC mimics a more abstract conditional Kolmogorov Complexity in our settings and measures how adapting to patterns in the training text helps to compress the query text.

9.1.4 Relation to the Likelihood Ratio Test

Introduce an empirically centered version $CCC'(B|A)$: extract $CCC(B'|A)$ from $CCC(B|A)$, where $B' : |B'| = |B|$ is generated by the same SED as A . Since B' is distributed as A and does not depend on B , homogeneity tests based on CCC and CCC' are equivalent. However, CCC' mimics the conditional version of the Likelihood version homogeneity test as in [59] trained on A .

Conjecture. $CCC'(B|A)$ approximates the most powerful Likelihood Ratio Test of Q, A homogeneity under our condition on sample sizes and validity of SED approximation for both Q, A

9.1.5 Slices of the query text

The only difference of CCC from U is canceling the uncontrollable $|Q_c|$ removal depending on the unknown ‘adaptation’ window of the UC used.

In our case studies we average sliced CCC of text $Q_i, i = 1, \dots, m = \lceil |Q|/L \rceil$, given the firmly attributed text A , dividing the *query text* Q into slices of equal length L . Universal compressors used are the same for all sizes of texts.

$$\overline{CCC(Q|A)} := \sum_{i=1}^m \frac{CCC(Q_i|A)}{m}, CCC(Q_i) = |Q_i|, \quad (72)$$

$$\overline{CC(Q)} := \sum_{i=1}^m \frac{CC(Q_i)}{m}, \quad (73)$$

$$\sigma(\overline{CCC}) = \sqrt{\sum_{i=1}^m (CCC(Q_i|A) - \overline{CCC(Q|A)})^2 / m(m-1)}, \quad (74)$$

$$\sigma(\overline{CC}) := \sqrt{\sum_{i=1}^m (CC(Q_i) - \overline{CC(Q)})^2 / m(m-1)}, \quad (75)$$

$$\bar{\sigma}(Q, Q', \kappa) := \sqrt{\sigma^2(Q, \kappa) + \sigma^2(Q', \kappa)}, \kappa := CC \text{ or } CCC. \quad (76)$$

We call the first two empirical quantities ‘Mean $CCC(Q)$ and Mean $CC(Q)$ ’ respectively. introduce t -statistics for independent Q, Q' :

$$t(CCC(Q, Q'|A)) := |\overline{CCC(Q|A)} - \overline{CCC(Q'|A)}| / \bar{\sigma}(Q, Q', CCC), \quad (77)$$

$$t(CC(Q, Q')) := |\overline{CC(Q)} - \overline{CC(Q')}| / \bar{\sigma}(Q, Q', CC). \quad (78)$$

CC-statistics for Q, A are called insignificantly different at some significance level, if the corresponding t does not exceed its critical value (which practically is chosen around 1.5-2).

Claim. Both our case studies in section 11 and statistical simulation show that the sliced CCC-approach has a good homogeneity discrimination power in this range for moderate $|Q|$ and much larger $|A|$ (see 11.5) in a surprisingly wide range of case studies with **insignificantly varying mean unconditional complexity CC** of compression.

9.1.6 CCC- sample size requirements

Statistical testing of the latter condition is straightforward due to the **asymptotic normality** results of the compression complexity for IID and MC sources described in [56] and normal plots for LT.

The very artificial proof in [56] on around 50 pages and brilliant direct probabilistic proof in [1] on ‘only’ 30 pages (valid only for symmetric

binary IID case) have very plausible extension for CCC which theoretically support a quite **unusual sample size relation** for UC-homogeneity testing: **sample size of the training text must dramatically exceed those of slices of a query text.**

Justification. The training test A being fixed, $VarCCC(Q_i|A)$ of independent copies $Q_i, i = 1, \dots, N$ of the query text Q, are of order of $|Q|$ due to almost renewal-type patterns acquiring and practically finite mean memory size of UC, while the mean increase in $CCC(Q|A)$ redundancy for *different distributions* of Q and A *as compared to their identity* is $o(|(A|Q)|^b)$ for any $b > 0$ for FAUC (accurate upper bound for adaptation period even for LZ78 is absent so far (perhaps, they can be obtained extending LZ-78 upper bounds for redundancy in [51]). Thus, the t-ratio is negligible under the asymptotics $|A| \rightarrow \infty, 0 < \varepsilon < |Q|/|A|$. [25] explains this informally as follows: if the training A and alternative style query text Q sizes are comparable, then two flaws in homogeneity testing happen: a UC adapts to both at the extra length cost $o(|(A|Q)|^b)$ for any $b > 0$, this extra amount of $CCC(Q|A)$ is hidden in the noise with $VarCCC((A|Q)|)$ of order $|A|Q|$. Second, the mean $CCC(Q|A)$ of larger slices of query texts have a **bigger bias** due to **self-adapting of UC to the slices' patterns.**

This makes sample size requirements and symmetry arguments in [10] also based on the conditional compression complexity although **ignoring assessment of statistical stability**, unappealing, and explains examples of CV05 misclassification shown in [46]. It can explain also the roots of early heated discussion around simpler development in [3], where the *sample size relation and statistical stability* issues were not addressed.

9.2 Naive explanation of CCC-consistency on toy example

We study here performance of CCC-attributor in two ways: i. justify it for very distinct distributions of query and training IID sequences and ii. show the results of simulation, when these distributions can be closer. Consider a training binary Bernoulli(1/100) sequence $X_1^{10000000}$ with $P(X = 1) = p = 1/100$ and the query Bernoulli(0.99) sequence Y_1^{1000} with the opposite distribution $P(Y = 0) = 1/100$ and compare the lengths of LZ-compressed sequences $X_1^{1001000}$ and $X_1^{10000000}Y_1^{1000}$. Note that the entropies h of X and Y are the same and thus both belong to $M(h), h = -p \log p - (1 - p) \log p$. Let us support discussion of asymptotic performance of CCC in section 9.1.6 by direct arguments.

The classical von Mises's results state that the number of rare patterns in a Bernoulli(p) sequence of length N consisting of r ones has the Poisson(λ) distribution, if $Np^r(1-p) = \lambda$ for large N (see [14], problem 11.26. The cardinality of patterns is understood there in a slightly different sense which does not influence our argument significantly).

Thus $X_1^{10000000}$ contains only the Poisson(1) distributed number of 111-patterns (i.e. only one such pattern in the mean) and *much less likely patterns with larger number of ones*. The additional length of compressed $X_1^{1001000}$ w.r.t. the length of compressed $X_1^{10000000}$ is due most likely to few occurrences of *large size patterns consisting mostly of zeroes* in the continuation of the sequence.

The length of LZ-compressed file is approximately $c \log c$ bits, where c is the number of distinct patterns in the initial string. The concatenated sequence $X_1^{10000000}Y_1^{1000}$ contains most likely **more than hundred new patterns** w.r.t. $X_1^{10000000}$ *consisting mostly of ones*, and thus the compressed $X_1^{10000000}Y_1^{1000}$ contains hundreds of additional bits w.r.t. compressed $X_1^{1001000}$ most likely.

Remark. Von Mises (see [14], section 13.7) and [56] prove the asymptotic normality of the patterns' cardinality in Bernoulli sequences which agrees with our empirical CCC-normality plots.

A MATLAB simulation (with the code written by D. Malioutov (MIT) using the commercial update of LZ78 for UNIX systems) compared the CCC of I.I.D. binary query strings of length N_2 generated first for the same randomization parameter p_1 as for the training string of length N_1 , and CCC for the second query string with the complementary randomization parameter $p_2 = 1 - p_1$ (*having the same unconditional CC*) in wider range of p_1 than that in our toy example. See tables in [27].

9.3 Methodology

◊ Firmly attributed corpora are usually referred to as *training texts* for training the compressor and the text under investigation is referred as the *query text*. *Query texts* may be disputed ones or those used for estimating the performance of attributors.

◊ In case studies the *equally sized slices* Q_i, T_i of *query and training texts*, $Q, T(k)$ for several slice sizes and calculate the averages over slices $\overline{CC(T(k))}, \overline{CCC(Q|T(k))}$ and their empirical standard deviations for each training text $T(k)$ to analyze authorship with the CCC- analysis. Comparing $\overline{CCC(T(k)|T)}$ of few *query texts* is also used sometimes keeping

Table 5: P-value for Venus vs Hero 1 homogeneity under training on Amores, Venus

	size 10000	size 5000	size 2700	size 2000
trained on Amores				
Venus vs Hero 1	0.00973	0.00113	$1 * 10^{-6}$	$2 * 10^{-10}$
Venus vs Hero 2	0.07148	0.03004	0.00421	0.00334
Hero 1 vs Hero 2	0.0274	0.0057	$6 * 10^{-6}$	$1 * 10^{-8}$
trained on Amores + Venus				
Hero 1 vs Hero 2	0.00671	0.00021	$2 * 10^{-7}$	$5 * 10^{-9}$

the *training text* T fixed.

◊ $\overline{CCC(Q(k)|T)}$ may be interpreted as empirical generalized distance similarly to the cross-entropy which it presumably approximates. Thus we can compare CCC of various training texts for a fixed query text and vice versa. This is as good as for the cross-entropy which is the main asymptotic tool of statistical discrimination.

◊ Equality of slice lengths makes comparison CCC, CC equivalent to comparing their corresponding normalized versions

$$CCr = |A_c|/|A| \quad (79)$$

and similarly defined CCCr.

◊ Empirical study shows that equal slice sizes do not make CCC-resolution essentially worse as a price for simplicity of the procedure. Pattern tables show that the mid-sized patterns make main contribution to the discrimination while the low-sized patterns are equally filled in for competing authors.

◊ A typical table 1 in [27] (displayed here as table 5) shows that the P-values of homogeneity testing are minimal for the minimal slice size we dared to consider: 2KBytes (around half-page). Further studies must show if even smaller sample size can be chosen without losing validity of SED approximation.

P-values evaluated by Normal tables rather than by Student table are somewhat lower-biased.

◊ CC-attribution turns out the same in most case studies when using various popular commercial UC: WinZip, BWT, etc., showing attractive empirical invariance of the CCC-method

◊ LZ-78 was studied in more detail due to its additional application described in the next section. G. Cunningham found that the CCC-resolution of LZ-78 is better when using bytes rather than bits as an alphabet. In case study 11.2 the corresponding t-value is 2.88 versus 1.42 for bits.

◊ If $|\overline{CC(T(k))} - \overline{CC(Q(k))}|$ is insignificant for a slice size than the same turns out true for all larger slice sizes, see figures 7, 13.

◊ Plotting CC values over slices serves for checking insignificance of the entropy rates difference between training and query texts and verifies the homogeneity of the style in the training texts itself.

◊ If Mean $CC(Q)$ is significantly different from Mean $CC(T)$ (which can be established using their asymptotic normality in [56], the author of the training text T is unlikely to be the author of Q . If **Means CC 's of Q and T are not significantly different**, then the smaller the Mean $CCC(Q|T)$, the stronger appears the evidence for the similarities in style between two texts, and we expect Mean CCC to be the smallest if trained on the *training text* written by the author of the query text.

◊ Thus the less is mean **Description Length**, the better is the evidence for authorship.

◊ The assumption of unconditional complexities of query and training texts to be approximately equal is illustrated by the extreme case of a long query text consisting of a repeated identical symbol. Its CCC is smallest for whatever training text.

◊ The program in PERL written by G. Cunningham implementing the CCC evaluation algorithm in [25] is published in [33] which is available from the website of SIBIRCON 2010.

9.4 Follow up analysis of most contributing patterns

Given two bodies of text, our goal is to find patterns that occur in one body significantly more than in the other, and to see what linguistic relevance these patterns have. The patterns we will examine are those that arise during the course of compression using the LZ-78 algorithm.

LZ-78 prepares the following by-product: in a binary LZ-tree of patterns constructed during compression we can evaluate cardinalities $n(\nu, A)$ of subtrees with given prefix ν for training text A : this is the cardinality of paths crossing ν or simply the number of vertices with prefix ν .

G. Cunningham wrote an economic code of LZ-tree storage and $n(\nu, A)$ evaluation using language PERL and algorithm from [27]. The ν -subtree is called **interesting** if its ‘t’-value is large for competing training texts A, A' .

$$‘t’ := \frac{|n(\nu, A) - n(\nu, A')|}{\sqrt{n(\nu, A)(c_1 - n(\nu, A))/c_1 + n(\nu, A')(c_2 - n(\nu, A'))/c_2}}, \quad (80)$$

where $c_i, i = 1, 2$, are total pattern cardinalities for A, A' which are well CCC-discriminated. Finally, symbols of English corresponding to the highly interesting patterns are tabulated.

[1] proves asymptotic independence of $n(\nu, A)$ for nonintersecting ν and different A which explains the meaning of our ‘t’. However, their vast multiplicity makes evaluation of statistical significance hard. Nevertheless, the tables of interesting patterns may contain unexpected ones for linguist studying comparative styles of competing authors.

To generate meaningful statistics, we divide each file F into *slices* F_1, \dots, F_n of a given number of bytes s . That is, F is the concatenation $F_1 \cup \dots \cup F_n$. Now we run the LZ-78 algorithm on each slice, making note of the maximal patterns: the entries in the dictionary that are not the prefix of any other dictionary entry. Since we represent the dictionary as a binary tree, these maximal patterns are exactly the leaves of the tree. Then we calculate a t-value for each pattern.

The asymptotical independence of $n(\nu, A)$ for disjoint patterns ν and different A is proved in [1] using the ‘Poissonization’. This explains the meaning of our ‘t’. However, the vast abundance of patterns makes evaluation of statistical significance hard. Nevertheless, the tables of interesting patterns may contain unexpected ones for linguist studying comparative styles of competing authors.

Here is a detailed description of the algorithm:

1. For each of the two files $F_i, i = 1, 2$:
 - (a) Create a histogram H_i , initializing it so that for each bit pattern ν we have $H_i(\nu) = 0$.
 - (b) Split the file into slices of the given size.
 - (c) For each slice:

- i. Run the LZ-78 algorithm on the slice, building a binary tree of patterns which is the dictionary.
 - ii. Then, for each maximal pattern ν of the dictionary (leaf of the tree), increment $H_i(\nu)$ by 1.
2. Define $c_i = \sum_{\nu} H_i(\nu)$.
 3. For each bit pattern ν in either H_i :
 - (a) Define $n_i(\nu) = \sum_{\mu} H_i(\nu \cup \mu)$, the number of times ν was a maximal pattern or the prefix of a maximal pattern in F_i .
 - (b) Compute the t value (note slight change of notation as compared to (80))

$$t = \frac{n_1(\nu) - n_2(\nu)}{\sqrt{\frac{n_1(\nu)(c_1 - n_1(\nu))}{c_1} + \frac{n_2(\nu)(c_2 - n_2(\nu))}{c_2}}} \quad (81)$$

- (c) Decode the bit string ν to a character string w (ASCII, UNICODE, etc., depending on the input).
- (d) Write ν , w , and t to a file.

9.5 Results

Three 100 KB files were randomly generated bit-by-bit, each bit independently of the previous bits. Each bit in file F_0 had a 90% chance to be a 0, while each bit in files F_1 and F_2 had a 10% chance to be a 0 with same entropy rates. When the CCC program was run using 4 KB slices, we found that $t(CCC(F_0, F_2|F_2)) = 147.6$ and $t(CCC(F_1, F_2|F_2)) = 0.6$.

Next, we ran the program to find the well-distinguished maximal patterns between F_0 and F_1 .

F_0 v. F_2	
Bit pattern	Absolute t-value
11111111	131.27
00000000	131.26
0000000000000000	72.04
1111111111111111	71.55
11111111111111111111	44.05
00000000000000000000	43.95
00010000	34.3308297
11101111	34.27392504

<i>F₁ v. F₂</i>	
Bit pattern	Absolute t-value
01111111111111111111111111111111	3.00
11111111111111111111111111111111	3.00
1001111111011111	2.89
1110110111111101	2.89
1011110111111111	2.76
1111111011110101	2.71
1111010110111111	2.71
1101111111111111111111110111111111	2.65
1111111101111111111111011111111111	2.65

Now we look at the CCC and distinguished pattern results for certain groupings of the Federalist Papers. We worked here only with Federalist Papers that are well-attributed. The file H is the concatenation of Federalist Papers 1, 6-9, and 11-13, all of which are well-attributed to Hamilton. The file M_1 is the concatenation of Federalist Papers 10, 14, and 37-40, while M_2 is the concatenation of Federalist Papers 41-46, all of which are well-attributed to Madison. We ran the CCC program using 4KB slices, and we found that $t(CCC(H, M_1|M_1)) = 4.9$, while $t(CCC(M_2, M_1|M_1)) = 0.2$. Below we summarize the well-distinguished maximal patterns, using the same files and 4KB slices.

<i>M₁ v. H</i>	
English string	Absolute t-value
lwoul	3.87
ouldl	3.78
pon	3.71
ent	3.31
uldl	3.27
isl	3.07
de	3.01
atel	3.00
onve	3.00
y the	3.00
tion t	3.00

M_1 v. M_2	
English string	Absolute t-value
N	3.68
rei	3.32
Stat	3.32
E	3.29
be_	3.26
_State	3.16
ue_	3.05
he St	3.00

10 Brief survey of micro-stylometry tools

We review only *context-free statistics of texts* taking aside also methods based on grammar. Context-free attributors are equally applicable to any language, even to the encoded messages which are not yet decoded such as wiretapped terrorist oral communications in possibly unknown or encoded language which can be used e.g. in tracking target terrorists in automatic regime or for fast discovery of abrupt change in stationary functioning of a system such as financial market or of user profile in a computer server. However, these methods are not always robust w.r.t. spelling errors and their resolution power may be inferior to semantic attributors.

One obstacle for implementing these methods is the *evolution and enrichment of styles* during professional careers of writers. So, unless we perform an analysis of the stylistic features of authors across time, we can only compare texts written at around the same time.

Also, authors can work in different *literary forms* (for instance, prose and verse) which may have different statistical properties. Therefore, appropriate *preprocessing* and segmentation into homogeneous parts must be applied to the texts to avoid heterogeneity of forms. Lack of plays segmentation delays their CCC-attribution. Misspellings and names need to be removed to preserve consistency. *Annotated texts* (e.g. verses with stressed vowels indicators) can be more useful resources for computer analysis than bare texts. Finally, reliable stylometry analysis should take into account all available information about a query text (e.g. *time of its preparation*).

The pioneering stylometric study [38, 39] was based on histograms of word-length distribution of various authors computed for 5 different text strings of length 1000 words from each author. These papers showed

significant difference of these histograms for different languages and also for some different authors (Dickens vs. Mill) using the same language. At the same time, histograms of Dickens were close to those of Thackeray in terms of their statistical variability estimated from repeated samples.

The second his paper describes the histograms for Shakespeare contemporaries commissioned and funded by A. Hemminway. This study demonstrated a significant difference of SC-histogram from those of all (including F. Bacon) contemporaries studied but one, calling attention to the striking practical identity of C. Marlowe's and SC histograms (The 'morning star' of Elizabethan poetry and drama Marlowe allegedly perished in May 1593 being 29 years old on bail from English inquisition awaiting the imminent death penalty, under extremely suspicious circumstances (see e.g. [44]), two weeks before the dedication was amended into the already published anonymously submitted poem claiming it to be the very first 'invention' of SC. This identity was shown by evaluating partial histograms for certain portions of the corpora studied and comparing their inter- and intra-deviations.

In an unpublished honors project (available by request), S. Li used a slight modification of Mendenhall's method for attributing popular poem ' 'Twas the night before Christmas' to H. Livingstone rather than to its official author C. Moore supporting the claim in Foster (2000).

Another distinction between the authors is in the numbers of English words they used: 8000 in Bacon's works vs. 31500 in SC including 3200 invented ones in SC which is more than the Bacon's, Jonson's and Chapman's joint contribution. However, G.M. Kruzhhkov in personal communication studied dynamics of acquiring new words in SC reinventing the well-known Heaps law. He shows that this **rate** in the SC is not the largest among his contemporaries. [57] attribution of a newly discovered poem ' Shall I die...' to SC implicitly presumes identity of rates of acquiring new words and forgetting others. Thus their approach appears questionable. Similar approaches based on Zipf and Heaps laws parameter estimation [37, 19], proved to discriminate languages but have insufficient resolution for discriminating authors.

Next to mention is the *Naive Bayes* (NB) classifier of [43] developed during their long and very costly work over the authorship attribution (Madison vs. Hamilton) of certain *Federalist papers* generously supported by the federal funding. After fitting appropriate parametric family of distributions (Poisson or negative binomial), they follow the Bayes rule for odds (*posterior odds is the product of prior odds times the likelihood ratio*), when multiplying the odds: Madison vs. Hamilton, by the sequence

of likelihood ratios corresponding to the frequencies of a certain collection of relatively frequent function words, obtaining astronomical odds in favor of Madison.

*This classifier presumes independence of function words usage, which is obviously **unjustified*** and ‘NB-likelihoods’ should not be taken seriously. Among many NB-applications is [23]’s attribution of Moliere plays to Corneille, attribution of parts of ‘Edward III’ to SC and Fletcher, and sorting out spam e-mails [11].

Some more popular attributors emerging after NB-approach based on the SVM (see [5]) and modeling language as an n-MC, see [47] should be also mentioned. Both are very computationally intensive, depend on choosing many parameters in a non-formalized way which makes their performance evaluation a hard problem.

Chronology of Plato works is studied with a new attributor based on sequences of long and short among last 5 vowels and survey of previous approaches to this problem are in [12].

An informal description in [13] of around 20-years-long study of several hundred ad hoc attributors (‘modes’) by many undergraduate students at Claremont McKenna College, CA, generously supported by the NSF, led them to choosing around 15 best among attributors for distinguishing SC from other corpora of that time. Elementary combinatorial arguments show that the statistical reliability of their choice is questionable due to the astronomical number of possible choices. It is also not clear whether the preprocessing design (removing names; comparing works written in around the same time to avoid evolution of style influence, etc., shown to be crucial in previous studies), correct evaluation of statistical significance of multiple decisions, and other important issues **were apparently ignored**. These concerns do not let us assess reliability of their inference so far although their segmentation of plays of Shakespeare’s contemporaries is useful.

Skipping discussion of other attributors, we move directly to the CCC-attributors which demonstrated even better performance *in average* in certain applications (apparently first introduced by Dmitry Khmelev in a not easily reachable Russian Proceedings, reproduced in one of Appendices to [22] before its tuning and improvement in [25, 27]. [10] gives a survey of numerous previous approaches: the classification and clustering of text libraries of comparable size using ‘similarity metrics’ mimicking information distances of Bennett et al inspired by analogy with Kolmogorov complexity (KC) and replacing KC by commercial universal compressors

satisfying certain properties. Symmetry of distance was an issue in these papers while **statistical assessment of attribution was completely ignored**, see [24, 3]. Moreover, the preprocessing stage seems to be missing in [10] which may explain their paradoxical claim that L. Tolstoy's work stays separately in the tree of the classical Russian authors. They apparently *forgot to remove substantial installments of French in L. Tolstoy* which has a different entropy rate.

The main distinction of our method of all the previous approaches is our compression of *many slices of the query text* enabling an **applied statistical analysis** of their conditional complexities in terms of their location centers and spread. In this way we can judge about statistical significance of mean CCC-differences similarly to [38, 39]. We show in [27] that **NCD-attributor of [24] fails** in attributing authorship cases, which are **significantly discriminated with our CCC-attributor**.

Since distribution of CCC over slices was only *empirically* established, we illustrated our results in [27] mainly by convincing histograms and plots showing consistency and approximate normality of the CCC-attributor. The latter was also used there for evaluating the P-value of attribution.

Finally, [15] and [20] apply alternative stylometry attributors to works of the 'author' we study in 3.4. [15] use a version of [42] attributor (vigorously criticized in [34]) to show implausibility of Sholokhov's authorship of a Nobel prize winning novel. The same method was used by their son to show that the 'History of Russia' by M. V. Lomonosov was falsified by G. Mueller. This claim is difficult to check because texts of that time are of questionable attribution. [20] showed that the mean sequence sizes in Sholokhov and his rival in authorship F. Kryukov proposed in [55] are significantly different. [2] agrees with [20] in implausibility of Kryukov's authorship using literary arguments. [35, 36] used very elaborate linguastatistical procedure to conclude that this disputed novel was apparently written by A. Serafimovich among their other attributing etudes.

11 Attribution of literary texts

The asymptotics supporting our approach for large samples by modeling language as a SED process may have dubious accuracy for moderate samples and **should be supported by attributing many literary texts with known authorship**. These are sections 11.1-2 and 11.5 (to some extent). Other sections deal with cases with unknown or disputed attribu-

tion. Sections 11.1-4 were processed by S. Brodsky with winzip, sections 11.4-5 – by I. Wickramasinghe with the same UC, the first examples in 11.5 were earlier processed by S. Li with BWT, see [25], 11.1-2 were also processed by G. Cunningham with LZ-78.

We use our sliced CCC-attributor (mimicking the original Kolmogorov’s idea), verifying insignificant variability of unconditional complexity and validity of CCC-Normal approximation in every case study.

11.1 Two translations of Shakespeare sonnets

In his pioneering statistical linguistics works, A.N. Kolmogorov singled out three sources of texts variability: their *information content, form and unconscious author’s style*. Since we are only interested in the latter, testing resolution power of attributors for texts with identical content and form is of special interest to us.

Thus around 20 different professional translations of the Shakespeare Sonnets into Russian is a good material for comparing designs of text preprocessing and attributors. Of course, the identity of the information content causes certain statistical *dependence between slices’ CCC* which makes the analysis accuracy slightly worse than for independent slices: the t-criterion of the mean difference between CCC has twice less number of degrees of freedom than that for independent slices.

Good compression of the Sonnets’ classical translations by Gerbel and Marshak due to a **regular structure** of text (about four times) can at the same time worsen CCC-discrimination for improper preprocessing designs. This was shown by comparing the CCC-attribution for two designs: without and *with preprocessing removing the regular structure*. The latter has larger t-value, although both are significant.

The standard preprocessing design: removing verse form-based carriage returns, spaces and capital letters, and cutting whole sonnets’ file into 70 equally sized slices of size 20006 bytes was applied.

We compared the inter-CCC $Inter_i$ of each slice **trained on the total alternative translation of all sonnets** with the intra-CCC $Intra_i, i = 1, \dots, 70$, of the same slice under alternative translation **trained on the remaining part of the same translation**.

The significant correlation between intra- and inter-CCC of slices (Pearson’s r is slightly less than 0.3) is due to the identity of the information content in both translations.

The ‘matching pairs’ t-test between 70 measurements is

$$t = (\bar{I}nter - \bar{I}ntra)/s_d = 4.69, s_d = StD(Inter - Inter). \quad (82)$$

Unconditional CCr for both translations are insignificantly different and only slightly less than 1/2.

Details can be sent by request.

Thus the translations by Gerbel and Marshak are firmly discriminated with CCC under standard preprocessing.

Approximate normality is seen from the normal plots of Mean inter- and intra-CCC in Figure 1.

11.2 Two books of Isaiah

A CCC-comparison of the two books Isaiah1 and Isaiah2 from the Bible was suggested by Prof. J. Ziv, Technion, who kindly connected us with Prof. M. Koppel, Bar Ilan University. Prof. Moshe Koppel provided us with the MS-Word files in Hebrew (which were converted by us into txt-files using Hebrew MS-Word routine) and informed us that in the 1980’a an Israeli named Radday did quite a bit of research on the Isaiah question and related Biblical authorship questions.

Our routine preprocessing including removing names and unnecessary punctuation was made by Andrew Michaelson, an alumnus of Maymonides Hebrew school, Brookline, MA, currently a NEU PhD student.

Using the same method as in previous section, G. Cunningham found:

i. the compressed length $|A_c^1|$ of Isaiah1 using LZ78. (201458 bits when reading octets, 284863 bits when reading bits instead).

ii. For each of 21 2Kb-sized slice Q of Isaiah1, the compressed length of (Isaiah1 with deleted Q).

iii. For each of 16 2Kb-sized slice P of Isaiah2, the compressed length of (Isaiah1 augmented by P).

iv. The intra-CCCs by subtracting ii. from $|A_c|$ of Isaiah1 and inter-CCC by subtracting $|A_c|$ from iii..

v. The empirical intra and inter means and variances separately.

vi. The t-values evaluated as in the previous section turned out to be 2.88 when reading octets with LZ-78; and 1.42, when reading bits instead. We regard the first one as better corresponding to the techniques used in our other applications.

Normal P-P Plot of VAR00001

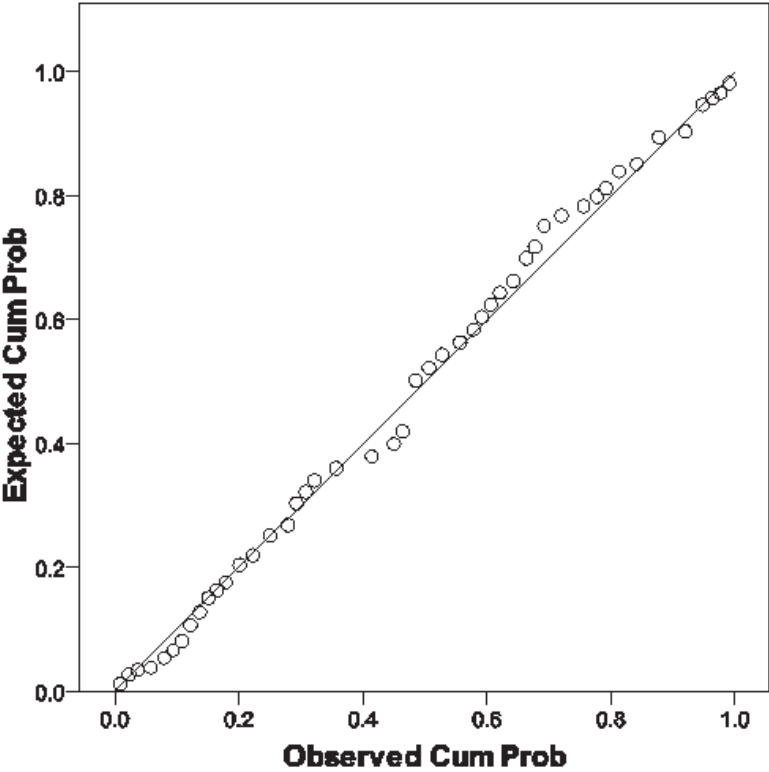


Figure 1: (a) Normal Plot: inter-CCC(slices of Gerbel |whole Marshak).

Normal P-P Plot of VAR00001

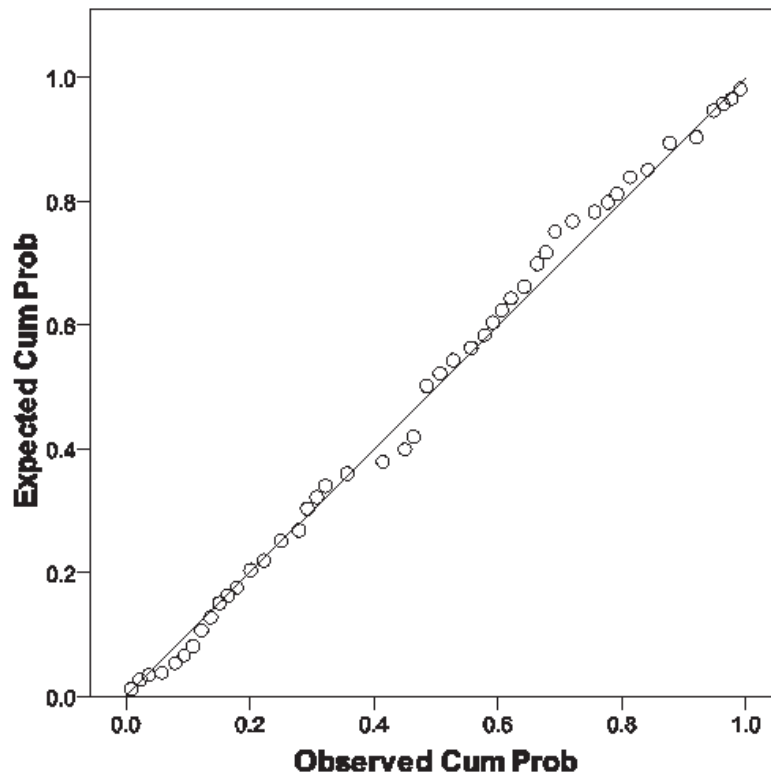


Figure 1: (b) Normal Plot: intra-CCC(slices of Marshak|remaining Marshak).

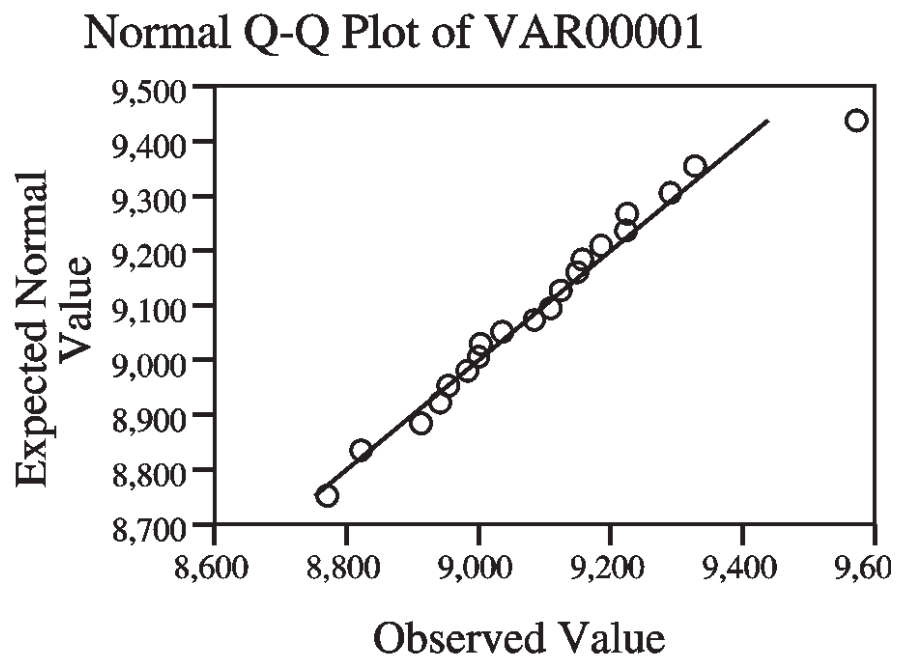


Figure 2: Normal Plots when reading octets with LZ-78: (a) intra-CCC(slices of Isaiah1 2|remaining Isaiah1).

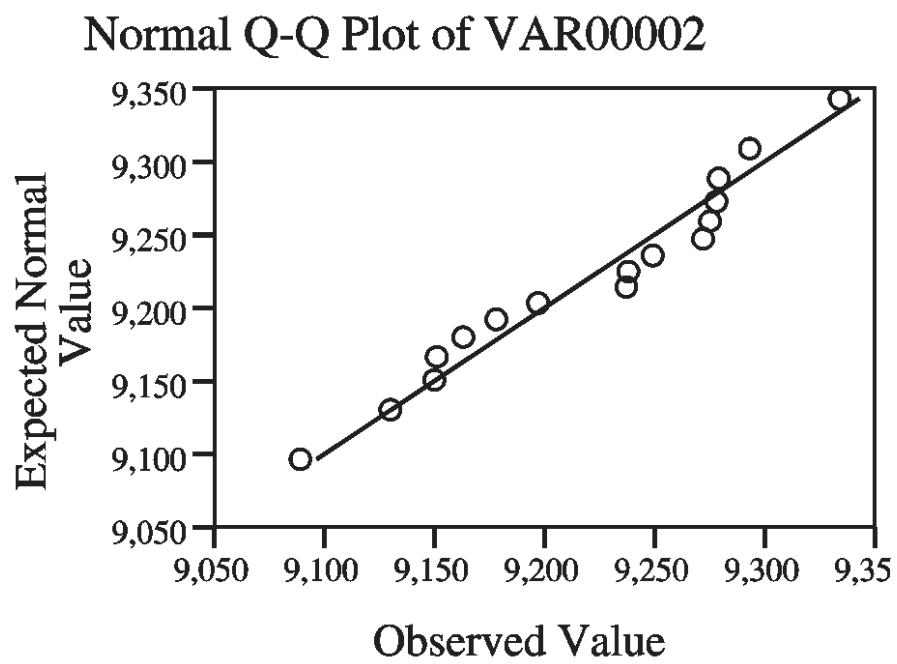


Figure 2: Normal Plots when reading octets with LZ-78: (b) inter-CCC(slices of Isaiah 2 |whole Isaiah1).

11.3 Two novels of the same author

The first short novel [7] describes what happened to him at the end of 20th century while in [8] the same author tries to mimic a boy of twelve's language telling about events in the middle 20th century. Thus styles of these two short novels are **intentionally different**. Nevertheless, the mean CCC-difference between inter- and intra CCC's for these two works of **the same author** is negligible (t-value is 0.062) in spite of a larger size of two works compared: 144 slices of 2000 bytes. This lack of discrimination (**t-value grows roughly as the square root of the number of slices, if other parameters are fixed**) shows that the CCC-analysis can successfully resolve authorship problems. The Pearson's $r = 0.058$ is not significant showing practical independence between the Inter and Intra CCC's. Approximate normality is seen from Figure 3.

11.4 Inhomogeneity of early Sholokhov's novel

The first Sholokhov's short novel 'Put'-dorozhen'ka' (abbreviated further as 'Put') was published first from 25/04/1925 till 21/05/1925 in Moscow newspaper 'Young Leninet's', issues 93-97, 99, 101-104, 106-114 (*at his age of 20*) and reprinted many times since. He had less than 4 years of primary war time education mostly at a Don village and brief Rostov 'prodnalog collecting' accounting courses. During his subsequent service, he was imprisoned on corruption charges for a short time and freed after allegedly forging his age as two years less to be immune from conviction. He soon flees the devastated Don for Moscow in late 1922. There he was employed for a considerable time by a senior secret police officer Mirumov involved in Stalin's program of '*preparing proletarian writers*'. Mirumov met and 'befriended' Sholokhov first most likely during their previous meeting in Rostov. According to [2], he might give Sholokhov (or the team of looters he belonged to, apparently headed by A. Serafimovich under the auspices of the secret police) manuscripts of a talented dissident Rostov newspaper editor, author of numerous articles and two short novels under the pseudonym Victor Sevsky from the circle of a famous poet Balmont. V. Sevsky was caught and liquidated by bolsheviks in Rostov jail (apparently in 1920). Sholokhov publishes the first short story in late 1924. He leaves Moscow for his native village the same year, marries a daughter of a semi-professional writer and stays there for several years (interrupted by his comparatively short visits to Moscow where he occasionally lives in Mirumov's apartment and devotes to him his first works) preparing 'Put', other 'Don stories' and by 1927 the first two parts of the famous novel 'Quiet flows Don' which

Normal P-P Plot of VAR00003

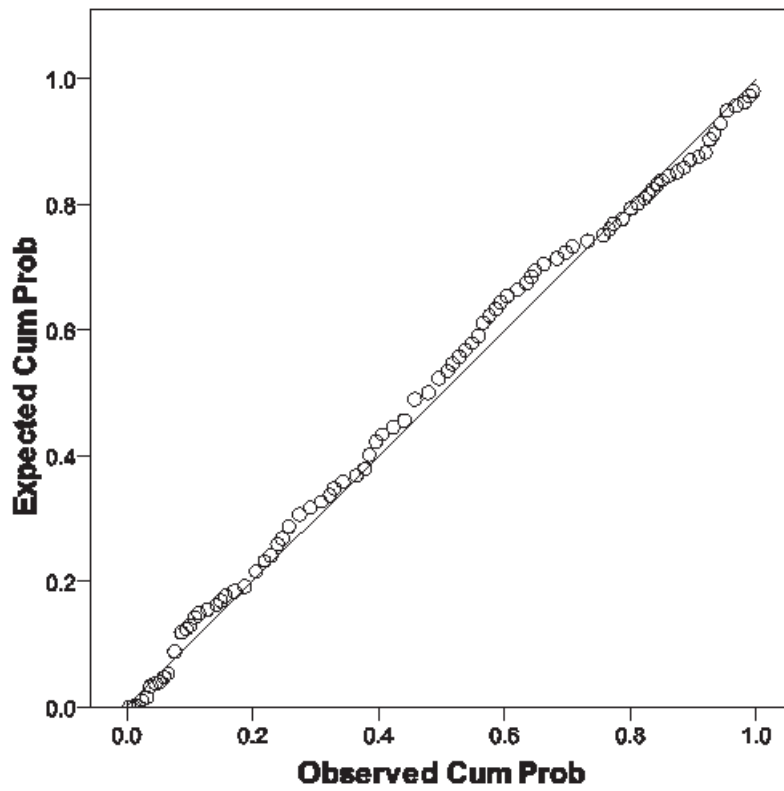


Figure 3: (a) Normal Plot: inter-CCC(slices of Brodsky 1 |whole Brodsky 2).

Normal P-P Plot of VAR00004

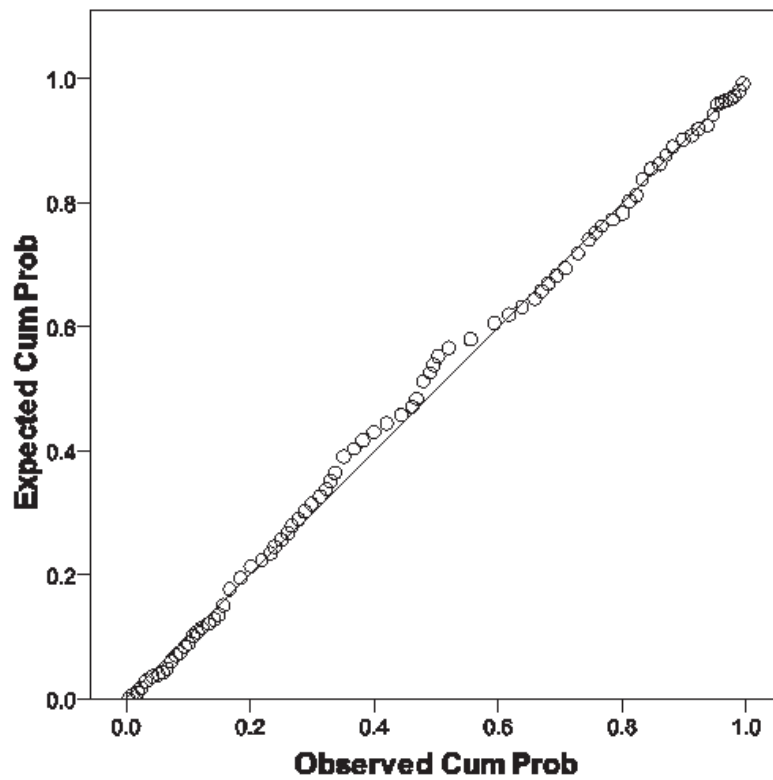


Figure 3: (b) Normal Plot: intra-CCC(slices of Brodsky 2|remaining Brodsky 2).

earned him later the Nobel prize. Any evolution of his style between the first and second parts of ‘Put’ of 12 pages each is extremely unlikely.

After removing names, we partitioned each part into 30 equal slices of 2000 bytes. The Mean Unconditional Complexities (\bar{C}) are statistically the same. The mean ‘Leave one out’ (intra)-CCC in each part was compared with the mean inter-CCC of each slice trained on another part. Their Standard Deviations are not significantly different. **The difference between Mean inter-CCC and mean intra-CCC turned out to be highly significant (exceeding four its Standard Deviations).**

Details are as follows: we computed thirty inter-CCC(slice of Part 2|whole Part 1) and thirty intra-CCC(slice of part 1|remaining part 1).

Mean inter CCC: $M_1 = 576.77$ Mean intra CCC: $M_2=559.43$, their difference is 17.34, StD (Mean inter CCC)= $s_1 = 2.49$, StD (Mean intra CCC)= $s_2 = 3.50$, finally, the StD of $M_2 - M_1$ is

$$s_d = \sqrt{(s_1^2) + s_2^2} = 4.30. \quad (83)$$

F-ratio < 2 permits using two-sample t-test with test statistic

$$t = (M_2 - M_1)/s_d = 4.03. \quad (84)$$

These large t-value and number 58 of degrees of freedom make the corresponding P-value, (i.e. the Student probability of this or higher CCC-deviation) of order 10^{-4} .

Remark. *In our computations we consider inter-CCC of different slices independent which seems to be a reasonable approximation. Intra-CCC may have slight correlation (for example, the sample correlation coefficient between first and last fifteen Part 1 intra-CCC’s is 0.156). However, we believe that this correlation would not have a significant impact on the t-value, see [41]. Figure 4 shows the Approximate Normality of the CCC’s distribution.*

Our two-sample t-test evaluation suggests that the two parts were written by different authors. This context-free conclusion coincides with that of unpublished sophisticated literary analysis of Bar-Sella.

Let us emphasize that **our analysis and unpublished literary analysis of Bar-Sella are based on different features of the text and thus complement and support each other.**

Normal Q-Q Plot of VAR00002

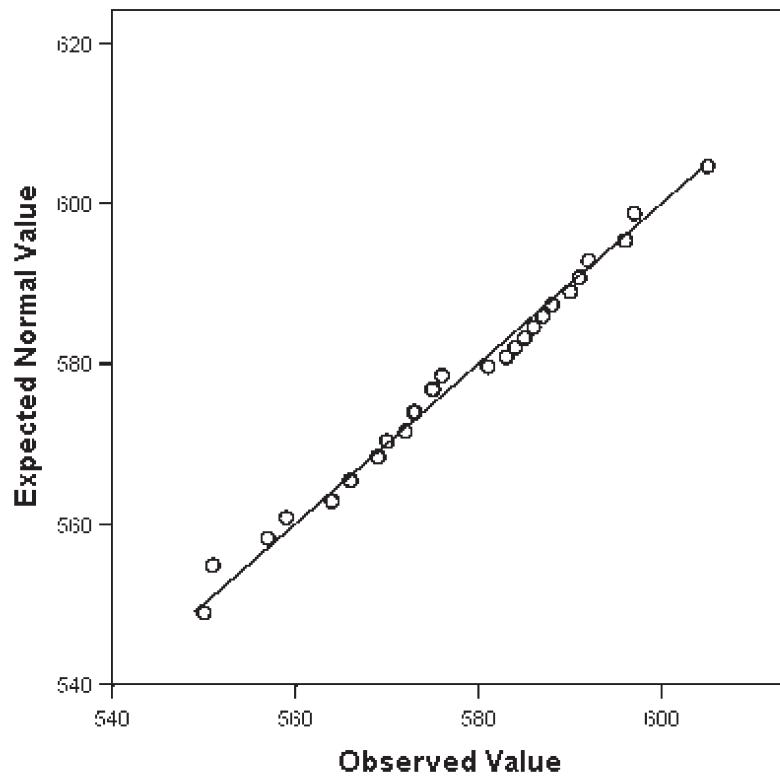


Figure 4: (a) Normal Plot: inter-CCC(slices of Part 2|whole Part 1).

Normal Q-Q Plot of V78

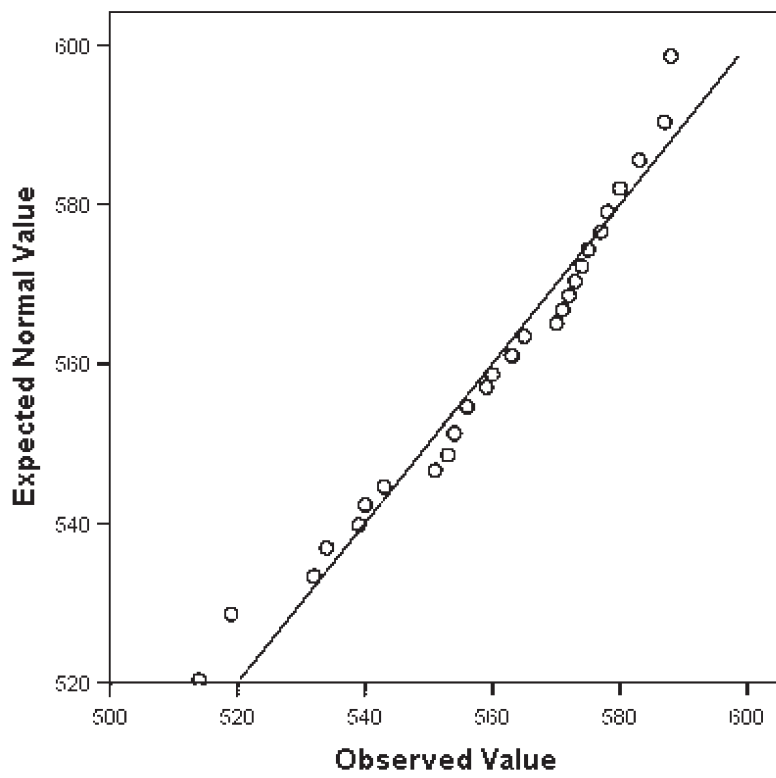


Figure 4: (b) Normal Plot: intra-CCC(slices of Part 1|remaining Part 1).

11.5 Attribution of Federalist Papers

11.5.1 The Federalist Papers

The Federalist Papers written by Alexander Hamilton, John Jay and James Madison appeared in newspapers in October 1787-August 1788 for persuading the citizens of the State of New York to ratify the U.S. Constitution. Seventy seven essays first appeared in several different newspapers all based in New York and then eight additional articles written by Hamilton on the same subject were published in a booklet form. Since then, the consensus has been that John Jay was the sole author of five (No. 2-5, No. 64) of a total 85 papers, that Hamilton was the sole author of 51 papers (Hf), that Madison was the sole author of 14 papers (Mf, No. 10,14,37-48) and that Madison and Hamilton collaborated on another three (No. 18-20). The authorship of the remaining 12 papers (Df, No. 49-58, 62,63) has been in dispute; these papers are usually referred to as the *disputed papers*. It has been generally agreed that the *Df-papers* were written by *either Madison or Hamilton*, without consensus on particulars. [43] and other stylometry attributors gave all Df's to Madison.

Our goal was answering the 3 questions: 1. Is CCC-attribution significant agreeing with previous decisions and also attributing all Mf to Madison? 2. What slice size is around optimal? 3. What training text size is sufficient?

Answers are: 'yes' on first questions: all Mf and all Df were attributed to Madison, all Hf were classified as Hamilton's, 2. minimal slice size 2Kb among the tested ones provides the best discrimination, 3. 13 Mf papers together with Madison's Helvidius papers of the total size around 280Kb is sufficient for significant attribution. Detailed tables are in [58]. **We present a typical tiny sample** of those results.

11.5.2 Training on one of Mf-papers

was not sufficient for reliable attribution. Namely, figure 1 shows that trained on one of Mf, the Mf- and Hf- CCC plots are overlapped inside their confidence intervals for all slices sizes.

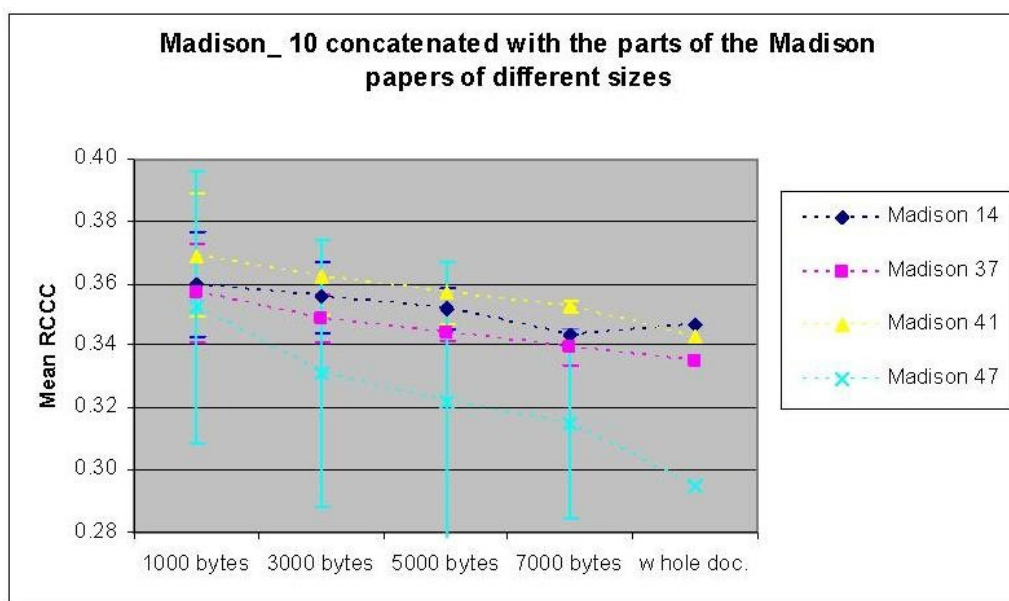


Figure 5: (a) CCC's when trained on one paper.

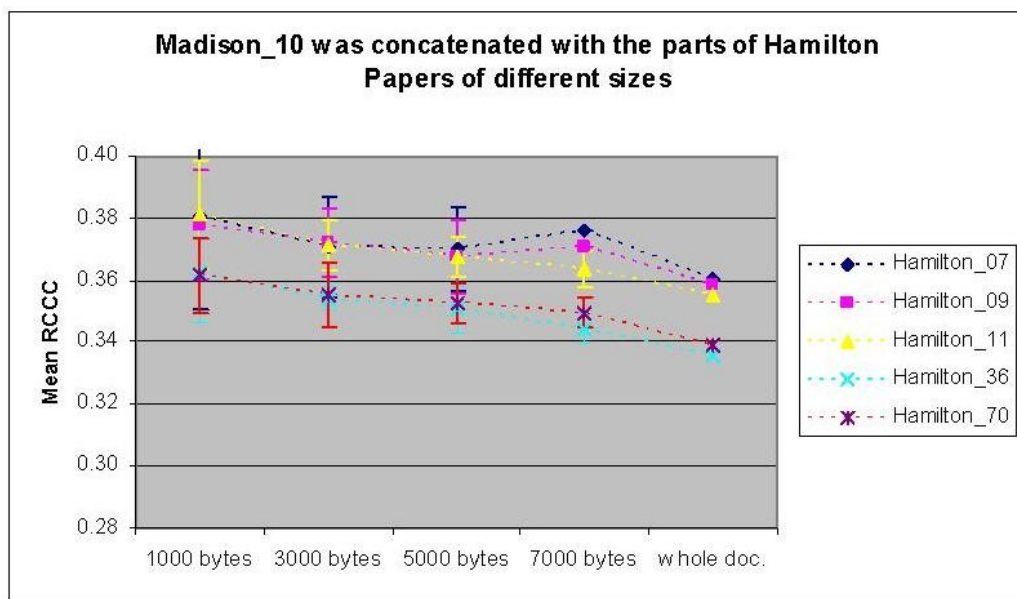


Figure 5: (b) CCC's when trained on one paper.

11.5.3 ‘Leave one out’ Mf as *Training Text*

Four of five Mf-essays (No. 10, 14, 37, 41, 47) were combined leaving one out. Five documents of the size of about 62,000-72,000 bytes after preprocessing were obtained. Our *query texts* are 12 disputed, 5 of the Hamilton essays (No. 07, 09, 11, 30, 70), 2 more of Madison essays (No. 46, 48) as well as the other Madison paper we left out when we combined them as the training set. Figures 5a-b show small variability of mean unconditional complexities and smaller mean CCC’s for query Mf than those of Hf.

- We *trained* the compressor separately on each of four-tuples
- We applied the compressor on the concatenated file xy_i , where $x \in M$ and $y \in \{\text{disputed papers}\}$ and y_i is the i^{th} part of the essay y
- We carried out this study by dividing the *disputed papers* into file sizes of 2000, 3000, and 5000 bytes.
- Mean *CCC* of disputed papers were compared with that for *query Mf- and Hf-texts* by evaluating P-values for the null hypothesis that the mean CCC of the latter are not more than those of a disputed paper.

We show only one typical table 3 [58] giving the P-values for the two sample t-test: $CCC(Df49|MF_i) \geq CCC(A_j|MF_i)$ trained on five choices of combined four Mf indicated in its columns with A_i indicated in its rows, when the slice sizes of the *query text* are 3 Kbytes.

The next figure 6 shows that *CCC*’ empirical distributions are close to Normal.

This and other numerous tables in [58] show insignificant difference of Df’s unconditional and Mf’s conditional complexities, and attribute most of Df to Madison. Still some exceptions require larger training text.

11.5.4 Training on all Mf and more Madison papers

We use here the same technique as before to study attribution for larger *training text*. The following documents were obtained by concatenating all the Mf’s leaving only one out. We combined the essays in ascending order of the number of the paper. The federalist papers used are No. 10, No.

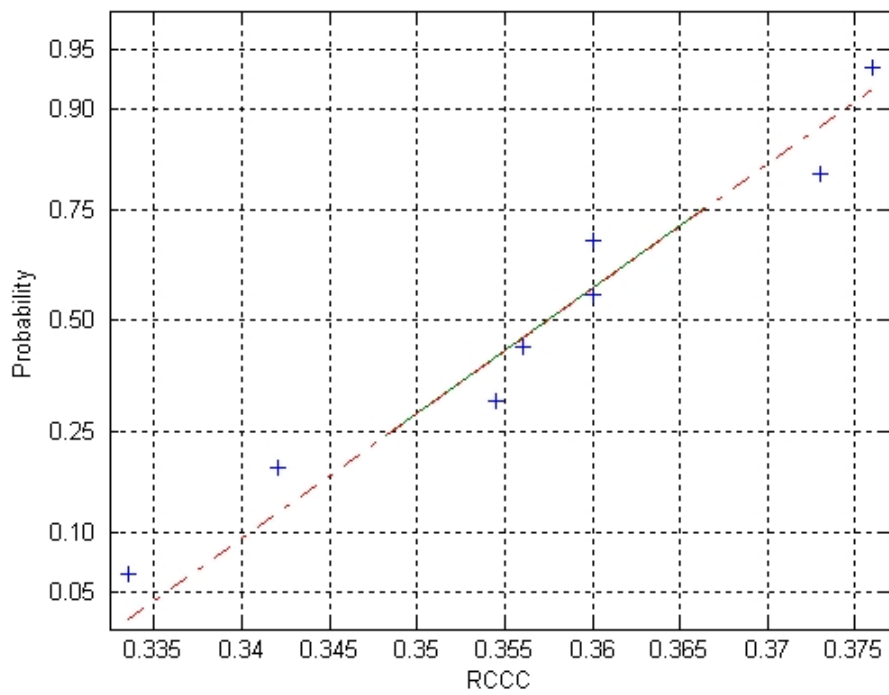


Figure 6: Normal Probability Plot of CCC for slices of Hamilton No. 70 of size 2000 bytes trained on Madison essays No. 10, No 14, No. 37 and No. 47.

Table 6: P-value of the two sample t-test
for *disputed paper No. 49*

Other documents	(a)	(b)	(c)	(d)	(e)
Hamilton 07	0.027	0.019	0.018	0.024	0.014
Hamilton 09	0.063	0.065	0.084	0.079	0.077
Hamilton 11	0.010	0.009	0.018	0.021	0.015
Hamilton 30	0.011	0.012	0.027	0.033	0.025
Hamilton 70	0.010	0.053	0.073	0.078	0.069
Madison left-(L)	0.209	0.078	0.200	0.157	0.141
Madison 46	0.215	0.486	0.373	0.371	0.433
Madison 48	0.177	0.342	0.378	0.383	0.400

14, No. 37 - No 48 which are written by Madison. Sizes of the *training text* varied from 208,000 to 216,500 bytes.

- (a1) : Concatenate all except No. 10
- (a2) : Concatenate all except No. 14
- (a3) : Concatenate all except No. 37
- (a4) : Concatenate all except No. 38

and so on.

The following Madison's documents written between 1787-1793, to avoid evolution of the author's style were used to enlarge *training texts*

- (s) : Concatenated four papers out of five (Number 1-4) called "Helvidius papers", written in reply to series by Hamilton called "Pacifcus papers" (24 Aug. - 14 Sep. 1793) on executive powers
- (t) : Concatenated eight papers from 1791-1792 Congress and republican opposition : (Mad 1 : Population and Emigration, *National Gazette*, Nov 21, 1791), (Mad 2 : consolidation, *National Gazette*, Dec. 5, 1791), (Mad 3 : Universal Peace, *National Gazette*, Feb. 2, 1792), (Mad 4 : Government of the United States, *National Gazette*, Feb 6, 1792), (Mad 5 : Spirit of Governments, *National Gazette*, Feb 20, 1792), (Mad 6 : A Candid State of Parties, *National Gazette*, Sep 26, 1792), (Mad 7 : Fashion, *National Gazette*, March 22, 1792), (Mad 8 : Property, *National Gazette*, March 29, 1792)

For collections of total size around 280Kb consisting of (a1)-(a4) and (s) of size 71,010 bytes as *training text*, **all mean CCC for Mf's and Df's are significantly lower than that of Hamilton**, making the attribution of Df certain.

11.6 Shakespeare controversy

11.6.1 Introduction

The controversy concerning authorship of the works ascribed to W. Shakespeare dates back several centuries due to the fact that rare documents related to his life are hard for many to reconcile with his authorship (see e.g.

<http://shakespeareauthorship.org/>).

Many jewels of English poetry, prose, statesmen, scientists have not accepted the official authorship version. Almost 2000 influential writers, scholars, actors, theater directors, etc. continue to be non-believers and signed the declaration of reasonable doubt to the British government demanding funding for studying the problem, see <http://www.doubtaboutwill.org/declaration>.

A **bibliography** of material relevant to the controversy that was compiled by Prof. J. Galland in 1947 is about **1500 pages** long (see [17]). A comparable work written today might well be at least several times as large. A substantial part of research moved to the Internet, since publishing works contradicting the official version in academic journals is practically unlikely.

The main problem for 'heretics' is that they do not agree on the alternative author. The most popular alternative candidate seems to move presently to be Christopher (Kit) Marlowe, sudden death of whom is disputed (see up to ten large books appeared in the last few years on this subject). The Hoffman's prize of around 1 000 000 English pounds awaits the person who will prove Marlowe's authorship of substantial part of the SC.

11.6.2 CCC-attribution of some Elizabethan poems

We studied the following versions of poems with corrected spelling errors:

- SC: Venus and Adonis (1593), Rape of Lucrece (1594) (we refer to these as Venus and Rape in this study).

- Kit Marlowe's: translation of Ovid's Elegies (Amores).
- Kit Marlowe's: a version of Hero and Leander (Hero 1) both published posthumously in 1598.
- Marlowe's smoother version of Hero and Leander (Hero 2).
- disputed anapest poem 'Shall I die...' earlier attributed in [57].

Kit's translation of Ovid's Elegies (Amores):

<http://www2.prestel.co.uk/reynolds/ovid.htm>,

Venus and Adonis (Venus): <http://etext.lib.virginia.edu/etcbin/toccer-new2?id=MobVenu.sgm&images=images/modeng&data=/texts/english/modeng/parsed&tag=public&part=all>

Hero and Leander (Hero1):

<http://darkwing.uoregon.edu/~rbear/marlowe1.html>

Hero and Leander (Hero2):

<http://www2.prestel.co.uk/reynolds/hero.htm>

Shall I die, shall I fly :

<http://www.shaksper.net/archives/1997/0390.html>

These versions with corrected spelling errors in original versions (produced by several publishers in two countries), were recommended to us by British linguist Peter Bull.

Comparatively very long Amores was used as training text which we concatenated with equally-sized slices of the other poems that were used as *query text*. Thus, the size of the training text was not an issue unlike our treatment of *Federalist Papers*. We studied attribution under different sizes of slices, keeping a reasonable number of slices for estimating *Std* of their CCC thanks to large sizes of the poems analyzed. Later we used also the concatenated text of the two poems Amores and Venus as a *training text*.

11.6.3 \overline{CC} for the poems

We calculated the \overline{CC} for each poem divided into slices of various sizes.

The unconditional complexities for all four poems are **surprisingly close** for any partitioning, which shows an extraordinary consistency of the authors' style. \overline{CC} decreases with the increasing slice size as we discussed in the previous section.

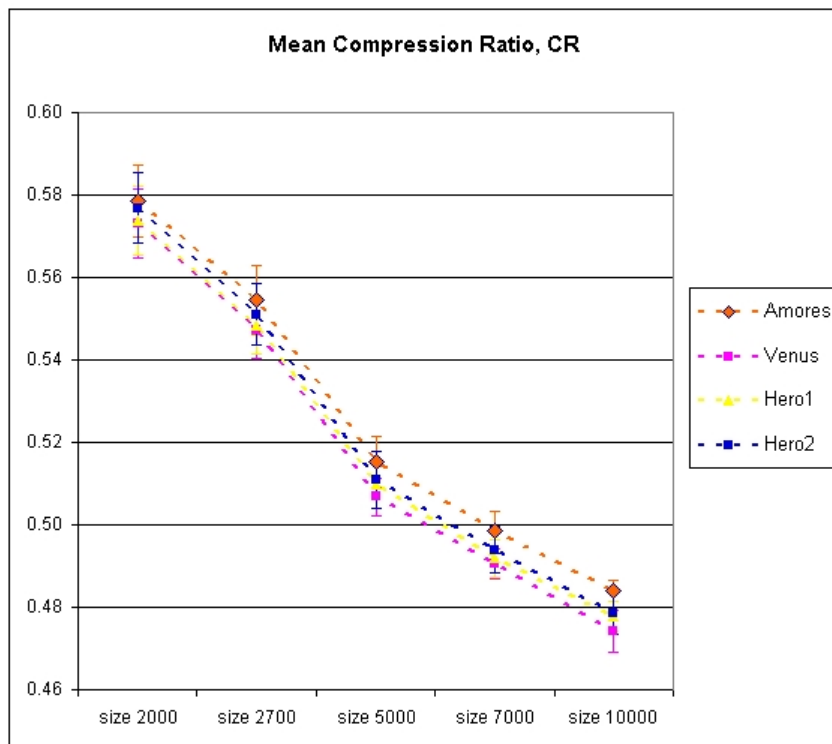


Figure 7: Mean Compression Ratio CC for Amores, Venus, Hero 1 and Hero 2.

11.6.4 Comparison of *CCC* for the poems

The plots show that in terms of *CCC*, Marlowe's translation of *Amores* (the first English translation written apparently during his stay in Cambridge around 1585 and published 'posthumously' 5 years later than *Venus*) helps compress *Venus* significantly better than his own *Hero and Leander* written allegedly at around the same time as *Venus* before his alleged 'untimely demise', registered by Marlowe in 1593 and published first separately in 1598 and then (the same year) together with its twice larger continuation ascribed to G. Chapman. *Amores* was printed in the Netherlands in 1598 and all its copies brought to England were immediately burnt by the orders of Marlowe's deadly foe archbishop Whitgift.

Kit and W. Shakespeare belonged to quite different layers of the society. According to Wikipedia, master of Cambridge University degree was given to Kit after unprecedented petition of the Privy Council, he was a high level spy working for and under patronage of two generations of Cecils ruling over Elizabethan England as Prime Ministers. Kit was employed in their covert operations in several countries and for educating a likely successor to the throne. Often, some of his patrons provided him with lodging in their estates. Any interaction with commoner W. Shakespeare associated with a competing theater is not documented and unlikely.

The normal probability plots shown support asymptotic normality of the *CCC* for slices.

Table 1 in section shows extremely low P-values for the two sample t-test of a 'natural' hypothesis 1.2

$$MeanCCC(Hero|Amores) \leq MeanCCC(Venus|Amores).$$

11.7 *Amores et al versus Rape of Lucrece*

The second work in SC '*Rape of Lucrece*' was prepared and published in haste (1594) thanks to an extraordinary success of unusually erotic for its time *Venus* which was reprinted around ten times during 1593!

Here we compared three versions of *Rape of Lucrece* with the poems we studied before: *Amores*, *Venus* and *Hero 1* using two different compressors *wzip* and *pkzip* and dividing our *query text* '*Rape of Lucrece*' into parts of size 5000 bytes . Essentially the same results were obtained for both compressors.

We see that *Venus* helps compressing *Rape of Lucrece* significantly better than others, the concatenated *training text* '*Amores and Venus*' helped even more significantly. Our *query text* '*Rape of Lucrece*' was

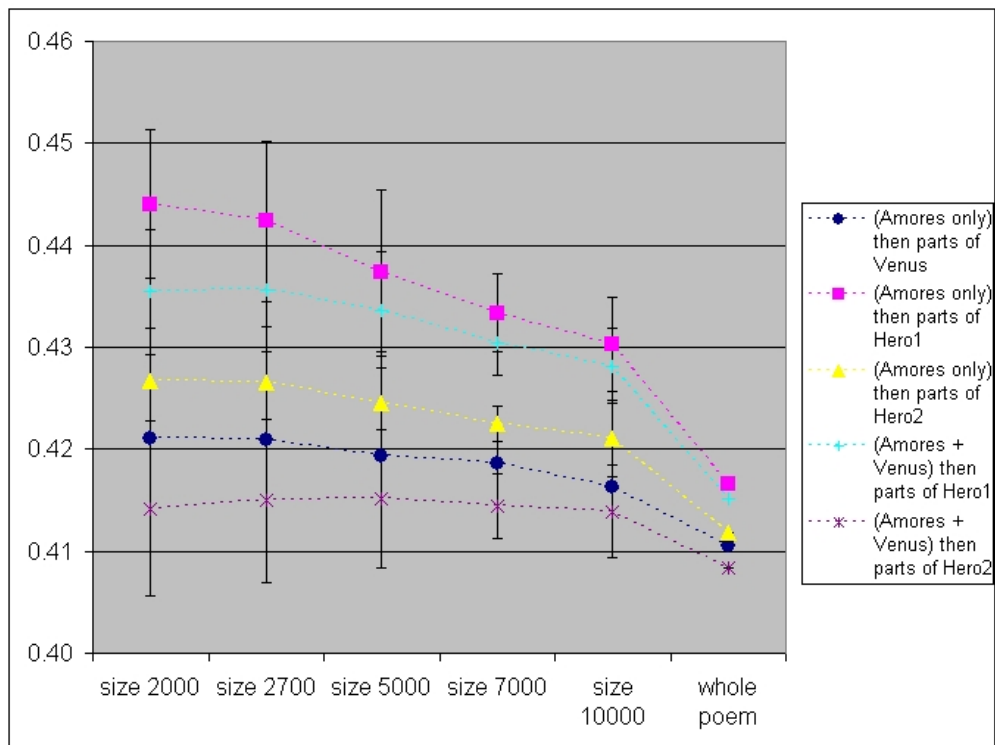


Figure 8: Mean CCC_r for the concatenated poems.

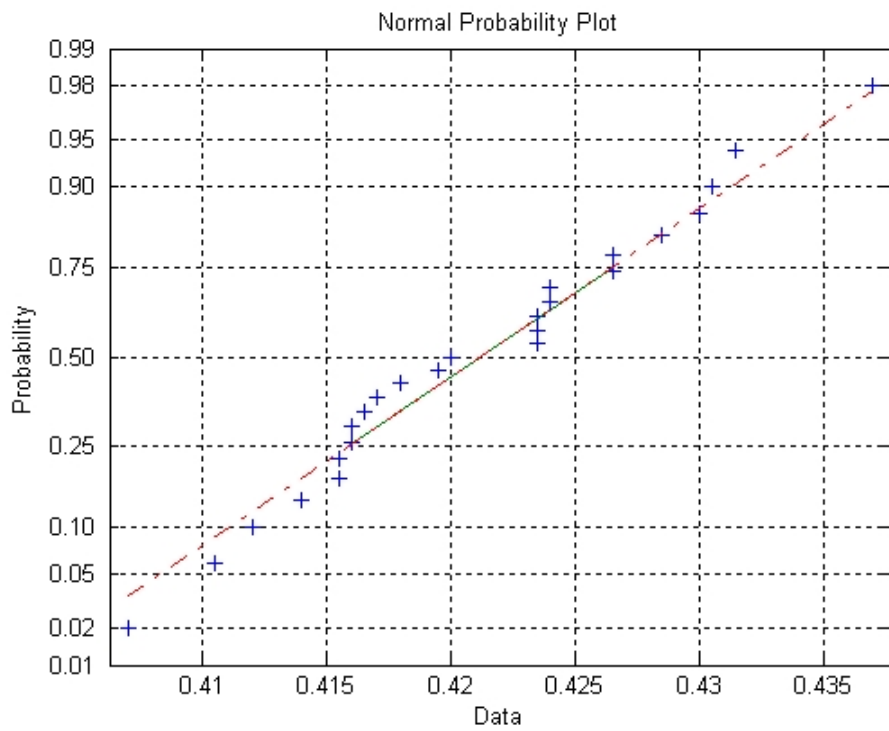


Figure 9: Normal probability plot for $CCCr$ of (a): Venus trained on Amores.

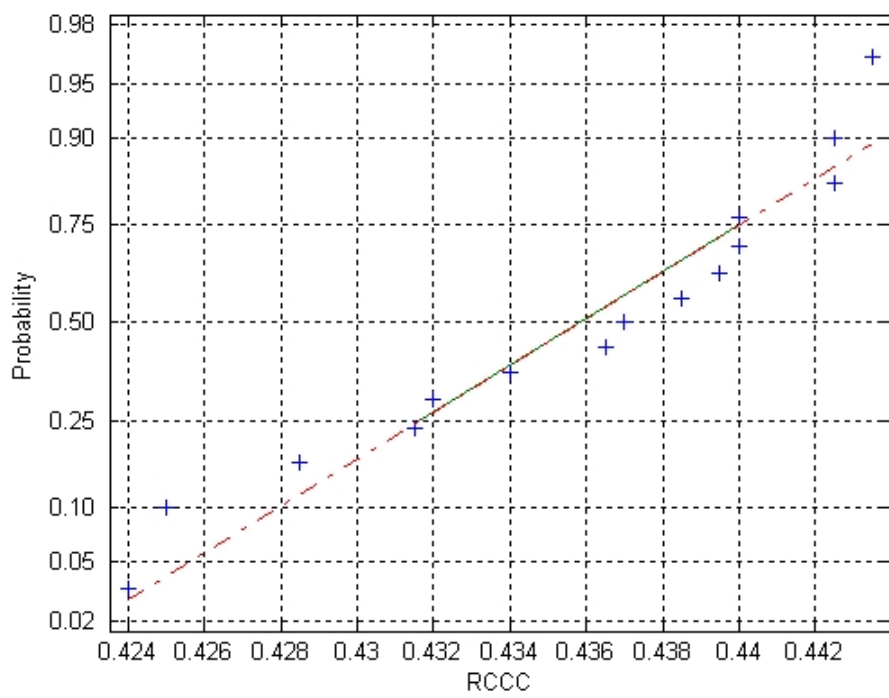


Figure 9 (b): Hero 1 trained on the concatenated text of Amores, Venus.

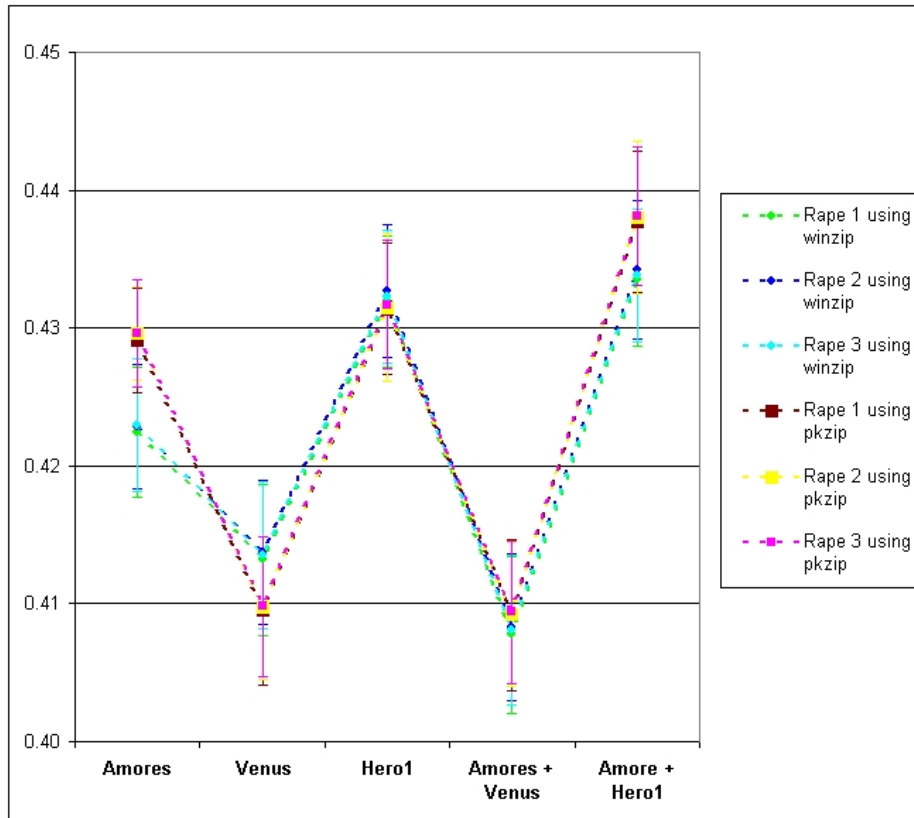


Figure 10: Mean, *StD* of *CCCr* for three versions of Rape of Lucrece with *training texts*: Amores, Venus, Amores and Venus, Amores and Hero1

fixed using different *training texts* different in size. Whereas Amores is 102,161 bytes, Venus is 51,052 bytes and Hero1 is 33,507 bytes after pre-processing.

One of explanations of the above results would be that styles of poems following each other almost immediately are closer than those of more timely Amores which eventually was the source for both, while the final editing of Hero took place several years later.

It is found in [25] that $\overline{CCc}(Hero2) < \overline{CCr}(Shall) < \overline{CCr}(Hero1)$, when trained on ‘Amores’. These results make the Marlowe’s authorship of both ‘Venus and Adonis’ and ‘Shall I die, shall I fly?’ likely.

11.8 Hero and Leander versus its continuation

We applied our method to compare the following poems,

- Hero1 (same as in section 3.2) vs HeroChapman 1, a continuation of Hero and Leander written by George Chapman
- Hero 1598(a version of Hero and Leander) vs HeroChapman 1598, a version of continuation of Hero and Leander written by G. Chapman

The following two plots show $CCCr$, when query texts are Hero 1, HeroChapman 1, Hero 1598 and HeroChapman 1598 with slice sizes 2.7Kb.

11.9 Comparison with poems Chapman i , $i = 1, 2, 3$

For style comparison, Peter Bull recommended three poems: Chapman $i = 1, \dots, 3$, namely ‘The Shadow of Night’, ‘Ovid’s Banquet of Sense’ and ‘The Tears of Peace’ written by G. Chapman around the same time. Their Mean CC are lower than those in Figure 7. We use also the poems from 5.3 as *training* for ‘query’ Chapman $i = 1, \dots, 3$ which were divided into parts of size 3000 bytes. Although the mean CCC are lower when both the *training* and the *query text* are firmly Chapman’s, a more detailed study is preferable.

Our results do not exclude that Chapman helped publishing the disgraced Kit’s ‘Hero’ by putting his name on its continuation, or Chapman edited both ‘Hero’ and its continuation. To distinguish between these alternatives, a further more detailed interdisciplinary comparative style analysis of Kit and G. Chapman is desirable.

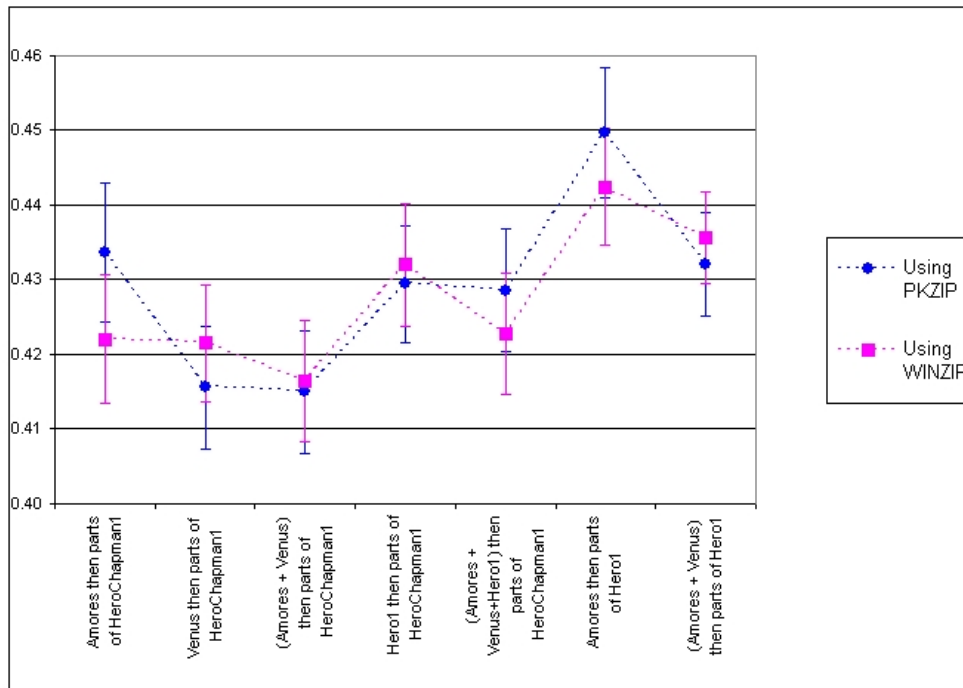


Figure 11: Mean, *Std* of *CCC* for Hero 1 and HeroChapman 1 and some training texts.

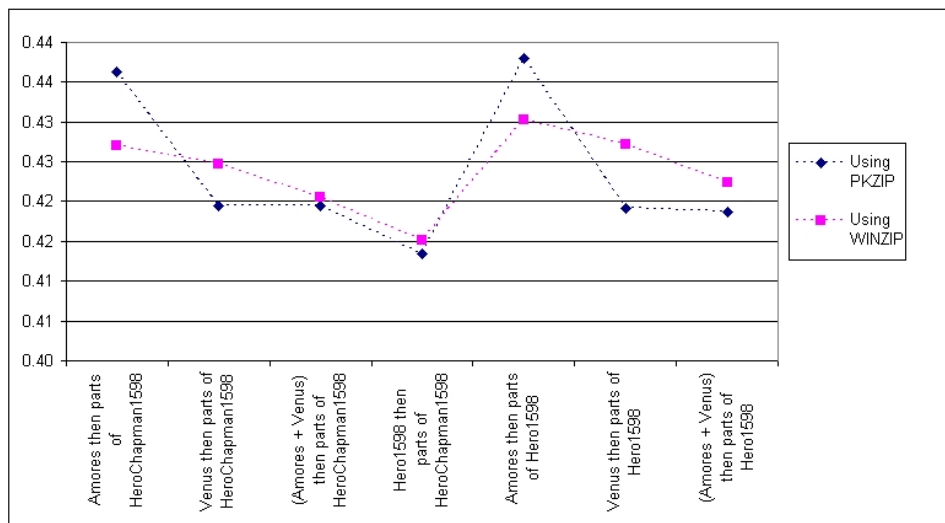


Figure 12: Mean *CCC* taking Hero 1598 and HeroChapman 1598 as disputed poems.

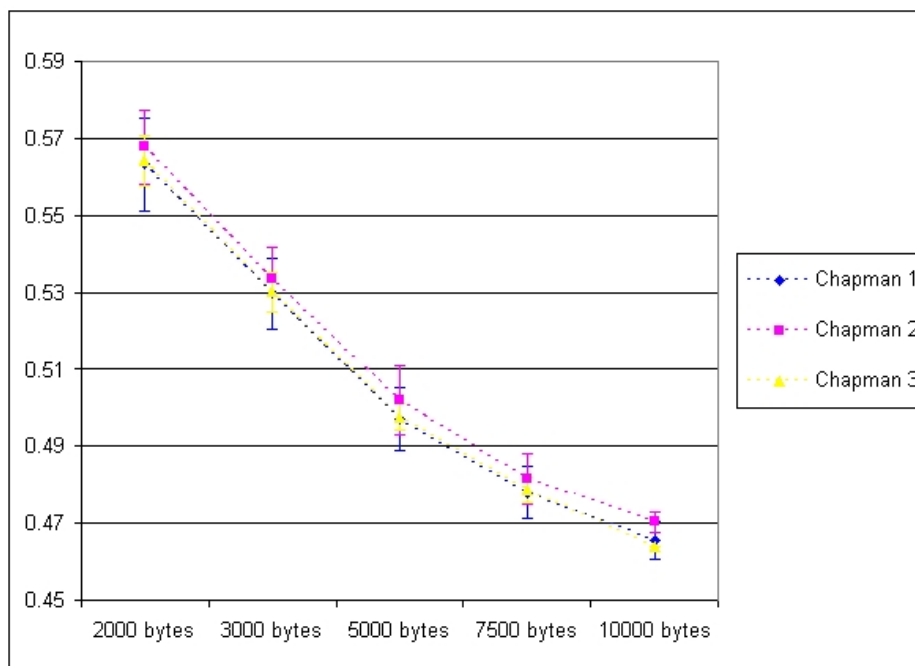


Figure 13: Mean CCr for several Chapman's poems.

Remark. [25] argues that the homogeneity P-value of 154 SC sonnets' fourteen lines versus the presence of anagrammed Marlowe's signatures in first two (four) of them is less than 0.0375 (respectively 2/1000).

Conclusion. Our (**primarily methodological**) survey of results obtained by me and my pupils for the last 5 years shows that CCC-testing is efficient in solving attribution problems in several languages. It uses the fundamental Shannon paradigm modeling sufficiently long LT as stationary ergodic random process and the Kolmogorov-Rissanen paradigm of statistical decisions based on the MDL principle **combining individual and statistical approaches** in the texts analysis.

The statistical assessment of our decisions is achieved via averaging which approximates the ensemble means. This is the only situation (to the best of my knowledge), where the approximation of the incomputable Kolmogorov complexity via commercial UC enables statistically viable and empirically feasible assessment evaluation.

Some other compressor-based approaches are inspired solely by an analogy with incomputable KC and ignore the Kolmogorov-Shannon-Rissanen statistical paradigm. Some of them introduce artificial restrictions such as symmetry taken from other fields by analogy and irrelevant in statistical context. This approach seems to me **pseudoscience misinterpreting the groundbreaking ideas of Kolmogorov-Shannon-Rissanen** and misleading readers.

Success of our statistical approach to LT analysis relies on close collaboration with linguists.

Acknowledgements. The author is grateful to Slava Brodsky, Gabriel Cunningham, Sufeng Li, Dmitry Malioutov, Andrew Michaelson, Stefan Savev, Sufeng Li and Irosha Wickramasinghe for their hard programming and/or processing work: Case studies 11.1-4 were processed by Slava Brodsky with decisive contribution of G. Cunningham in 11.1-2, who wrote a code for finding patterns most contributing to attribution implementing essentially my algorithm [27]. Another NEU student Michaelson made preproceeding for section 11.2. Statistical simulation of section 9.2 and U-discrimination for IID sources

was done respectively by D. Malioutov and S. Savev. The author benefitted a lot from many fruitful discussions with B. Ryabko about lossless UC.

Northeastern University's small grant enabled financial support for G. Cunningham. Peter Bull, Jacob Ziv and Zeev Bar Sella suggested appropriate applications. Moshe Koppel, Z. Bar Sella and Peter Bull sent

us original texts for analysis.

References

- [1] D. Aldous and P. Shields: A Diffusion Limit for a Class of Randomly Growing Binary Trees, *Probab. Th. Rel. Fields*, **79**, 509-542, 1988.
- [2] Z. Bar-Sella: *Literaturnyi kotlovan : proekt "Pisatel Sholokhov"*, Russian Humanitary University (Rossiiskii gos. gumanitarnyi universitet, in Russian), 2005.
- [3] D. Benedetto, E. Caglioti and V. Loreto: Language Trees and Zipping. *Physical Review Letters*, **88**, No. 4, 28 January 2002, p. 048702.
- [4] C.H. Bennet, P. Gács, M. Li, P.M.B. Vitányi and W. Zurek: Information Distance. *IEEE Trans. Inform. Theory*, **IT-44:4**, 1407–1423, 1998.
- [5] R. Bosch and J. Smith: Separating hyperplanes and authorship of the Federalist papers. *Amer. Math. Monthly*, **105(7)**, 601–608, 1998.
- [6] C. Brinegar: Mark Twain and the Quintus Curtis Snodgrass Letters. *Jour. American Statistical Association*, **58(301)**, 85-96, 1963.
- [7] Slava Brodsky: *Bredovy soup*, Limbus Press, Moscow, (in Russian), 2004.
- [8] Slava Brodsky: *Funny children stories*, Manhattan Academia, NY (in Russian), 2007.
- [9] N. Chomsky: Three models for the description of language, *IRE Trans. Inform. Theory*, **2:3**, 113-124, 1956.
- [10] R. Cilibrasi and P. Vitányi: Clustering by Compression, *IEEE Transaction of Information Theory*, **IT-51:4**, 1523–1545, 2005.
- [11] M. Corney: *Analyzing E-mail Text Authorship for Forensic Purposes*, Master Thesis, Queensland Uni. Tech. Australia, 2003.
- [12] P. Diaconis and J. Salzman: Projection pursuit for discrete data. *Probability and Statistics: Essays in Honor of David A. Freedman*, **2**, 265 - 288, IMS Collections, 2008.

- [13] W.J. Elliott and R.J. Valenza. And then there were none: winning the Shakespeare claimants, *Computers and the Humanities*, **30**, 191-245, 1996.
- [14] W. Feller: *Introduction to Probability Theory and its Applications*, volume 1, 3rd edition, Wiley, N.Y., 1968.
- [15] V.P. Fomenko and T.G. Fomenko: Author's invariant of Russian literary texts. Who was the author of 'Quiet flows Don?', *Appendix 3 in Fomenko, A.T. 'Methods of statistical analysis of historical texts'*, **2**, Kraft and Leon, Moscow (In Russian), 1999.
- [16] D. Foster: *Author unknown*. H. Holt, N.Y., 2000.
- [17] W. Friedman and E. Friedman: *The Shakespearean Ciphers exposed*, Cambridge University Press, 1957.
- [18] A. Gavish and L. Lempel: Match-Length Functions for Data Compression. *IEEE Trans. Inform. Th.*, **42-5**, 1375-1380, 1996.
- [19] A. Gelbukh and G. Sidorov: Zipf and Heaps Laws Coefficients Depend on Language, *Springer Lecture Notes in Computer Science* No. 2004, 332-335, 2001.
- [20] G. Kjetsaa and S. Gustavsson: *Authorship of Quiet Don*. Solum, Norway, 1986.
- [21] A.N. Kolmogorov: Three approaches to the quantitative definition of information, *Problems of information transmission*, **1**, 3-11, 1965.
- [22] O. Kukushkina, A. Polikarpov and D. Khmelev: Text Authorship attribution using letters and grammatical information, *Problems of information transmission*, **37(2)**, 172-184, 2001.
- [23] D. Labbe: *Corneille dans l'ombre de Moliere?* Les Impressions Nouvelles, Paris-Bruxelles, 2004.
- [24] M. Li, X. Chen, X. Li, B. Ma and p. Vitaniy: The similarity metric. *IEEE Transaction of Information Theory*, **IT-50:12**, 3250-3264, 2004.
- [25] M.B. Malyutov: OP&PM, Review of methods and examples of Authorship Attribution of texts. *Review of Applied and Industrial Mathematics*, TVP Press, **12**, No.1, 2005, 41-77 (In Russian), 2005.

- [26] M.B. Malyutov: Authorship Attribution of texts: a review. *Springer L. Notes in Comp. Sci. 4123*, R. Ahlswede et al eds, 362-380, 2007.
- [27] M.B. Malyutov, C.I. Wickramasinghe and S. LI : Conditional Complexity of Compression for Authorship Attribution, *SFB 649 Discussion Paper No. 57, Humboldt University, Berlin*, 2007.
- [28] M.B. Malyutov and Slava Brodsky: The MDL - procedure for attributing the authorship of texts: *OP&PM: Review of Applied and Industrial Mathematics* TVP Press, **16**, No.1, 25-34, 2009.
- [29] M.B. Malyutov: The MDL - principle in attributing the authorship of literary texts: *Proceedings of the Dobrushin International Conference, Moscow, IITP, July 2009*.
- [30] M.B. Malyutov: The MDL - principle in testing homogeneity between styles of literary texts: *Proceedings of the WITMSE - 2009, Tampere University of Technology, August 2009*.
- [31] M.B. Malyutov: The MDL - principle in testing homogeneity between styles of literary texts: a review. *OP&PM: Review of Applied and Industrial Mathematics*, TVP Press, **17**, No.3, 2010, (In Russian).
- [32] M.B. Malyutov: Recovery of sparse active inputs in general systems: a review. *Proceedings, 2010 IEEE region 8 international conference on computational technologies in electrical and electronics engineering SIBIRCON-2010, Irkutsk, Russia, 1*, 15-23, 2010.
- [33] M.B. Malyutov and G. Cunningham: Pattern discovery in LZ-78 texts homogeneity discrimination. *Proceedings, 2010 IEEE region 8 international conference on computational technologies in electrical and electronics engineering SIBIRCON-2010, Irkutsk, Russia, 1*, 23-28., 2010.
- [34] A.A. Markov. On an application of statistical method, *Izvestia Imper. Acad. of Sciences*, Series VI, **X**, No. 4, p. 239, 1916, (in Russian).
- [35] M.A. Marusenko, R.H. Piotrowski and Yu.V. Romanov: NLP and Attribution of Pseudonymic Texts: Who is Really the Author of the 'Quiet Flows the Don'. *SPECOM'2004: 9th Conference Speech and Computer*, Saint-Petersburg, Russia, September 20-22, 2004, 423-427, 2004.

- [36] M.A. Marusenko, B.A. Bessonov, L.M. Bogdanova, M.A. Anikin and N.E. Myasoedova. *Search for the lost author, attribution etudes*, Filological Department, Sankt Petersburg University, (in Russian), 2001.
- [37] V.P. Maslov and T.V. Maslova. On Zipf law and rank distributions in linguistics and semiotics, *Mat. Zametki*, **80:5**, 728-732, 2005.
- [38] T.A. Mendenhall: The characteristic curves of composition, *Science*, **11**, 237-249, 1887.
- [39] T.A. Mendenhall: A mechanical solution to a literary problem. *Popular Science Monthly*, **60**, 97-105, 1901.
- [40] N. Merhav: The MDL principle for piecewise stationary sources, *IEEE Trans. Inform. Th.*, **39-6**, 1962-1967, 1993.
- [41] D.S. Moore, G.P. McCabe and B. Craig: *Introduction to the Practice of Statistics*, 6th edition, Freeman, 2008.
- [42] N.A. Morozov: Linguistic spectra... stylometry etude, *Izvestia Imper. Acad. of Sciences, Russian language series*, XX, **4**, 1915, (in Russian).
- [43] F. Mosteller. and D. Wallace: *Inference and Disputed authorship: The Federalist papers*, Addison-Wesley, 1964.
- [44] Ch. Nicholl: *The Reckoning*, Chicago University Press, 1992.
- [45] J. Rissanen: Universal coding, Information, Prediction and Estimation. *IEEE Trans. Inform. Th.*, **30-4**, 629-636, 1984.
- [46] J. Rocha, F. Roselland and J. Segura: The Universal Similarity Metric does not detect domain similarity, *arXiv:q-bio.QM/0603007* v1 6 Mar 2006
- [47] R. Rosenfeld: A Maximum Entropy Approach to Adaptive Statistical Language Modeling. *Computer, Speech and Language* **10**, 187-228, 1996.
- [48] B. Ryabko: Twice - universal codes, *Problems of information transmission*, **20**, No 3, 24-28, 1984.
- [49] B. Ryabko: Prediction of random sequences and universal coding. *Problems of information transmission*, **24**, No 3, 3-14, 1988.

- [50] B. Ryabko and J. Astola: Universal Codes as a Basis for Time Series Testing *Statistical Methodology*, **3**, 375-397, 2006.
- [51] S. Savari: Redundancy of the Lempel-Ziv Increment Parsing Rule. *IEEE Trans. Inform. Th.*, **43-1**, 9-21, 1997.
- [52] G. Shamir and N. Merhav: Low-Complexity Sequential Lossless Coding for Piecewise Stationary Memoryless Sources. *IEEE Trans. Inform. Th.*, **45-5**, 1498-1519, 1999.
- [53] C. Shannon: A Mathematical Theory of Communication, *Bell System Tech. J.*, 27, 379-423, 623-656, 1948.
- [54] C. Shannon: Communication Theory of Secrecy Systems. *Bell System Tech. J.*, **28**, 656-715, 1949.
- [55] A. Solzhenitsyn: *Stremya Tikhogo Dona*, YMCA Press, 1976 (in Russian).
- [56] W. Szpankowski: *Average Case Analysis of Algorithms on Sequences*, Wiley, N.Y., 2001.
- [57] R. Thisted and B. Efron: Did Shakespeare write a newly discovered poem? *Biometrika*, **74**, 445-455, 1987.
- [58] C.I. Wickramasinghe: *PhD dissertation*, Mathematics Department, Northeastern University, Boston, MA, 2005.
- [59] J. Ziv: On classification and universal data compression. *IEEE Trans. on Inform. Th.*, **34:2**, 278-286, 1988.