

The MDL-PRINCIPLE in ATTRIBUTING AUTHORSHIP of TEXTS

Mikhail Malyutov

Math. Dept., Northeastern University, Boston, MA 02115

ABSTRACT

We study a new *context-free* computationally simple stylometry-based attributor: the mean sliced *conditional compression complexity (CCC)* of *literary texts* which is inspired by the incomputable Kolmogorov conditional complexity. Whereas other stylometry tools can occasionally almost coincide for different authors, our CCC-attributor introduced in Malyutov (2005) is asymptotically strictly minimal for the true author, if the query texts are sufficiently large but much less than the training texts, universal compressor is good and sampling bias is avoided. This classifier simplifies the Ryabko and Astola (2006) homogeneity test (partly based on compression) **under insignificant difference of unconditional complexities of training and query texts** which can be verified using its asymptotic normality proved in Szpankowski (2001) and elsewhere for IID and Markov sources and normal plots for real literary texts. It is *consistent* under large text approximation as a *stationary ergodic sequence* which follows from the *lower bound for the minimax compression redundancy of piecewise stationary strings* (Merhav (1993)) and from our elementary combinatorial arguments and simulation for IID sources. The *t-ratio* use measuring how many standard deviations are in the mean difference between slices' mean CCCs enables evaluation of its P-value of statistical significance. It is based on the *asymptotic normality* of slices' CCC 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 processed by Sufeng Li, Irosha Wickramasinghe, Slava Brodsky, Gabriel Cunningham and Andrew Michaelson: 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 SCCC-relations between certain Elizabethan poems. Two different S. Brodsky's novels *deliberately written in different styles* and various Madison's papers showed insignificant mean CCC-difference as the useless Vitanyi-Cilibrasi test did in ALL cases studied.

Keywords: compression length, authorship attribution, homogeneity testing, asymptotic normality.

1 Discrimination with Universal Compressors

C. Shannon (1948, 1949) created a comprehensive theory of information transmission based on Kolmogorov’s statistical theory. In particular, **given** a distribution on an alphabet, the mean length of the Shannon-Fano compression of the IID string with elements from this alphabet attains asymptotically the Shannon’s entropy lower bound for the length (complexity) of compression. A.N. Kolmogorov (1965) developed a complexity theory of an **individual string** such that for large strings *belonging to a statistical ensemble* their mean complexity approximates their entropy, and sketched (for IID input) the first so-called **universal compressor** (UC) which adapts 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 $\mathbf{L} = \{L_n : \mathbf{B}^n \rightarrow \mathbf{B}^\infty, n = 1, 2, \dots\}$ is (weakly) **universal**, if for any $P \in \mathbf{P}$ and $\epsilon > 0, \mathbf{B} = \{0, 1\}$, it holds:

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

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. A profoundly smarter method implementing much fuller the above-mentioned Kolmogorov’s idea for compressors took more than ten years to emerge in two Lempel-Ziv (LZ) compressor constructions (1977-78). 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. Wyner and Ziv proved that LZ-78 is an UC implying

$$\lim_{n \rightarrow \infty} P(|L_n(x)|/|x| \rightarrow h) = 1 \quad \text{as} \quad |x| \rightarrow \infty \quad (2)$$

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 Shannon (1949), where SED strings were first singled out as popular models of natural language. By nineties, versions of LZ became everyday tools in computer practice. Rissanen’s pioneering publication on the **Minimum Description Length principle** (MDL) in 1978 (continued in his paper (1984)) and Ziv

(1988) 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 Ryabko and Astola (2006).

1.0.1 Ryabko-Astola and U-statistics

Define $|A|$ and $|A_c|$ as the lengths of respectively **binary** string A and its compression A_c .

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

The Ryabko and Astola homogeneity of two strings test statistic T is

$$T = h_n^*(S) - |A_c| - |Q_c|, \quad (3)$$

where the empirical Shannon entropy 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. 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 T rather than equally computationally intensive Rosenfeld's (1996) Likelihood methods based on n -MC training is questionable. For shorter LT accuracy of SED model may be insufficient, 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 Chomsky (1956).

Consider $U(Q, A) = |S_c| - |A_c| - |Q|$. Quantity $U(Q, A)$ *mimics the Ryabko and Astola statistic T* . In $U(Q, A)$ we *replace their empirical Shannon 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 T 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 seemed to hold also for $U(Q, A)$ in some range of $|A|/|Q|$ due to the **lower bound for the minimax mean UC-compression redundancy of piecewise-stationary sources (Merhav (1993))** which is logarithmic in $(|Q| + |A|)$.

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 excessively large 'transition 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, but it is not applicable in our LT studies.

1.0.2 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 are respectively

$$CCC(B|A) = |S_c| - |A_c|. \quad (4)$$

The *CCC* mimics an abstract conditional Kolmogorov Complexity in our settings and measures how adapting to patterns in the training text helps to compress the query text. It presumably 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 .

The only difference of CCC from U is canceling the $|Q_c|$ removal which prevents the aforementioned inconsistency of U - statistic.

We average sliced *CCC* of text $Q_i, i = 1, \dots, m = \lfloor |Q|/L \rfloor$, given the firmly attributed text A , dividing the *query text* Q into slices of equal length L and used the same UC for all sizes of texts.

$$\overline{CCC(Q|A)} := \sum_{i=1}^m \frac{CCC(Q_i|A)}{m} := \sum_{i=1}^m \frac{CC(Q_i)}{m}. \quad (5)$$

We call the last two empirical quantities ‘*Mean CCC(Q)* and *Mean CC(Q)*’, respectively.

Claim. Both our case studies and statistical simulation in section 3 show that the sliced CCC-attribution has a good homogeneity discrimination power in this range for moderate $|Q|$ in a surprisingly wide range of case studies with **insignificantly varying mean unconditional complexity CC** of compression.

Statistical testing of the latter condition is straightforward due to the **asymptotic normality** results of the compression complexity described in Szpankowski (2001). Its very plausible extension for CCC would theoretically support a quite **unusual sample size relation** for UC-attributing

authorship: **sample size of the training text must dramatically exceed those of slices of a query text.** 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 $|Q|$, while the mean increase in $CCC(Q|A)$ redundancy for *different distributions* of Q and A *as compared to their identity* seems to be $o(|(A|Q)|^b)$ for any $b > 0$ (accurate upper bound even for LZ78 is absent so far (see some LZ-78 upper bounds in Savari (1997)), the lower bound in Merhav(1993) is only $O(\log(|(A|Q)|))$). Thus, the t-ratio is negligible under the asymptotics $|A| \rightarrow \infty, 0 < \epsilon < |Q|/|A|$. Malyutov (2005) explains this informally as follows: if the training A and alternative style query text Q sizes are comparable, then two flaws 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 Cilibraşi and Vitányi (2005) (CV05) also based on the conditional compression complexity although **ignoring assessment of statistical stability**, unappealing, and explains examples of CV05 misclassification shown in Rocha et al (2006). It can explain also the roots of early heated discussion around simpler development in Benedetto et al (2002), where the *sample size relation and statistical stability* issues were not addressed.

Due to space limitation, we skip sections: **Brief survey of micro-stylometry tools, Methodology, Simulation study of CCC-attributor, Extended LZ index and many exciting examples of Attribution of literary texts** which are described in detail in my larger paper in Russian under review in 'Problems of Information Transmission', MalBrod09 and in MWL07.

1.1 Follow up Analysis

LZ-78 generates the binary tree of all patterns found in LT: thus for every pattern ν we can evaluate frequency of the cases when ν is a **prefix** of the further text which is the cardinality of the subtree rooted in ν .

G. Cunningham implemented in Perl language my algorithm (MWL07) of economic LZ-tree construction and evaluating cardinalities of interesting subtrees. Subtree rooted in ν is called interesting, if 't-value' for its cardinalities $n(\nu, A)$, is large for competing candidates for authorship.

$$t' = (n(\nu, A) - n(\nu, A')) / \sqrt{[n(\nu, A)(c_1 - n(\nu, A)) / c_1 + n(\nu, A')(c_2 - n(\nu, A')) / c_2]}, \quad (6)$$

Table 1: Most ‘interesting patterns for Federalist papers

Binary pattern	t-value	Patterns in English
011100100110010101101001	4,08	rei
001000000110010001100101	3,62	de
0110100001100101001000000101001101110100	3,43	he St
01100001011010010110111001110011	3,38	ains
01100101011100100110000101101100	3,38	eral
01110100011010000110111101110010	3,28	thor
011010000010000001110111	3,15	h w
01100101011011100110010001100101	3,15	ende
01101100011001010010000001100001	3,15	le a
0111010101100100	3,15	ud
001000000110000101101110011001000010000001110010	3,14	and r
01110100011010000110010101101101	3,12	them
011001100110010101100100	3,12	fed
011001110110111000100000	3,12	gn

where $c_i, i = 1, 2$, are total patterns cardinalities for competing candidates. Finally, the tables of English patterns corresponding to interesting binary patterns are tabulated.

Any solid judgement about corresponding P-values is hard due to vast multiplicities of not independent patterns. Still the tables like the shown one for Federalist papers may be useful source of information for linguists.

REFERENCES

- Benedetto, D. Caglioti, E. and Loreto, V. (2002): Language Trees and Zip-ping. *Physical Review Letters*, **88**, No. 4, 28 January 2002, p. 048702.
- Bennett, C.H., Gács, P., Li, M., Vitányi, P.M.B., Zurek, W. (1998): Information Distance. *IEEE Trans. Inform. Theory*, **IT-44:4**, 1407–1423.
- Chomsky, N. (1956). Three models for the description of language, *IRE Trans. Inform. Theory*, **2:3**, 113-124.
- Cilibrasi, R. and Vitanyi, P. (2005): Clustering by Compression, *IEEE Transaction of Information Theory*, **IT-51:4**, 1523–1545
- Kolmogorov A.N. (1965): Three approaches to the quantitative definition of information, *Problems of information transmission*, **1**, 3–11
- Kukushkina, O., Polikarpov, A. and Khmelev, D. (2001): Text Authorship attribution using letters and grammatical information, *Problems of information transmission*, **37(2)**, 172-184

- Li, M., Chen, X., Li, X. Ma, B. and Vitaniy, P. (2004): The similarity metric. *IEEE Transaction of Information Theory*, **IT-50:12**, 3250–3264.
- Malyutov, M.B. (2005): Review of methods and examples of Authorship Attribution of texts. *OP&PM: Review of Applied and Industrial Mathematics*, TVP Pres, **12**, No.1, 2005, 41-77 (In Russian).
- Malyutov, M.B. and Brodsky, S. (2009): MDL - principle for authorship attribution of texts. *OP&PM: Review of Applied and Industrial Mathematics*, TVP Press, **16**, No.1, 25-34 (In Russian). .
- Malyutov, M.B., Wickramasinghe, C.I. and Li, S. (2007): Conditional Complexity of Compression for Authorship Attribution, *SFB 649 Discussion Paper No. 57, Humboldt University, Berlin*.
- Merhav, N. (1993): The MDL principle for piecewise stationary sources, *IEEE Trans. Inform. Th.*, **39-6**, 1962-1967.
- Rissanen, J. (1984): Universal coding, Information, Prediction and Estimation. *IEEE Trans. Inform. Th.*, **30-4**, 629-636.
- Rocha, J., Rosella, F. and Segura, J. (2006). The Universal Similarity Metric does not detect domain similarity, arXiv:q-bio.QM/0603007 v1 6 Mar 2006
- Ryabko, B. and Astola, Y. (2006). Universal Codes as a Basis for Time Series Testing *Statistical Methodology*, **3**, 375-397.
- Savari, S.(1997). Redundancy of the Lempel-Ziv Increment Parsing Rule. *IEEE Trans. Inform. Th.*, **43-1**, 9-21.
- Shannon, C. (1949): Communication Theory of Secrecy Systems. *Bell System Tech. J.*, **28**, 656-715.
- Szpankowski, W. (2001): *Average Case Analysis of Algorithms on Sequences*, Wiley, N.Y.
- Ziv, J. (1988): On classification and universal data compression. *IEEE Trans. on Inform. Th.*, **34:2**, 278-286.