Compression enables homogeneity test of stationary sources

Mikhail Malyutov

Outline

Everyone has an experience of using universal compressors (UC) such as Zip for storing and transmitting files. We develop the theory and their application for statistical analysis. Let $\mathbf{B} = \{0,1\})$, $\mathbf{x^N} \in \mathbf{B}^N = (x_1, ..., x_N)$ be a stationary ergodic random binary (**training**) string (text, SET) distributed as $P_0 = P$.

Arbitrary UC (see [15]) maps source strings $\mathbf{x^N} \in \mathbf{B^N}$ into compressed binary strings $\mathbf{x_c^N}$ of approximate length $|\mathbf{x}_c^N| = -\log P(\mathbf{x}^N)) = L^N$ thus *generating the approximate Loglikelihood of source* $\mathbf{x}^N$ – the main inference tool about P.

Consider a **query** binary SET $\mathbf{y}^M$ distributed as P_1 and test whether the homogeneity hypothesis $P_0 = P_1$ contradicts the data or not. Let us partition $\mathbf{y}^M$ into several **slices** $\mathbf{y}_i, i = 1, \ldots, S$, of identical length n divided by 'brakes' - strings of relatively small-length δ to provide approximate independence of slices (brakes of length 2k are sufficient for k-MC). Introduce concatenated strings $\mathbf{C}_i = (\mathbf{x}^N, \mathbf{y}_i)$. Define $CCC_i = |C_i| - |\mathbf{x}^N|$. CCC-statistic is $\overline{CCC}$ = average of all CCC_i. We meet $\overline{CCC}$ in study of very small error probabilities when finding few inhomogeneous among vast number of homogeneous inputs [10]. Homogeneity of two texts is tested by the test statistic $\mathcal{T} = \overline{CCC}/s$, s is the standard deviation of $CCC_i, i = 1, \ldots, S$. Extensive experimentation with real and simulated data [7, 8, 9, 13] showed that $\mathcal{T}$ well discriminates between homogeneity and its absence in spite of the lack of knowledge about P_0, P_1. Here we prove its consistency, asymptotic normality and CCC tail optimality approximating that of the Likelihood Ratio Test (LRT) in full generality under certain natural assumption about the sizes of the training string and query slices.

Validity of our assumption in applications depends both on the compressor used and the source distribution. All results are new. All previous attempts [1, 14] of proving asymptotic normality under the null hypothesis were extremely involved technically and applied to *one* compressor and IID source of dubious importance in applications. The main advantages of CCC-test are: i. its applicability for arbitrary UC and long memory sources, where the likelihood is hard to evaluate and ii. its computational simplicity (as compared to statistics from [15] and first part of [13]) enabling processing of multi-channel simultaneously on line like change-point detection, see [10].

Preliminaries

Conditions on regular stationary ergodic distributions (SED) of strings are as in [15]. SET are approximated by n-Markov Chains.

Continuing and correcting the von Mises study of a string randomness, Kolmogorov [6] outlined a compressor adapting to an unknown IID distribution and gave a sketch of a version of Theorem 1 below for IID sources, connecting for the first time the notions of *complexity and randomness*. First practically implementable UC LZ-77/78 were invented in 12 years and became everyday tools of computer work after ten more years.

Introduce $L^n = -\log P(x^n)$ and the entropy $h^n = H(P) = -\sum P(x^n) \log P(x^n)$.

Warning. Without mentioning, we actually consider conditional probabilities and expectations of functions of the query slice with regard to the training text.

A compressor is called UC (in a weak sense), if it adapts to an **unknown** SED distribution, namely if for any $P \in \mathbf{P}$ and any $\epsilon > 0$, it holds:

$$\lim_{n\to\infty} P(x \in \mathbf{B}^n : |x_c^n| - L^n \le n\epsilon) = 1.$$

Equivalently, UC attain for $P \in \mathbf{P}$ asymptotically the Shannon entropy lower bound:

$$\lim_{n\to\infty} P(||\mathbf{x^n}_c|/|\mathbf{x^n}| \to h) = 1 \quad as \quad |x| \to \infty$$

established in the works of C. Shannon in 1948-1949, where SED strings were first singled out also as appropriate models of natural language.

J. Rissanen's pioneering publication on the **Minimum Description Length** principle [12] and [15], (goodness of fit and homogeneity testing) initiated applications of UC to statistical problems for SED sources continued in a series of recent papers of B. Ryabko with coauthors.

Kraft inequality. Lengths of any uniquely decodable compressor satisfy the following inequality : $\Sigma_{\mathbf{B}^n} 2^{-|\mathbf{x_c^n}|)} \le 1$.

Introduce diversion (cross entropy) $D(P_1||P_0) = \mathbf{E}_1 \log(P_1/P_0)$ and consider goodness of fit tests of P_0 vs. P_1.

'Stein lemma' for SED [15](Proved first for IID case by H. Cramer in 1938). *If $D(P_1||P_0) \ge \lambda$ and any $0 < \varepsilon < 1$, then the error probabilities of LRT satisfy simultaneously*

$$P_0(L_0 - L_1 > n\lambda) \le 2^{-n\lambda}$$

and

$$\lim P_1(L_0 - L_1 > n\lambda)) \ge 1 - \varepsilon > 0.$$

No other test has both error probabilities less in order of magnitude.

Theorem 1 [15]. Consider test statistic $T = L_0^n - |\mathbf{x}_c^n| - n\lambda$. Then nonparametric goodness of fit test $T > 0$ has the same asymptotics of the error probabilities as in the Stein lemma.

Main results

'Quasiclassical Approximation' assumption (QAA). The sizes of the training string N and query slices n grow in such a way that the joint distribution of $CCC_i, i = 1, \dots, S$, converges in Probability to $P_1(L_0^n(\mathbf{y}))$: $N \to \infty$ and $n \to \infty$ is sufficiently smaller than N.

The intuitive meaning of this assumption is: given a very long training set, continuing it with a comparatively small query slice with alternative distribution P_1 does not affect significantly the *encoding rule*. The typical theoretical relation between lengths as discussed in [13, 11] is $n \le const \log N \to \infty$.

In practice, an appropriate slice size is determined by empirical optimization.

Define entropy rates $h_i = \lim h_i^n, h_i^n = \mathbf{E}L_i^n, i = 0, 1$.

Under QAA and $h_1 \ge h_0$ the three following statements are true.

Theorem 2: Consistency.*The mean CCC is strictly minimal as $n \to \infty$, if $P_0 = P_1$.*

Generate an artificial (n)-sequence $\mathbf{z^n}$ independent of $\mathbf{y^n}$, $\mathbf{z^n}$ distributed as P_0 and denote by CCC_0 its CCC.

In [9, 13] such test was obtained by training the slice on the remaining portion of the same text. Also assume that the 'brakes' negligible sizes are such that the joint distribution of S slices of size n converge to their product distribution in Probability.

Theorem 3. *Suppose P_1, P_0 are SED, $D(P_1||P_0) > \lambda$ and we reject homogeneity, if the 'conditional version of the Likelihood Ratio' test $\mathcal{T} = \overline{CCC} - \overline{CCC}_0 > n\lambda$. Then the same error probability asymptotics as for LRT is valid for this test.*

Let us study the central range of CCC-distribution and now assume in addition to QAA that P_1 *is contiguous w.r.t.* P_0, i.e. the sets of P_0 measure 0 have also P_1 measure 0. This natural for LT in the same language assumption for k-MC means that transitions impossible in P_0 are equally impossible in P_1. We assume also that P_0 - distribution of L^n is asymptotically Normal (AN).

Theorem 4. *AN holds also for $CCC_i, i = 1, \ldots, S$, under P_1 and all assumptions made (Le Cam lemma [4]). Statistic $\mathcal{T}$ has asymptotically central/non-central Student distribution respectively under P_0 and P_1 with S-1 degrees of freedom.*

Sketch of AN justification

In IID cases AN is proved in [1, 14] on several dozen pages of hard reading. Our approach for general UC and SED outlined below is shorter.

J. Ziv's claim (personal communication): Since $\mathbf{y}^n$ is almost incompressible, the compressed file right hand tail with infinite/long memory is (respectively, converges to in Divergence) IID(1/2). Thus lengths of compressed texts is an asymptotically sufficient statistic containing all information in the training SET while compressed texts themselves is a white noise!

We have UC: $\mathbf{y}^n \to Z_n := (m(n), \mathbf{z}^m)$, The joint distribution of the random vector Z_n is $\mathcal{L}_m$. Given the training text, Z_n is a deterministic invertible function of $\mathbf{y}^n$. Thus entropy $H(\mathcal{L}_m) = h^n$.

According to the Ziv's claim, $h^n/m(n) \to 1$ in Probability. 'In Probability' will further mean in Probability of $x^N, \mathbf{y}^n$, where for some $g(\cdot) :\to \mathbf{N}, n < g(N)$.

Assumption A**: 'Second thermodynamics law'**. UC is such that $h^n/m(n)$ growth is monotone in Probability. The limiting joint distribution $\mathcal{L}_{m(\mathbf{x}^n)}$ maximizes entropy $H(\mathbf{x}^n))$ for *fixed variance* of h^n in Probability.

Corollary. $h^n/m(n)$ is a non-decreasing submartingale under A bounded by 1, it is AN in Probability for large n after removing fitted small parabolic drift, all random IID parameters of the Bernoulli variables z_i are AN due to the well-known entropy maximization of a shrinking distributions on $[0, 1]$.

Now, $m(n)$ is the first exit time, when the AN sequence $s^i = \sum_1^i h(z_j)$ hits high level h^n. Its AN is well-known, see [3].

Preceding work

Kolmogorov's complexity and randomness relation is the cornerstone in our constructions (see theorem 1). On the eve of final grave illness, he made a sketch of complementing this, in parallel to far from Mathematics Solomonov and Cheitin, by an Abstract Theory of Kolmogorov Complexity (KC-AT). KC-AT inspired D. Khmelev to introduce the kernel of our CCC method outside of statistical paradigm. For SED string $\mathbf{x}^N$ and a fixed UC, $|\mathbf{x}_c^N|$ is

an approximation to conditional KC. However, if $\mathbf{x}^N$ is a Genome part, $|\mathbf{x}_c^N|$ is several times more, than $|\mathbf{x}^N|$ for UC zip and is in no way approximation to conditional KC. I am unaware of nontrivial sufficiently rich cases in addition to statistical models, where *incomputable* KC can be approximated from below and above by *computable functions*, at least theoretically. Thus, replacement of KC with quantity evaluated with UC as in [2] needs justification. Far from elementary statistical UC theory *must* be applied for strings approximated by statistical model such as LT.

A survey of related developments (before my first working version of CCC was introduced in [7]) is in [2]. All of them follow Khmelev adding artificial transformations as in [2] irrelevant in statistical context and only worsening accuracy of analysis; their replacement of KC with quantity evaluated with UC is not justified. Thus validity of their applications is doubtful. Their classifier poorly discriminates between LT in same language (see [8]) and its output depends on texts entropies not mentioned in [2]. Their claim that L. Tolstoy stands alone among Russian writers is caused most likely by inadequate texts preparation for analysis: they did not remove large portions of French in Tolstoy having different entropy rate.

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] R. Cilibrasi and P. Vitanyi, Clustering by Compression, *IEEE Transaction of Information Theory*, **IT-51**, 1523–1545, 2005.

[3] A. Gut and J. Steinebach, Convergence rates and precise asymptotics for renewal counting processes and some first passage problems, *UUDM Report 2002:7*, available at http://www.math.uu.se/research/pub/preprints.html

[4] Ya. Hajek and Z. Shidak, *Theory of rank tests*, Academic Press, 1967.

[5] D.V. Khmelev: Complexity approach to the text authorship attribution, *Abstracts, Conference 'Russian Language', Philology Dept., Moscow State University*, 426-427, 2001.

[6] Kolmogorov A.N.: Three approaches to the quantitative definition of information, *Problems of information transmission*, **1**, 3–11, 1965.

[7] Malyutov, M.B.: Review of methods and examples of Authorship Attribution of texts. *Review of Applied and Industrial Mathematics*, TVP Press, **12**, No.1, 41–77, 2005 (In Russian).

[8] Malyutov, M.B., Wickramasinghe, C.I. and Li, S.: Conditional Complexity of Compression for Authorship Attribution, *SFB 649 Discussion Paper No. 57*, Humboldt University, Berlin, 2007.

[9] Malyutov, M.B. and Brodsky, S.: MDL - procedure for attributing the authorship of texts: *Review of Applied and Industrial Mathematics*, TVP Press, **16**, No.1, 2009, 25–34, 2009 (In Russian).

[10] M. Malyutov, Recovery of sparse active inputs in general systems: a review, in *Proceedings, International Conference on Computational Technologies in Electrical and Electronics Engineering, IEEE Region 8, SIBIRCON 2010*, 15 - 22, available via IEEXplore.

[11] N. Merhav: The MDL principle for piecewise stationary sources, *IEEE Trans. Inform. Th.*, **39-6**, 1962-1967, , 1993.

[12] J. Rissanen, Universal coding, Information, Prediction and Estimation. *IEEE Trans. Inform. Th.*, **30-4**, 629-636, 1984.

[13] B.Ryabko, J. Astola and M. Malyutov, Compression-Based Methods of Prediction and Statistical Analysis of Time Series: Theory and Applications, Tampere International Center for Signal Processing. TICSP series No. 56, 2010.

[14] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley, N.Y., 2001

[15] J. Ziv, On classification and universal data compression, *IEEE Trans. on Inform. Th.*, **34:2**, 278-286, 1988.