Compression based homogeneity testing

M. Malyutov

Northeastern University, Boston,

SAPS-8, Le Mans, March 22, 2011

Let $\mathbf{B} = \{0, 1\})$, $\mathbf{x^n} \in \mathbf{B}^n = (x_1, ..., x_n)$ be a stationary ergodic random binary (**training**) string distributed as $P_0 = P$.

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

Consider a **query** binary ergodic string $\mathbf{y}^N$ distributed as P_1 and test whether the homogeneity hypothesis $P_0 = P_1$ contradicts the data or not. Let us partition $\mathbf{y}^N$ into several **slices** $\mathbf{y}_i, i = 1, \ldots, S$, of identical length divided by 'brakes' - strings of relatively small-length 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.

The ratio of $\overline{CCC}$ and their standard deviation is our homogeneity test statistic $\mathcal{T}$.

Extensive experimentation with real and simulated data (see Ryabko, Astola and Malyutov, 2010)) showed its high resolution for SED P_0, P_1. This test is nonparametric w.r.t. arbitrary stationary ergodic data distributions.

We prove consistency, asymptotic normality and optimality of exponential tails of $\mathcal{T}$ in full generality under certain natural assumption about the sizes of the training string and query slices.

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

Diversion (cross entropy) $D(P_1||P_0) = \mathbf{E}_1 \log(P_1/P_0)$

Goodness of fit tests of P_0 vs. P_1.

'Stein lemma' for SED (Ziv, 1988). *If $D(P_1||P_0) \geq \lambda$, L_i logLikelihood of P_i and any $0 < \varepsilon < 1$, then simultaneously*

$$P_0(L_1 - L_0 > n\lambda) \leq 2^{-n\lambda} \tag{1}$$

and

$$\lim P_1(L_1 - L_0 > n\lambda)) \geq 1 - \varepsilon > 0. \tag{2}$$

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

Ziv's theorem for P_0 known and P_1 unknown (Ziv, 1988). 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.

QuasiClassical assumption Training string size N_0 and query slices size N_1 grow s.t. the distribution of CCC_i converges in Probability to $P_1(L_0^n(\mathbf{y}))$ as $N_0 \to \infty$ and $N_1 \to \infty$ is sufficiently smaller than N_0.

Intuitive meaning: 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. Typical theoretical relation $N_1 \leq const \log N_0$.

Under QAA and $h_1 \geq h_0$, the following statements are true.

Theorem 1: Consistency. *The mean CCC is strictly minimal as $n \to \infty$, if $P_0 = P_1$.*

Proof. $\mathbf{E}_1(CCC) - \mathbf{E}_0(CCC) = \sum(P_0(x) - P_1(x)) \log P_0(x) = -h_0^n + \sum P_1(x) \log P_1(x)(P_0/P_1) = h_1^n - h_0^n + D(P_1||P_0)$. Proof now follows from (5) and positivity of divergence unless $P_0 = P_1$.

Generate an artificial $N_1 = n$-sequence $\mathbf{z^n}$ independent of $\mathbf{y^n}$, $\mathbf{z^n}$ distributed as P_0 and denote by CCC^0 its CCC. Also assume that $n = k(m + \delta^n)$ and the 'brakes' negligible sizes are such that the joint distribution of k slices of size m converge to their product distribution in Probability.

Theorem 2: Tails of mean CCC and ML tests. *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 $T' = \overline{CCC} - \overline{CCC}_0 > n\lambda$. Then (3), (4) are valid for this test.*

Proof sketch. Under negligible brakes and independent slices, their probabilities multiply.

For transparency: replace the condition under summation to a similar one for the whole query string:

instead of $P_0(T' > 0) = \sum_{\mathbf{y},\mathbf{z}:\overline{CCC}-\overline{CCC}_0>\lambda} P_0(\mathbf{y})P_0(\mathbf{z})$,

we write: $CCC - CCC_0 > n\lambda$ which is approximated in Probability P_0 by $L_n(\mathbf{y}) - |\mathbf{z}| > n\lambda$.

Thus

$P_0(T' > 0) \leq \sum_{\mathbf{z}} \sum_{P_0 \leq 2^{-n\lambda-|\mathbf{z}|}} P_0(\mathbf{y})P_0(\mathbf{z}) \leq 2^{-n\lambda} \sum_{\mathbf{z}} 2^{-|\mathbf{z}|} = 2^{-n\lambda}$

Informally again,
$\lim P_1(T' > 0) = \lim P_1(n^{-1}(|\mathbf{y}| - |\mathbf{z}|) > \lambda = D(P_1||P_0) + \varepsilon, \varepsilon > 0.$

$|\mathbf{y}|/n$ is in Probability P_1 around $-\log P_0(\mathbf{y}) = \mathbf{E_1}(-\log(\mathbf{P_0}(\mathbf{y}))) + \mathbf{r}$,
$|\mathbf{z}|/n$ is in Probability P_0 around $-\log P_0(\mathbf{z}) = h_0^n + r'$.

As in the Consistency proof, all the principal deterministic terms drop out, and we are left with the condition $r < \varepsilon + r'$ which probability converges to 1

since both r, r' shrink to zero in the product $(\mathbf{y}, \mathbf{z})$-Probability as $n \to \infty$.

Assume: 1. QAA and P_1 is contiguous w.r.t. P_0

transitions impossible in P_0 are equally impossible in P_1.

2. P_0 distribution of L^n is asymptotically Normal with Mean m_n and Variance σ_n^2.

Usually both are linear in n up to a slowly varying function like logarithm which is natural for compression mimicking the renewal process.

J. Ziv's claim: compressed file right hand tail with infinite memory converges to IID(1/2) in Divergence.

Implies asymptotic normality universally in our setting.

Theorem 3.*Asymptotic Normality under P_1 and all assumptions made holds (Le Cam lemmas, Hajek, Shidak). Statistic $\mathcal{T}$ has asymptotically Student distribution under P_0 and non-central Student under P_1 with k degrees of freedom.*

Sketch of AN proof. UC: $\mathbf{x}^n \to [Y_n := (m, \mathbf{y}^m)$, All UC compress in optimal way as the size $N_0 \to \infty$, $h^n/m(n) \to 1$ as $n \to \infty$.

Assumption A_1: 'Second thermodynamics law'. UC is s.t. this slope growth is monotone in Probability.

Corollary. $h^n/m(n)$ is a non-decreasing supermartingale under A_1 bounded by 1, it is AN in Probability for large n after extracting fitted trend.

Assumption A_2. *The limiting joint distribution of parameters p of $\mathbf{y}^m$ is IID for large n in Probability.*

Thus entropy $h(\mathbf{y}^m)$ as a sum of IID summands is AN for large n in Probability. $m(n)$ is a function of this sum, namely $m(n) = h^n/h(\mathbf{y}^m)$. Thus distribution of $m(n)$ is AN in view of the well-known δ-method.

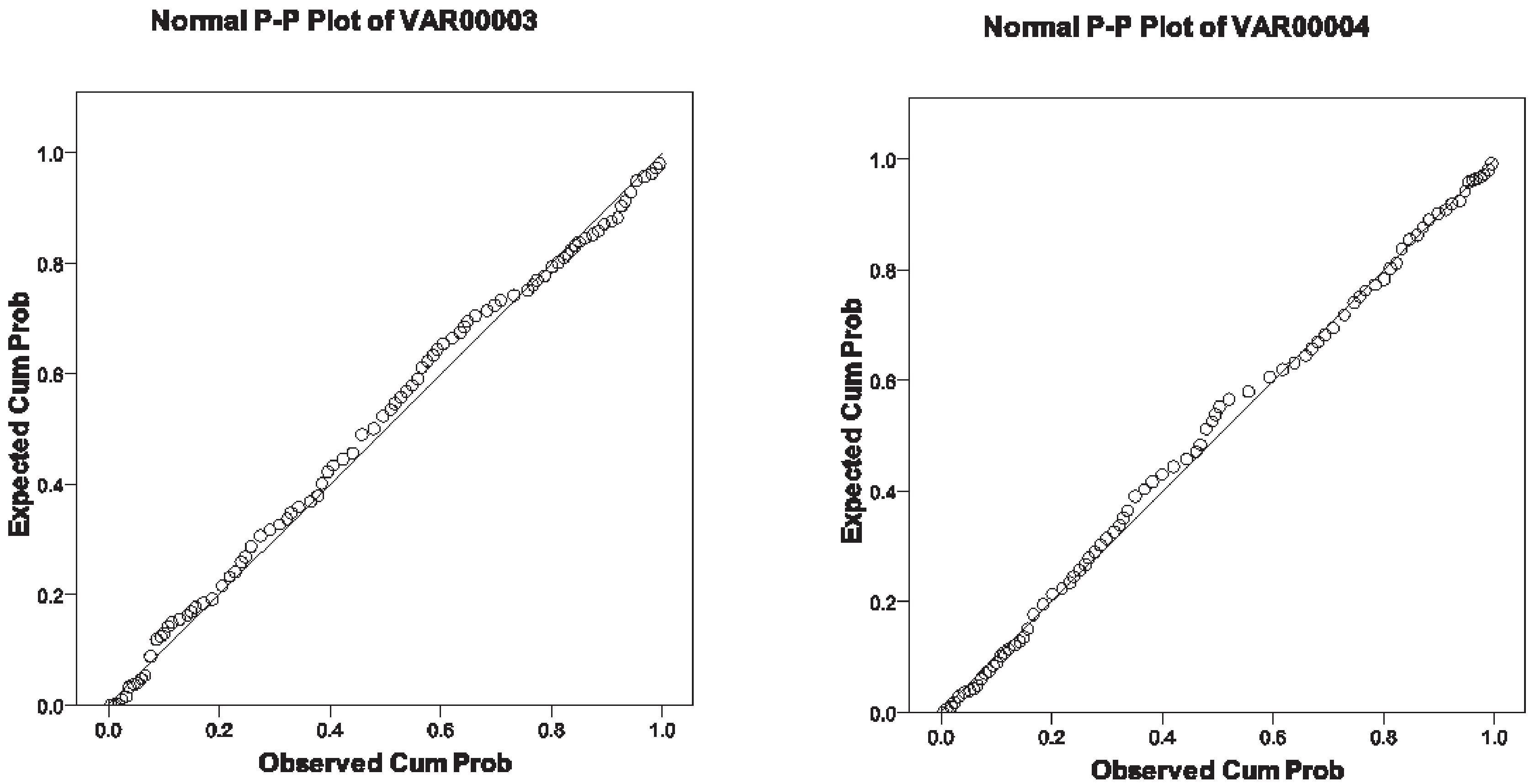

Figure 1: (a) Normal Plot: inter-CCC(slices of Brodsky 1 |whole Brodsky 2). (b) Normal Plot: intra-CCC(slices of Brodsky 2|remaining Brodsky 2).

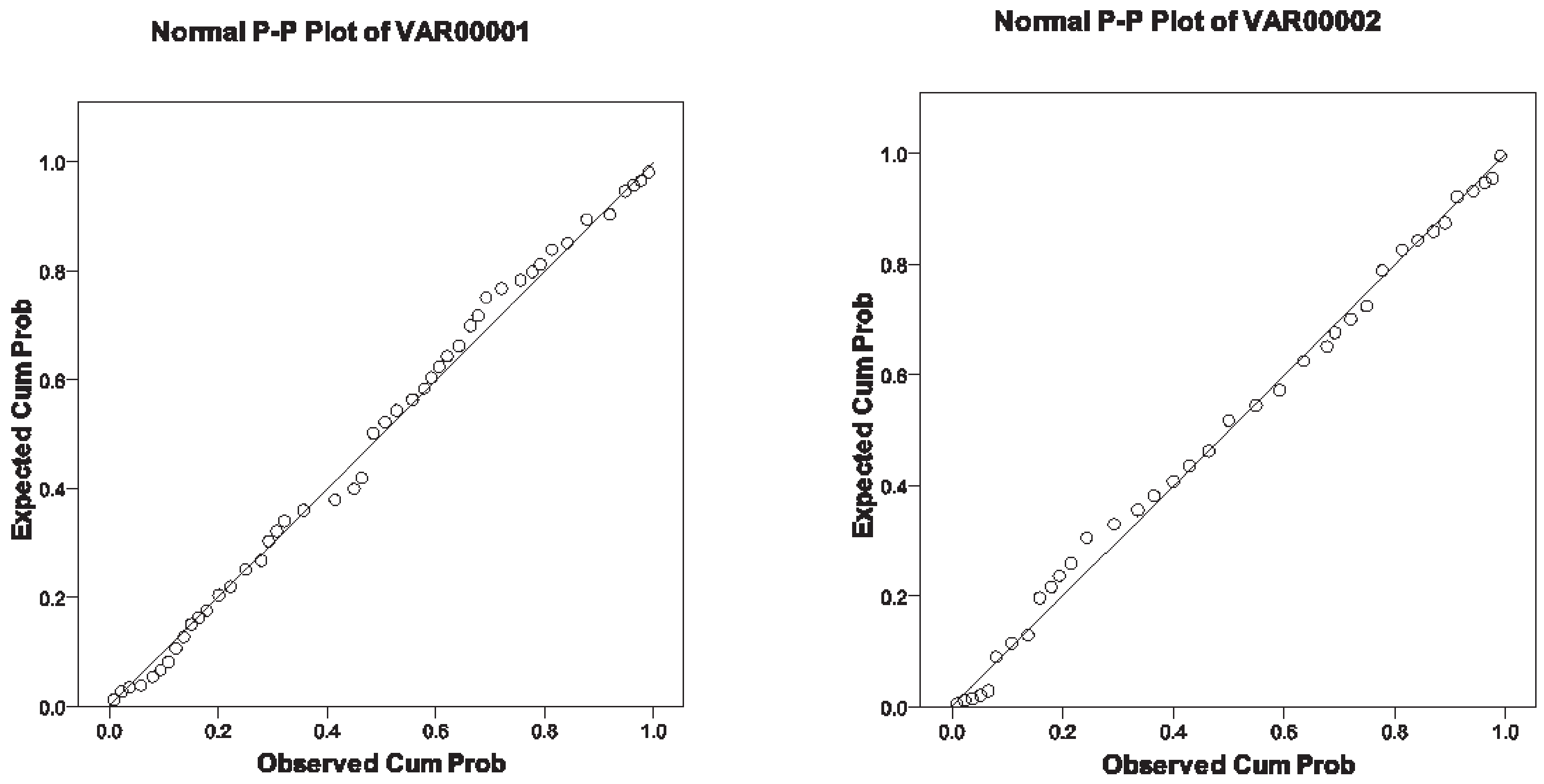

Figure 2: (a) Normal Plot: inter-CCC(slices of Gerbel |whole Marshak). (b) Normal Plot: intra-CCC(slices of Marshak|remaining Marshak).

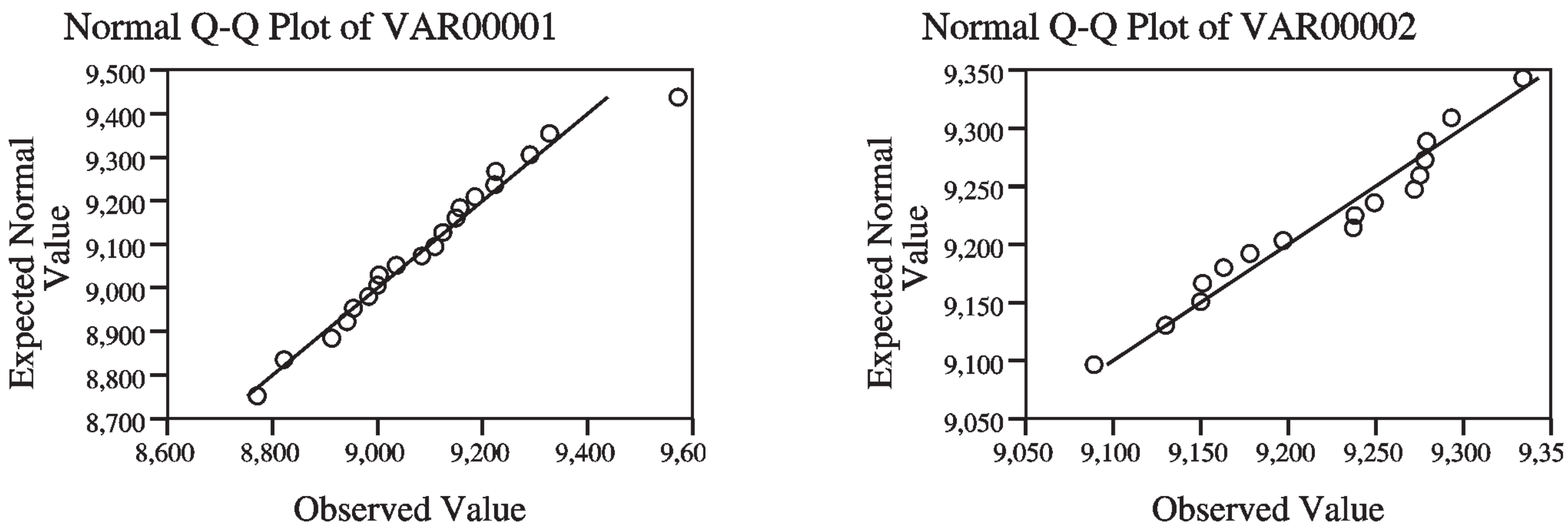

Figure 3: Normal Plots when reading octets with LZ-78: (a) intra-CCC(slices of Isaiah1 2|remaining Isaiah1). (b) inter-CCC(slices of Isaiah 2 |whole Isaiah1)

Alternative methods Some other approaches replacing Shannon-Rissanen's theory implied statistical assessment of UC-based classification decisions inspired by the Kolmogorov complexity with artificial restrictions taken from other fields by analogy and irrelevant in statistical context seem to me **pseudoscience ignoring groundbreaking ideas of Kolmogorov-Shannon-Rissanen** and misleading readers.

One of these methods, widely popularized Cilibrasi and Vitanyi (2005) showed no discrimination power in CCC-processed examples. Their ignorance in statistical aspects of compression lead to their claim that L. Tolstoy stands alone among Russian writers: they did not remove large portions of French having different entropy rate.

Expected Multi-Channel Applications Applications of multi-channel change-point detection in colored noise (possibly different among channels):

i. Detecting the *change-point* in users' profiles in a large computer network possibly caused by unauthorized intrusion into the system.

ii. Monitoring natural language texts, or the phone call traffic in 'hot' areas for their profiles matching those of special interest.

iii. Mine detection using routine road profile monitoring with relevant sensors.

References I

- D. Aldous and P. Shields, A Diffusion Limit for a Class of Randomly Growing Binary Trees, *Probab. Th. Rel. Fields*, **79**, 509-542, 1988.
- R. Cilibrasi and P. Vitanyi, Clustering by Compression, *I*EEE Transaction of Information Theory, **IT-51**, 1523–1545, 2005.
- 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.

References II

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

Acknowledgements: Author grateful to J. Ziv and V. Spokoiny for discussion.