Time series homogeneity test via training VLMC

M. Malyutov
Northeastern University, Boston,
SAPS-9, Uni. of Maine, Le Mans 3/12/2013

Intro.

- Modeling random processes as full n - Markov Chains (MC) can be inadequate, if n is small, and over-parameterized for large n.
- If say, the cardinality of the base state space is four, n=10, then the number of parameters is around 3100000.
- The popular since sixties Box-Jenkins ARIMA approach in quality control is inadequate in linguistics, genomics and proteomics, security, etc, where comparatively long **non-isotropic contexts are important for prediction** leading to huge memory size of the full n-Markov Chain (MC).

- Popularity of sparse Variable memory Length MC (VLMC), is increasing fast since its invention by **J. Rissanen** in 1983 via sparse **stochastic suffix tree construction** with algorithm **'Context'** for compression and – in 21st century, for classification aims in genomics and proteomics.
- VLMC idea: the probability of each symbol only depends on a **finite part of the total past n-string**. The **length of this relevant 'context' is a function of the past itself**. This can drastically **cut the number of parameters** of the full n-MC.
- J. Ziv (2011) shows: If the training string cannot be treated as a realization of a stationary ergodic process (as in Genomics and Proteomics), then the algorithms worked out for processing both training and query strings as realizations of VLMC are robust to departures from stationarity and even to lack of randomness.

- Testing proximity between proteins (Balding, Bush et al, 2008) used a messy **test BB** on stochastic suffix trees generated by 'Context'.
- If a universal compressor such as zip (LZ - 78) compresses efficiently the LONG training string (such as literary text), then the homogeneity **CCC - test** with its theory developed by us two years ago is a computationally simple efficient substitute for the test BB.
- Approximated **Likelihood Ratio test** for query vs. simulated training strings **given the 'frozen' stochastic suffix tree of the training string** is proposed here.

- Our test VLMClr is the Studentized sum of empirical loglikelihood ratios between the query slices and simulated training string continuation of the same length. We prove **exponential tails optimality** and **asymptotic normality of our test** similarly to our study of the CCC-test.
- An iterative procedure of improving an estimate of the steady state distribution of sparse stochastic suffix tree of the training string and complexity of the iterative procedure is outlined.
- One of major additional advantages of VLMClr over CCC is its more straightforward use for the follow up estimation of contexts contributing the most to the discrimination between strings distributions (styles of authors or proteomic sequences) which were previously shown to be distinct. This is crucial for convincing linguists or biologists, who are generally skeptical about statistical string processing.

Federalist papers discrimination : Madison vs Hamilton

Combine all 14 Madison's article into one file and use it as the training data. The cutoff number n is set to be 15 (sequence of at most 15 *English letters or space* decide the next letter).

Run the 'Context' software in R (Mächler and P. Bühlmann, 2004) for training VLMC of Madison. Divide Hamilton papers into several slices of equal size, find the log-likelihood of each query (Hamilton) slice. T-test rejects style homogeneity of the two authors for selected three slice sizes with t-values from 3 to 4. No. of Contexts is around 2400 as compared to $(27)^{15}$.

Follow up: For each context found in training VLMC of each author, calculate its mean number of occurrences. Cut Madison/Hamilton data into respectively 9/6, 14/9 and 20/14 slices to compare results stability. Finally, we calculate the t-value for occurrence differences for each VLMC context, order them and find the most significant.

Sparse VLMC over alphabet A ('letters') is a very special case of n-MC. n is the maximal length of **contexts**. A context

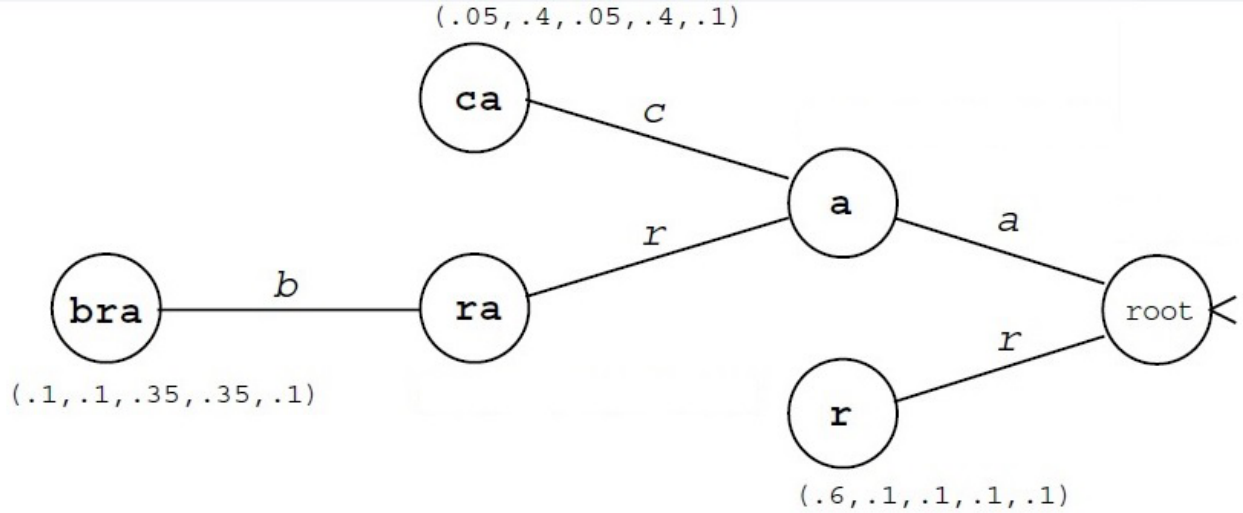
$$C(x_0) = x_{-1}, \ldots, x{-}k, k \leq n := x_{-1}^{-k}, x_i \in A \tag{1}$$

(to a current state x_0) is a subsequence of the past states x_{-1}^{-n} of the **minimal length** such that the conditional probability satisfies:

$$P(x_0)|x_{-1}^{-m}) \equiv P(x_0)|x_{-1}^{-k}) \forall m > k. \tag{2}$$

For large n, VLMC is sparse, if the total number of contexts $\mathcal{C}(n)$ is polynomial in n, informally, if $\mathcal{C}(n) << 2^n$. VLMC can be viewed as probability suffix tree, or as an 2×2^n super-matrix P of transition probabilities between contexts outlined further.

Example of stochastic context tree is on the next slide.


(.05,.4,.05,.4,.1)
ca
c
a
r
a
b
bra
ra
root
r
r
(.1,.1,.35,.35,.1)
(.6,.1,.1,.1,.1)

Part 1. Direct problems. For a context C, denote by $P_2(C|x_{-n}^{-1})$ the sum of probabilities below over all paths leading to the end of C:

$$P(x_0, x_1, \ldots, x_\tau | x_{-1}, \ldots, x_{-n}) := \Pi_{i=0}^{\tau} P(x_i | C(x_i)) \tag{3}$$

τ is the first hitting time of the end of a context C.

P_2 is a sparse $2^n \times 2^n$ transition matrix between **'peacock tails'** = sets of n-sequences ending with a context. We regard them as new 'states' of embedded 'MC with distributed transition time'.

If all the P entries are $> \delta > 0$ (total positivity) then the minimal entry of P_2 is also positive ($> \epsilon > 0$).

We define the steady state (stationary) distribution $Q(C, C \in \mathcal{C})$ over all L contexts as the solution to the equation:

$$QP_2 = P_2. \tag{4}$$

Proposition 1 (Ergodic theorem, Perron-Frobenius form).

The solution to (4) exists and is unique, if all the P entries are $> \delta > 0$ (total positivity).

Proof follows from the arguments in our Remark 2.

Conjecture. Proposition 1 is valid, under replacing the total P positivity with a weaker condition of irreducibility and aperiodicity of P in n-MC as in classical MC theory is (actually, total positivity of P_2^m for some m is sufficient to verify!).

Remark 2. Our proposition can be used to improve a Q - approximation Q_1 on the same set of contexts as the true value by using the inequality

$$||Q - Q_1 P|| \leq \lambda ||Q - Q_1|| \tag{5}$$

with $||\cdot||$ being the sum of absolute values and $\lambda < 1$ - is a certain constant somewhat similar to the second eigenvalue of the operator P.

Proof. $||Q - Q_1 P|| = ||(Q - Q_1)P|| = ||(Q - Q_1)(P - \epsilon)|| \leq$
$\sum \sum ||Q - Q_1||(P - \epsilon) \leq (1 - \epsilon)||Q - Q_1||$

The complexity of this multiplication is $O(L^k N^l)$ for some positive k, l, m rather than $2^{mn} L^l$, if we modify multiplication rule naturally for many identical entries.

Part 2. Inverse problem.

- The consistency of the Rissanen's 'Context' algorithm aimed at estimation of all contexts and corresponding transition probabilities is proved e.g. in Rissanen (1983), Bühlmann et al, Galves and Löherbach (2008).
- Test BB of homogeneity between finite VLMC's describing protein families was considered by Balding, Bush et al using rather complicated statistical tests probabilistic suffix trees.
- Simultaneously, the elementary nonparametric CCC-test based on Conditional Complexity of Compression was studied with mathematical theory developed by us recently as justification of its around 100 successful applications to authorship attribution of Literary Texts (LT), see Ryabko, Astola and Malyutov, 2010.

- VLMClr homogeneity test methodologically is similar to the CCC-test. Its mathematical theory is more straightforward than that for the CCC-test. VLMC training stage in the Rissanen's approach to VLMC-based compression is used under exponential mixing property assumption on VLMC.
- The major fancy feature of the CCC-test was necessity to cut the query into small slices to avoid the UC self-adaptation to the query which would prevent discrimination unless the slice size is sufficiently small.
- The work in progress similar to Belloni, 2011, – the incremental upgrade of the VLMC-parameters, when VLMClr is used online for fast detection of abrupt changes in the statistical string profile while avoiding false alarms in case of a slow drift of the stationary regime.

- The major feature of the CCC-test and VLMClr is cutting the query into slices. For CCC, slices must be small to avoid the UC self-adaptation to the query which would prevent discrimination unless the slice size is sufficiently small.
- The main advantage of the likelihood approach based on the VLMC training is that you can choose larger slices (**no adaptation takes places**) with the only aims to include majority of its contexts and estimate the variance of the likelihood increments.
- Let us estimate the VLMC probability tree of the large stationary ergodic 'training' string T. Then it is concatenated first with a query string Q and second, with a string S simulated from the training distribution of the same size as Q (for the latter, see algorithm in Mächler and Bühlmann, 2004.

- Then we find log-likelihoods of both Q and S using the derived probability model of the training string. Thus the novelty of VLMC based approach as compared to the CCC-methodology is possibility of likelihood evaluations instead of the lengths evaluation (which only approximates log-likelihoods under certain conditions) avoiding the problem of compressor retraining (adapting) on the query string.
- One of major **additional advantages of VLMClr over CCC** is its more straightforward use for the follow up estimation of contexts contributing the most to the discrimination of styles of authors which were previously shown to be distinct. This is crucial for convincing linguists who are generally skeptical about statistical LT processing.

- For this aim, we propose cutting both the training and query LT into several slices for estimating the mean frequencies and their empirical variances by their direct count which approximate steady state probabilities and variances of contexts.
- A better approximation to the stationary distribution of contexts can be iteratively derived by multiplying our empirical distribution with P matrix as is described in the Direct part.
- We expect more transparent results of this follow up as compared to those obtained with LZ-78 in our joint paper with G. Cunningham, 2010.

- Taking into account that the normalized frequency of occurrences of a context (with its frequency more than some threshold) in LT of size n is Asymptotically Normal (AN) with variance $\sigma_i^2/$ and σ_i can be estimated via these frequencies in slices; $i = 0$ for training LT and one for query. Then the normalized difference between frequencies for 0 and 1 cases is AN with variance as sum of the above variances.
- We decided to exclude the contexts with frequencies less than a threshold in our first applications in spite of possibility to loose some unstable but potentially significant information contained in rare contexts.
- Thus you can find p-value for their equality and order these p-values starting with the minimal ones.

- Choose one of texts as Training, estimate the VLMC for both the loglikelihoods $L_Q(k), L_T(k)$ of k-th Query and generated Training slices, $R = L_Q(k) - L_T(k)$, given the VLMC model for the Training. Compute the empirical variances V_Q, V_T of these and the t-statistic $t(k) = \bar{R}/\sqrt{((V_Q + V_T)/k)}$ with $k - 1$ df. Find $k*$ from the condition that $t(k^*)$ is maximal. Then p-value of homogeneity is evaluated for the t-distribution with $k^* - 1$ df.

- For every VLMC context evaluate its multiplicities in k(Tr) Training and k(Qu) Query slices of the SAME LENGTH, their corresponding empirical means m(Tr), m(Qu) and empirical variances V(Tr), V(Qu).
$T(k) = (m(Tr) - m(Qu))/\sqrt{(V(Tr)/k(Tr) + V(Qu)/k(Qu))}$.
Choose k(Tr) and k(Qu) s.t. $t(k*(Tr), k*(Qu))$ is maximal and slice sizes are equal. Slice size must be several times more than the context size. Order $T*$ for different contexts.

- **Madison vs Hamilton**. The VLMC significantly different contexts appear in all 9/6, 14/9 and 20/14 slices with p-value < 0.01:
- Patterns that Madison uses more frequently than Hamilton:

  *bo , *el , *on*t , *on*th , *th , ay*b , ay*be , bot , both , by , by* , by*o , by*t , d* , d*on , de* , der* , e* , ed*b , ese* , eside , ewe , f* , fore* , g*the* , han* , he*n , ix , ixe , kscgr* , lst , lt* , nd*be , orm
- Patterns that Hamilton uses more frequently than Madison:

  *at , *at* , *nat , *ther , *this* , *to , *to* , *up , *wo , ces , ct , dic , duc , e*ar , e*to* , erac , es*of* , eso , ies , ilit , ity* , lit , nati , nation , ne , om , ont , ontr
- In our discrimination we used the software developed by Mächler and described in his popular tutorial with Bühlmann.

- D. Balding, P. A. Ferrari, R. Fraiman, M. Sued, Limit theorems for sequences of random trees, Test DOI 10.1007/s11749-008-0092-z, Springer, (2008).
- A. Belloni and R. I. Oliveira, Approximate group context tree: applications to dynamic programming and dynamic choice models, arXiv.org ¿ stat ¿ arXiv:1107.0312, 2011.
- M. Mächler and P. Bühlmann, Variable Length Markov Chains: Methodology, Computing, and Software, Journal of Computational and Graphical Statistics, Volume 13, Number 2, 2004, 435 - 455.
- J.R. Busch, P.A. Ferrari, A.G. Flesia, R. Fraiman, S.P. Grynberg and F. Leonardi (2008), Testing statistical hypothesis on random trees and applications to the protein classification problem, The Annals of Applied Statistics 2009, Vol. 3, No. 2, 542563.

- Malyutov, M.B., Authorship attribution of literary texts: a review, *Review of Applied and Industrial Mathematics*, 2005, TVP Press, V. **12**, No.1, P. 41–77 (In Russian).
- Malyutov, M.B. and Cunningham, G., LZ-78 generated patterns in texts inhomogeneity, Proceedings, International Conference on Computational Technologies in Electrical and Electronics Engineering, IEEE Region 8, SIBIRCON 2010, 1, 15-22, available via IEEXplore..
- Malyutov, M.B.: Compression Based Homogeneity Testing. Doklady of RAS, **443**, 4, 1–4 (2012).

- J. Rissanen, A universal data compression system, IEEE Trans. Inform. Theory, vol. 29, Number 5, pp. 656664, 1983.
- Ryabko, B., Astola, J., Malyutov, M.B.: Compression-Based Methods of Prediction and Statistical Analysis of Time Series: Theory and Applications, Tampere, TICSP series No. 56 (2010)
- Ziv, J., On Classification and Universal Data Compression, *IEEE Trans. on Inform. Th.*, . **34:2**, . 278-286, 1988.
- J. Ziv, On Finite Memory Universal Data Compression and Classification of Individual Sequences, IEEE Trans. on Inf. Th., **54:4**, 2008, 1626-1636.

- A. Galves and E. Löherbach. Stochastic chains with memory of variable length, Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday, TICSP, 117-134, 2008.
- J. Ziv. An Axiomatic Approach to the notion of Similarity of individual Sequences and their Classifcation, CCP 2011: 3–7.