Time series homogeneity test via VLMC training

Xin Li

Northeastern University

Introduction

- 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 3.1 million.
- 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).

Introduction

- Popularity of sparse Variable memory Length MC (VLMC), is increasing rapidly after **J. Rissanen** constructed in 1983 **stochastic suffix tree** by algorithm **'Context'** for compression and proved its consistency under stationarity with exponential mixing..
- The VLMC main 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 constructing suffix tree can be used for more robust similarity tests without stationarity and even without randomness assumptions.

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, **an illustrative example of stochastic context tree is on the next slide**.

Figure

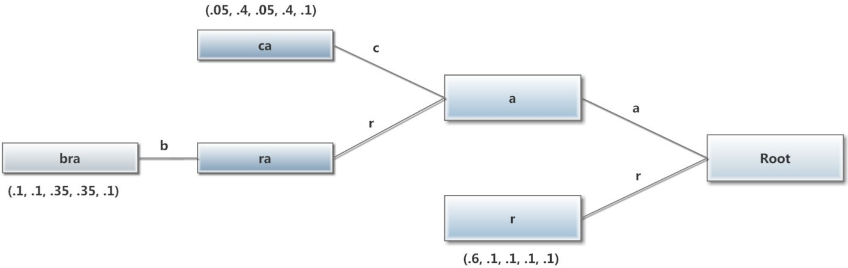


Figure: Illustrative example of stochastic context tree with distributions of the root given contexts written under the leaves of the tree

- Testing proximity between non-stationary proteins used either likelihood comparisons (G. Bejeranol, 2003), or equally unjustified messy **test BB** on stochastic suffix trees (Balding, Bush et al, 2008) generated by ‘Context’ or alternative algorithm PST of ‘Probability Stochastic Tree’.
- If a universal compressor such as zip (LZ - 78) compresses efficiently the LONG presumably stationary 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 test BB and likelihood-based tests.
- 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 log-likelihood 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.
- We find the frequencies of all contexts in slices of training and query texts.
- 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 different regions of data strings) which were previously shown to be distinct. This is crucial for convincing linguists or biologists, who are generally skeptical about statistical string processing.

- 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.
- The authorship of 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.
- All previous stylometry attributors have given all Dfs to Madison.

Our goal was answering the 3 questions:
1. Is VLMC- methodology attributing all Mf to Madison and
2. rejects significantly identity of the Hf style to that of Mf ?
3. What contexts are most statistically different in Mf and Hf?
Answers are: yes on first questions: Mf were attributed to Madison, Hf and Mf identity of styles was rejected.

First, we 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 (thus at most 15 *letters* decide about the next letter)

Table: Variable Length Markov Chain Training Result:

alphabet	'*abcdefghijklmno pqrstuvwxyz'
number of alphabet	27
number of letters	228744
maximal order of Markov chain	13
context tree size	3365
number of leaves	2353
AIC	644816

Intra VLMC test

We use each slice of Madison's data as query string and use the remaining 8 slices as training string. We want to compute the log-likelihood of each query string.

Table: Inter-loglikelihood output

-27863.56	-28047.04	-27236.75	-26559.74
-24995.70	-27173.49	-26209.20	-25182.81
-25622.52			

Hamilton's article has 152496 letters. We cut the letters into 6 slices so that each slice contains 25416 letters (25416x6=152496)
Last, do the inter-VLMC test. Use the total training result of Madison to predict the log-likelihood of each slice of Hamilton.

Table: Intra-loglikelihood output

-28552.20	-28462.64	-28511.57	-28234.03
-27227.31	-26510.97		

The mean value of these 9 log-likelihood is -27916.45 The variance of these 9 log-likelihood is 721562.6
Use the formula to do test:

$$t = \frac{l_1 - l_2}{\sqrt{var_1/n_1 + var_2/n_2}}$$

Plug in the numbers and we get the t-value 2.690809

Finally, to check consistency of our discrimination we also did a comparison between Madison itself. (we suppose to have a t-value around 0)
The inter VLMC of Madison vs Madison log-likelihood result:

Table: Inter-loglikelihood output

-27725.26	-27341.10	-26948.34	-26352.83
-23933.95	-26703.40	-25770.18	-24609.77
-25525.76			

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.

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.

Nasdaq data

We use historical **Nasdaq data** on multivariate daily returns for almost 498 days from April 4th 2011 to March 27th 2013 collected from *finance.yahoo.com* and converted into log-returns. We reduce the dimensionality of single-day returns via MatLab version of the principal component analysis (PCA) and compress the data set to the sequence of first (either two or three) Principal Components (PC) describing a major part of the data variability, see figure 3. We fit their VLMC stochastic model and apply it for discrimination between statistical properties of different parts of the data.

Fiture

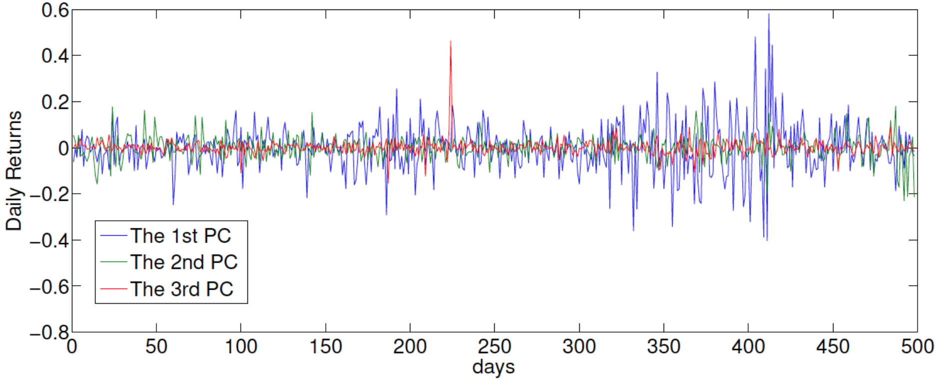


Figure: Three PC of daily returns

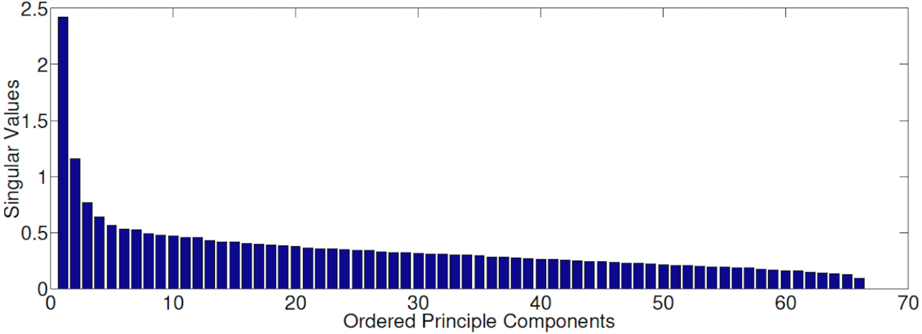


Figure: Singular values of daily returns

Results for 3 PCs

First, the range of each PC is divided into three equal intervals (bins) . Triples of PC-values are compressed to triples of integers from set $\{1,2,3\}$ according to their belonging to corresponding bins and their triples are labeled with 26 English letters from A to Z or the * symbol.

Only 8 of all 27 symbols were observed in the whole sequence.
The homogeneity t-test between 1-150 and 301-420 (quiet and volatile regions) trained on 301-420, cutting into 12 slices. The t-score is 0.1419433.

Table: Variable Length Markov Chain Training Result:

alphabet	'bdejkntw'
number of alphabet	8
number of letters	120
maximal order of Markov chain	2
context tree size	7
number of leaves	5
AIC	315.7

Table: Inter-loglikelihood output

-5.109994	-8.113694	-11.494689	-5.622557
-5.109994	-5.199606	-5.020382	-8.624520
-6.135120	-5.109994	-8.022345	-4.507818
-5.109994	-5.622557	-4.374287	

Table: Intra-loglikelihood output

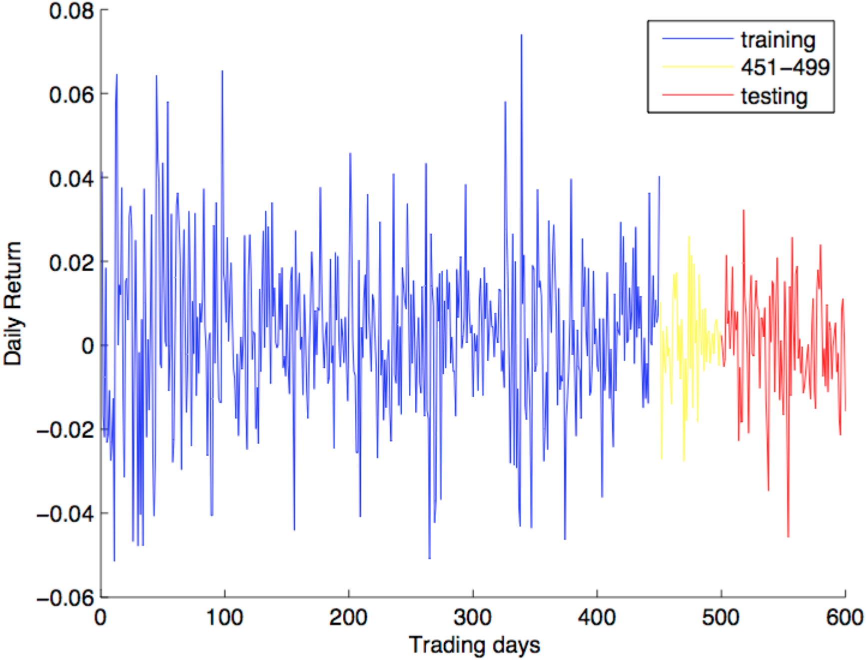
-6.224733	-6.135120	-5.622557	-5.622557
-10.284486	-10.012552	-10.144724	-10.347058
-7.736400	-6.498026	-9.982415	-9.603383

In the 2-PC case, the quiet region has the pattern L (indicating that the first PC-value is located in the second bin and the second PC-value is located in the third bin) and H (indicating that the first PC-value is located in the third bin and the second PC-value is located in the second bin) while the volatile region have the pattern B (indicating that the first PC-value is located in the first bin and the second PC-value is located in the second bin) and Q (indicating that the first PC-value is located in the fourth bin while the second PC-value is located in the second bin).

In the 3-PC case, the quiet region has the pattern N (indicating that all three PC-values are located in the second bin) while the volatile region has the pattern E (indicating that the first PC-value is located in the first bin while the second and third PC-values are located in the second bin).

In this subsection, we will make a comparison between our VLMC method and the GARCH model ([1], [2]) applied to two different sets of financial data.

The first data set we use is the daily log-return data of APPLE Inc. starting from Jan. 2nd, 2009 (Figure 4). By observation, we pick the volatile region (the first 450 days returns)and the quiet region (the 500th to 600th days returns) to make a comparison. We first fit the data with the GARCH(1,1) modeled using the MATLAB(R2011a) GARCH toolbox.


0.08
0.06
0.04
0.02
0
−0.02
−0.04
−0.06
Daily Return
0
100
200
300
400
500
600
Trading days
training
451−499
testing

$$
\begin{aligned}
y_t &= C + \epsilon_t && (3)\\
\epsilon_t &= \sigma_t z_t && (4)\\
\sigma_t^2 &= \kappa + G_1 \sigma_{t-1}^2 + A_1 \epsilon_{t-1}^2 && (5)
\end{aligned}
$$

Let $\hat{\alpha}_1$ and $\hat{\beta}_1$ be the estimator for GARCH(1) and ARCH(1) in the first model. Similar notation can be defined for $\hat{\alpha}_2$ and $\hat{\beta}_2$. From the results, we have $z_1 = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\sqrt{\sigma_{\alpha_1}^2 + \sigma_{\alpha_2}^2}} \doteq 2.1554$, and $z_2 = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\sigma_{\beta_1}^2 + \sigma_{\beta_2}^2}} \doteq -1.6971$. The p-values obtained are p_1=0.0311 and $p_2 = 0.0897$.

We apply the same data on VLMC. The homogeneity t-test between 1-450 and 500-600 (quiet and volatile regions) trained on 1-450 shows that the t-value is -16.02058. Thus, the p-value $p < 0.00001$. This p-value by VLMC is much smaller than the Z-score by GARCH.

We also use the **first** principal component of Nasdaq daily log-return data for comparison with GARCH. Again, let $\hat{\alpha}_1$ and $\hat{\beta}_1$ be the estimator for GARCH(1) and ARCH(1) in the first model. Similar notation can be defined for $\hat{\alpha}_2$ and $\hat{\beta}_2$. From the results, we have $z_1 = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\sqrt{\sigma^2_{\alpha_1} + \sigma^2_{\alpha_2}}} \doteq -1.1798$, and $z_2 = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\sigma^2_{\beta_1} + \sigma^2_{\beta_2}}} \doteq 2.1554$. And thus, the p-values are $p_1 = 0.2381$ and $p_2 = 0.0311$.

We divide the range of the first PC of Nasdaq daily log-returns into 27 bins. Each bin is labeled with 26 English letters from A to Z and symbol *. The sequence of the first PC of daily log-returns is converted into a sequence of symbols. The homogeneity t-test between 1-150 and 301-420 (quiet and volatile regions) trained on 301-420 shows that the t-value is -7.048379. Thus, the p-value is $p < 0.000001$ This p-value by VLMC is much smaller than the p-value by GARCH.

Helium emissions and seismic events

An approximately **10-year–long set of Helium emissions data from three deep Armenian wells Kadaran, Ararat and Surenavan**, the **earthquake dates** in their vicinity shown in our figure 4 was sent to us by Dr. E.A. Haroutunian (Inst. for Informatics and Automat. Problems, Armenian Acad. Sci.) for our robust analysis. In [4] they showed separately for each well that the Wilcoxon statistical test distinguishes between quiet region of the plot and that preceding strong earthquakes. Wilcoxon test was derived under independence assumption of samples which does not hold in this application. Thus our problem was to check if VLMCIr can distinguish between the above regions. Instead of separate study of data from the three wells we used PCA–compressed data. The earthquake days from the observations start were 529, 925, 1437, 1797, 1997, 2470, 2629 and 2854. The singular value plot (figure 5) suggests using either one or 2 PCs.

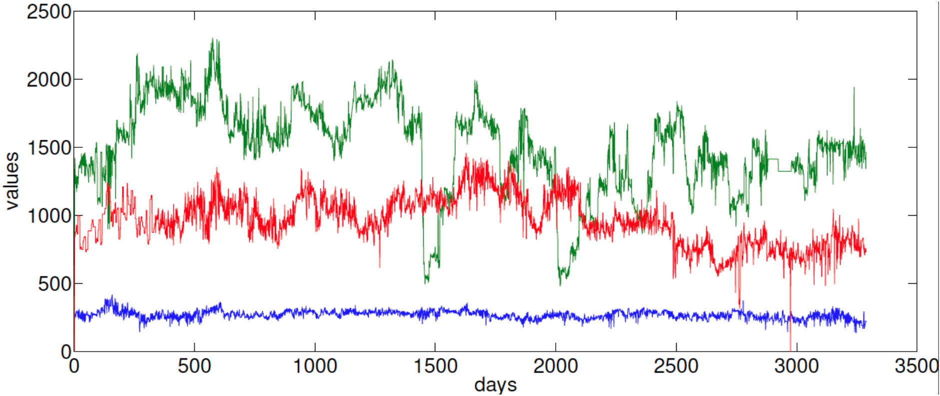


Figure: Helium emissions data from three deep Armenian wells

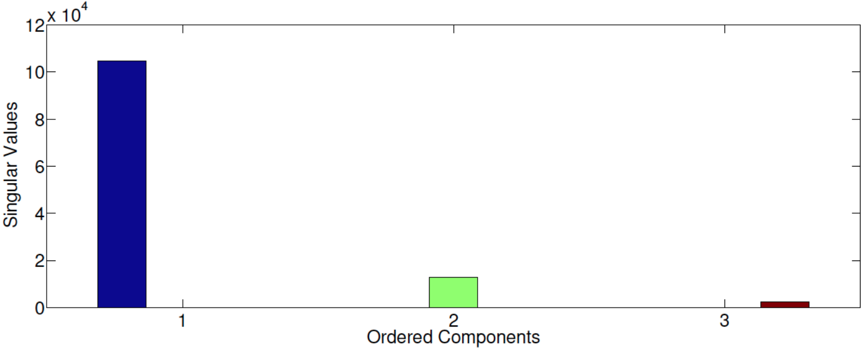


Figure: Singular Values

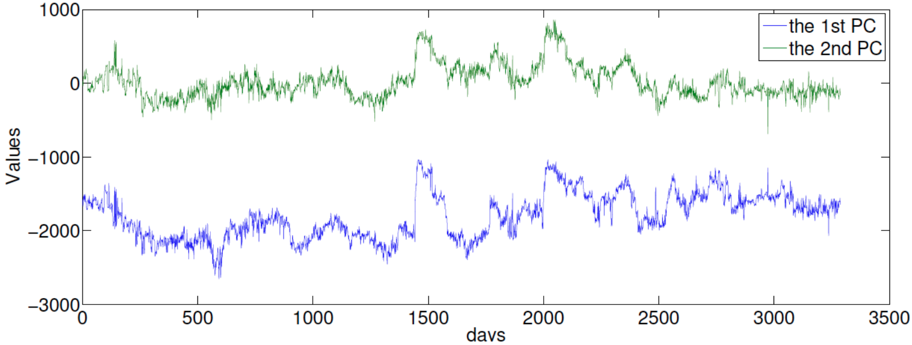


Figure: Top two principle components

In one–PC case we replaced the continuous PC values with 27 letters from A to Y describing inter $k/(27), 0 \leq k \leq 27$ –locations of observations. The ‘Context’ gave us the following parameters of the stochastic context tree:

Table: Variable Length Markov Chain Training Result:

alphabet	'abfghlmnqrstw'
number of alphabet	17
number of letters	400
maximal order of Markov chain	3
context tree size	28
number of leaves	21
AIC	1964

The quiet region 1-400 has the letters "hijklmnopqr", while the volatile region before and after earthquake 429-568 has the letters "opqrstuvwxy", it is impossible to train either region to predict the other one. It is also not necessary to do that because one can easily distinguish different regions by observing that the quiet region has the first PC value up above a level "n" (corresponding to a value -1815.101), and the volatile region has the first PC value down below a level "s" (corresponding to a number -2174.363).

In 2–PCs case, the range of each PC is divided into 5 equal bins. PC-values are compressed to 5 integers according to their belonging to bins and their pairs are labeled with 25 English letters from A to Y. Training the quiet region between 1-400 will provide the following training result:

Table: Variable Length Markov Chain Training Result:

alphabet	'abfghlmnqrstw'
number of alphabet	13
number of letters	400
maximal order of Markov chain	3
context tree size	22
number of leaves	15
AIC	1026

The homogeneity t-value between 1-400 (quiet region) and 429-528 (before earthquake) is 4.4 which means that these two region are quite different. By calculating t-value for each context, we get the context that distinguishes the most between these two regions. "l" (the first PC value in the third bin and the second PC value in the second bin) is the typical pattern of volatile regions before earthquake and "b" (the first PC value is in the first quartile and the second PC value is in the second quartile) is the typical pattern of quiet region.
The homogeneity t-test between 1-400 (quiet region and 529-568 (*after* earthquake) is 0.8, which means that we find not much difference between the quiet region and the region after earthquake. In addition, we find an interesting letter "c"(it means that the first PC is located in the first bin and the second PC is in the third bin) which can be an indicator for quiet times to follow because, when each "c" appears, there were at least 100 quiet days beyond it in the future.

References

D. Balding, P. A. Ferrari, R. Fraiman, M. Sued, Limit theorems for sequences of random trees, , arXiv.org $>$ stat $>$ arXiv:math/0406280, 2007.

G. Bejerano, Automata Learning and Stochastic Modeling for Biosequence Analysis, PhD dissertation, Hebrew University, Jerusalem, 2003.

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, Vol. 13, No. 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, Vol. 3, No. 2, 2009, 542–563.

Malyutov, M.B., Authorship attribution of literary texts: a review, *Review of Applied and Industrial Mathematics*, TVP Press, **12**, No.1, 2005, pp. 41–77 (In Russian).

References

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.

Cover, T.M. and Thomas, J.A. Elements of Information Theory, second edition, Wiley, Hoboken, 2006.

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..

Haroutunian, E.A., Safarian, I.A., Petrossian, P.A., Nersesian,H.V. Earthquake precursor Identification on the Base of Statistical Analysis of Hydrogeochemical Time Series, Mathematical problems of Computer Science, 18. 33-39, 1997.

Reimer, G.M. Use of Soil-Gas Helium Concentrations for Earthquake Prediction: Limitations Imposed by Diurnal Variation, Journal of Geophysical Research, Vol. 85, No. B6, 3107–3114, 1980.

Malyutov, M.B.: Compression Based Homogeneity Testing. Doklady of Russian Acad. Sci., **443**, 4, 427–430, 2012.

References

- Mosteller, F. and Wallace, D. Inference and Disputed Authorship: The Federalist papers, Addison-Wesley, 1964.
- J. Rissanen, A universal data compression system, IEEE Trans. Inform. Theory, Vol. 29, No. 5, 1983, pp. 656 – 664.
- B. Ryabko, J. Astola, M.B. Malyutov: Compression-Based Methods of Prediction and Statistical Analysis of Time Series: Theory and Applications, Tampere, TICSP series No. 56, Tampere Tech Uni., 2010.
- A. Galves and E. Loecherbach. Stochastic chains with memory of variable length, Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday, TICSP series, Tampere Tech. Uni., 117–134, 2008.
- C. I. Wickramasinghe, The Relative Conditional Complexity of Compression for Authorship Attribution of Texts, *PhD dissertation*, Mathematics Department, Northeastern University, Boston, MA, 2005.
- A Note on the Compaction of long Training Sequences for Universal Classification – a Non–Probabilistic Approach: arxiv.org/abs/1102.5482, 2012.

References

GARCH toolbox User's guide, The MathWorks, Inc, 2002.

John C. Hull, Options, Futures, and Other Derivatives, Eighth Edition, Prentice Hall, 2011.