DFA Method for the Analysis of Long-Range Correlations: Application to Statistical Linguistics

L. B. Ivanitskyi and O. S. Kushnir

Optoelectronics and Information Technologies Department,
Ivan Franko National University of Lviv, 107 Tarnavsky Street, 79017 Lviv, Ukraine

lubomyr.ivanitskiy@gmail.com, o.s.kushnir@lnu.edu.ua

Long-range correlations are present in various time series related to many complex systems [1]. To analyze them quantitatively, one can make use of a so-called fluctuation analysis (FA) technique. In brief, it analyzes the mean-square fluctuation $F(w) \sim w^{\alpha}$ as a function of time window size w and finds the exponent α that quantizes presence (or absence) of the long-range correlations in a series, and character of those correlations. In case of non-stationary series with variable statistical moments, i.e. trends available in the series, the FA is insufficient and must be replaced by a detrended fluctuation analysis (DFA) [2, 3]. In spite of a great attention of researchers to this field in the recent decades, we believe that some methodical, technical and even principal moments of the DFA method still need their clarification. In particular, this concerns the case if DFA is applied to such fields as, e.g., a statistical linguistics [4]. Analysis of fluctuations in linguistic systems has a number of peculiarities. In particular, the appropriate time series in many cases include only 1's and 0's, depending on whether the condition of availability of a linguistic element (a given letter, n-gram or word, a word of some length, etc.) at a certain 'time' position in a text is true or false. Moreover, the linguistic time series are 'sparse' in the sense that 1's occur mainly with the relative frequencies as small as $f = 0.01$ or less.

Using Python, we have developed a number of computer programs: (1) an 'extracting' program that assigns a time series to a given symbolic linguistic sequence, (2) a program for generating stochastic (noisy) time series with prescribed exponent values α (where the case $\alpha = ½$ corresponds to a white noise with no long-range correlations), which is based upon a standard Fourier-filtering technique, and (3) a program for analyzing time series and calculating α for the cases of DFA-n of different orders n, with a so-called double passing (see [1]). Continuous time series with the terms varying in the regions [–1; 1] and [0; 1] have been studied, as well as discrete series that involve 0's and 1's or –1's and 1's.

Below we describe in brief the main points investigated by us and the appropriate conclusions.

1. Our analyzing program reproduces with sufficient accuracy the exponents α introduced by the generating program, at least in the region 0.2÷1.2 tested by us.
2. The mean α value for the case of white noise depends very weakly on the series size L in the region $10^{10}÷10^{21}$, while the standard deviation $\Delta\alpha$ (i.e., an error of

estimating α) decreases with increasing L according to the power law $\Delta\alpha(L) \sim L^{-a}$, with a being close to the value ½ that follows from the central limit theorem (cf. also with the data [5]). Since we have $\Delta\alpha$ less than 0.01 beginning from $L_{\min} \approx (1\div5)\cdot10^4$, practical difficulties associated with preventing finite-size effects and enabling reliable DFA data can occur for many linguistic systems, which are often shorter than $L_{\min}$.

3. The influence of fitting methods built-in in Python (linear fitting in log-log scale or nonlinear power-law fitting $F(w) = Aw^{\alpha}$ (see [6, 7]) used to derive α with DFA-n (n = 0, 1 and 2) has been studied on a set of 100 binary time sequences (white noise; $L = 5\cdot10^4$; the frequency $f = ½$ for 1's). The data incline to a counter-intuitive conclusion: the fitting method affects insignificantly the results of the scaling analysis and, moreover, the linear fitting yields in somewhat lower square deviations $\Delta\alpha$. Almost the same results are obtained for the case of correlated binary series $\alpha = 0.3\div1.8$. Notice also that our results confirm a known fact that the method DFA-0 is in no case valid for the non-stationary series with $\alpha > 1$.
4. In case of 'sparse' discrete time series with no long-range correlations (the time length $L = 5\cdot10^4$ and the frequency region $f = 10^{-5}\div5\cdot10^{-3}$ for 1's), the nonlinear fitting of the DFA-0,1,2 data becomes superior, because the appropriate α's are closer to the theoretical value ½ for the white noise. Furthermore, these series reveal a crossover phenomenon at $w_{co} = 50\div6000$, with $w_{co} \sim f^{-0.8}$. It is interesting that this crossover has nothing to do with the known competition of trends and fluctuations (see [3]). As a result, one has to reckon with the following problem for the linguistic sequences: since both of the minimal and maximal time windows are limited within the DFA ($w_{\min} > w_{co}$ and $w_{\max} < 0.1L$), there can be a situation when the optimal text-window region $w_{\min} < w < w_{\max}$ for the analysis does not exist at all. The reasons for limitations put on the smallest windows deserve further studies.
5. Even with no accounting for sparseness of the time series and in case of the simplest discrete sequences of 0's and 1's taken with the same frequencies, the FA method appears to be inferior in many respects to any of DFA-n. This implies that the DFA has clear preferences over the FA when being applied to symbolic sequences dealt with in the statistical linguistics.

References

1. J. W. Kantelhard. In: Mathematics of Complexity and Dynamical Systems, Ed. by R. A. Meyers (New York: Springer), 463 (2011).
2. C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons and H. E. Stanley. *Nature* **356**, 168 (1992).
3. Kun Hu, P. Ch. Ivanov, Zhi Chen, P. Carpena and H. E. Stanley. *Phys. Rev. E* **64**, 011114.
4. W. Ebeling and A. Neiman. *Physica A* **215**, 233 (1995).
5. P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. V. Coronado and J. L. Oliver. *Phys. Rev. E* **79**, 035102(R) (2009).
6. A. Clauset, C. R. Shalizi and M. E. J. Newman. *SIAM Rev.* **51**, 661 (2009).
7. M. L. Goldstein, S. A. Morris and G. G. Yen. *Eur. Phys. J. B* **41**, 255 (2004).