Information Theory and Computational Statistics in Signal Processing and Analysis — Optimized Implementations in R

Eduarda T. C. Chagas, Alejandro C. Frery
Universidade Federal de Alagoas

etcc@ic.ufal.br

Abstract

A time series is a set of data obtained sequentially through observations over a period of time, not precisely partitioned in equal spaces of time, resulting from the operation of a system. Such system may be comprised of any mixture of deterministic and stochastic components. It is of paramount importance to detect and characterize structural dependence patterns in such data sets.

Being a solid area in Statistics, the analysis of time series has a applications in the most diversified areas, such as bank data analysis, vehicular network characterization, seismic data, among many others.

There are several ways to perform the analysis of these data, however the vast majority of them are composed of language libraries dedicated to the development of mathematical calculations, which require a minimum knowledge of them. In this way, the project aims to provide a friendly graphical tool with efficient, fast and good numerical quality functionalities that allows for interactive and exploratory analysis of the data through techniques derived from Information Theory. The fundamental requirements are portability with respect to operating systems and hardware architectures, and the use of FLOSS (Free / Open Source Software) tools.

Time series $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ are transformed in sequences of symbols by the technique of Bandt & Pompe approach [1]. Each symbol is a nonparametric mapping between groups of N values and ordinal patterns. We then analyze frequency distribution of the possible $N!$ symbols.

We then extract quantifiers, such as entropy, stochastic distances to an equilibrium distribution, and the statistical complexity from the histogram. In particular, the Shannon entropy is the metric used to measure the system disorder that gave rise to the $\boldsymbol{x}$ data; it is computed them from $\boldsymbol{h} = (h_1, \ldots, h_{N!})$, the histogram of the proportions, as

$$H(\boldsymbol{h}) = -\sum_{i=1}^{N!}(h_i \log h_i), \tag{1}$$

where, by convention, $-\infty 0 = 0$.

A common reference distribution is the uniform over the $N!$ symbols: $\boldsymbol{u} = (1/N!, \ldots, 1/N!)$. The Jensen-Shannon distance measures how close or far the underlying dynamics lie with respect to a process without information:

$$D(\boldsymbol{h}, \boldsymbol{u}) = \sum_{i=1}^{N!}\left(h_i \log \frac{h_i}{u_i} + u_i \log \frac{u_i}{p_i}\right), \tag{2}$$

where $u_i = 1/N!$.

We also calculate the statistical complexity, which seeks to find structures of interaction and dependence between the elements of the series:

$$C(\boldsymbol{h}, \boldsymbol{u}) = H(\boldsymbol{h})D(\boldsymbol{h}, \boldsymbol{u}). \tag{3}$$

Each time series can be thus represented by a point $(H(\boldsymbol{h}), C(\boldsymbol{h}, \boldsymbol{u}))$ in a compact subset of $\mathbb{R}^2$: the Entropy-Complexity plane. By means of such a tool it is possible to discover the nature of the series, determining if this is a chaotic, stochastic or deterministic sequence [2].

Although only the Shannon entropy and the Jensen-Shannon divergence were cited, the project provides other measures of entropy [3] and stochastic distances [4].

Our system promotes the analysis of the underlying dynamics of time series. It enriches such analysis by a number of descriptors, aiding a variety of applications of the Bandt & Pompe symbolization approach, for example, in the recognition of patterns of behavior in vehicular networks [5] and in the discrimination between stochastic and chaotic phenomena [6].

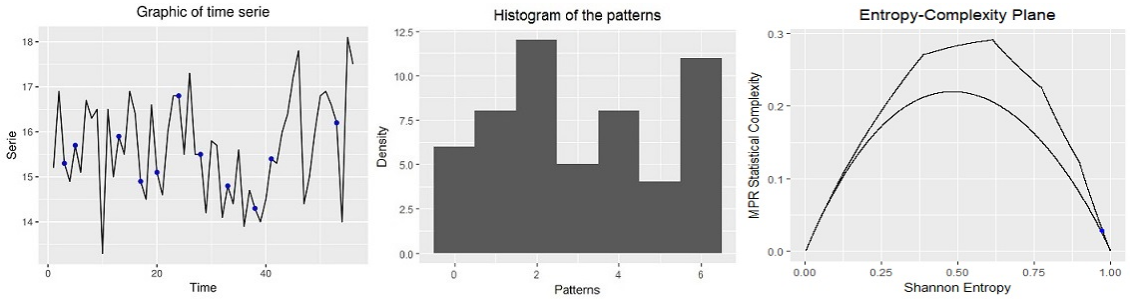

Figure 1: Graphical representation of the analysis of a time series of annual production of barley per acre.

Keywords: Time series, information theory, `R` language.

References

[1] C. Bandt and B. Pompe. Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, 88:174102–1–174102–4, Apr 2002.

[2] O. A. Rosso, H. A Larrondo, M. T. Martín, A. Plastino, and M. A. Fuentes. Distinguishing noise from chaos. *Physical Review Letters*, 99:154102, 2007.

[3] M. Salicrú, M. L. Mendéndez, and L. Pardo. Asymptotic distribution of (h, ϕ)-entropy. *Communications in Statistics - Theory Methods*, 22(7):2015–2031, 1993.

[4] L. Pardo. *Statistical Inference Based on Divergence Measures*. Number 185 in Statistics, textbooks and monographs. Chapman & Hall/CRC, Boca Raton, 2006.

[5] A. L. L. Aquino, T. S. G. Cavalcante, E. S. Almeida, A. C. Frery, and O. A. Rosso. Characterization of vehicle behavior with information theory. *The European Physical Journal B: Condensed Matter and Complex Systems*, 88(10):257–269, Oct 2015.

[6] M. G. Ravetti, L. C. Carpi, B. A. Gonçalves, A. C. Frery, and O. A Rosso. Distinguishing noise from chaos: objective versus subjective criteria using Horizontal Visibility Graph. *PLOS One*, 9(9):1–15, 2014.