An improved estimator of Shannon entropy with applications to systems with memory

Juan De Gregorio, David Sánchez and Raúl Toral
IFISC (CSIC-UIB), Instituto de Fsica Interdisciplinar y Sistemas Complejos, Campus Universitat de les Illes Balears, E-07122 Palma de Mallorca, Spain

Many systems can be described with Markovian models in which the future state of the system only depends on its present state. While in some cases this is enough to predict the evolution of the system, in other cases it is necessary to take into account also the past states of the system. Which is the minimum number of past states needed in order to faithfully determine the probability of a future state? We call this number the memory (or order) m. To find m is not a simple task given that it depends on a large number of conditional probabilities but it turns out that a simpler answer to this question can be found from an analysis of the Shannon entropy.

Estimating the Shannon entropy of a given sample is still an open problem. The naive estimator, which is calculated simply replacing the probabilities by their respective frequencies, is biased and this can result in an extreme underestimation of the entropy [1], especially in the undersampled regime where the number of possible outcomes is similar or larger than the number of observations. There have been many attempts to improve this estimator [2]. Here, I will present an estimator that generally improves the one proposed by Chao-Shen [3] using a combination of a Horvitz-Thompson adjustment [4] and a correction to the probabilities to account for missing elements in the sample. Our estimator allows us to address strong correlations and is particularly useful to study systems with memory. As an example, in figure 1 we show a plot of Shannon entropy per block of size n for a particular case of a Markovian, binary system with transition probabilities chosen randomly, alongside the results obtained with the naive estimator (green), Chao-Shen's (blue) and our proposed estimator (red), calculated from a sequence of 10^4 realizations generated numerically. It can be seen that our estimator overlaps the exact one.

Using an ordinal pattern approach, we have applied this method to the determination of the minimum memory required to describe lexical statistics of texts in different languages and we have seen that, despite the different characteristics of each language, all of them can be described with a model of memory $m = 2$. We have also applied our method to the study of daily precipitation in different worldwide locations.

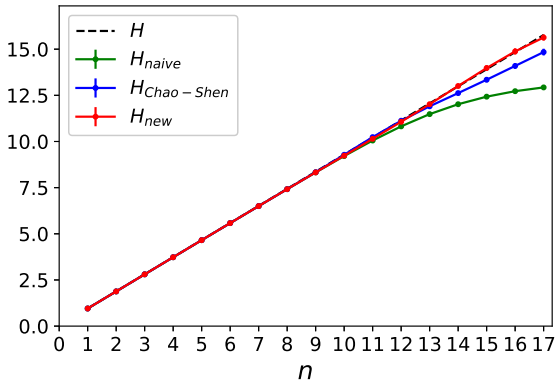

Fig. 1. Exact Shannon entropy per block of size n (dotted line) for a particular case of a Markovian, binary system with transition probabilities chosen randomly. From this setting, a sequence of 10^4 realizations is generated numerically from which we calculate the naive entropy estimator (green), Chao-Shen's (blue) and our proposed estimator (red).

[1] L. Paninski, *Estimation of entropy and mutual information*, Neural Comput. **15**, 11911253 (2003).

[2] J. Hausser, K. Strimmer *Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks*, J. Mach. Learn. Res. **10**, 14691484 (2009).

[3] A. Chao, T. J. Shen *Nonparametric estimation of Shannons index of diversity when there are unseen species in the sample*, Environ. Ecol. Stat. **10**, 429443 (2003).

[4] D. G. Horvitz, D. J. Thompson *A generalization of sampling without replacement from a finite universe*, Journal of the American Stat. Assoc. **47**, 66385 (1952).