Estimating Entropy of Correlated Discrete Sequences: Performance Analysis and a New Estimator

Juan De Gregorio[1], David Sánchez[1], and Raúl Toral[1]
[1]Institute for Cross-Disciplinary Physics and Complex Systems IFISC (UIB-CSIC),
Campus Universitat de les Illes Balears, E-07122 Palma de Mallorca, Spain

The Shannon entropy of a random variable, which measures its intrinsic uncertainty, is widely used in a variety of fields, such as statistical physics, biology, neuroscience, cryptography and linguistics, among many others.

Estimating the entropy of discrete sequences can be challenging due to limited available data. Moreover, there is currently no known unbiased estimator for the entropy [1], making the task even more difficult, especially in an undersample regime, in which the size of the sequence N is smaller than the number of possible outcomes L.

While numerous entropy estimators have been proposed in the literature (refer to, e.g., [1, 2]), their performance when considering the bias and standard deviation vary significantly depending on the specific system under study and the size of the available data. Most of these entropy estimator are designed particularly considering that the sequence is generated by independent events. To address possible correlations within the sequence, we propose a new entropy estimator that takes into account the order in which the elements appear in the sequence, as well as a Horvitz-Thompson correction [3] to address the issue of potential missing outcomes in a short sequence ($N < L$) [4].

Since entropy estimators are typically evaluated and compared only considering independent sequences [5], we have conducted a detailed analysis of the performance of some of the mostly used entropy estimators, when applied to correlated data. Specifically, we present the results for i) binary Markovian sequences (Fig. 1) and ii) Markovian systems in the undersample regime. In addition, we have also included our new estimator into this analysis (red crosses in Fig. 1) and we have found that it performs remarkably well in terms of the bias although showing a large dispersion.

[1] Paninski, L., *Estimation of entropy and mutual information*, Neural Computation **15**, 1191-1253, 2003.

[2] Chao, A.; Shen, T., *Nonparametric estimation of Shannons diversity index when there are unseen species in sample*, Environmental and Ecological Statistics **10**, 429-443, 2003.

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

[4] De Gregorio, J.; Sánchez, D.; Toral, R., *An improved estimator of Shannon entropy with applications to systems with memory*, Chaos, Solitons & Fractals **165**, 112797, 2022.

[5] Contreras Rodrguez, L.; Madarro-Cap, E.J.; Legn-Prez, C.M.; Rojas, O.; Sosa-Gmez, G., *Selecting an Effective Entropy Estimator for Short Sequences of Bits and Bytes with Maximum Entropy*, Entropy **23**, 561, 2021.

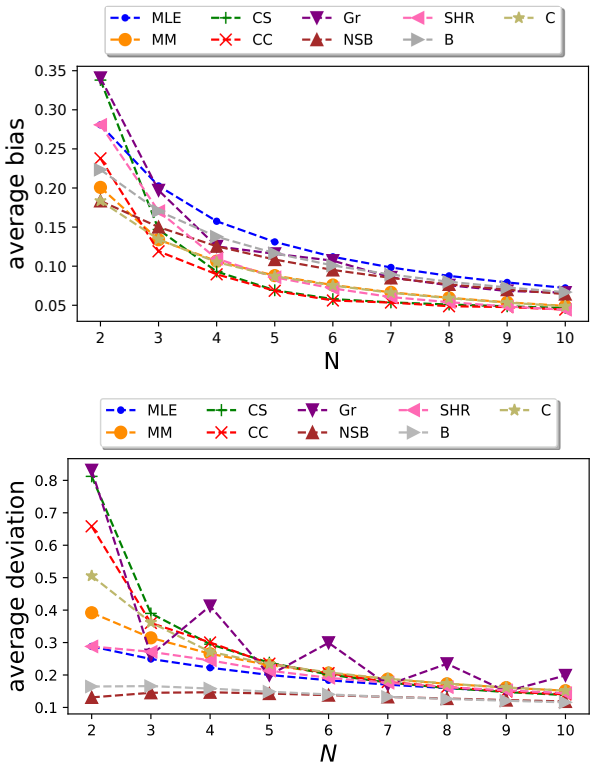

Fig. 1. Average bias (top) and deviation (bottom) of the entropy estimators when applied to Markovian, binary sequences for different sequence size N. The red line corresponds to our proposed estimator.