Estimation of Shannon entropy with applications to systems with memory

Raul Toral[1], David Sanchez, Juan de Gregorio
[1]*IFISC, Institute for Cross Disciplinary Physics and Complex Systems, Palma De Mallorca, Spain*

The estimation of entropy for very short discrete sequences is a significant challenge with broad applications in fields such as statistics, linguistics, ecology, and neuroscience. Although numerous attempts have been made to develop an unbiased estimator with minimal mean squared error, the performance of proposed estimators varies widely depending on the system being studied and the available data size. In this study, we analyzed commonly used entropy estimators for two types of Markovian systems: binary sequences and undersampled Markovian systems. Our approach centers on the block Shannon entropy, where data is grouped into blocks of a specified size, and the entropy is derived from the probability of each block. The maximum likelihood estimator for entropy works well when the sequence length is much greater than the number of possible blocks. However, many entropy estimators in the literature are limited to independent systems and have not been applied to correlated systems. Here we have extended this analysis to Markov chains, which have many applications across disciplines. To address these issues, we propose a new method to determine the order of a Markov chain that best represents a given dataset. Our method is model-independent and sufficiently precise, even in the undersampled regime. We introduce an estimator that can be used for any observable, such as entropy, to provide information on correlated systems. Importantly, the method is valid for both Markovian and non-Markovian systems. We compare the performance of our proposed estimator with those designed for independent systems and find that it significantly increases the accuracy of entropy estimation for systems with memory. We tested our method on both numerically generated and real-world datasets and observed promising results. These successful results will certainly encourage further applications of the proposals discussed in this work. Additionally, our improved estimator can be applied to any model of correlations (not necessarily a Markov chain) and is thus of interest in situations where entropy is fundamental to understand the system's properties.

References

[1] J. D. Gregorio, D. Sanchez, R. Toral, An improved estimator of Shannon entropy with applications to systems with memory, Chaos Sol. Fractals 165, 112797 (2022).