Eur. Phys. J. Special Topics **226**, 2191–2204 (2017)

DOI: 10.1140/epjst/e2016-60397-x

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

Dynamical complexity of short and noisy time series

Compression-Complexity vs. Shannon entropy

Nithin Nagaraj[1,a] and Karthi Balasubramanian[2,b]

[1] Consciousness Studies Programme, National Institute of Advanced Studies, Indian Institute of Science Campus, Bengaluru 560 012, India
[2] Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India

Received 30 August 2016 / Received in final form 20 November 2016
Published online 31 January 2017

Abstract. Shannon entropy has been extensively used for characterizing complexity of time series arising from chaotic dynamical systems and stochastic processes such as Markov chains. However, for short and noisy time series, Shannon entropy performs poorly. Complexity measures which are based on lossless compression algorithms are a good substitute in such scenarios. We evaluate the performance of two such *Compression-Complexity Measures* namely Lempel-Ziv complexity (LZ) and Effort-To-Compress (ETC) on short time series from chaotic dynamical systems in the presence of noise. Both LZ and ETC outperform Shannon entropy (H) in accurately characterizing the dynamical complexity of such systems. For very short binary sequences (which arise in neuroscience applications), ETC has higher number of distinct complexity values than LZ and H, thus enabling a finer resolution. For two-state ergodic Markov chains, we empirically show that ETC converges to a steady state value faster than LZ. *Compression-Complexity measures* are promising for applications which involve short and noisy time series.

1 Introduction

Claude Shannon introduced the idea of 'entropy' as a quantitative measure of information in 1948 [1] when he was building a mathematical theory of communication. The notion of entropy had already been proposed in thermodynamics (Clausius, 1965) and in statistical physics (Boltzmann and Gibbs, 1900s). Shannon entropy of a discrete random variable is defined as:

$$H(\chi) = -\sum_{i=1}^{M} p_i \log_2(p_i) \quad \text{bits/symbol}, \tag{1}$$

[a] e-mail: nithin@nias.iisc.ernet.in
[b] e-mail: b_karthi@cb.amrita.edu

where χ is the random variable with M possible events and the probability of occurrence of the i-th event is given by $p_i > 0$. The maximum value of the concave function $H(prob.)$ is achieved for a uniform random variable with all events equally likely ($H = \log_2(M)$ bits).

Apart from playing a fundamental role in communications, information and coding theory, Shannon entropy is also used to characterize the complexity of a time series. Low entropy of a time series indicates low complexity (less randomness and hence more structure) whereas a higher value of entropy of a time series would imply a higher complexity (more randomness and hence less structure). This is because, Shannon entropy characterizes the degree of compressibility of an input sequence. Today, Shannon entropy (or H), and some of its related information theoretic measures (such as mutual information, conditional entropy etc.), continue to be widely used as measures of dynamical complexity in several applications. It is used in biomedical applications [2], for e.g., as a pattern classification tool in heart rate variability analysis [3]; to measure structural and dynamical complexity of networks [4] and communication complexity [5]; for biological sequence analysis in bioinformatics [6,7]; in econometric/financial time series analysis [8–10]; and not to miss out on the various entropic forms in physics [11]. This is by no means an exhaustive list, but only serves as indicative of the diverse domains in which Shannon entropy is applied.

However, Shannon entropy (H) has serious drawbacks when the time series under consideration is short and noisy. In this work, we point out these limitations and propose the use of *Compression-Complexity* measures to overcome these limitations of Shannon entropy for characterizing dynamical complexity of short and noisy time series. *Compression-Complexity* measures shall be defined as complexity measures based on lossless compression algorithms. This is the subject matter discussed in Sections 2 and 3 of this paper.

Signals that are seen in real world are never completely random in nature, though they may be stochastic in origin. In several instances, these signals behave as information sources that may be modelled as Markov or hidden Markov processes. Markov chains, named after Andrei Andreievich Markov (1856–1922), is a type of random process which has the property that the current state of the system depends only on its immediate past state[1] and not on the sequence of past states prior to that. The transition from one state to another state is captured by transition probabilities. Markov chains have played a vital role for modeling in statistical mechanics. Dating back to the urn models for mixing of D. Bernoulli (1769), Laplace (1812) and Ehrenfest (1907), these are simple examples of Markov chain models (known as random walks).

Many real world systems behave like Markov sources that produce signals that may be recreated using finite chain Markov process models. *E.g.*, the patterned structure of heart-beat intervals [12–15], base compositions of DNA sequences [16–19], decomposition and recognition of speech [20–22], language scripts modelling [23–25], information sources in communication systems [26–28], trend prediction of stock indices [29] and analysis of share prices [30], can all be mathematically viewed as Markov processes/chains. Hence, a study of the performance of complexity measures on data produced from Markov chains would be a good indication of its performance on real world signals. In section 4, we simulate a 2-state Markov chain and evaluate the performance of *Compression-Complexity* in characterizing its complexity.

We conclude with future research directions in the last section.

[1] This is the definition of a 1-order Markov process.