



Entropy and Complexity: Analysis of Symbolic Sequences



R. Steuer L. Molgedey W. Ebeling M. A. Jiménez-Montaño

Institut für Physik, Humboldt-Universität zu Berlin
<http://summa.physik.hu-berlin.de> steuer@physik.hu-berlin.de

I. Introduction: Symbolic Dynamics

One commonly used method for time series analysis is the concept of symbolic dynamics. The basic idea is to convert the measured data into a corresponding sequence of symbols and thus giving a coarse-grained description of the investigated system. The resulting sequence of symbols could then be examined by the concepts of complexity and Shannon entropy. However, the success of this approach will mainly depend on the choice of the partition. While the construction of a generating partition mostly fails in the case of empirical data, there are still some criteria on how such a mapping should be performed. Here we will discuss the application of these concepts focusing on:

- Effects of choosing different partitions.
- Comparison of the numerically obtained values with theoretical estimates for certain well-known model systems.

II. The Shannon n-gram Entropies

Definition: As the basic ingredients we have

- Symbolic Sequences of length N , consisting of successive symbols (letters) from a finite alphabet $A = \{A_1, A_2, \dots, A_k\}$. Substrings are termed n -words or n -blocks.

...1001010110101101011010...

→ n-Block

- The maximal possible number of different n -blocks occurring in the sequence is k^n . Both, finite size effects and structure within the sequence will reduce this number considerably.
- Assuming stationarity, any n -block will occur with a probability $p_i^{(n)}$.
- The information content of each n -block is $I_i^{(n)} = -\log p_i^{(n)}$.

$$\text{The Entropy: } H_n = -\sum p_i^{(n)} \log p_i^{(n)}$$

The entropies H_n are a measure of predictability and give the average information contained in a word of length n . For our purpose we will largely rely on the conditional entropies h_n , defined as:

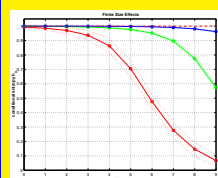
$$h_n := H_{n+1} - H_n$$

The conditional entropies h_n give the average information necessary to predict the next letter, given the preceding n letters. Note that $h_{n+1} \leq h_n$. A quantity of particular interest is the limit h for large n .

$$h := \lim_{n \rightarrow \infty} h_n$$

Finite Size Effects: The calculation of entropies is seriously affected by systematic errors due to the finite size of the samples. The Shannon entropy gets systematically underestimated.

$$\langle H_n^{observed} \rangle \approx H_n - \frac{M_n - 1}{2N \log \lambda} \quad (1)$$



The Plot shows the average conditional entropies (h_n) of a binary random sequence for different sequence lengths $N = 100, 1000$ and 10000 . The theoretical value is $h_n = 1$ for all n . All deviations are due to finite size effects. The dashed lines indicate the standard deviation estimated out of 1000 trials. For further details see [3] and references therein.

III. Entropy Analysis of Scalar Time Series

The application of the entropy concept requires a symbolic representation of the real valued data x_i . This is obtained by introducing a partition $P = \{P_1, \dots, P_\lambda\}$, which divides the phase-space Γ into λ disjoint sets, each of which is labelled with a symbol A_i out of the alphabet A . Consequently, the time evolution of the system is translated into a sequence of symbols. Considering the entropy $h_n(P)$ as a function of the partition P , one gets the Kolmogorov-Sinai entropy h_{KS} as the supremum over all possible partitions.

$$h_{KS} := \sup_{(P)} \lim_{n \rightarrow \infty} h_n(P)$$

Pesin identity: Kolmogorov-Sinai entropy h_{KS} can be no larger than the sum of the positive Liapunov exponents λ_i^+ . For certain systems the equality holds:

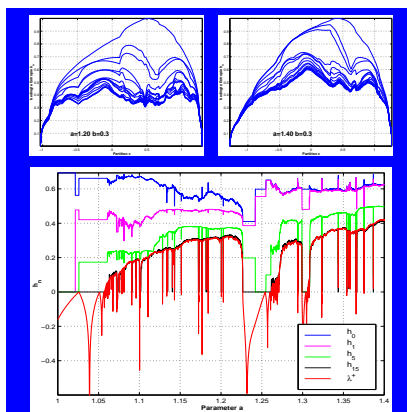
$$h_{KS} = \sum \lambda_i^+$$

Generating Partitions: In the case of deterministic and time-discrete systems $f: R^m \rightarrow R^m$ any given n -word identifies a certain region Γ_n in phase-space. For a generating partition P_n this region is supposed to shrink further and further for increasing n (dynamical refinement). Each infinite symbolic sequence corresponds to an individual point in phase-space. In this case the supremum over all possible partitions may be omitted.

The optimal partition: For practical purposes we will simply define the optimal partition as the partition that most effectively the randomness of the original data. That is, the best partition is the one that gives the largest complexity estimate.

III. Example: The Hénon Map

To exemplify these concepts we will discuss the 2-dimensional Hénon Map $x_{n+1} = 1 - ax_n^2 + bx_{n-1}$. We will not use the generating partition given by Grassberger et al. (1985), but instead we will simply estimate the conditional entropy as the maximal value for a binary partition. The plot confirms a strong dependence of the conditional entropy from the choice of the partition parameter.

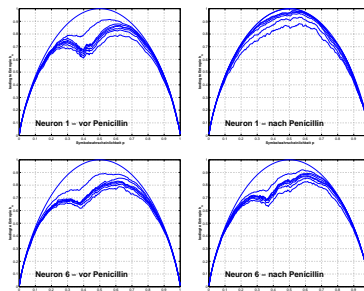


upper plot: The conditional entropies h_0 to h_{12} (from above) as a function of the binary partition threshold c . lower plot: The conditional entropies and the positive Liapunov exponent λ^+ for the dissipative Hénon Map.

IV. Application to Neuronal Spike Trains

The data: The measurement of the investigated interspike intervals goes back to Rapp et al. [2]. The data consists of seven single-unit records obtained from cortical neurons of a rat before and after the application of penicillin directly on the surgically exposed cerebral cortex. This procedure is believed to be a (rather imperfect) model of human focal epilepsy. For further details see [2] and references therein.

The conditional entropies: All data was mapped on binary symbol sequences. The conditional entropies h_n were estimated as a function of the threshold parameter c . In figure 3 however we used the corresponding symbol probability p to obtain a certain invariance to data transformations.



The conditional entropies h_0 to h_7 (from above) as a function of symbol probability. Left: Neuron 1 and 6 before the application of penicillin. Right: After the application of penicillin.

Result: Starting with neuron 1 one observes that the estimated conditional entropies strongly depend on the choice of the partition. After the application of penicillin this structure has vanished. The plot looks like one of a random sequence. However, this is no systematic feature of the investigated spike trains. For neuron 6 the structure is present before and after the application of penicillin.

Surrogate Sequences:

The validity of results obtained from finite time series is usually verified by considering surrogate sequences with known statistical properties.

The most straightforward way in this case is to build a markovian process with memory m having the same transition probabilities as the original sequences. The plot shows the conditional entropy of neuron 1 before penicillin treatment and the corresponding results for markovian surrogates up to order $m = 3$. The errorbars denote the standard deviations of the surrogate ensembles. In the realm of good statistics, that means up to a word length of $n \approx 5$, the time series is consistent with a third order markovian process.

Further Literature

- (1) Komplexe Strukturen: Entropie und Information W. Ebeling, J. Freund, F. Schweitzer; B.G.Teubner (1998)
- (2) The Algorithmic Complexity of Neural Spike Trains Increases During Focal Seizures P.E. Rapp et al.; The Journal of Neuroscience, Vol. 14 No.8 (1994)
- (3) Finite Sample Effects in Sequence Analysis H. Herzel, A. O. Schmitt, W. Ebeling; Chaos, Solitons and Fractals, Vol.4 No. 1 (1994)
- (4) Entropy and optimal Partition for Data Analysis R. Steuer, L. Molgedey, W. Ebeling and M. A. Jiménez-Montaño; EPJ B, submitted (2000)