Eur. Phys. J. B **19**, 265–269 (2001)

THE EUROPEAN PHYSICAL JOURNAL B

EDP Sciences

Entropy and optimal partition for data analysis

R. Steuer[1,a], L. Molgedey[1], W. Ebeling[1], and M.A. Jiménez-Montaño[2]

[1] Institute of Physics, Humboldt-University, 10115 Berlin, Germany
[2] Department of Physics and Mathematics, Universidad de las Americas-Puebla, Cholula 72820, Mexico

Received 1st September 2000

Abstract. The concept of symbolic dynamics, entropy and complexity measures has been widely utilized for the analysis of measured time series. However, little attention as been devoted to investigate the effects of choosing different partitions to obtain the coarse-grained symbolic sequences. Because the theoretical concepts of generating partitions mostly fail in the case of empirical data, one commonly introduces a homogeneous partition which ensures roughly equidistributed symbols. We will show that such a choice may lead to spurious results for the estimated entropy and will not fully reveal the randomness of the sequence.

PACS. 05.45.Tp Time series analysis

1 Introduction

Empirical and experimental work usually consists to a great deal in acquiring records of real numbers. The main task then is to extract the features of the investigated system from that time series and, hopefully, be able to take a glance at the laws which governs them. One commonly used method for this purpose is the concept of symbolic dynamics. The basic idea is to convert the measured time series into a corresponding sequence of symbols and thus giving a symbolic representation of the investigated system. Concepts to analyze such sequences were already given 1951 by C. Shannon in his seminal paper "Predictions and Entropy of Printed English". Since then, Shannon's approach was applied to a wide range of topics, including biosequences and many other information carriers [1–6].

In the first section we will give a brief introduction to the concepts of symbol sequence analysis. In the second section we will review the application of these concepts to scalar time series. One common approach is to introduce a coarse-grained description of the sequence by partitioning the continuous phase space into a finite number of cells. We will discuss the application of these concepts, using the logistic map as a well known example. In particular we will investigate the effects of using different partitions and compare the results to earlier obtained theoretical values.

Finally we will apply these methods to the analysis of neural spike trains, going back to measurements of Rapp *et al.* [6]. For this purpose we need to consider the systematic bias and finite length effects in our entropy approximations. The validity of the results will be tested using ensembles of surrogate sequences.

[a] e-mail: steuer@physik.hu-berlin.de

2 The Shannon n-gram entropies

Let S be a sequence of length N composed of symbols (letters) from a finite alphabet of λ letters. Substrings of n letters are termed n-words or n-blocks. Assuming stationarity, any n-word i is expected to occur with the well-defined probability $p_i^{(n)}$ at any arbitrary site in the sequence. Following Shannon, the block entropies of words of length n (n-gram entropies) are given by

$$H_n = -\sum p_i^{(n)} \log p_i^{(n)}. \qquad (1)$$

The summation has to be carried out over all words with $p_i^{(n)} > 0$. The entropies H_n measure the amount of information contained in a word of length n or, equivalently, the average information necessary to predict a subsequence of length n. Thus one may introduce the conditional entropies h_n as the average information necessary to predict the next symbol, given the preceding n symbols, by

$$h_n = H_{n+1} - H_n. \qquad (2)$$

The definition of the h_n is supplemented by $h_0 := H_1$. Note that the interpretation of the conditional entropies h_n implies the inequality

$$h_{n+1} \leq h_n. \qquad (3)$$

A quantity of particular interest is the entropy of the source defined as the limit of the conditional entropies h_n for large n.

$$h := \lim_{n\to\infty} h_n = \lim_{n\to\infty} \frac{H_n}{n} \cdot \qquad (4)$$

The limit entropy h (or Kolmogorov-Sinai entropy) is the average amount of information necessary to predict the next symbol when being informed about the complete prehistory of the system. Since a positive Kolmogorov-Sinai entropy implies the existence of a positive Lyapunov exponent, it is an important measure of chaos. The speed of convergence of the differential entropies to their limit h can be taken as a measure of correlations [7–9].

3 Entropy analysis of scalar time series

A direct application of the entropy concept requires a symbolic representation of the real value data x_t.

This is achieved by introducing a (finite) partition P, which divides the full continuous phase-space Γ into λ disjoint sets. Each set is labelled with a symbol (or letter) A_i out of the alphabet A. The resulting symbol sequence now gives a coarse-grained description of the time evolution of the dynamical system. Applying the concepts of the first section on the symbolic sequences one gets the conditional entropies $h_n(P)$ with respect to the partition P. In the case of deterministic and time-discrete systems $f : R^m \to R^m$ each n-word identifies a region Γ_n in phase-space,

$$\Gamma_n = A_1 \cap f^{-1}(A_2) \cap \ \ldots \ \cap f^{-(n-1)}(A_n) \qquad (5)$$

with $f^{-(i-1)}(A_i)$ denoting the $(i-1)$th backward iterate of the partition corresponding to the ith letter. For an appropriate choice of the partition the region Γ_n is supposed to shrink further and further for increasing n (dynamical refinement). The Kolmogorov-Sinai entropy h is given by the limit of the conditional entropies $h_n(P)$ for finer and finer partitions or equivalently, as the supremum over all possible partitions P

$$h = \sup_{\{P\}} \lim_{n\to\infty} h_n(P). \qquad (6)$$

For a generating partition $P_{\rm g}$ the limit for finer and finer partitions may be avoided. A partition is called generating if the dynamical refinement for increasing n divides the phase space into arbitrarily fine regions, that is each (infinite) symbol sequence corresponds to an individual point in phase-space. In this case the mapping between the (infinite) symbolic sequence and the (infinite) scalar time series is unique.

Even though generating partitions are known for several systems [10], in most cases a direct application of these concepts fails due to the obstacle of constructing such a partition for a given system. For practical purposes we will simply define the best partition as the the partition that most effectively reveals the randomness of the original data as already suggested in [6].

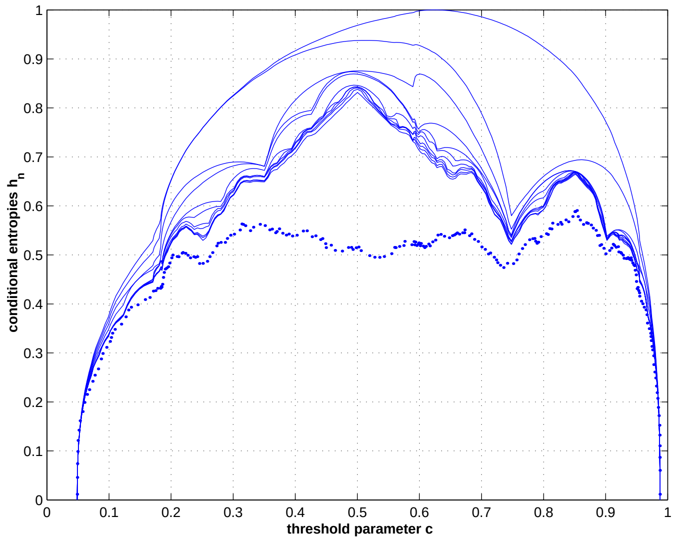

Fig. 1. The conditional entropies h_0 to h_{10} for the logistic map at $r = 3.95$ in descending order as a function of the threshold parameter c and sequence length $N = 100\,000$. The dotted values indicate the conditional entropy h_{10} calculated with $N = 1000$.

However we shall note the relationship of Kolmogorov-Sinai entropy h to the Lyapunov exponents λ. In most cases h is equal to the sum of positive Lyapunov exponents λ^+ (Pesin identity).

$$h = \sum \lambda_i^+. \qquad (7)$$

4 The logistic map

As perhaps one of the best studied system in nonlinear dynamics, the logistic map needs no special introduction.

$$x_{n+1} = f(x_n) = rx_n(1 - x_n) \qquad r \in [0, 4]. \qquad (8)$$

We will use it to exemplify the concept of a coarse grained description and will benefit from the fact that most properties are known analytically. A generating partition is given by the critical point $c = 0.5$.

$$x_n \in [0, c] \to S_n = 0 \qquad x_n \in (c, 1] \to S_n = 1. \qquad (9)$$

The resulting symbolic dynamics at the period accumulation point $r_\infty = 3.5699...$ has already been studied in detail by several authors [11]. Now we neglect our knowledge of the generating partition and estimate the conditional entropies h_n for $c \in [0, 1]$ and $r = 3.95$. As expected and already observed in [12] the higher order entropies attain their maximal value for a partition with $c = 0.5$ (see Fig. 1). With respect to the maximum entropy we will call this an optimal binary partition. As observed in Figure 1 the accuracy of the estimated entropies h_n is seriously affected by systematic errors due to the finite length N of the sequence. How these difficulties can be dealt with has

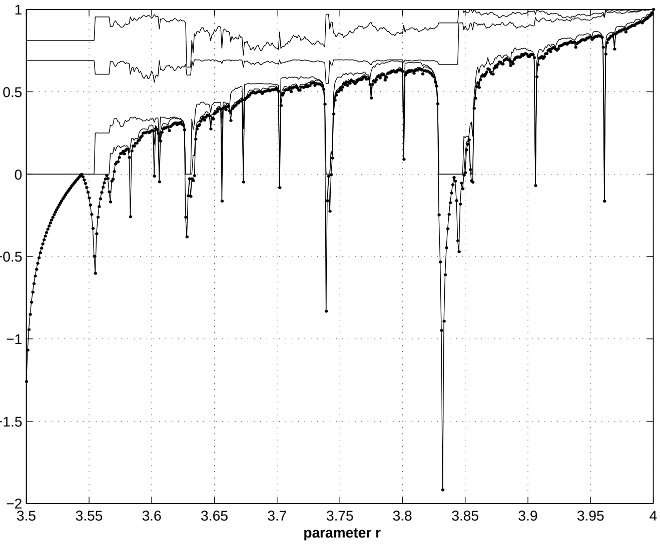

Fig. 2. The conditional entropies h_0, h_1, h_5 and h_{10} (from above) and the Lyapunov exponent λ (lowest curve) *versus* parameter r.

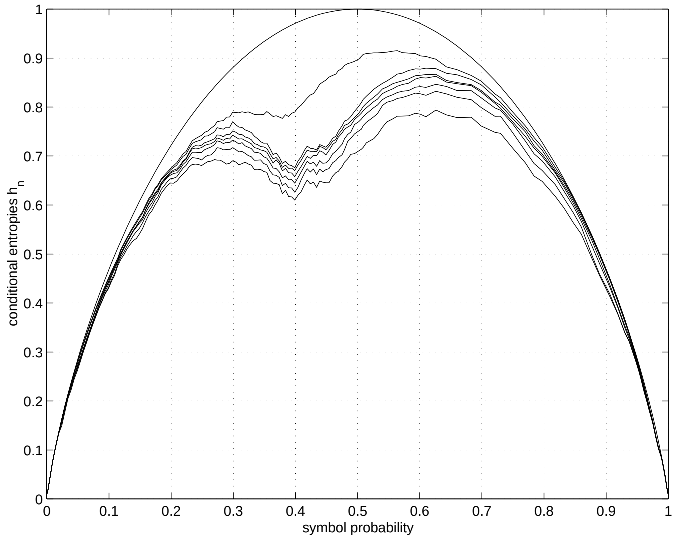

Fig. 3. Neuron 1 before penicillin treatment: The conditional entropies h_0 to h_7 in descending order as a function of the symbol '0' probability p.

already been investigated in previous work [3]. Note that for increasing n the conditional entropies h_n converge towards the positive Lyapunov exponent λ^+. In Figure 2 this is visualized by plotting h_n for increasing n *versus* parameter r [13].

5 The analysis of neural spike trains

As an application we will discuss time series obtained from interspike interval trains, going back to measurements of Rapp *et al.* [6]. The data consists of seven single-unit records of length $N = 1000$ obtained from cortical neurons of a rat before and after the application of penicillin. All data was mapped on binary symbol sequences dependent on the threshold parameter c. Instead of plotting the conditional entropies h_n *versus* the threshold parameter we switch to the corresponding symbol probability. This will yield a certain invariance to simple data transformations like $f(x) = x\sqrt{|x|}$. Starting with neuron 1 before penicillin treatment one observes that the estimated conditional entropies h_n are strongly dependent on the choice of the binary partition (see Fig. 3). After the application of penicillin the observed structure has vanished as seen in Figure 4. The entropy plot looks very much like that of a random sequence. We shall note however that this is no systematic feature before and after penicillin treatment but could also be found *vice versa* (see Tab. 1). Before we proceed we should stress two points. By maximizing the conditional entropy of a binary sequence with respect to the partition threshold c we do not meet Kolmogorov's criterion for the supremum of the conditional entropies for all possible partitions Secondly, we do not claim that these estimated entropies should tend towards the sum of positive Lyapunov exponents. What we aim at are simple rules about how the choice of a partition should be performed to most effectively reveal the structure of a given sequence.

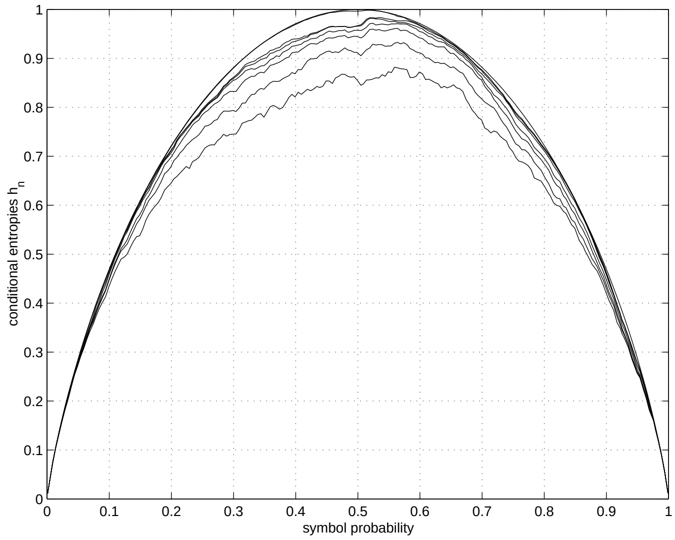

Fig. 4. Neuron 1 after penicillin treatment: The conditional entropies h_0 to h_7 in descending order as a function of the symbol '0' probability p.

5.1 Finite size effects and surrogate sequences

The approximation of the Kolmogorov-Sinai entropy requires to consider longer and longer words. However, on experimental data, this is limited due to finite length effect. Therefore the optimal partition should maximize h_n

Table 1. The conditional entropies h_n before and after penicillin treatment, denoted as (before $\to$ after) for all seven investigated neurons. The entropy was estimated for a binary partition with respect to h_3 being maximal. The first row denotes the symbol '0' probability p corresponding to the threshold parameter.

	Neuron 1	Neuron 2	Neuron 3	Neuron 4	Neuron 5	Neuron 6	Neuron 7
p	$0.63 \to 0.52$	$0.53 \to 0.52$	$0.48 \to 0.47$	$0.5 \to 0.48$	$0.37 \to 0.49$	$0.64 \to 0.61$	$0.43 \to 0.54$
h_0	$0.95 \to 1.00$	$1.00 \to 1.00$	$1.00 \to 1.00$	$1.00 \to 1.00$	$0.95 \to 1.00$	$0.94 \to 0.96$	$0.99 \to 1.00$
h_1	$0.90 \to 1.00$	$1.00 \to 1.00$	$1.00 \to 0.97$	$1.00 \to 1.00$	$0.89 \to 1.00$	$0.88 \to 0.92$	$0.95 \to 0.99$
h_2	$0.88 \to 0.99$	$0.99 \to 0.99$	$0.98 \to 0.97$	$1.00 \to 0.99$	$0.86 \to 0.99$	$0.86 \to 0.89$	$0.95 \to 0.99$
h_3	$0.87 \to 0.98$	$0.99 \to 0.99$	$0.97 \to 0.97$	$1.00 \to 0.99$	$0.85 \to 0.99$	$0.83 \to 0.88$	$0.95 \to 0.99$
h_4	$0.86 \to 0.97$	$0.99 \to 0.99$	$0.96 \to 0.96$	$1.00 \to 0.98$	$0.85 \to 0.98$	$0.83 \to 0.88$	$0.95 \to 0.98$

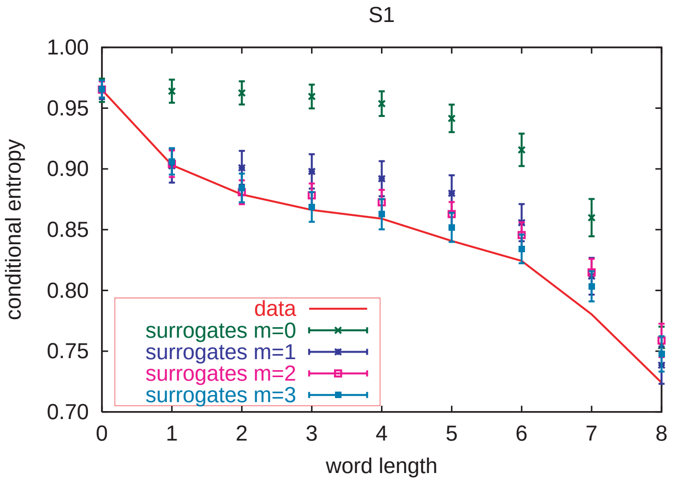

Fig. 5. The conditional entropy of neuron 1 before penicillin treatment and corresponding surrogate sequences in the case of an optimal binary partition

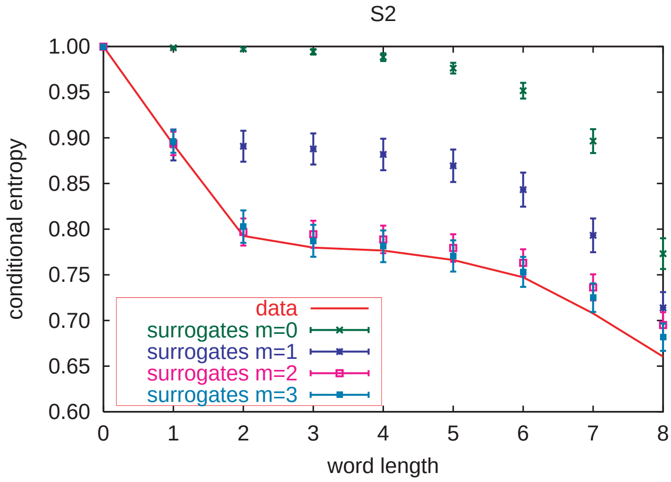

Fig. 6. The conditional entropy of neuron 1 before penicillin treatment and corresponding surrogate sequences, partitioned with equal symbol frequency.

for a large, but finite, word length $n+1$. Large means here as large as possible with small finite length effects.

In order to deal with the finite length effects, we built for each partition a series of ensembles of surrogate sequences with the identical length as the original data. The sequences of the series $m = 0, 1, \ldots$ were constructed by a Markovian process with memory m having the same transition probabilities $p(A_{m+1}|A_m, \ldots, A_1)$ as the original sequence. That means the sequences of order $m = 0$ are Bernoulli sequences with the same mean symbol frequencies as the original sequence. The first order surrogate sequence $m = 1$ corresponds to a first order Markov process with the the same transition probabilities as the original sequence. In Figures 5 and 6 the conditional entropies h_n of the original and surrogate sequences for two different partitions are shown. The errorbars show the standard deviations of the surrogate ensembles. The deviations are due to finite size effects. We used that particular parameter m as the value for optimizing the partition, where the conditional entropies of the original sequence were the first time within the confidence intervals of the surrogate sequences having a memory of m. Hence, in the concrete case h_3 was used for finding the optimal partition.

6 Conclusion

When applying the concepts of symbolic dynamics to measured time series special diligence should be devoted to the choice of the partition. As we have demonstrated a homogeneous partition might lead to spurious results for the estimated conditional entropies. We therefore suggest to maximize the entropies given a certain length of the alphabet. This method is easily generalized to three or more symbols. What has yet to be considered is the comparability of entropies stemming from different encodings with increasing alphabet length. Still, choosing the partition according to a maximized entropy gives a better tool to differentiate sequences than the usually used homogeneous partition.

This work was supported by the DFG (Sfb 555 - Project A5). M.A. J.-M. thanks CONACyT (Project 32201-E) for partial support.

References

1. C. Shannon, Bell Systems Tech. **30**, 50 (1951).
2. W. Ebeling, M.A. Jiménez-Montano, Math. Biosci. **52**, 53 (1980).
3. H. Herzel, A.O. Schmitt, W. Ebeling, Chaos Solitons Fractals **4**, 97 (1994); Phys. Rev. E **50**, 5061 (1994).
4. A.O. Schmitt, W. Ebeling, H. Herzel, Biosystems **37**, 199 (1996).
5. W. Ebeling, M.A. Jiménez-Montaño, T. Pohl, in *Entropy Measures, Maximum Entropy Principles and emerging applications*, edited by Karmeshu (Springer Berlin, 2000).
6. P.E. Rapp, I.D. Zimmerman, E.P. Vining, N. Cohen, A.M. Albano, M.A. Jiménez-Montaño, J. Neurosci. **148**, 4731 (1994).
7. A.N. Kolmogorov, Problemy Peredachi Inform. **1**, 3 (1965); IEEE Transactions Inform. Theory **14**, 14 (1968).
8. Ya.B. Sinai, Dokl. Akad. Nauk. USSR **124**, 768 (1959); *ibid.* **125**, 1200 (1959).
9. J.-P. Eckmann, D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).
10. A. Politi, F. Christiansen, Phys. Rev E **51**, R3811 (1995).
11. K. Karamanos, G. Nicolis, Chaos Solitons Fractals **10**, 77 (1999).
12. J.P. Crutchfield, N.H. Packard, Physica D **7**, 201 (1983).
13. H. Atmanspacher, J. Kurths, H. Scheingraber, R. Wackerbauer, A. Witt, Open Syst. Inform. Dynamics **1**, 269 (1992).