Doctopic: Nonlinear science ARTICLE IN PRESS PLA:24440

Physics Letters A ••• (••••) •••–•••

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Permutation entropy based time series analysis: Equalities in the input signal can lead to false conclusions

Luciano Zunino [a,b,*], Felipe Olivares [c], Felix Scholkmann [d,e], Osvaldo A. Rosso [f,g,h]

[a] Centro de Investigaciones Ópticas (CONICET La Plata – CIC), C.C. 3, 1897 Gonnet, Argentina
[b] Departamento de Ciencias Básicas, Facultad de Ingeniería, Universidad Nacional de La Plata (UNLP), 1900 La Plata, Argentina
[c] Instituto de Física, Pontificia Universidad Católica de Valparaíso (PUCV), 23-40025 Valparaíso, Chile
[d] Research Office for Complex Physical and Biological Systems (ROCoS), Mutschellenstr. 179, 8038 Zurich, Switzerland
[e] Biomedical Optics Research Laboratory, Department of Neonatology, University Hospital Zurich, University of Zurich, 8091 Zurich, Switzerland
[f] Instituto de Física, Universidade Federal de Alagoas (UFAL), BR 104 Norte km 97, 57072-970, Maceió, Alagoas, Brazil
[g] Instituto Tecnológico de Buenos Aires (ITBA) and CONICET, C1106ACD, Av. Eduardo Madero 399, Ciudad Autónoma de Buenos Aires, Argentina
[h] Complex Systems Group, Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Av. Mons. Álvaro del Portillo 12.455, Las Condes, Santiago, Chile

A R T I C L E I N F O

Article history:
Received 8 February 2017
Received in revised form 29 March 2017
Accepted 30 March 2017
Available online xxxx
Communicated by C.R. Doering

Keywords:
Time series analysis
Permutation entropy
Equalities
Spurious temporal correlations

A B S T R A C T

A symbolic encoding scheme, based on the ordinal relation between the amplitude of neighboring values of a given data sequence, should be implemented before estimating the permutation entropy. Consequently, equalities in the analyzed signal, *i.e.* repeated equal values, deserve special attention and treatment. In this work, we carefully study the effect that the presence of equalities has on permutation entropy estimated values when these ties are symbolized, as it is commonly done, according to their order of appearance. On the one hand, the analysis of computer-generated time series is initially developed to understand the incidence of repeated values on permutation entropy estimations in controlled scenarios. The presence of temporal correlations is erroneously concluded when true pseudorandom time series with low amplitude resolutions are considered. On the other hand, the analysis of real-world data is included to illustrate how the presence of a significant number of equal values can give rise to false conclusions regarding the underlying temporal structures in practical contexts.

1. Introduction

Permutation entropy (PE) is becoming a popular tool for the characterization of complex time series. Since its introduction almost fifteen years ago by Bandt and Pompe (BP) in their foundational paper [1], it has been successfully applied in a wide range of scientific areas and for a vast number of purposes. Without being exhaustive, applications in heterogeneous fields, such as biomedical signal processing and analysis [2–10], optical chaos [11–15], hydrology [16–18], geophysics [19–21], econophysics [22–25], engineering [26–29], and biometrics [30] can be mentioned. The PE is just the celebrated Shannon entropic measure evaluated using the ordinal scheme introduced by BP to extract the probability distribution associated with an input signal. This ordinal symbolic method, based on the relative amplitude of time series values, naturally arises from the time series (without any model assumptions) and inherits the causal information that stems from the temporal structure of the system dynamics. The relative frequencies of ordinal or permutation patterns, that quantify the temporal ranking information in the data sequence, need to be firstly calculated. Because of its definition via ordinal relationships, the way to handle equal amplitude values may have significant consequences when estimating the ordinal patterns probability distribution. In the case of variables with continuous distributions, ties can be simply ignored because they are very rare. However, experimental data digitized with relatively low amplitude resolutions could have a non-negligible number of equalities and, consequently, the PE estimations may be significantly affected by the procedure to consider them. Equal values in the time series are very usually ranked according to their temporal order. The other recipe, suggested by BP [1], is to break ties by adding a small amount of noise. This second alternative has been rarely implemented. In this paper, we characterize the effect that the presence of equalities in the data sequence has on the PE estimations when the former, most used, approach is adopted. Through numerical and real-world data analysis, we demonstrate that the PE estimated values are biased

* Corresponding author at: Centro de Investigaciones Ópticas (CONICET La Plata – CIC), C.C. 3, 1897 Gonnet, Argentina.
E-mail addresses: lucianoz@ciop.unlp.edu.ar (L. Zunino), olivaresfe@gmail.com (F. Olivares), Felix.Scholkmann@gmail.com (F. Scholkmann), oarosso@gmail.com (O.A. Rosso).

http://dx.doi.org/10.1016/j.physleta.2017.03.052
0375-9601/© 2017 Elsevier B.V. All rights reserved.

as a consequence of the presence of equal values, and more regular dynamics than those expected can be erroneously concluded. We consider that this finding is relevant for a more appropriate interpretation of the results obtained when the PE is used for characterizing the underlying dynamics of experimentally acquired observables. In particular, the effect of ties should be especially considered when using the PE for comparing the regularity degree of two or more experimental datasets digitized with different amplitude resolutions. Our main motivation is to warn future PE users about the importance to take this limitation into account for avoiding potential misunderstandings. Obviously, all related quantifiers estimated using the BP symbolic representation, *i.e.* with the symbolic method that considers the temporal ranking information (ordinal or permutation patterns) of the time series, can be also affected by this issue. We can enumerate permutation statistical complexity [31,32], permutation directionality index [33], symbolic transfer entropy [34], Tsallis permutation entropy [35], Rényi permutation entropy [36,37], conditional entropy of ordinal patterns [38], permutation min-entropy [39], multiscale permutation entropy measures [40], permutation Hurst exponent estimator [41] and time-scale independent permutation entropy [42], among many others. It is worth mentioning here that Bian et al. [43] have proposed the *modified* permutation-entropy (mPE) as an interesting alternative for dealing with equal values. Mapping equal values to the same symbols, these authors have shown that the mPE allows for an improved characterization of heart rate variability signals under different physiological and pathological conditions. However, Bian and co-authors' approach has a different physical interpretation and can not be considered as a *generalized* permutation entropy. For instance, the mPE does not reach its maximum value for a totally random signals (white noise) as this actually happens for the standard PE. Also the *weighted* permutation entropy (WPE) has been introduced by Fadlallah et al. [44] a couple of years ago as an improved PE by incorporating amplitude information. Through this weighted scheme, better noise robustness and distinctive ability to characterize data with spiky features or having abrupt changes in magnitude have been achieved. As it will be shown below, the presence of ties also has a significant incidence on WPE estimated values.

The remainder of the paper is organized as follows. In Section 2, the PE is introduced. A testbed analysis on computer-generated time series is included in Section 3 in order to understand the incidence of the presence of equal values on the PE estimations. Section 4 presents a couple of applications to illustrate this drawback in practical situations. Finally, in Section 5, the main conclusions reached in this work are summarized.

2. Permutation entropy

The PE has been introduced by BP as a natural complexity measure for time series [1]. It is the Shannon entropy of the ordinal symbolic representation obtained from the original sequence of observations. The idea behind ordinal pattern analysis is to consider order relations between values of time series instead of the values themselves. This ordinal symbolization is distinguished from other symbolic representations principally due to important practical advantages. Namely, it is conceptually simple, computationally fast, robust against noise, and invariant with respect to nonlinear monotonous transformations. Furthermore, the BP ordinal method of symbolization naturally arises from the time series, avoids amplitude threshold dependencies that affect other more conventional symbolization recipes based on range partitioning [45], and, perhaps more importantly, inherits the causal information that stems from the dynamical evolution of the system. As stated by Amigó et al. [46], "ordinal patterns are not symbols *ad hoc* but they actually encapsulate qualitative information about the temporal structure of the underlying data." Because of all these advantages, the BP approach has been commonly implemented for revealing the presence of subtle temporal correlations in time series [39,47–54].

Next, we summarize how to estimate the PE from a time series with a toy numerical example. Let us assume that we start with the time series $X = \{4, 1, 6, 5, 10, 7, 2, 8, 9, 3\}$. To symbolize the series into ordinal patterns, two parameters, the *embedding dimension* $D \geq 2$ ($D \in \mathbb{N}$, number of elements to be compared with each other) and the *embedding delay* τ ($\tau \in \mathbb{N}$, time separation between elements) should be chosen. The time series is then partitioned into subsets of length D with delay τ similarly to phase space reconstruction by means of time-delay-embedding. The elements in each new partition (of length D) are replaced by their ranks in the subset. For example, if we set $D = 3$ and $\tau = 1$, there are eight different three-dimensional vectors associated with X. The first one $(x_0, x_1, x_2) = (4, 1, 6)$ is mapped to the ordinal pattern (102). The second three-dimensional vector is $(x_0, x_1, x_2) = (1, 6, 5)$, and (021) will be its related permutation. The procedure continues so on until the last sequence, (8, 9, 3), is mapped to its corresponding motif, (120). Afterward, an ordinal pattern probability distribution, $P = \{p(\pi_i), i = 1, \ldots, D!\}$, can be obtained from the time series by computing the relative frequencies of the $D!$ possible permutations π_i. Continuing with the toy example: $p(\pi_1) = p(012) = 1/8$, $p(\pi_2) = p(021) = 1/4$, $p(\pi_3) = p(102) = 3/8$, $p(\pi_4) = p(120) = 1/8$, $p(\pi_5) = p(201) = 0$, and $p(\pi_6) = p(210) = 1/8$. The PE is just the Shannon entropy estimated by using this ordinal pattern probability distribution, $S[P] = -\sum_{i=1}^{D!} p(\pi_i) \log(p(\pi_i))$. Coming back to the example, $S[P(X)] = -(3/8)\log(3/8) - (1/4)\log(1/4) - 3(1/8)\log(1/8) \approx 1.4942$. It quantifies the *temporal structural diversity* of a time series. If some ordinal patterns appear more frequently than others, the PE decreases, indicating that the signal is less random and more predictable. This allows to unveil hidden temporal information that helps to achieve a better understanding of the underlying mechanisms that govern the dynamics. Technically speaking, the ordinal pattern probability distribution P is obtained once we fix the embedding dimension D and the embedding delay time τ. Taking into account that there are $D!$ potential permutations for a D-dimensional vector, the condition $N \gg D!$, with N the length of the time series, must be satisfied in order to obtain a reliable estimation of P [55,56]. For practical purposes, BP suggest in their seminal paper to estimate the frequency of ordinal patterns with $3 \leq D \leq 7$ and embedding delay $\tau = 1$ (consecutive points). It has been recently shown that the analysis of the PE as a function of τ may be particularly helpful for characterizing experimental time series on a wide range of temporal scales [32,57]. By changing the value of the embedding delay τ different time scales are being considered because τ physically corresponds to multiples of the sampling time of the signal under analysis. For further details about the BP methodology, we recommend Refs. [57–59]. It is common to normalize the PE, and therefore in this paper, a normalized PE given by

$$\mathcal{H}_S[P] = S[P]/S_{\max} = S[P]/\log(D!) \tag{1}$$

is implemented, with $S_{\max} = \log(D!)$ the value obtained from an equiprobable ordinal pattern probability distribution. Defined in this way, $\mathcal{H}_S$ ranges between 0 and 1. The maximum value is obtained for a totally random stochastic process (white noise) while the minimum value is reached for a completely regular (monotonically increasing or decreasing) time series.

Since the BP approach symbolizes the series replacing the observable value by its corresponding rank in the sequence, the occurrence of equal values deserves a special handle. In the case of two elements in the vector having the same value, they are very often ranked by their temporal order. For example, a vector (1, 4, 1), would be mapped to (021). This is the most com-