Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Improving quality of sample entropy estimation for continuous distribution probability functions

Janusz Miśkiewicz *

Institute of Theoretical Physics, University of Wrocław, pl. M. Borna 9, 50-204 Wrocław, Poland
Department of Physics and Biophysics, Wrocław University of Environmental and Life Sciences, ul. Norwida 25, 50-375, Wrocław, Poland

HIGHLIGHTS

- Sample entropy of the system described by continuous probability distribution function.
- Two algorithms for Shannon and Tsallis entropies are proposed and their features discussed.
- The possible problems of traditional histogram based sample entropy estimations are discussed.
- The advantages of the proposed algorithms are discussed.
- The Matlab implementations of presented algorithms are provided.

ARTICLE INFO

Article history:
Received 6 August 2015
Received in revised form 14 December 2015
Available online xxxx

Keywords:
Entropy
Sample entropy
Data analysis

ABSTRACT

Entropy is a one of the key parameters characterizing state of system in statistical physics. Although, the entropy is defined for systems described by discrete and continuous probability distribution function (PDF), in numerous applications the sample entropy is estimated by a histogram, which, in fact, denotes that the continuous PDF is represented by a set of probabilities. Such a procedure may lead to ambiguities and even misinterpretation of the results. Within this paper, two possible general algorithms based on continuous PDF estimation are discussed in the application to the Shannon and Tsallis entropies. It is shown that the proposed algorithms may improve entropy estimation, particularly in the case of small data sets.

1. Introduction

The aim of the paper is to discuss the sample entropy estimation for the system described by continuous probability distribution function (PDF). The investigated algorithms are based on continuous PDF estimation methods and applied to Shannon and Tsallis entropies.

The information entropy was introduced by Shannon in the paper [1] and developed later in the book [2]. The main advantage of the Shannon entropy (SE) is that it allows to measure the information content of a given sequence. This feature combined with the present ease of processing information made the SE a very useful and popular research tool. Despite the fact that SE was defined in 1948 as the tool of information theory [3] it is still an important research method in broad range of scientific activity fields starting from such a physics core like high elementary particle physics [4], standard physical models like harmonic oscillator [5], neural networks [6] up to the interdisciplinary applications: biophysical [7], econophysical [8] and others.

* Correspondence to: Institute of Theoretical Physics, University of Wrocław, pl. M. Borna 9, 50-204 Wrocław, Poland.
E-mail address: janusz.miskiewicz@ift.uni.wroc.pl.

http://dx.doi.org/10.1016/j.physa.2015.12.106
0378-4371/

Although SE was defined for discrete PDF as

$$SE = -\sum_i p_i \log p_i \tag{1}$$

where p_i is the probability of the ith event, it was generalized into the continuous case

$$SE = -\int_{-\infty}^{\infty} \rho(x) \log \rho(x) \mathrm{d}x \tag{2}$$

where $\rho(x)$ is the density probability function. Unfortunately, the continuous SE differs from the discrete case in the limit by a potentially infinite offset [9]. If the continuous probability distribution ρ is discretized into bins of size δ then by the mean value theorem:

$$\int_{-\infty}^{\infty} \rho(x)\mathrm{d}x = \lim_{h\to 0} \sum_{i=-\infty}^{\infty} \rho(x_i)\delta.$$

So the discrete approximation of SE of continuous distribution can be expressed:

$$\begin{aligned} SE_\delta &= -\lim_{\delta\to 0} \sum_{i=-\infty}^{\infty} \rho(x_i)\delta \log(\rho(x_i)\delta) \\ &= -\int_{-\infty}^{\infty} \rho(x) \log \rho(x)\mathrm{d}x - \lim_{\delta\to 0} \sum_{i=-\infty}^{\infty} \rho(x_i)\delta \log \delta. \end{aligned} \tag{3}$$

Observe that in the limit $\delta \to 0$ the logarithm goes to infinity, therefore, the relation between continuous entropy and its discrete approximation is:

$$SE = -\int_{-\infty}^{\infty} \rho(x) \ln \rho(x)\mathrm{d}x = \lim_{\delta\to 0}(S_\delta + \log \delta). \tag{4}$$

Due to Eq. (4) the estimate of continuous distribution in the simplest form can be defined [10] as:

$$\hat{SE} = -\sum_i p_i \log p_i + \log \delta. \tag{5}$$

It is true that SE is mainly used considering "discrete systems". There are various reasons for this situation. Firstly, in the case of huge enough data set the estimation based on Eq. (5) allows to achieve high-quality results. The second (and, in fact, the main) reason is the popularity of the histogram technique, which is easily achievable and implemented in all popular data analysis packages.

The histogram method requires additional comments. It generates a discrete set of probabilities, which can be deceptively easy substituted into Eq. (1). The popularity of the histogram has such an effect that even in the analysis of clearly continuously distributed variable the naive estimator based on Eq. (1) is often used instead the corrected version given by Eq. (5). Of course, in several cases discretization of a continuous distribution function is entirely justified by some well-established categorization e.g. in demography, when analysed population is often divided into age groups according to wealth, their sex or ability to work etc. However, in the case of a low number of data points e.g. in the analysis of globalization process based on macro-economy data the reliable histogram cannot be obtained and an alternative method—the Theil index [11] have to be used to approximate entropy of a macroeconomy system [12–15]. Another problem is related to the research methodology. In data categorization, the domain of the distribution function is replaced by a discrete index. Consequently, the domain is lost in the analysis. Finally, the histogram algorithm outcome strongly depends on the chosen number of bins. The illustration of such a naive estimation is presented in Fig. 1, where SE was calculated by the direct application of the histogram i.e. by substituting the tabulated probabilities of an arbitrarily chosen bins into Eq. (1), as it is often done. Fig. 1 presents the Shannon entropy estimated by histograms with a different number (h) of bins. The data set contains 10^5 points, sampled from (i) the uniform, (ii) the normal ($\mu = 0, \sigma = 1$), (iii) the log-normal ($\mu = 0.5, \sigma = 0.25$) Eq. (6), and (iv) the Weibull ($A = 1, B = 2$) Eq. (7) distribution. The definitions and detailed discussion of Log-normal and Weibull distributions properties can be found in various textbooks e.g. Refs. [16–20], but for the convenience of the reader the probability distribution functions of the Log-normal and Weibull distribution are defined by Eqs. (6) and (7) respectively.

$$P(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right); \quad x > 0. \tag{6}$$

$$P(x|A,B) = \begin{cases} \dfrac{B}{A}\left(\dfrac{x}{A}\right)^{B-1} \exp\left(-\left(\dfrac{x}{A}\right)^{B}\right) & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{7}$$

where $A > 0$ is the scale parameter and $B > 0$ is the shape parameter.

Please cite this article in press as: J. Miśkiewicz, Improving quality of sample entropy estimation for continuous distribution probability functions, Physica A (2016), http://dx.doi.org/10.1016/j.physa.2015.12.106