Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Physica A 376 (2007) 158–164
PHYSICA A
www.elsevier.com/locate/physa

Revisiting sample entropy analysis

R.B. Govindan[a,*], J.D. Wilson[a], H. Eswaran[b], C.L. Lowery[b], H. Preißl[b,c]

[a]*Graduate Institute of Technology, University of Arkansas at Little Rock, AR 72204, USA*
[b]*Department of Obstetrics and Gynecology, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA*
[c]*MEG-Center, University of Tübingen, 72076 Tübingen, Germany*

Received 25 August 2006
Available online 15 November 2006

Abstract

We modify the definition of sample entropy (SaEn) by incorporating a time delay between the components of the block (from which the densities are estimated) and show that the modified method characterizes the complexity of the system better than the original version. We apply the modified SaEn to the standard deterministic systems and stochastic processes (uncorrelated and long range correlated (LRC) processes) and show that the underlying complexity of the system is better quantified by the modified method. We extend this analysis to the RR intervals of the normal and congestive heart failure (CHF) subjects (available via www.physionet.org) and show that there is a good degree of separation between the two groups.

Keywords: Time series analysis; Fluctuation phenomena; Random processes; Noise; Brownian motion; Complex systems

1. Introduction

Quantification of a complex signal is an essential task in understanding the mechanism governing the dynamics of the system. Hence, such a quantification demands employment of reliable methods. One of the most commonly used methods to quantify the complexity of a given system is approximate entropy (ApEn) [1]. As there are methodological pitfalls in ApEn [2,3], another measure, sample entropy (SaEn) was proposed [2]. However, besides correcting some of the biases in the ApEn, the current definition of SaEn is still inadequate to characterize the complexity of the process completely. Yet another measure introduced in this context is multiscale entropy (MSE) [3,4]. In this approach, the signal is partitioned into disjoint windows of size τ and the data are averaged inside each time window. MSE is defined as the variation of the SaEn as a function of τ. From signal processing perspective, averaging the data within a time window τ is equivalent to down-sampling the data by a factor of τ to reduce the high frequency components. Alternatively, such an averaging procedure is equivalent to the low pass filter with τ determining the filter cut-off. Naively one would expect any complexity measure to yield a lower value for the averaged data compared to the non-averaged data. As different physiological processes operate at different frequency regimes, after a particular value of τ

*Corresponding author.
E-mail address: rbgovindan@git.ualr.edu (R.B. Govindan).

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2006.10.077

which may be related to the characteristic frequency of the system, MSE will reach a *pseudo convergence*. It has been shown that MSE is different in normal and pathological states of the heart [3,4]. Note that in the limit $\tau \to N$ (*length of the data*), MSE will tend to zero and hence the convergence shown in the earlier works [3,4] should be considered as *pseudo convergence*. In the case of continuous systems, the same signal sampled at two different sampling rates would show different behavior in the MSE analysis and this has been demonstrated in a recent work [5].

Here we modify the definition of SaEn by incorporating time delay δ between the successive components of the block and show that this modified method characterizes the complexity of the dynamics better than its original form proposed by Richman [2]. First we apply our proposed modified method to three test cases: (a) Sine wave sampled at two different frequencies; (b) limit cycle of Van der Pol and the Roessler systems and (c) chaotic Roessler system. We then extend this analysis to stochastic processes which have physiological interests. Finally, we apply our algorithm to heart beat (RR) intervals of human cardiac signals.

The paper is organized as follows: In Section 2 we discuss the methodology of the SaEn and the proposed modification and in Section 3 the application of the modified method to the numerically simulated deterministic systems and stochastic processes. In Section 4, we discuss the application of the method to the RR intervals of normal and congestive heart failure (CHF) cases and draw conclusions in Section 5.

2. Methodology

Let us consider a data set $x(i)$, $i = 1$ to N, sampled at discrete interval of time Δt. For discrete systems, like RR intervals, stochastic process (e.g. white noise), $\Delta t = 1$, whereas for the continuous systems, Δt is inverse of the sampling frequency of the process. In SaEn analysis, $x(i)$ is divided into overlapping blocks (vectors) of size (dimension) m with a natural time gap of one (sample unit) between the successive components of the block. This procedure is similar to the reconstruction of attractor from time series [6,7]. For instance, the ith m-dimensional block $\mathbf{V}(i)$ will be $\mathbf{V}(i) = \{x(i),\ x(i+\delta), \ldots, x(i+[m-1]\delta)\}$, where δ is the time delay between the successive components of the vector and is set to unity in complexity analysis. The reconstructed vectors are the state space (m-dimensional) representation of the dynamics of the system. SaEn is defined as the logarithmic difference between the probability (density) of occurrence of the vector $\mathbf{V}(i)$ within a chosen distance r in (m)-dimension and the probability of occurrence of the vector $\mathbf{V}(i)$ within the same chosen distance r in ($m+1$)-dimension. To estimate the density $\rho^m(r)$ of the m-dimensional state space, we consider a ball of radius r around the ith vector $\mathbf{V}(i)$ and compute the fraction of the reconstructed vectors fall into this radius. Then, we repeat calculation for all the vectors ($i = 1$ to number of vectors). This procedure is repeated for the vectors reconstructed in $m+1$ state space to estimate the density $\rho^{m+1}(r)$ in $m+1$ dimension. The density of the state space is defined as the measure of the fraction of the reconstructed vectors which fall within the chosen radius. Based on the above discussion, SaEn is defined as follows:

$$SaEn(m, r) = \log[\rho^m(r)/\rho^{m+1}(r)]. \quad (1)$$

The densities $\rho^m(r)$ and $\rho^{m+1}(r)$ are estimated for the same number of vectors (i.e., for the number of vectors in the $m+1$ dimension), to avoid the bias in the estimation [2]. In most of the applications, these densities are estimated by the Grassberger–Procaccia integrals [7–9]. As mentioned in the introduction, dynamical correlations, in some cases, are not associated with dynamics (for example, the same signal sampled at two different frequencies), and in some other cases, are associated with the intrinsic dynamics of the system (for example long range correlated (LRC) processes), will not allow the correct quantification of the underlying complexity of the system. In order to overcome the difficulties due to these type of correlations we introduce a time delay δ between the successive components of the block. We choose δ as the time point at which autocorrelation function falls below $1/e$ [10]. The time delay will take a value greater than unity in the presence of the dynamical correlations. With this modification the above expression for SaEn will reshape as follows:

$$SaEn(m, r, \delta) = (1/\delta)\log[\rho^m(r)/\rho^{m+1}(r)]. \quad (2)$$

In the density estimation, in order not to account for the temporally (auto-) correlated points, for each center i we discard $i+\delta$ number of points, similar to the Theiler correlation in the estimation of the correlation