Biomedical Signal Processing and Control 14 (2014) 141–147

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Parametric estimation of sample entropy in heart rate variability analysis

Md Aktaruzzaman, Roberto Sassi*

Dipartimento di Informatica, Università degli Studi di Milano, Italy

ARTICLE INFO

Article history:
Received 5 May 2014
Received in revised form 2 July 2014
Accepted 19 July 2014
Available online 9 August 2014

Keywords:
Sample entropy
Autoregressive models
Parametric estimates
Heart rate variability

ABSTRACT

In this paper, a detailed study on the possibility and significance of performing a parametric estimation of sample entropy (SampEn) is proposed. SampEn is a non-linear metric, meant to quantify regularity of a time series. It is widely employed on biomedical signals, especially on heart rate variability. Results relevant to approximate entropy, a related index, are also reported.

An analytical expression for SampEn of an autoregressive (AR) model is derived first. Then we study the feasibility of a parametric estimation of SampEn through AR models, both on synthetic and real series. RR series of different lengths are fitted to an AR model and then expected values of SampEn ($SampEn_\mu$) are estimated.

Values of SampEn, computed from real beat-to-beat interval time series (obtained from 72 normal subjects and 29 congestive heart failure patients), with $m=1$ and $r=0.2$, are within the standard range of $SampEn_\mu$ more than 83% (for series length $N=75$) and 28% (for $N=1500$) of the cases. Surrogate data have been employed to verify if departures from Gaussianity are to account for the mismatch.

The work supports the finding that when numerical and parametric estimates of SampEn agree, SampEn is mainly influenced by linear properties of the series. A disagreement, on the contrary, might point those cases where SampEn is truly offering new information, not readily available with traditional temporal and spectral parameters.

1. Introduction

Heart rate variability (HRV) analysis is an important tool for evaluating cardiac autonomic regulation [1]. About 30 years ago, Pincus [2] developed a family of statistics, called approximate entropy (ApEn), to measure series regularity. Many potential applications [3–5] of this method can be found in medical research literature, in particular for detecting and testing the regularity of HRV data. Successively, to address some manifest limitations of ApEn (Pincus himself [6] reported ApEn to be a biased statistic), Richman and Moorman [7] introduced Sample Entropy (SampEn), which was also applied successfully to a wide range of problems [8–10].

In the last 20 years [11], ApEn and SampEn had been the most commonly used methods to quantify the regularity of biological data. Both metrics estimate the differential entropy rate of the series [12,13]. However, SampEn: (i) is less prone to practical inconsistency, as it requires less lengthy series to converge to the final value and (ii) is relatively less biased even for not-so-long series [7]. Regarding the term "inconsistency", Pincus [12] considered the problem of assessing if a stochastic process A was more regular than process B, by means of computing ApEn. He defined "consistent" those processes for which ApEn of A was always larger (or smaller) than ApEn of B, for any value of the parameters on which the metric depends, (*i.e.*, m and r, see Section 2.1). Here we use the term "practical consistency" to refer to the fact that ApEn (or SampEn) of series S_A is larger than ApEn (or SampEn) of series S_B for a broad range of the parameters values.

ApEn and SampEn are extremely sensitive to data length (N), particularly for very short data sets, *i.e.*, $N \leq 200$ [11]. Hence, their estimates may be far away from what expected using longer series. Unfortunately, short series are generally used in real applications. So, issues of convergence may appear when estimating the regularity of short data. A related problem arises in spectral analysis, where long time stationary series are required to achieve lower variance of the estimates. Hence, since the works of Bishop [14] and Kay and Marple [15], parametric spectrum analysis is commonly performed on short RR series [1] of 3–5 min, which are reasonably stationary.

* Corresponding author at: Dipartimento di Informatica, Università degli Studi di Milano, Via Bramante 65, 26013 Crema, Italy. Tel.: +39 02 503 30056.
E-mail addresses: md.aktaruzzaman@unimi.it (M. Aktaruzzaman), roberto.sassi@unimi.it (R. Sassi).

http://dx.doi.org/10.1016/j.bspc.2014.07.011
1746-8094/

We therefore verify in this work, if, in analogous circumstances, a SampEn computation based on a parametric representation of the series might convey new information.

The first step of a parametric approach is to select the most appropriate family of model. For HRV signal, the most commonly used models are moving average (MA), autoregressive (AR), and autoregressive moving average (ARMA). Identification of AR models has been explored largely in the literature, it requires solving simpler equations than those required for ARMA models, and AR models are maximum-entropy models (among those sharing the same autocorrelation function). Thus, AR models are mostly used for HRV maximum-entropy spectral estimation.

In this work, we will first discuss the conditions under which a parametric estimation of SampEn (and ApEn) is possible. We will limit our attention to linear AR models. Pincus [2] and then Lake [13] already tackled the problem of deriving analytical formulas of ApEn and SampEn for an AR process. Following the suggestion in [2], in this work, we have first extended the analytical expression of ApEn. Then, we have also derived an analytical expression for SampEn and tested these predictions on simulated series and real HRV data, obtained from Holter recordings.

2. Materials and methods

2.1. ApEn

ApEn measures the likelihood that runs of patterns that are close remain close at the next incremental comparisons. The determination of this statistic is dependent on the prior selection of two unknown parameters: the length (m) of compared runs, also called templates, constructed from the series and a filtering threshold (r), *i.e.*, the tolerance of mismatch between the corresponding elements of the templates.

Given a time series $\{u[i]|1 \le i \le N\}$ of N data points, the calculation of ApEn [2] is as follows:

1. Form templates $V_m[j] = \{u[j], \ldots, u[j+m-1]\}$ of size m, for $1 \le j \le N-m+1$;
2. Define the distance between $V_m[j]$ and $V_m[i]$: $d(V_m[j], V_m[i]) = \max_{0\le k\le m-1} |u[j+k] - u[i+k]|$;
3. Let A_j^m be the number of templates $V_m[i]$ such that $d(V_m[j], V_m[i]) \le r$, where $1 \le i \le N-m+1$, and $C_j^m(r) = A_j^m/(N-m+1)$;
4. Define $\Phi^m(r) = \sum_{j=1}^{N-m+1} \log C_j^m(r)/(N-m+1)$;
5. Increase m by 1 and repeat steps 1 to 4;
1. Finally, $\mathrm{ApEn}(m, r) = \lim_{N\to\infty} [\Phi^m(r) - \Phi^{m+1}(r)]$ and it's corresponding statistics, $\mathrm{ApEn}(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r)$, is estimated for finite series.

2.2. SampEn

SampEn is similar but improved version of ApEn. The calculation of SampEn [7] is performed by the following steps:

1. Let A_j^m be the number of templates $V_m[i]$ such that $d(V_m[j], V_m[i]) \le r$, where $1 \le i \ne j \le N-m$, and $C_j^m(r) = A_j^m/(N-m-1)$;
2. Let A_j^{m+1} be the number of templates $V_{m+1}[i]$ such that $d(V_{m+1}[j], V_{m+1}[i]) \le r$, where $1 \le i \ne j \le N-m$, and $C_j^{m+1}(r) = A_j^{m+1}/(N-m-1)$;
3. Define $A_m(r) = \sum_{j=1}^{N-m} C_j^m(r)/(N-m)$ and $A_{m+1}(r) = \sum_{j=1}^{N-m} C_j^{m+1}(r)/(N-m)$, then
4. $\mathrm{SampEn}(m, r) = \lim_{N\to\infty} [\log A^m(r) - \log A^{m+1}(r)]$ and, for a finite series, $\mathrm{SampEn}(m, r, N) = \log A^m(r) - \log A^{m+1}(r)$.

2.3. AR process

An AR process of order M can be expressed as

$$x[n] = -\sum_{i=1}^{M} a_i x[n-i] + w[n]$$

where a_i are real coefficients and $w[n]$ is a white Gaussian noise (WGN) with mean zero and variance σ_w^2. The parameters of the model and the autocovariance function values γ_k, are linked by the Yule–Walker's equations

$$\begin{pmatrix} 1 & a_1 & a_2 & \cdots & a_M \\ a_1 & 1+a_2 & a_3 & \cdots & 0 \\ a_2 & a_1+a_3 & 1+a_4 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_M & a_{M-1} & a_{M-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \cdots \\ \gamma_M \end{pmatrix} = \begin{pmatrix} \sigma_w^2 \\ 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}. \quad (1)$$

The m consecutive values, $X_m[n] = \{x[n], \ldots, x[n+m-1]\}$, are multivariate normal on $\mathbb{R}^m$, with Normal joint probability density $f(X_m) = \mathcal{N}(0, \Sigma_m) = \exp(-X_m^T \Sigma_m^{-1} X_m/2)/[(2\pi)^m \det(\Sigma_m)]^{1/2}$ and Toeplitz covariance matrix

$$\Sigma_m = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{pmatrix}.$$

The values γ_m, for $m \le M$, are defined by Eq. (1). When $m > M$, further elements in Σ_m are still dictated by the Yule–Walker's equation $\gamma_k = -\sum_{i=1}^{M} a_i \gamma_{k-i}$.

Denoting $\rho_k = \gamma_k/\gamma_0$ the autocorrelation coefficient, the variance $\sigma_y^2 = \gamma_0$ of the series generated by the AR process is

$$\sigma_y^2 = \sigma_w^2(1 + a_1\rho_1 + \cdots + a_M\rho_M)^{-1} = \sigma_w^2 c, \quad (2)$$

where $c = (1 + a_1\rho_1 + \cdots + a_M\rho_M)^{-1}$.

2.4. ApEn and SampEn theoretical values for $N \to \infty$

The analytical expression of $\mathrm{ApEn}(m = 1, r)$ for a stochastic (thus also for an AR) process was given by Pincus in [2]. Let

$$Q_m = \int_{x[m]-r}^{x[m]+r} \cdots \int_{x[1]-r}^{x[1]+r} f(\Xi_m) \mathrm{d}\xi_1 \cdots \mathrm{d}\xi_m$$

be the probability that the values X_m lie within the hypercube of side $2r$, where $f(X_m)$ is the multivariate probability density of the ergodic stochastic process. Then

$$\mathrm{ApEn}_{\mathrm{TH}}(1, r) = \int\int_{\mathbb{R}^2} \mathrm{f}(X_2) \log\left(\frac{Q_1}{Q_2}\right) \mathrm{d}x[1]\mathrm{d}x[2].$$

This equation can be extended to derive a general analytical expression of $\mathrm{ApEn}(m, r)$ for any m as

$$\mathrm{ApEn}_{\mathrm{TH}}(m, r) = \int_{\mathbb{R}^{m+1}} \cdots \int \mathrm{f}(X_{m+1}) \log\left(\frac{Q_m}{Q_{m+1}}\right) \mathrm{d}X_{m+1}. \quad (3)$$

where $\mathrm{d}X_m = \mathrm{d}x[1]\mathrm{d}x[2]\cdots\mathrm{d}x[m]$,

Following a similar approach, a theoretical value for SampEn of an AR process can be derived from the definition. In fact, the probability of matching two templates of size m within error tolerance r