Sample Entropy Parametric Estimation for Heart Rate Variability Analysis

M Aktaruzzaman, R Sassi

Dipartimento di Informatica, Università degli Studi di Milano, Italy

Abstract

Aims: *Sample Entropy (SampEn) is a powerful approach for characterizing heart rate variability regularity. On the other hand, autoregressive (AR) models have been employed for maximum-entropy spectral estimation for more than 40 years. The aim of this study is to explore the feasibility of a parametric approach for SampEn estimation through AR models. We re-analyze the Physionet paroxysmal Atrial Fibrillation (AF) database, where RR series are provided before and after an AF episode, for 25 patients. In particular, we selected short RR series, close to AF episodes, to fit an AR model. Then, theoretical values of SampEn, based on each AR model, were analytically derived (SE$_{th}$) and also estimated numerically (SE$_{syn}$). The value of SampEn (SE$_{rr}$), computed on the 50 RR series with r=0.2×STD, m=1 and N=120, were within the standard range of SE$_{syn}$ in 30 cases (39 for SE$_{th}$). This figure increased to 82% of cases, if shorter series were selected (N=75), and if RR series were replaced by surrogates with Gaussian amplitude distribution. Interestingly, without removing ectopic beats, every estimate of SampEn considered was significantly different between pre- and post- AF (SE$_{rr}$: p=0.02; SE$_{syn}$: p=0.0024; SE$_{th}$: p=0.023). When an AR model is appropriate and theoretical estimates differ from numerical ones, a parametric approach might enlighten additional information brought by SampEn.*

1. Introduction

Heart rate variability (HRV) analysis has become an important tool for evaluating cardiac autonomic regulation [1]. Pincus [2] at first developed a family of statistics, called approximate entropy (ApEn) to measure system complexity. This statistics has been used for measuring the regularity of HRV before the spontaneous onset of paroxysmal atrial fibrillation (AF) since 1999 [3].

ApEn was shown to be a biased statistics [4] and, to overcome this limitation, Richman and Moorman [5] introduced SampEn. The advantages of the latter over ApEn are (i) it converges rapidly, (ii) it is less prone to inconsistency and (iii) it is relatively less biased even for finite data sets [5]. The statistics SampEn has been applied successfully to a wide variety of time series analysis [6–8].

Estimation of these families of statistics requires a prior selection of the unknown parameters m and r. To the best of our knowledge, there is no universal rule for the selection of these free parameters. The recommended value of r in the range [0.1 0.2]×STD has been shown to be applicable to a wide variety of signals [2,5,9–11]. The value of m depends actually on the length of the series and it should be kept small ($m = 1$) for short series (length$\leq$ 120 points).

Pincus [12], later Lake [13], showed that both ApEn and SampEn are related to (differential) entropy rate: a central concept of information theory. Autoregressive (AR) models have been employed for maximum-entropy spectral estimation for more than 40 years [14] and already in [9,13] analytical formulas of ApEn and SampEn for this specific class of processes were provided. In this work, we extend and verify these analytical predictions with the aim of understanding if parametric estimation through AR models of SampEn is possible and practically sensible. To do this, we verify the convergence of SampEn of the process and finally explore the feasibility of parametric approach for SampEn estimation of HRV during paroxysmal AF.

2. Methods

2.1. Autoregressive processes

AR processes are commonly used to model time series. An AR process of order p is a linear combination of previous p samples and additive white Gaussian noise with mean 0 (zero) and variance σ_w^2. It can be expressed as

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

where $a_i, i = 1, 2, \cdots, p$ are the coefficients of the AR model and $w(n)$ is the injective white noise.

The joint probability density of the m consecutive values $X_m[n] = (x[n+m], \cdots, x[n+1])$ is multivariate normal on $\mathbb{R}^m$, with $X_m \sim \mathcal{N}(0, \Sigma_m)$. The Toeplitz covariance matrix Σ_m is completely defined by the coefficients of the AR model and the variance σ_w^2. When $m > p$, further elements in Σ_m are still dictated by the Yule-Walker's equation $\rho_k = -\sum_{i=1}^{p} a_i \rho_{k-1}$.

2.2. Theoretical values for $N \to \infty$

For a stochastic process (thus also AR), the analytical expression of ApEn($m = 1, r$) has been given in [9] as

$$\mathrm{AE}_{th}(1,r) = \iint_{\mathbb{R}^2} \mathrm{f}(X_2)\left\{\log\left[\int_{x_1-r}^{x_1+r} \mathrm{f}(X_1)\mathrm{d}\xi_1\right] \right.$$
$$\left. - \log\left[\int_{x_2-r}^{x_2+r}\int_{x_1-r}^{x_1+r} \mathrm{f}(X_2)\mathrm{d}\xi_1\mathrm{d}\xi_2\right]\right\}\mathrm{d}x_1\mathrm{d}x_2$$

where $\mathrm{f}(X_m) = \mathcal{N}(0, \Sigma_m)$ is the joint normal stationary probability of $(x[n+m], \cdots, x[1])$ on $\mathbb{R}^m$. This equation can be extended to derive a general expression of ApEn for any m and r as

$$\mathrm{AE}_{th}(m,r) = \underset{\mathbb{R}^{m+1}}{\int\cdots\int} \mathrm{f}(X_{m+1})\log\left(\frac{P_m}{P_{m+1}}\right)\mathrm{d}X_{m+1}, \tag{1}$$

where $\mathrm{d}X_m = \mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_m$, and

$$P_m = \int_{x_m-r}^{x_m+r}\cdots\int_{x_1-r}^{x_1+r} \mathrm{f}(X_m)\mathrm{d}\xi_1\cdots\mathrm{d}\xi_m. \tag{2}$$

Following a similar procedure, it is possible to derive a theoretical value for SampEn, based on its definition. In fact, the probability that there is a match of templates of length m (i.e. the maximum absolute difference between the corresponding elements of any two templates is r) is given by equation (2) where $\mathrm{f}(X_m) = \mathcal{N}(0, 2\Sigma_m)$. This follows by the fact that the difference of any two templates, $X_m[i] - X_m[j] \sim \mathcal{N}(0, 2\Sigma_m)$. Hence

$$\mathrm{SE}_{th}(m,r) = \log(\mathrm{P}_m) - \log(\mathrm{P}_{m+1}). \tag{3}$$

Please notice that the theoretical values in (1) and (3) depend on both m and r, but not on N, as they are given in the limit $N \to \infty$.

2.3. Theoretical values for N and $m \to \infty$

Lake [13] derived theoretical expressions for ApEn and SampEn from the definition of (differential) entropy rate, in the limit $m \to \infty$. If r is chosen independently of the standard deviation of the sequence (STD), it can be shown that they are respectively

$$\mathrm{AE}_L(r) = \log(\sigma_\mathrm{w}) + \frac{1}{2}\left[\log(2\pi) + 1\right] - \log(2\mathrm{r})$$
$$\mathrm{SE}_L(r) = \log(\sigma_\mathrm{w}) + \frac{1}{2}\log(4\pi) - \log(2\mathrm{r}).$$

On the other hand, if r is a percentage of STD, then the expressions become

$$\mathrm{AE}_L(r) = \log\left(\frac{\sigma_\mathrm{w}}{\sigma_\mathrm{y}}\right) + \frac{1}{2}\left[\log(2\pi) + 1\right] - \log(2\mathrm{r}) \tag{4}$$
$$\mathrm{SE}_L(r) = \log\left(\frac{\sigma_\mathrm{w}}{\sigma_\mathrm{y}}\right) + \frac{1}{2}\log(4\pi) - \log(2\mathrm{r}) \tag{5}$$

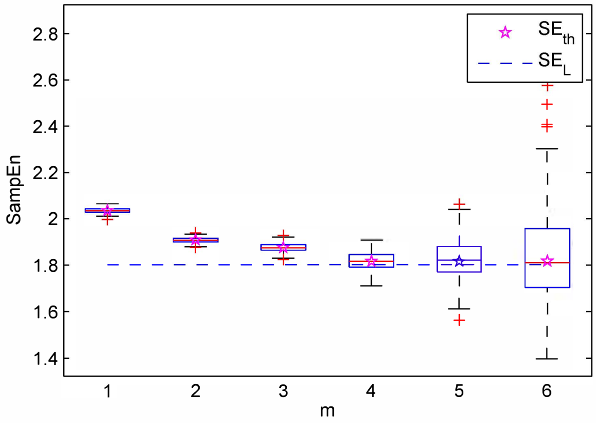

Figure 1. SampEn of the model of coefficients $\{1, -0.8, 0.46, 0.02, -0.33\}$ for different values of m, $r = 0.2 \times$ STD and $N = 5000$. SE$_L$ is constant for all values of m because it depends only on r. The probability density of numerical estimation from 200 times simulation is represented by boxplot. SE$_{th}$ varies with m and it is always inside the standard range of numerical estimation. Although SE$_L$ differs from SE$_{th}$ and SE$_{syn}$ for $m < 4$, they converge to a common value for any $m \geq 4$.

where σ_y is the standard deviation of the series obtained from the AR process.

Let ρ_k denotes the autocorrelation coefficient of the AR process at lag k. Then from the Yule Walker's equation, the variance of the process is

$$\sigma_\mathrm{y}^2 = \sigma_\mathrm{w}^2(1 + \mathrm{a}_1\rho_1 + \cdots + \mathrm{a}_\mathrm{m}\rho_\mathrm{m})^{-1} = \sigma_w^2 c,$$

where $c = 1/(1 + a_1\rho_1 + \cdots + a_m\rho_m)$. Replacing σ_y in equations (4) and (5) by $\sigma_w\sqrt{c}$

$$\mathrm{AE}_L(r) = \frac{1}{2}\left[\log(2\pi) + 1\right] - \log(2\mathrm{r}\sqrt{\mathrm{c}})$$
$$\mathrm{SE}_L(r) = \frac{1}{2}\log(4\pi) - \log(2\mathrm{r}\sqrt{\mathrm{c}})$$

So, if r is fixed, Lake's estimates depend only on the variance of the prediction error σ_w^2. On the other hand, if r varies with STD, they depend on the coefficients of the model (thus the order p), but not on σ_w^2.

2.4. Expected values for fixed N and m

The theoretical expressions for entropies given in sections 2.2 and 2.3 are independent on the series length N. When working on finite series, a convergence question may arise. In fact, the numerical estimates obtained from short series, using the algorithms proposed in [4] and [5], might be still far from reaching the expected values. On

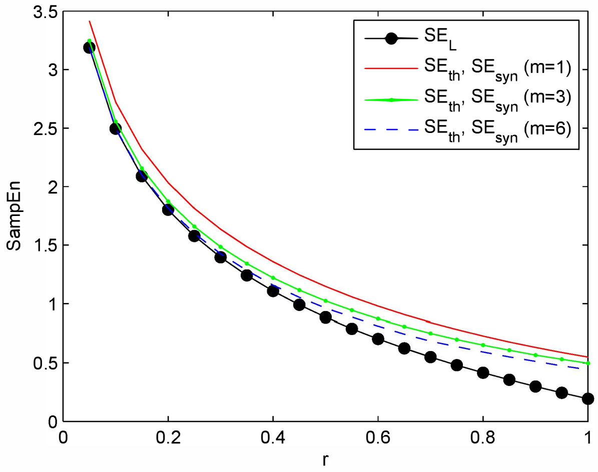

Figure 2. Convergence of SampEn (model as in Figure 1) with different m and r; $N = 5000$. SE$_{syn}$ approximately coincides with SE$_{th}$, for any m and r. They also converge to SE$_L$ as m tends to the model order $p = 4$, with smaller r ($r \leq 0.2$).

the other hand, real applications are meant to work on finite size-series.

An operative procedure is to resort to a certain number of Montecarlo simulations of the AR model. In each run, a synthetic series of a certain length N is produced, and then numerical values of ApEn and SampEn are numerically assessed for a specific couple of m and r. Finally, the values of the entropy rates are used to estimate the probability density function of the statistics, from which average values and standard deviation can be computed.

In our study, expected values of SampEn SE$_{syn}(m, r, N)$ were obtained following this procedure (200 runs) and computing their average value.

2.5. AR-process SampEn convergence

To investigate the convergence of SE$_L$, SE$_{th}$ and SE$_{syn}$ of an AR process, an arbitrary AR model of order 4 was considered. SE$_{th}$, SE$_L$ and SE$_{syn}$ of the model were determined as described in sections 2.2 to 2.4. The convergence with $m = 1, \cdots, 6$ for constant and different values of r, is shown in Figure 1 and 2, respectively. Models with different orders are instead considered in Figure 3.

2.6. Parametric approach on real RR series

Both ApEn and SampEn measures were used successfully to detect the reduction of complexity before paroxysmal AF [3, 15]. To investigate the feasibility of parametric SampEn estimation, we re-analyzed the paroxysmal AF database of 25 patients on Physionet, where RR series were provided before and after an episode of AF. To remove artifacts in the series, we performed two levels of pre-processing. In the first, only extreme artifacts were removed, defined as those RR which lie outside the range [Q$_1$ Q$_3$]$\pm 3 \times$ IR. Here, Q$_1$ and Q$_3$ are first and third quartiles, respectively and IR=Q$_3$-Q$_1$. In the second stage of pre-processing, the ectopic beats were removed. Only those RR, which were within 20% variation of the previous accepted RR, were included (the first accepted RR value for each series must lie within the IR).

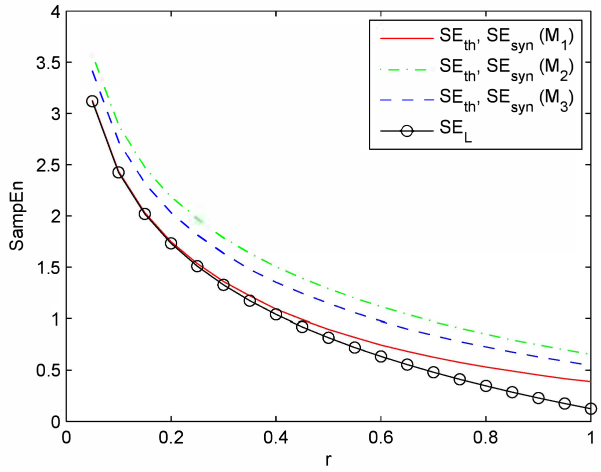

Figure 3. Convergence of SampEn for 3 different models (M$_1$ coefficients: $\{1, -0.77\}$, M$_2$: $\{1, -0.04, 0.67\}$ and M$_3$: $\{1, 0.55, 0.24, 0.39\}$) with $m = 1$ and $r = 0.05, \cdots, 1$. The values of SE$_{th}$ and SE$_L$ differ for different models, while SE$_{th}$ and SE$_{syn}$ approximatively coincide. Since m=1 (which is the order of M$_1$), the closest convergence of SE$_{th}$, SE$_L$ and SE$_{syn}$ is found for M$_1$.

Short RR series of length 120 points just immediately before (pre-AF) and after an episode of AF (post-AF) were selected to fit an AR model. Then, SE$_{th}$ of each AR model was derived. At the same time, SE$_{syn}$ of the model was estimated. Finally, paired t-test was done on pre- and post-AF estimations to see if they are distinguishable.

3. Results

After editing, the numerical values of SampEn (SE$_{rr}$) of the original RR segments, with $m = 1$, $r = 0.2 \times$ STD and series length $N = 120$, were within the standard range of SE$_{syn}$ for 30 cases out of 50. This is shown in Figure 4. Although, SE$_{th}$ was in agreement with SE$_{syn}$ for 39 cases.

To investigate if non-Gaussianity was the cause of the disagreement between SE$_{rr}$ and SE$_{syn}$, surrogated data of the original RR segments were constructed in such a way that the original distribution of RR values was replaced by a Gaussian distribution of same STD, preserving the order

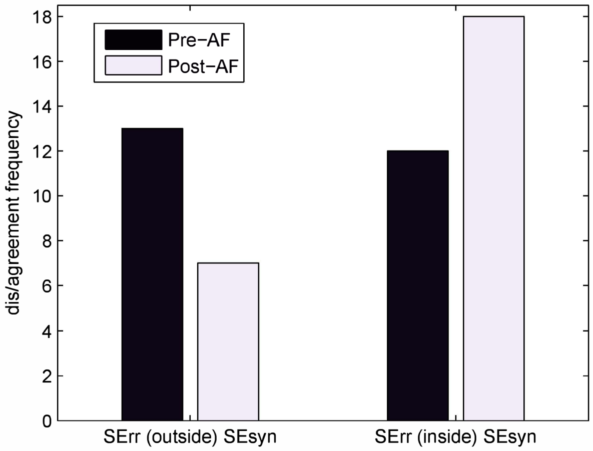

Figure 4. Frequency of dis/agreement (left/right) of SE_{rr} with SE_{syn} before (Pre-AF) and after AF (Post-AF) for series of length $N = 120$ points. Out of 25 patients, SE_{rr} is within the standard range of SE_{syn} for 12 and 18 patients, respectively, during pre-and post-AFs.

of ranks. Then, if shorter surrogate series ($N = 75$) were considered, the figure of agreement rose to 82%.

Finally we observed that the editing method employed largely affects the results, as ectopic beats are more probable before an AF episode [3]. Interestingly, every estimates of SampEn considered was significantly different between pre-and post-AF series (with SE_{rr}: p=0.020; SE_{syn}: p=0.0024; SE_{th}: p=0.023) when only extreme artifacts are removed and keeping the ectopic beats. However, this is not the case, if ectopic beats are removed (as admitted also in [3, 15]).

4. Conclusions

The important finding of this study is that parametric estimation of SampEn is possible. Numerically computed values of SampEn, SE_L, SE_{th} and SE_{syn} all converge to a common value for an AR process (if proper values for N and m are employed). When an AR model is appropriate and when theoretical estimates differ from numerical ones, this approach might provide additional information (*i.e.* non-Gaussianity or non-stationarity of the RR series).

Moreover, parametrically-estimated values of SampEn support the statement that there is reduction of complexity in HRV before the onset of AF. While this is in agreement with [3, 15], it also confirms that the complexity-reduction is mostly due to the presence of ectopic beats. In this study, the database employed contained only of small number of patients, and further studies with larger populations are needed.

References

[1] Saul JP, Arai Y, Berger RD, Lilly LS, Colucci WS, Cohen RJ. Assessment of autonomic regulation in chronic congestive heart failure by heart rate spectral analysis. Am J Cardiol 1988;61:1292–1299.

[2] Pincus SM. Approximate entropy as a measure of system complexity. Proc Natl Acad Sci 1991;88:2297–2301.

[3] Vikman S, Mäkikallio TH, Yli-Mäyry S, Pikkujämsä S, Koivisto AM, Reinikainen P, Juhani-Airaksinen KE, Huikuri HV. Altered complexity and correlation properties of R-R interval dynamics before the spontaneous onset of paroxysmal atrial fibrillation. J Am Heart Assoc 1999; 100:2079–2084.

[4] Pincus S. Approximate entropy as a complexity measure. Chaos Interdiscipl J Non linear Sci 1995;5:110–117.

[5] Richman JS, Moorman JR. Physiological time series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 2000;278:2039–2049.

[6] Lake DE, Richman JS, Griffin MP, Moorman JR. Sample entropy analysis of neonatal heart rate variability. Am J Physiol Regul Integr Comp Physiol 2002;283:789–787.

[7] Al-Angari HM, Sahakian AV. Use of sample entropy approach to study heart rate variability in obstructive sleep apnea syndrome. IEEE Trans Biomed Eng 2007;54:1900–1904.

[8] Goya-Esteban R, Marques de Sá JP, Rojo-Álvarez JL, Barquero-Pérez O. Characterization of heart rate variability loss with aging and heart failure using sample entropy. Comput Cardiol 2008;35:41–44.

[9] Pincus SM, Goldberger AL. Physiological time-series analysis: what does regularity quantify. Am J Physiol Heart Circ Physiol 1994;266:1643–1656.

[10] Signorini MG, Sassi R, Lombardi F, Cerutti S. Regularity patterns in heart rate variability signal: the approximate entropy approach. Proc Int Conf IEEE Med Biol Soc 1998; 20:306–309.

[11] Signorini MG, Ferrario M, Marchetti M, Marseglia A. Nonlinear analysis of heart rate variability signal for the characterization of cardiac heart failure patients. Conf Proc IEEE Eng Med Biol Soc 2006;1:3431–3434.

[12] Pincus SM, Huang WM. Approximate entropy: Statistical properties and applcations. Commun Statist Theor Method 1992;21:3061–3077.

[13] Lake DE. Renyi entropy measures of heart rate gaussianity. IEEE Trans Biomed Eng 2006;53:21–27.

[14] Ulrych TJ, Bishop TN. Maximum entropy spectral analysis and autoregressive decomposition. Rev Geophys Space Phys 1975;13:183–200.

[15] Tuzcu V, Nas S, Börklü T, Ugur A. Decrease in the heart rate complexity prior to the onset of atrial fibrillation. Europace 2006;8:398–402.

Address for correspondence:

M. Aktaruzzaman
Dipartimento di Informatica, Università degli Studi di Milano,
Via Bramante 65, Crema, 26013, Italy.
md.aktaruzzaman@unimi.it