BIAS ON THE APPROXIMATE ENTROPY OF RR TIME SERIES CAUSED BY THE ECG SAMPLING FREQUENCY

M.A. Garcia-Gonzalez, M. Fernandez-Chimeno and J. Ramos-Castro

Electronic and Biomedical Instrumentation Group, Electronic Engineering Department, Technical University of Catalonia (UPC), Barcelona, Spain

magarcia@eel.upc.es

Abstract: Approximate Entropy (ApEn) is a measure of the complexity of time series. This work shows that the choice of the sampling frequency (f_s) of the electrocardiogram (ECG) from where RR time series are derived affects the ApEn of these time series. The main effect of f_s is the presence of a bias in the ApEn that can be very high in certain applications (especially when dealing with RR time series with low variance). The bias decreases monotonically by increasing f_s. Moreover, the bias decreases by increasing the standard deviation (SD) of the time series, although the decrease is not monotonic but behaves as a damped oscillatory pattern.

Introduction

Approximate Entropy (*ApEn*) is a measure that quantifies the regularity of a time series. This index has been developed to classify complex systems with underlying deterministic chaos or random processes [1].

Approximate Entropy can be computed in a general way as:

$$ApEn(N,m,k) = ApEn(m,k) = \Phi^m(k) - \Phi^{m+1}(k) \quad (1)$$

where N is the number of samples of the time series, m is the window's length and k is the threshold. Moreover:

$$\Phi^m(k) = \frac{\sum_{i=1}^{N-m+1} \log\left(C_i^m(k)\right)}{N-m+1} \quad (2)$$

$$C_i^m(k) = \frac{N_p}{N-m+1} \quad (3)$$

where N_p is the number of vectors $X(j)$ (j from 1 to N-m+1) whose Euclidian distance to $X(i)$ is lower than k·SD, being $X(i)$ a vector with m samples defined as $X(i)=[x(i), x(i+1),\ldots,x(i+m-1)]$ and SD the standard deviation of the time series. $x(i)$ is the time series whose *ApEn* is being computed.

There is scientific evidence that under certain pathologies, the *ApEn* of the RR time series is different when comparing to that of healthy subjects. The main advantage of *ApEn* is its ability to quantify the complexity of time series (as is the case of the RR time series in the analysis of heart rate variability) with a low number of samples.

The motivation and properties of *ApEn* are described from time series obtained with mathematical models (so with time series whose resolution and signal-to-noise ratio can be considered for practical purposes as infinite). However, the *ApEn* is routinely applied to actual time series without taking into account practical limitations as the noise or the resolution. These restrictions can create both bias and uncertainty in the estimation of the *ApEn*

In the case of heart rate variability studies, the resolution of the RR time series is an unavoidable limitation that can be modelled as additive noise. The standard deviation of the noise in each RR sample associated with the finite resolution can be computed as:

$$\sigma(RR)_{f_s} = \frac{1}{\sqrt{6}\cdot f_s} \quad (4)$$

where f_s is the sampling frequency of the ECG from where the RR time series has been obtained. This noise arises from the difference of two white processes with uniform distribution so the noise associated with the finite resolution cannot be considered white. Merri *et al* characterises more in depth this noise in [2].

Since this limitation is always present in heart rate variability studies and the sampling frequency of the ECG is not always as high as desirable, this work aims to evaluate how the resolution of the RR time series afects the estimation of *ApEn*.

Materials and Methods

Equations (1) to (3) define a family of estimators for the Approximate Entropy. In this work we will assume that the number of samples of the time series is N=300 and, as suggested by Pincus in [1], that m (the length of the vectors) is 2.

In order to estimate the bias and the uncertainty in *ApEn* some simulations of fractional Brownian noise (fBm) have been carried on because the complexity of fBm time series changes with the Hurst exponent (H). By changing H and the SD of the time series, several conclusions can be inferred on the effect of the f_s in the estimation of *ApEn*.

ISSN: 1727-1983

A fBm generator provides time series with a nearly infinite resolution. In order to add the limitation of the f_s the same methodology that was exploited in [3] has been employed: a simulated RR time series with infinite resolution is cumulatively summated in order to obtain the R time series (a time series with the time location of the QRS complexes taking the convention that the first QRS complex is at zero time). Next, a random constant (τ) is added to each sample of the R time series (τ is uniformly distributed between $-1/(2 \cdot f_s)$ and $1/(2 \cdot f_s)$). Then, the resolution of the R time series is reduced by rounding each sample to the nearest multiple of $1/f_s$. Finally, the RR time series with finite resolution is obtained by differentiating the rounded R time series. By adding different τ to the original R time series, several RR time series with finite resolution can be obtained from an infinite resolution RR time series. In order to estimate the bias and uncertainty in *ApEn* associated to this loss of resolution, the *ApEn* in the original (infinite resolution) must be calculated ($ApEn_t$) as well as those of the finite resolution time series ($ApEn_m$).

The fast algorithm for the estimation of *ApEn* employed in this work is described in [4]. The algorithm used to generate fBm time series is described in [5]. With both tools, the following simulations have been carried on:

First simulation: This simulation aims to obtain the bias and uncertainty on *ApEn* with $k = 0.2$. The simulated time series have 300 samples and the selected sampling frequency is 128 Hz. The simulation has generated 100 realizations of fBm time series whose $ApEn_t$ ranges from 0.05 to 1.18. The output of the generator has zero mean and a variance of 1.00 so each realization has been converted in 50 pseudo-RR time series by adding 1000 ms (mean heart rate of 60 bpm) and changing the standard deviation from 1 ms to 200 ms. This permits us to study the effect of the standard deviation on the bias. For each one of the 50 pseudo-RR time series, 100 realizations with finite resolution have been obtained by using the procedure previously described. Then, for each $ApEn_t$ and each simulated SD, 100 finite resolution time series have been obtained and the bias of *ApEn* for this time series has been estimated as:

$$b(SD) = ApEn_t - \frac{1}{100}\sum_{i=1}^{100} ApEn_m(i) \qquad (5)$$

where $ApEn_m$ is the estimated *ApEn* of the finite resolution time series. In this formula we stress that the bias can depend on the SD of the time series. In a similar way, the uncertainty of *ApEn* is estimated as:

$$u(SD) = \sqrt{\frac{1}{99}\sum_{i=1}^{100}\left(ApEn_m(i) - \frac{1}{100}\sum_{j=1}^{100} ApEn_m(j)\right)^2} \qquad (6)$$

As we will see in the results section, the bias is in general very much greater than the uncertainty. The following simulations have the purpose to observe how the bias changes with f_s and k.

Second simulation: As in the previous simulation, 100 realizations of fBm with 300 samples have been employed. For each one, 50 pseudo-RR time series have been obtained with SD ranging from 1 ms to 200 ms. Because the uncertainty is negligible in front of the bias, only one finite realization for each tested f_s has been obtained with $\tau = 0$. The considered sampling frequencies have been 128 Hz, 250 Hz, 500 Hz, 1 kHz and 2 kHz.

Third simulation: This simulation has the purpose to observe the effect of the threshold on the bias of the *ApEn*. We have obtained 5000 realizations of pseudo-RR time series with different SD. From each realization, a finite resolution time series has been obtained with a sampling frequency of 128 Hz ($\tau = 0$). *ApEn* has been obtained for k changing from 0.10 to 0.30 in 0.05 steps.

Actual RR time series can not be considered as purely fBm. In order to observe if the bias follows a law similar to that obtained from simulation, we have studied RR time series from the *normal sinus rhythm RR time interval database* that can be downloaded from [6]. These time series are derived from electrocardiographic recordings sampled at 128 Hz. We have analyzed 2342 RR time series without artifacts each consisting in 300 samples. For each time series, we have calculated $ApEn_m$ ($k = 0.2$). We have obtained 100 infinite resolution candidate time series by adding uniform noise between $-1/(2 \cdot f_s)$ and $1/(2 \cdot f_s)$ to the cumulated sum of the RR time series and taking the differentiation next. The bias has been obtained in a similar way as in (5) although we have interchanged $ApEn_t$ by $ApEn_m$ and the sign of the expression.

Results

Figure 1 shows the bias on *ApEn* obtained for the first simulation. There is not a monotonically behaviour of the bias with the increase of the standard deviation. Instead, the bias of *ApEn* oscillates (though the amplitude of the oscillation decreases with increasing SD) with a certain period that is, approximately, 39 ms and, from the other simulations, can be formulated as:

$$T(b(SD(ms))) = \frac{1000}{f_s \cdot k} \qquad (7)$$

where T is the period of the oscillation, k the threshold and f_s the sampling frequency. The value of the maximum and minimum in each cycle depends on the value of the approximate entropy ($ApEn_t$).

The obtained uncertainty is very much lower than the bias as shown in figure 2. Mean uncertainty ± standard deviation are displayed respect to the SD of the time series.

ISSN: 1727-1983

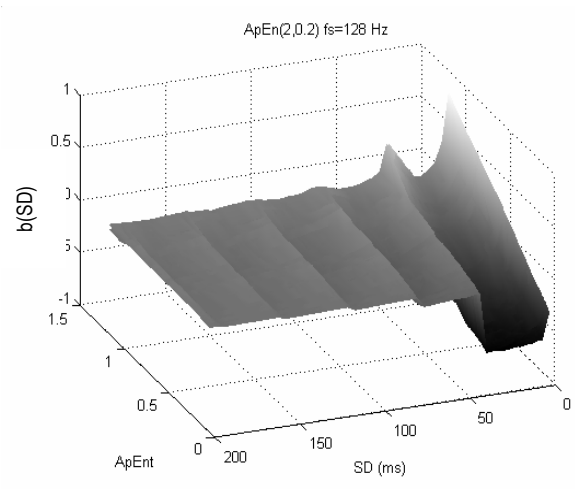

Figure 1: Bias on approximate entropy obtained by the first simulation (f_s = 128 Hz)

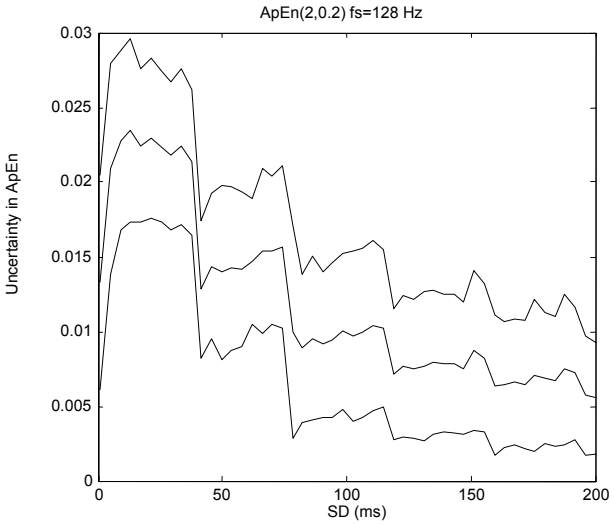

Figure 2: Mean and mean ± SD of the uncertainties in *ApEn* obtaining in the first simulation (f_s = 128 Hz)

The second simulation provides information on what is the minimal sampling frequency that must be employed in order to keep the bias lower than a certain limit. Results show that the higher the sampling frequency, the lower the bias is. Nevertheless, the same oscillatory effect still holds with a period as defined in (7). Moreover, the bias for a certain standard deviation can be related with that obtained in the case of f_s = 128 Hz by:

$$b(SD)_{f_s} \approx b\left(SD \cdot \frac{f_s}{128Hz} \right)_{128Hz} \qquad (8)$$

As an example: the bias for a time series with standard deviation of 30 ms and f_s = 500 Hz will be approximately equal to that of a time series with equal *ApEn* sampled at 128 Hz and with a standard deviation of 117 ms.

Figure 3 shows the obtained bias for the different sampling frequencies. Note that the abscise axis is the product of the standard deviation of the time series and the sampling frequency. For a constant product of sampling frequency and SD, each dot is the bias for a certain $ApEn_t$.

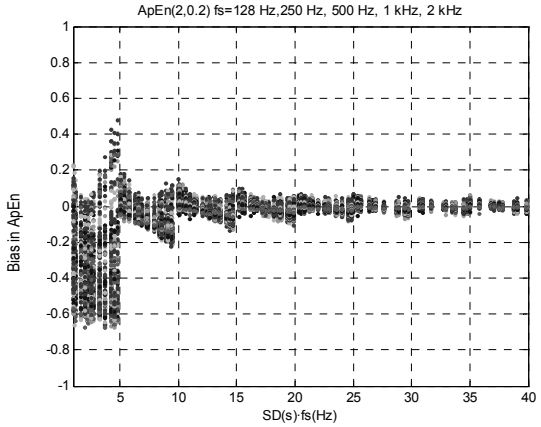

Figure 3: Bias in *ApEn* obtained by the second simulation for different sampling frequencies

The third simulation has the purpose of studying the effect of *k* in the bias of *ApEn*. Figure 4 shows this bias for the considered thresholds. Now, the abscise axis is the product of the threshold and the standard deviation of the time series because the following approximation holds:

$$b(SD)_k \approx b\left(SD \cdot \frac{k}{0.2} \right)_{0.2} \qquad (9)$$

and for any sampling frequency and *k*:

$$b(SD)_{k,f_s} \approx b\left(SD \cdot \frac{k}{0.2} \cdot \frac{f_s}{128Hz} \right)_{0.2,128Hz} \qquad (10)$$

and $b(x)_{0.2,128\ Hz}$ can be obtained in figure 1 for several $ApEn_t$.

Finally, figure 5 shows the bias in *ApEn* obtained for actual RR time series compared with the results of the first simulation because the *normal sinus rhythm RR time interval database* is sampled at 128 Hz. The estimations of the bias are similar to that obtained by simulation. The accumulation of circles in the positive half of the bias for SD lower than 50 ms is due to the fact that the approximate entropy of the time series is near 1.0 while the first simulation considered a wider range of *ApEn* (see figure 1).

Discussion

The results show that the sampling frequency causes a bias in the approximate entropy and, at a lower degree, an uncertainty. The bias depends on the standard deviation of the RR time series, on the sampling frequency of the ECG, on the threshold and on the *ApEn* by itself. Surprisingly, in certain conditions, the bias can

ISSN: 1727-1983

rise by increasing the SD of the time series. This implies that when measuring time series with identical complexity different *ApEn* can be obtained due to a slight change in SD. As an example: Let's suppose a time series whose approximate entropy (infinite resolution) is 0.92. The estimated *ApEn* when sampled at 128 Hz with k = 0.2 will be 0.96 if SD = 74 ms and 0.85 if SD = 78 ms.

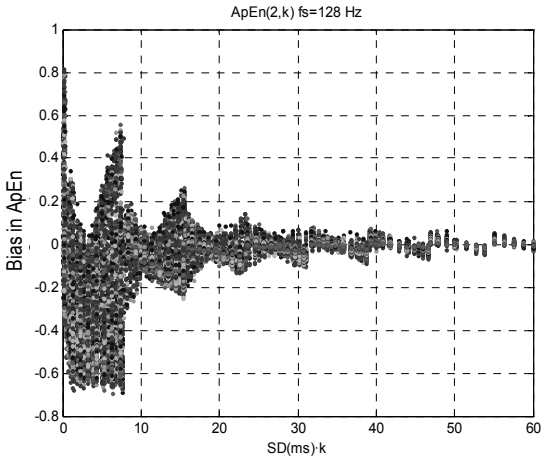

Figure 4: Bias in *ApEn* obtained by the third simulation for different thresholds (f_s = 128 Hz)

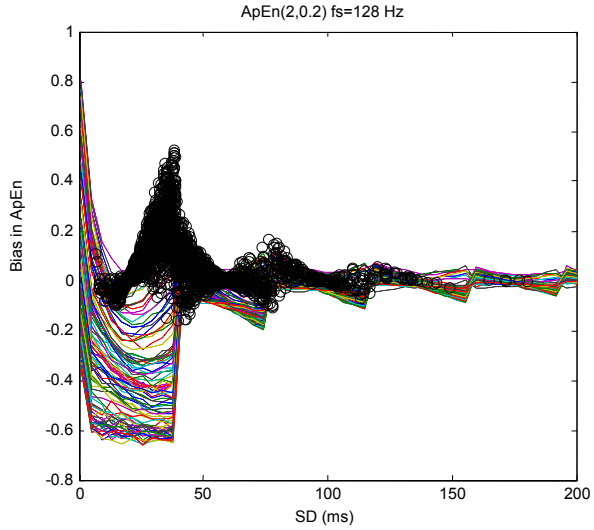

Figure 5: Bias in *ApEn* obtained in the normal sinus rhythm RR interval database (black circles) compared with the bias of the first simulation (coloured lines)

From the previous example, the following question arises: what is the minimum standard deviation of a time series in order to ensure that the bias in *ApEn* is lower than a certain limit ± e? In the case of f_s = 128 Hz, k = 0.2 and e = 0.05, the minimum SD is 160 ms (considering all the studied approximate entropies). This limit reduces to 20.5 ms when the sampling frequency is 1 kHz (see expression (10)). Several studies consider a sampling frequency of 1 kHz as a sufficiently higher value. Nevertheless, if the standard deviation is lower than 20.5 ms (a very usual case in pathological subjects or in exercise tests) the *ApEn* should not be reported with more than a decimal figure due to its bias.

The solution to these problems should be the correction of the bias. Nevertheless, it is not an immediate process because there is not a bijective correspondence between the bias and the approximate entropy. In the usual practice, we obtain the $ApEn_m$ because we have knowledge of the finite resolution time series. To this value, several values of $ApEn_t$ can be considered as candidates for the correction of the bias.

Finally, the effect of the number of samples has not been assessed in this work and probably also affects the bias. We only have considered a time span of 5 min (if the heart rate is close to 60 bpm).

Conclusions

The sampling frequency that it is used to acquire electrocardiograms affects the estimation of the approximate entropy of RR time series. The major effect is a bias that depends on the sampling frequency, on the standard deviation of the time series (not monotonically), on the threshold and on the approximate entropy of the signal. Time series with the same complexity but with slightly different standard deviation can present values of approximate entropy quite different. The error in the approximate entropy when using a sampling frequency of 1 kHz and a threshold of 0.2 can be higher than 0.05 if the standard deviation of the time series is lower than 20.5 ms.

References

[1] PINCUS S.M. (1991): 'Approximate entropy as a measure of system complexity', *Proc Natl Acad Sci USA*, **88**, pp. 2297-2301

[2] MERRI M., FARDEN D.C., MOTTLEY J.G. and TITLEBAUM E.L. (1990):'Sampling Frequency of the Electrocardiogram for Spectral Analysis of the Heart Rate Variability', *IEEE Transactions on Biomedical Engineering* , **37**, pp.99-106

[3] GARCÍA-GONZÁLEZ M.A., FERNÁNDEZ-CHIMENO M. and RAMOS-CASTRO J. (2004): 'Bias and uncertainty in heart rate variability spectral indices due to the finite ECG sampling frequency', *Physiological Measurement*, **25**, pp. 489-504

[4] FUNSENG Y., BO H. and QINGYU T. (2001): 'Approximate Entropy and its application in biosignal analysis', in AKAY M. (Ed.) (2001): 'Nonlinear Biomedical Signal Processing Volume II Dynamic Analysis and Modeling', (IEEE Press, New York), pp. 72-91

[5] PAXSON V. (1997):'Fast, Approximate Synthesis of Fractional Gaussian Noise for Generating Self-Similar Network Traffic', *Computer Communication Review,* **27**, pp. 5-18

[6] http://www.physionet.org/physiobank/database/nsr2db/

ISSN: 1727-1983