Tolerance to Spikes: a Comparison of Sample and Bubble Entropy

George Manis[1], Roberto Sassi[2]

[1]Deptartment of Computer Science and Engineering, University of Ioannina, Greece
[2]Dipartimento di Informatica, Università degli Studi di Milano, Italy

Abstract

Bubble Entropy is a new metric aiming at the quantification of the entropy of a series, the most important property of which is the total elimination of the scale parameter (e.g., r when computing Sample Entropy) and the low dependency on the length of the runs compared (e.g., the m parameter).

In this paper, we compare the tolerance of Bubble Entropy to spikes and compare it to that of Sample Entropy. We use RR series publicly available on Physionet and compute Sample and Bubble Entropy before and after artificially added spikes. We add N spikes of value $a \cdot \mathrm{std}$, where a is a parameter and std the standard deviation of the signal. We compute then the relative error (absolute error / known value).

According to our experiments, Bubble Entropy exhibits a remarkable tolerance to spikes. In all our experiments, Bubble Entropy presents a low relative error. Compared to Sample Entropy, the error reported by Bubble Entropy is always smaller and sometimes remarkably ($p \ll 0.001$, Wilcoxon rank-sum test). The mean relative error for Sample Entropy ranges from 0.084 to 0.435, while the mean relative error for Bubble Entropy from 0.069 to 0.122. Sample over Bubble Entropy relative errors ratios reach up to 3.57.

1. Introduction

The value of using non linear methods in the quantification of properties of Heart Rate Variability (HRV) has been proved in the last decades by their extensive use, both in clinical practice and research. An analytical report on the use of these methods can be found in [1]. Among the others, techniques meant to assess the entropy of the HRV series had a relevant spot, with Approximate (ApEn) [2] and Sample Entropy (SampEn)[3][4] being those which were mostly used and cited. The research community seems, day by day, to trust Sample Entropy more than Approximate Entropy, even though the advantages of the latter is not wise to be ignored.

The number of papers using and citing Sample Entropy is huge. In an attempt to show its validity and wide acceptance, we refer here to a few which belong in diverse scientific fields and not in the intersection of computing and cardiology or, even generally, in biomedical engineering. Sample Entropy has been used, amongst others, in earthquake data analysis [5], in stock market manipulation studies [6], in mechanical engineering [7] and as a measurement for climate complexity [8].

A remarkable disadvantage of entropy based metrics is their dependence on parameters which should be selected manually and empirically by the user. In particular, both ApEn and SampEn depend on the number m of consecutive points of the series considered (subseries) and on a scale parameter r (the tolerance employed in the comparisons of subseries). Even though typical values for both Approximate and Sample Entropy are usually used and seem to be widely accepted, the dependence of entropy based metrics on parameters is still there and can not be ignored.

Recently *Bubble Entropy* has been proposed [9] as an attempt to introduce an entropy based metric "without parameters", or with minimum dependence on them. While still depending on parameters, Bubble Entropy has successfully eliminated the necessity of the parameter r and has limited the importance of the parameter m. Until now, Bubble Entropy has been successfully tested in the discrimination of recordings acquired from healthy and from congestive heart failure patients [9] and for identifying electrocardiotocograms of fetuses with normal and abnormal pH values at birth [10].

Currently, we are working on identifying the advantages and disadvantages of the method and the characteristic of the recordings on which it is beneficial to be applied. In this paper we test the tolerance of the method to the existence of spikes in heart rate variability series. In order to have a meaningful and comparative view with another method, we selected, not surprisingly, to compare it with Sample Entropy.

In the rest of the paper, we first present Bubble Entropy, in a detail sufficient for the reader to understand the basic idea, the validity and the definition. Focus is given to

the limited dependence of Bubble Entropy on the parameters. Next, we present our experimental results, showing that Bubble Entropy is significantly tolerant to spikes, more tolerant than Sample Entropy. The last section states the main conclusions of this work.

2. Bubble Entropy

In this section, we first give some background information in order to show how Bubble Entropy was evolved from existing common entropy based methods. Next we describe the method itself.

A reasonable starting point for the elimination of the parameter r is symbolic analysis [11]. In symbolic analysis, the range of possible values of a time series is split into a number of equal sized intervals. Each interval is given a symbolic name, e.g. A, B, C, etc. Each point of the time series belongs in one of those intervals and is named after this interval. Entropy can be computed based on these symbolic names. This can be done in one dimensional or in an m dimensional space.

Symbolic series: BCCACBDC

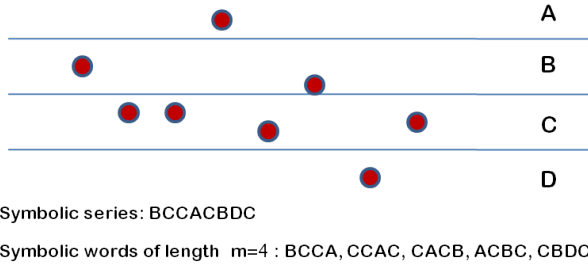Symbolic words of length m=4 : BCCA, CCAC, CACB, ACBC, CBDC

Figure 1. Symbolic analysis: example of a symbolic series and of symbolic words of length $m = 4$

In figure 1 we can see an example. The three lines separate the space into four subspaces. Each subspace is named as A, B, C or D. The first point belongs in subspace B, the second in C, the third in C, so the symbolic sequence $BCCACBDC$ is formed. If we embed the series into an $m = 4$ dimensional space, we can define symbolic words. The first four symbols form the word $BCCA$, symbols in positions 2 to 5 form the word $CCAC$ and so on.

In the computation of entropy described above there is no r parameter, since there is no reason to define the distance between two vectors. We have two relations between symbolic names: identical and different.

However, in this way the r parameter has not been really eliminated. Actually, it has been hided or, even better, it has been replaced by another parameter: the size of the intervals or (equivalently) the number of the intervals.

In order to describe the space with symbols without dividing it into predefined or user defined subspaces, we first embed the time series in an m dimensional space. Then

for each vector (point in the m dimensional space), we define the subspaces, not based on thresholds with predefined positions (lines as in figure 1), but using the ranks of the elements ("lines" defined by the values of the elements in each vector, please see figure 2).

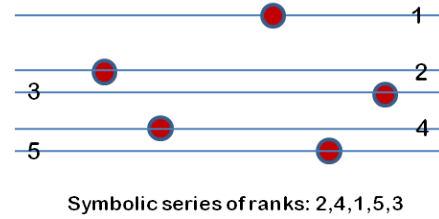

Symbolic series of ranks: 2,4,1,5,3

Figure 2. Ranks form a symbolic series of length $m = 5$

Next, we give symbolic names to each element of the vector depending on which subspace it belongs. In this way, each vector is transformed to a symbolic word of length m. Instead of using letters as symbols, we select to use numbers as symbols as shown in figure 2.

One can notice here two things: a) in each symbolic word, only the symbols $1 \ldots m$ appear and each symbol appears in a word only once; b) the produced symbols express the order (rank) of each element, i.e. the position it will have if we sort the vector.

Since a sorting procedure is involved and since the effort invested in the sorting procedure expresses, in a way, some kind of complexity, we select to base our further computation on this complexity. The most appropriate algorithm to give us a clear measure of the sorting effort is *Bubble Sort*. We employ Bubble Sort and name the method after it.

When computing Bubble Entropy, for each vector of the m dimensional space we count the number of swaps required by Bubble Sort to sort this vector. By doing so, we achieve to transform the given time series into a series of numbers of sorting steps, a series expressing complexity. We compute the entropy on this series.

Similarly to Sample Entropy:
- Bubble entropy is defined as a conditional entropy, i.e. what is reported is the increase of entropy when moving from an m-dimensional space to the $m+1$ dimensional space
- We use the Rényi definition of entropy of order 2:

$$E_{\text{Rényi}} = \frac{\log \sum_{j=1}^{m} p_j^{\alpha}}{(1 - \alpha) \log m}. \qquad (1)$$

Intuitively, a second order entropy metric give emphasis on sharp changes, something important in heart rate variability signals, at least more emphasis than a linear definition, like the one proposed by Shannon. A third order entropy or an entropy of higher order would emphasize sharp

Table 1. Sample / Bubble Entropy Relative Error

Parameters	Sample	Bubble	Sample / Bubble	p-value
α=10, N=1	0.118	0.076	1.55	$\ll 0.001$
α=10, N=2	0.161	0.102	1.58	$\ll 0.001$
α=10, N=3	0.178	0.120	1.48	$\ll 0.001$
α=20, N=1	0.084	0.069	1.21	$\ll 0.001$
α=20, N=2	0.134	0.115	1.17	$\ll 0.001$
α=20, N=3	0.178	0.101	1.76	$\ll 0.001$
α=30, N=1	0.139	0.070	1.99	$\ll 0.001$
α=30, N=2	0.285	0.099	2.88	$\ll 0.001$
α=30, N=3	0.435	0.122	3.57	$\ll 0.001$

changes even more, something that it is not necessarily desirable.

Pseudo-code for the computation of Bubble Entropy follows:

- *step 1:* Compute Entropy in m dimensional space:
    - *step 1.1:* embed the signal into m dimensional space
        - *step 1.1.1:* for each vector compute the number of swaps required by Bubble Sort to sort it
    - *step 1.2:* construct a series with the computed number of swaps
    - *step 1.3:* use Rényi definition of entropy to compute the amount of entropy in m-dimensional space
- *step 2:* Compute Entropy in $m+1$ dimensional space following the procedure described in step 1
- *step 3:* Report the difference of Entropy computed in steps 1 and 2

3.   Tolerance to Spikes

For our experiments we used the RR series contained in the *Fantasia* dataset [12]. This is a publicly available dataset which can be freely downloaded from Physionet [13]. It consists of 40 recordings, 20 of which have been acquired from young subjects of age between 21 and 34 years old, and 20 from elderly subject all between 68 and 85 years. All subjects were healthy. All subjects remained in a resting position and in sinus rhythm during recording. They were watching the movie *Fantasia* (Disney, 1940) to help maintain monitoring conditions consistent among subjects.

From each RR signal of the Fantasia dataset we randomly selected 300 consecutive RR intervals and computed Sample and Bubble Entropy on them. Then, we randomly added N spikes of value $a \cdot \text{std}$, where std is the standard deviation of the signal. We computed again Sample and Bubble Entropy and estimated the relative error (absolute error / known value). We repeated the experiment 100 times and reported the mean relative error (*i.e.*, average of 4000 signals).

Results are shown in table 1 for $a = 10, 20, 30$ and for $N = 1, 2, 3$. Commonly employed parameters have been selected for Sample Entropy ($m = 2$, $r = 0.2$) and $m = 10$ for Bubble Entropy. As we discussed in [9], the value of m is not critical when computing Bubble Entropy, as long as m is reasonably large.

In the table, the value before the slash is the relative error of Sample Entropy and the value after the slash is the corresponding error of Bubble Entropy. In all cases Bubble Entropy presents a low relative error, sometimes much lower than that of Sample Entropy ($p \ll 0.001$, Wilcoxon rank-sum test). The mean relative error for Sample Entropy ranges from 0.084 to 0.435, while the mean relative error for Bubble Entropy from 0.069 to 0.122. Sample over Bubble Entropy relative errors ratios reach up to 3.57 for the values of α and N we tested.

To explain this outcome, we must notice that the effect of a spike in the computation of Bubble Entropy does not modify significantly the number of the expected swaps. The maximum number of swaps is the swaps necessary to sort a vector whose elements are in descending order:

$$n_{max} = \frac{m(m-1)}{2}. \tag{2}$$

The expected number of swaps when sorting a vector of size m is:

$$n_{exp} = \frac{1}{2}\frac{m(m-1)}{2} = \frac{m(m-1)}{4}. \tag{3}$$

If we devide by m, we get the expected swaps per element:

$$n_{elem} = \frac{m(m-1)/4}{m} = \frac{(m-1)}{4}. \tag{4}$$

This is the upper bound of the number of swaps which a spike is expected to add or remove to the total number of swaps necessary to sort the vector. This amount is not significant compared to the expected number of swaps, and it will be uniformly distributed, something that will not influence significantly the final computed value of Bubble Entropy.

On the other hand, when we use Sample Entropy:
- spikes influence the value of r. Please recall that r depends on the standard deviation of the signal. Standard deviation is highly affected by spikes. The value of r participates in every check for similarity;
- all checks for similarity which include spikes are expected to fail.

4.   Conclusions

In this paper we study the tolerance of Bubble Entropy to spikes in heart rate variability signals. Bubble Entropy embeds the signal into an m dimensional space and then

sorts each vector of the m dimensional space using the Bubble Sort algorithm. The number of necessary swaps produces a new series for which the Rényi Entropy is computed. The same procedure is followed in the $m+1$ dimensional space. The difference of the entropy in these two spaces is reported as Bubble Entropy.

We used publicly available signals and we artificially added spikes in them. Then we compared the relative error produced in the values of Bubble Entropy and Sample Entropy, due to these spikes. Our experiments showed that Bubble Entropy was proved more tolerant to the existence of spikes.

The application of Bubble Entropy [9, 10] has shown that it is a promising method which can successfully extract valuable information from HRV signals. The method should be further tested in practice as a prognostic and diagnostic index using available datasets and clinical scenarios.

References

[1] Sassi R, Cerutti S, Lombardi F, Malik M, Huikuri HV, Peng CK, Schmidt G, Yamamoto Y. Advances in heart rate variability signal analysis: joint position statement by the e-Cardiology ESC Working Group and the European Heart Rhythm Association co-endorsed by the Asia Pacific Heart Rhythm Society. Europace September 2015;17(9):1341–1353.

[2] Pincus SM. Approximate entropy as a measure of system complexity. Proc Natl Acad Sci March 1991;88(6):2297–2301.

[3] Lake DE, Richman JS, Griffin MP, Moorman JR. Sample entropy analysis of neonatal heart rate variability. Am J Physiol Reg I September 2002;283(3):R789–R797.

[4] Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. American Journal of Physiology Heart and Circulatory Physiology 2000;278(6):H2039–H2049.

[5] Min L, Guang M, Sarkar N. Complexity analysis of 2010 Baja California earthquake based on entropy measurements. In Second International Conference on Vulnerability and Risk Analysis and Management (ICVRAM) and the Sixth International Symposium on Uncertainty, Modeling, and Analysis (ISUMA). 2014; .

[6] Reddy YV, Sebastin A. Measuring stock price manipulation using entropy analysis: A conceptual framework. The ICFAI Journal of Applied Finance 2006;12(5):14–28.

[7] Yao P, Xue J, Zhou K, Wang X. Sample entropy-based approach to evaluate the stability of double-wire pulsed MIG welding. Mathematical Problems in Engineering 2014;.

[8] Shuangcheng L, Qiaofu Z, Shaohong W, Erfu D. Measurement of climate complexity using sample entropy. International Journal of Climatology 2006;26(15):2131–2139.

[9] Manis G, Aktaruzzaman MD, Sassi R. Bubble entropy: an entropy almost free of parameters. Transactions on Biomedical Engineering 2017;.

[10] Manis G, Sassi R. Relation between fetal HRV and value of umbilical cord artery pH in labor, a study with entropy measures. In Proc. of 30th IEEE International Symposium on Computer-Based Medical Systems. 2017; .

[11] Citi L, Guffanti G, Mainardi L. Rank-based multi-scale entropy analysis of heart rate variability. In Computing in Cardiology, volume 41. 2014; 597–600.

[12] Iyengar N, Peng CK, Morin R, Goldberger AL, Lipsitz LA. Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. The American Journal of Physiology October 1996;271(4 Pt 2).

[13] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000 (June 13);101(23):e215–e220.

Address for correspondence:

George Manis

University of Ioannina, Dept. of Computer Science and Engineering, P.O. Box 1186, Ioannina, 45110, Greece.
manis@cs.uoi.gr