

BOOTSTRAPPING THE EMPIRICAL BOUNDS ON THE VARIABILITY OF SAMPLE ENTROPY IN 24-HOUR ECG RECORDINGS FOR 1 HOUR SEGMENTS

Sebastian Zurek¹, Waldemar Grabowski¹, Marcin Kosmider², Szymon Jurga²
Przemyslaw Guzik³, Jaroslaw Piskorski¹

¹ Institute of Physics, University of Zielona Gora
Zielona Gora, Poland

² Faculty of Medicine and Health Sciences, University of Zielona Gora
Zielona Gora, Poland

³ Department of Cardiology - Intensive Therapy, Poznan University of Medical Sciences
Poznan, Poland

sebzur@codelab.pl, wgrabowski@gmail.com, kosmo@codelab.pl, szymon.jurga@vp.pl
pguzik@ptkardio.pl, jaropis@zg.home.pl

Received: 15 May 2018; Accepted: 28 June 2018

Abstract. We investigate the variability of one of the most often used complexity measures in the analysis of the time series of *RR* intervals, i.e. Sample Entropy. The analysis is carried out for a dense matrix of possible r thresholds in 79 24h recordings, for segments consisting of 5000 consecutive beats, randomly selected from the whole recording. It is repeated for the same recordings in random order. This study is made possible by the novel NCM algorithm which is many orders of magnitude faster than the alternative approaches. We find that the bootstrapped standard errors for Sample entropy are large for *RR* intervals in physiological order compared to the standard errors for shuffled data which correspond to the maximum available entropy. This result indicates that Sample Entropy varies widely over the circadian period. This paper is purely methodological and no physiological interpretations are attempted.

MSC 2010: 92B25, 92B15, 68W99

Keywords: sample entropy, complexity, bootstrap, heart rhythm variability

1. Introduction

Sample Entropy is one of the most often used complexity-based measures used in Heart Rate Variability analysis [1–3]. It is based on the Grassberger-Proaccia [4] approach to calculating the correlation dimension, which was utilized by Steven Pincus to define the complexity parameter Approximate Entropy (ApEn) with the aim of estimating the Kolmogorov-Sinai (KS) entropy for certain processes and also quantifying the complexity inherent in biological processes. This approach was also used for quantifying the complexity inherent in biological processes, mainly the *RR* inter-

vals time series, which is the time series of the time distances between consecutive R waves in the QRS complexes [1, 5–7]. This approach was further refined by Richman and Moorman [7] to remove the bias which is inherent in ApEn as well as improving the definition of the parameter in probabilistic terms (SampEn involves exact conditional probabilities, while ApEn only uses approximate ones) and making the parameter more reliable for short recordings.

Since then, SampEn has been used in an enormous number of theoretical, physiological and clinical studies. The applications range from evaluating emotions [8], to HRV [9], to predicting the life of Li-ion batteries [10]. The methodology is usually very similar - the variance of the analyzed time series is calculated, and then the similarity threshold is selected in relation to the standard deviation (i.e. square root of the calculated variance). The usual SampEn algorithm is applied to the time series and the results are processed with the use of standard statistical methods.

The segments for which the analysis is carried out is usually quite short - from a few hundred to a few thousand points [11]. This is due to the relatively high computational burden of the method which makes it impracticable to calculate for long time series. A recent development in the form of the NCM algorithm makes it possible to rapidly calculate this parameter for very long recordings, including 24-hour ECG recordings. This algorithm does not rely on hardware, like multiprocessing for instance, but rather relies on exploiting the mathematical structure of the algorithm to speed it up, without introducing any approximations [11].

As mentioned above, the analyzed segments of ECG are quite short and they are selected quite randomly from all the possible segments in a 24-hour recording. Thus, it is interesting to see what the variability of the SampEn parameter is throughout the 24 hours, especially in relation to a well-defined level which in this paper is assumed to be the *RR* intervals time series in random order.

For this reason, in the present paper, our goal is to derive empirical bounds for SampEn variability in a physiological time series of *RR* intervals, study their dependence on the threshold r , and relate it to natural reference level, which is the SampEn of the recordings in random order. Another, equally important goal of this paper is to check whether it is possible to construct a bootstrapping approach to comparing the values of Sample Entropy over a range of values with the use of the NCM algorithm.

2. Sample Entropy

In the present section, we describe the salient properties of Sample Entropy to the extent necessary to carry on the analysis in the paper. Details may be found in [7] and references therein.

The algorithm of Sample Entropy may be examined from many points of view as presented in [11]. One of them, which proved to be helpful in algorithmic work, is the differences between curves that are built from correlation sums, defined by the following functions:

$$C^m(r) = L_m^{-1} \sum_{i=1}^{L_m} C_i^m(r). \quad (1)$$

The elements inside the sum in equation (1) are defined as follows

$$C_i^m(r) = (L_m - 1)^{-1} \sum_{j=1, j \neq i}^{L_m} \Theta(r - |\vec{v}_m(i) - \vec{v}_m(j)|), \quad (2)$$

where Θ is the *Heaviside function* defined in the following way

$$\Theta(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \quad (3)$$

The norm in the Heaviside function is the *maximum coordinate difference* [5] and is defined as

$$|\vec{v}_m(i) - \vec{v}_m(j)| = \max_{k=1,2,\dots,m} (|u(i + (k-1)\tau) - u(j + (k-1)\tau)|). \quad (4)$$

and vectors v are constructed in the following way

$$\vec{v}_{m,\tau}(i) = [u(i), u(i + \tau), u(i + 2\tau), \dots, u(i + (m-1)\tau)] \quad (5)$$

i.e. vector $\vec{v}_{m,\tau}(i)$ consists of m consecutive points in the studied dataset. The r parameter in the above formulas is the threshold, which is usually defined in relation to the standard deviation of the time series, usually between 0.1 and 0.2 [5].

An alternative way of looking at the C_i^m parameter is the following: it is the probability that a vector $v_m(j)$ is within distance r of *another* vector $v_m(i)$. This can be expressed as:

$$C_m^i = \frac{B_i}{N - m + 1} \quad (6)$$

Where B_i is the number of “matches”, or, in other words, the count of how many times the Heaviside function defined by eq. (3) is non-zero [7].

Sample entropy is then defined as

$$\text{SampEn} = -\ln \frac{C^{m+1}(r)}{C^m(r)}, \quad (7)$$

which can be interpreted as the spread between adjacent $\ln(C^m(r))$ curves in the first interpretation [11] or as the negative log of conditional probability that two sequences within a maximum distance of r for m points remain within this distance for $m + 1$ points [7].

The first interpretation uses similarity while the other one uses repeatability - both lead to the conclusion, that SampEn may be treated as a measure of complexity. It can be argued that an estimator of this type estimates the KS entropy for a limited range of well-defined processes [5].

3. The NCM algorithm

The NCM algorithm is based on both the *look-up* table and vectorization techniques. The procedure utilizes triangular matrices \hat{N} with the following elements:

$$n_{ij} = ||u_i - u_{i+(j+1)\cdot\tau}|| \quad (8)$$

where u is the component of the time series (e.g. RR interval) and τ is the time lag. For a dataset containing N points, the dimension of the matrix is $N - \tau - 1$ rows and $(N - 1)/\tau - 1$ columns, which is a rather large matrix, but it is enough to set a limited sub-block for the calculation. Because of the symmetry of the matrix, elements whose indices do not meet the condition $i + (j + 1) \cdot \tau \leq N - \tau - 1$ are set to zero, which results in halving the number of summations in the correlation sum algorithm (eq. (2)). The next effect of the assumed approach is removing redundant operations in calculating C^m by reducing the number of operations for a maximum norm from m^2 to m . The last large speed improvement is obtained by calculating the first occurrence of 0 from an arithmetic operation instead of a loop, with a zero at

$$(n_{i,j}^m - b)/a, \quad (9)$$

for every $\hat{n}_{i,j}^m$ and with the set of filter values defined as a linear function $\hat{r}f(\hat{i})$ with

$$a = -\frac{r_{max} - r_{min}}{n - 1}, \quad b = r_{max}. \quad (10)$$

There are, obviously, other optimizations customarily used in algorithmic techniques. The full description of the algorithm may be found in [11].

4. Bootstrapping

Bootstrapping is a method of testing or measuring a quantity which relies on random sampling with replacement. It allows one to estimate the sampling distribution of statistics in cases in which the underlying distribution is unknown. The basic premise of this method is that the repeated sampling with replacement from the sample parallels repeated measurement from the population. Bootstrapping can be used for constructing tests of hypotheses or to construct estimators without the use of standard statistical inference, thus avoiding the complications of pivotal variables [12].

5. Materials and methods

The calculations were carried out for *RR* intervals extracted from the 24-h Holter ECG recordings of 79 healthy volunteers, 41 men, aged 35 ± 7.4 years. All subjects were in sinus rhythm (55-90 beats per minute) and the average number of heart beats of sinus origin was 106432. The subjects had not suffered from any chronic or acute disease up to 3 months before the examination. The participants were not taking any medications and were not involved in any kind of endurance training.

The ECG lines from which the *RR* intervals were extracted were sampled at 200 Hz. No interpolation algorithm was used to extract the R-waves. The resting heart rate of the subjects was between 50 and 90 beats per minute. The R-waves were first automatically identified by the Zymed system and then reviewed by a trained technician. No *RR* time series contained over 1% ectopic beats or artefacts.

To establish the experimental bounds of the variability of the *RR* intervals, each recording was divided into 5000 beats (approximately 1-hour) segments. These segments were selected randomly, with replacement from the whole recording so as to meet the bootstrapping technique requirements [12]. By repeated sampling, we establish the data-based regions corresponding to three standard errors of the normal distribution, beyond which the distributions drop to zero.

The 24-h Holter ECG contains data from both wake and sleep states. In this paper, however, we do not try to reach any physiological conclusions. Our aim is to show how to construct bootstrapped standard errors with the use of NCM for a series over an extended amount of time. *RR* intervals from 24-h are among the longest such series and the seep-wake difference does not affect any of the methods constructed in this paper.

The SampEn was calculated for all the segments, for a dense range of r values – from 0.01 to 300 ms with the step of 0.3 ms. This approximately corresponds to the fractional change of SD from $8.3 \times 10^{-5}SD$ to $2.5SD$ by $2.5 \times 10^{-4}SD$. This step was carried out with the use of the NCM algorithm. If the main purpose of the study was physiological rather than methodological, using the fractional change of SD from the start would be a more correct approach. The Python code for the implementation used in this paper can be found in [13].

The next step consisted of shuffling the 79 recordings to random order and repeating the above procedure.

We present the results as graphs and some representative numerical results.

6. Results

The profile of the SampEn parameter obtained for a single individual is presented in Figure 1. The yellow part of the figure corresponds to the segments in physiological order, the green elements are segments for the same recordings after shuffling. Table 1 contains a selection of representative SampEn values for characteristic

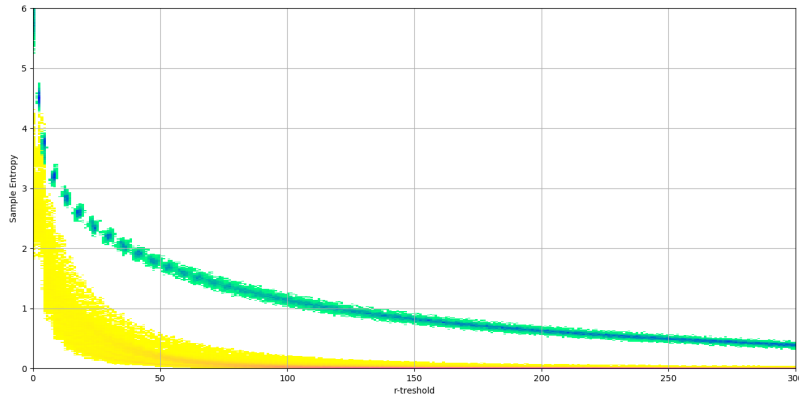


Fig. 1. The values of SampEn for a range of r values for 1-hour segments from a healthy, 36 year old male. The yellow area corresponds to the RR intervals in physiological order, the green area corresponds to the same data after shuffling

r values for the single recording presented in Figure 1 - they are expressed as ranges (minimum to maximum values) measured over segments of 0.6 ms width on the r axis, centered on the r value.

Table 1

SampEn ranges (minimum to maximum value) for characteristic r values

r [ms]	50	100	150
physiological order	(0.02 – 0.59)	(~ 0 – 0.19)	(~ 0 – 0.11)
random order	(1.60 – 1.88)	(0.99 – 1.24)	(0.84 – 1.06)
r [ms]	200	250	300
physiological order	(~ 0 – 0.08)	(~ 0 – 0.03)	(~ 0 – 0.01)
random order	(0.54 – 0.70)	(0.41 – 0.56)	(0.30 – 0.42)

As can be seen, the shuffled distribution lies far above the physiological distribution and is much less variable. The boundaries of each region for each r value may be treated as corresponding to standard errors of the mean value, since they drop to 0 outside. On the basis of the analysis, we can say that the difference between the Sample Entropy derived from physiologically and randomly distributed RR s is highly statistically different for all values of threshold r . This is true of all the recordings analyzed. It is also interesting to note that the SampEn values are very variable for each value of r resulting in a substantial spread, and that this spread also depends on r . This is not the case for the shuffled values, which are well localized along the curve, or all values of r .

The profile for SampEn for the whole group is presented in Figure 2 and Table 2 - the interpretation of the regions is the same.

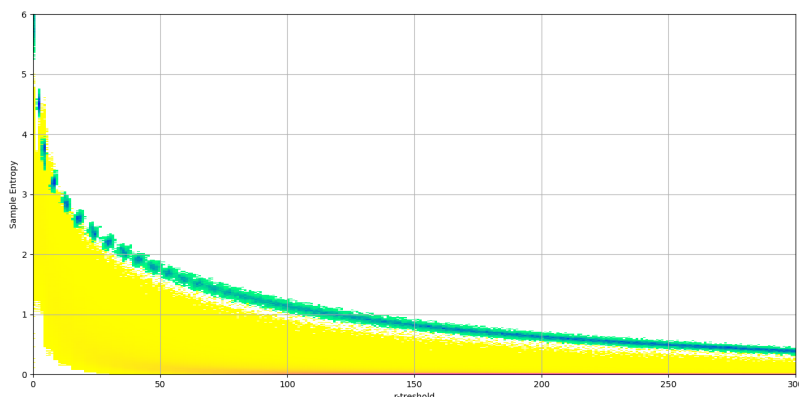


Fig. 2. The values of SampEn for a range of r values for 1-hour segments from the whole group of 79 healthy individuals - the rest of the description as in Figure 1

SampEn ranges (minimum to maximum value) for characteristic r values for the whole analyzed group

Table 2

r [ms]	50	100	150
physiological order	($\sim 0 - 1.51$)	($\sim 0 - 0.89$)	($\sim 0 - 0.67$)
random order	(1.59 - 1.83)	(0.90 - 1.21)	(0.80 - 1.11)
r [ms]	200	250	300
physiological order	($\sim 0 - 0.61$)	($\sim 0 - 0.38$)	($\sim 0 - 0.23$)
random order	(0.55 - 0.73)	(0.42 - 0.55)	(0.30 - 0.41)

In both these figures and tables, the spread of SampEn values for physiological order is also much greater than that of random order, and, similarly to the above case, both regions are well separated, except for values of r below 10 ms, where some overlap can be observed. This overlap happens in the region that is not used for physiological interpretations (it is under 0.1 SD in our case). The two most probable reasons for this phenomenon are the fact that all RR intervals time series become less and less self-similar as r drops, so SampEn should tend to the same values as r decreases. The other reason for this may be the fact that we have used absolute values of r rather than values in fractions of SD. If the time series are generated by the same or very similar processes, using r values in ms should spread the lines along the r axis. If we had used the fractions of SD, the overlap would most probably begin for values which are even smaller. Thus, in general, all the observations made for the individual case also apply here.

7. Conclusions

In this paper we have demonstrated that the NCM algorithm makes it possible to derive the bootstrapping distribution of the Sample Entropy parameter on a dense

grid of threshold r values. We find this distribution for the RR intervals time series in both physiological order and after shuffling. We found that the areas of physiological and random processes are well separated which means that in a formal statistical test we get a significant result rejecting the hypothesis that the regions overlap. Another result we noted was the fact that in the physiological case, the region covered by the SampEn value spreads depending on the r value, and in the random case the bootstrapped boundaries are much less sensitive to the threshold.

The main methodological result is the procedure for comparing regions of a Sample Entropy value parametrized by r . This may be extended to other domains, like time and comparing similar segments of ECG for instance, which may be very important for studying processes with well-defined boundaries, like the ECG for example from polysomnographic studies, where the moment of falling asleep and waking up or transitions between phases of sleep can be identified by a trained professional. The methods presented in this paper allow one to compare similar (in a broad sense) segments.

References

- [1] Pincus, S.M., & Goldberger, A.L. (1994). Physiological time-series analysis: What does regularity quantify? *Am J Physiol (Heart Circ Physiol)*, 266, H1643-1656.
- [2] Lu, S., Chen, X., Kanters, J., Solomon, I., & Chon, K. (2008). Automatic selection of the threshold value r for approximate entropy. *IEEE Trans Biomed Eng*, 55, 1966-1972.
- [3] Stam, C.J. (2005). Nonlinear dynamical analysis of eeg and meg: Review of an emerging field. *Clin Neurophysiol*, 116, 2266-2301.
- [4] Grassberger, P., & Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D*, 9, 189-208.
- [5] Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *PNAS*, 88, 2297-2301.
- [6] Pincus, S.M., Cummins, T.R., & Haddad, G.G. (1993). Heart rate control in normal and aborted-SIDS infants. *Am J Physiol Regulatory Integrative Comp Physiol*, 264, 638-646.
- [7] Richman, J.S., & Moorman, J.R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol (Heart Circ Physiol)*, 278, H2039-2049.
- [8] Nardelli, M., Valenza, G., Cristea, I.A., Gentili, C., Cotet, C., David, D., Lanata, A., & Sciling, E.P. (2015). Characterizing psychological dimensions in non-pathological subjects through autonomic nervous system dynamics. *Front. Comput. Neurosci.*, 9, 37.
- [9] Lake, D.E., Richman, J. S., Griffin, M.P., & Moorman, J.R. (2002). Sample entropy analysis of neonatal heart rate variability. *Am J Physiol Regul Integr Comp Physiol*, 283, R789-R797.
- [10] Hu, X., Ebenli, S., Jia, Z., & Egardt, B. (2014). Enhanced sample entropy-based health management of li-ion battery for electrified vehicles. *Energy*, 64, 953-960.

- [11] Zurek, S., Guzik, P., Pawlak, S., Kosmider, M., & Piskorski, J. (2012). On the relation between correlation dimension, approximate entropy and sample entropy parameters, and a fast algorithm for their calculation. *Physica A*, 391, 6601-6610.
- [12] Hardle, W., & Simar, L. (2003). *Applied Multivariate Statistical Analysis*. Berlin: Springer Verlag.
- [13] Zurek, S. (2018, May 14). NCM-algorithm n.d. Retrieved from <https://github.com/sebzur/NCM-algorithm>.