35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Assessing Sample Entropy of Physiological Signals by the Norm Component Matrix Algorithm: Application on Muscular Signals during Isometric Contraction

Paolo Castiglioni, Sebastian Żurek, Jaroslaw Piskorski, Marcin Kośmider, Przemyslaw Guzik, Emiliano Cè, Susanna Rampichini, and Giampiero Merati

Abstract— **Sample Entropy (SampEn) is a popular method for assessing the unpredictability of biological signals. Its calculation requires to preliminarily set the tolerance threshold *r* and the embedding dimension *m*. Even if most studies select *m*=2 and *r*=0.2 times the signal standard deviation, this choice is somewhat arbitrary. Effects of different *r* and *m* values on SampEn have been rarely assessed, because of the high computational burden of this task. Recently, however, a fast algorithm for estimating correlation sums (Norm Component Matrix, NCM) has been proposed that allows calculating SampEn quickly over wide ranges of *r* and *m*.**

The aim of our work is to describe the structure of SampEn of physiological signals with different complex dynamics as a function of *m* and *r* and in relation to the correlation sum. In particular, we investigate whether the criterion of "maximum entropy" for selecting *r* previously proposed for Approximate Entropy, also applies to SampEn; and whether information from correlation sums provides indications for the choice of *r* and *m*. For this aim we applied the NCM algorithm on electromyographic and mechanomyographic signals during isometric muscle contraction, estimating SampEn over wide ranges of *r* (0.01≤ *r*≤ 5) and *m* (from 1 to 11).

Results indicate that the "maximum entropy" criterion to select *r* in Approximate Entropy cannot be applied to SampEn. However, the analysis of correlation sums alternatively suggests to choose *r* that at any *m* maximizes the number of "escaping vectors", i.e., data points effectively contributing to the SampEn estimation.

I. Introduction

Entropy measures the degree of unpredictability of a time series. For physiological signals, frequently of limited duration, entropy is often assessed by the "Approximate Entropy", *ApEn* [1], or "Sample Entropy", *SampEn*, [2] estimators. *ApEn* and *SampEn* construct segments of *m* consecutive samples and represent each segment as a point in a space of *m* dimensions. When some of these points are grouped together, i.e., when they fall within a sphere of sufficiently small radius *d*, then the data segments are considered to be similar to each other. The distance *d* is commonly expressed as the fraction *r* of the standard deviation *SD* of the time series, being $d = r \times SD$. *ApEn* and *SampEn* calculate the conditional probability that segments similar in the *m* dimensional space remain similar when the number of dimensions increase by one (in other words, when the length of the data segments increases to *m+1*). Entropy is then estimated as the negative natural logarithm of this probability. *ApEn* and *SampEn* differ in the way the number of similar segments is calculated. It has been shown that the estimation bias is substantially lower for *SampEn* when the length *N* of the time series is relatively short, or the threshold *r* is particularly small [2].

The values of *r* and *m* should be set before estimating *ApEn* or *SampEn*. When *ApEn* was originally proposed, synthesized deterministic and stochastic signals suggested to choose *r* between 0.10 and 0.25 and to set *m*=2 [3, 4]. This indication has been generally accepted and most entropy studies of real physiological signals calculated *ApEn* and *SampEn* with *m*=2 and *r*=0.20. However, this choice has been recently criticized. Regarding *ApEn*, an arbitrary choice of *r* may lead to contradictory results even if the selection remains within the 0.10-0.25 range [8]. Some authors suggested that the choice depends on the signal dynamics, the optimal *r* being larger for signals with faster dynamics, like neural signals, than with slower dynamics, like heart rate [5-7]. They proposed to choose the *r* value which maximizes the *ApEn* estimate, an approach that removes ambiguities in assessing complexity of synthesised processes [6]. However, this approach requires the estimation of a whole profile of *ApEn* as a function of *r*. This is often a very time consuming calculation, which limits the analysis of *ApEn(r)* to very few *m* values. Moreover, it is unclear whether this approach can be extended to the selection of *r* in *SampEn*, because a relative maximum in *SampEn(r)* may not be present even if it appears in *ApEn(r)* [5]. As to *SampEn*, its sensitivity to *r* turned out to be a major problem in a specific application: the detection of atrial fibrillation in very short heart rate series. In this case information on signals unpredictability was alternatively obtained from the quadratic sample entropy, index derived from *SampEn* but which depends less than *SampEn* on the choice of *r* [9].

Until now detailed analysis of the features of *ApEn(r)* and *SampEn(r)* as a function of *m* have been limited by the required high computational burden. Recently, however, some of us presented a very fast algorithm for evaluating

P. Castiglioni is with Don C.Gnocchi Foundation, via Capecelatro 66, 20148 Milan, Italy (e-mail: pcastiglioni@ dongnocchi.it).

S Żurek, J. Piskorski and M. Kośmider are with Institute of Physics, University of Zielona Góra, Zielona Góra, Poland. P. Guzik is with Dept.of Cardiology - Intensive Therapy and Internal Diseases, Poznan University of Medical Sciences, Poznań, Poland.

E. Cè, S. Rampichini and G. Merati are with Dept. of Biomedical Sciences for Health, University of Milan, Italy.

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

the correlation sum, the Norm Component Matrix (NCM) algorithm [10]. By exploring the relationships between correlation sums and entropy, NCM also allows fast estimations of *ApEn* and *SampEn* over a wide range of *m* and *r* values.

The aim of the present study is therefore to describe in details the *SampEn* structure as a function of *r* and *m* for physiological signals with different complex dynamics. In particular, we will explore the relations between correlation sums and *SampEn*; we will verify whether the "maximum entropy" criterion for selecting *r* can be also adopted for *SampEn*; and whether information from correlation sums provides indications for selecting *r* at any given *m*. This will be done by applying the new NCM algorithm on electromyographic and mechanomyographic signals in volunteers during isometric muscle contraction, and by comparing correlation sums and *SampEn* over a wide range of *r* and *m* values.

II. Methods

A. SampEn Analysis by the NCM Algorithm

Given a signal of *N* samples, $\{x_i\}$ with $i=1,...N$, let's call $\{u_i\}$ the time series with zero mean and unit variance derived from $\{x_i\}$. Fixed a time lag τ, the vectors

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

describe data segments of *m* samples. There are

$$L_m = N-(m-1)\tau \quad (2)$$

of such vectors. The *maximum coordinate difference* distance *d* between two vectors, $|\vec{v}_{m,\tau}(i) - \vec{v}_{m,\tau}(j)|$, is [2]:

$$d = \max_{k=1,2,\ldots,m} (|u(i+(k-1)\tau) - u(j+(k-1)\tau)|) \quad (3).$$

The vector $\vec{v}_{m,\tau}(i)$ is similar to $\vec{v}_{m,\tau}(j)$ if *d* is lower than a predefined threshold *r*. Excluding $\vec{v}_{m,\tau}(i)$ itself, the number of vectors similar to $\vec{v}_{m,\tau}(i)$ is

$$S_i^m(r) = \sum_{j=1, j\neq i}^{L_m} \Theta(r-d) \quad (4)$$

with Θ the *Heaviside* function:

$$\Theta(x) = \begin{cases} 1 & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases} \quad (5).$$

$S_i^m(r)$ can be expressed as fraction of the number of all the vectors of size *m* ($\vec{v}_{m,\tau}(i)$ excluded):

$$C_i^m(r) = (L_m - 1)^{-1} S_i^m(r) \quad (6).$$

The correlation sum is the average of $C_i^m(r)$ over all the vectors of size *m*:

$$C^m(r) = L_m^{-1} \sum_{i=1}^{L_m} C_i^m(r) \quad (7).$$

The NCM algorithm is the fastest method for calculating the correlation sum. It builds the look-up table

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

of size $(N - \tau - 1) \times [(N-1)/\tau - 1]$. The symmetry of this matrix and the cumulative character of (7) speed up the calculations. The computation of $C^m(r)$ between r_{min} and r_{max} is replaced by an arithmetic comparison of the norms (3) with *r* in (4), allowing a very dense *r* sampling. Details can be found in [10]. *SampEn* is derived from the correlation sum:

$$SampEn(m,r) = -ln\ (C^{m+1}(r)/C^m(r)) \quad (9).$$

We set τ=1, as usually done in *SampEn* estimations. Moreover, we average $S_i^m(r)$ in (4) over all the L_m vectors:

$$S^m(r) = L_m^{-1} \sum_{i=1}^{L_m} S_i^m(r) \quad (10).$$

In this way we derive the average number of "escaping" vectors *E*, i.e., the number of vectors that escape the neighborhood *r* when the segment length rises from *m* to *m*+1, as

$$E(m,r) = S^m(r) - S^{m+1}(r) \quad (11).$$

Substituting (6) in (11) we obtain:

$$E(m,r) = C^m(r)(L_m - 1) - C^{m+1}(r)(L_{m+1} - 1) \quad (12).$$

From (2), when *N* is much greater than $m \times \tau$,

$$E(m,r) \cong N \times (C_m(r) - C_{m+1}(r)) \quad (13).$$

Eq (13) indicates that the number of escaping vectors from one embedding dimension to the next one is proportional to the difference in the correlation sums of the two embedding dimensions.

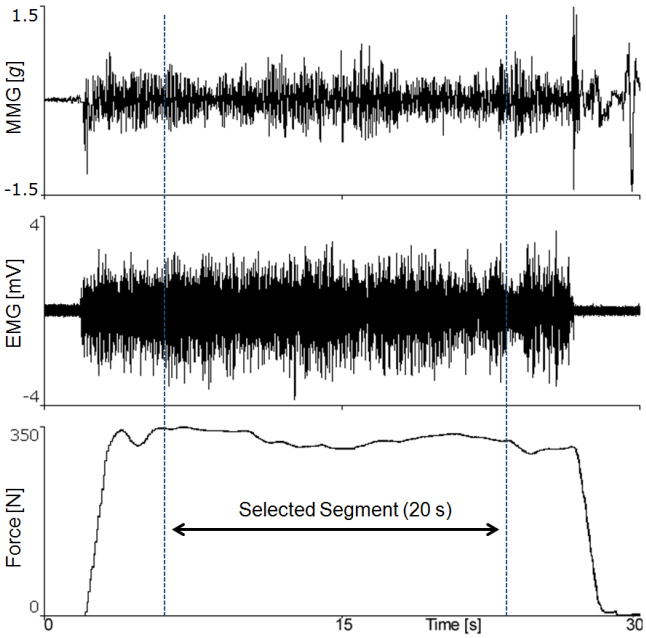

Figure 1. Example of data selection for entropy analysis. From top to bottom: mechanomyogram (MMG), electromyogram (EMG) and force during isometric contraction of ~25 s at 80% of maximal contraction force. A 20 s period of stable contraction is selected for the analysis.

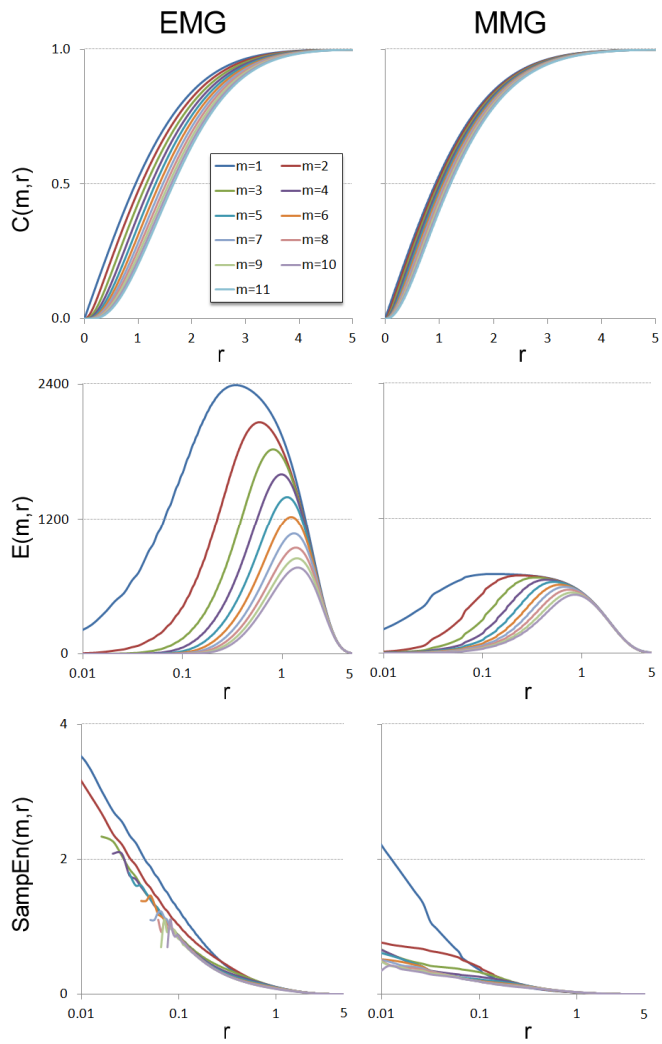

Figure 2. Entropy analysis for the EMG and MMG segments of figure 1. From top to bottom: correlation sum *C(m,r)*, function of the embedding dimension *m* and threshold *r*; number of "escaping vectors" *E(m,r)*; sample entropy *SampEn(m,r)*. Note the log scale for the horizontal axis of *E(m,r)* and *SampEn(m,r)*.

B. Data Collection and Analysis

Recordings were obtained in 8 male healthy volunteers. First, the force produced during an isometric maximal volitional contraction of the dominant arm was measured. Then, each subject performed an isometric contraction of the dominant arm for 25 s, with contraction force close to 80% of his maximal force. In this phase, the electromyogram (EMG) and mechanomyogram (MMG) were simultaneously measured with electrodes and accelerometers placed on the biceps of the dominant arm. MMG has a lower frequency content than EMG, and may reflect characteristics resonance frequencies of the muscle fibers. Recordings were sampled at 2 KHz. Segments of 20 s duration with stable contraction force were visually selected for *SampEn* analysis (see an example in figure 1). Correlation sums of EMG and MMG were estimated by the NCM algorithm for *r* between 0.01 and 5, and *m* between 1 and 11. *E(m,r)* and *SampEn(m,r)* were derived from the correlation sums.

III. Results

Figure 2 shows an example of *SampEn(m,r)* analysis in one subject. The correlation sums increase with *r* between 0 and 1, with lower values for larger *m*. This trend can be easily understood. When *r* is very small ($r \sim 0$), no neighbor points fall within the tolerance distance for almost all the segments and $C^m(r)$ is close to 0. When *r* is very large (e.g., r=5), the neighborhood of most vectors contains almost all the segments and $C^m(r)$ converges to 1. When *m* increases, the distance between couples of vectors may only increase or remain the same, explaining why $C^{m+1}(r) \leq C^m(r)$. EMG and MMG have similar $C^m(r)$ values at the lowest and highest *r*, but the two bundles of curves differ, being larger the dispersion for EMG. Differences also appear in *E(m,r)*, number of vectors escaping the neighborhood *r*. At the lowest *r*, *E(m,r)* is close to 0 because very few vectors have neighbors. Thus, very few vectors escape the neighborhood when *m* increases. *E(m,r)* is close to 0 also for the larger *r* because the tolerance is so wide that almost all the vectors remain similar when *m* increases. Therefore a low *E(m,r)* indicates that only a small fraction of the original *N* data effectively contribute to the estimation of *SampEn*. This happens when *r* is too small because vectors do not have neighbors, when *r* is too large because only vectors with the wider increases in *d* reflect the signal irregularity. The *E(m,r)* maximum, E_{MAX}, identifies univocally a tolerance *r*, hereafter indicated r_{MAX}, as a tradeoff between the number of vectors with neighbors, and sensitivity to detecting unpredictability.

In figure 2, r_{MAX} increases with *m* for both EMG and MMG. *SampEn* profiles do not show any clear maximum in *r* at any embedding dimension *m*. This is in contrast with *ApEn*, where maxima were repeatedly reported for a large variety of signals with very different dynamics [5-8]. However, differences between EMG and MMG appear also in *SampEn*. In particular, EMG estimates are "noisy" or not computable at the lower *r* when $m \geq 3$, probably reflecting the low number of points effectively contributing to the estimate.

TABLE I. Maximum of *E(m,r)*, E_{MAX}, and Corresponding Threshold *r*, r_{MAX}: Average on the Whole Group (N=8)

	EMG		MMG	
m	r_{MAX}	E_{MAX}	r_{MAX}	E_{MAX}
1	0.47	3231	0.26	906
2	0.72	2440	0.32	868
3	0.92	2046	0.40	826
4	1.07	1758	0.49	782
5	1.19	1525	0.61	733
6	1.29	1343	0.68	704
7	1.37	1197	0.74	675
8	1.43	1073	0.79	648
9	1.49	979	0.89	599
10	1.55	895	0.93	574

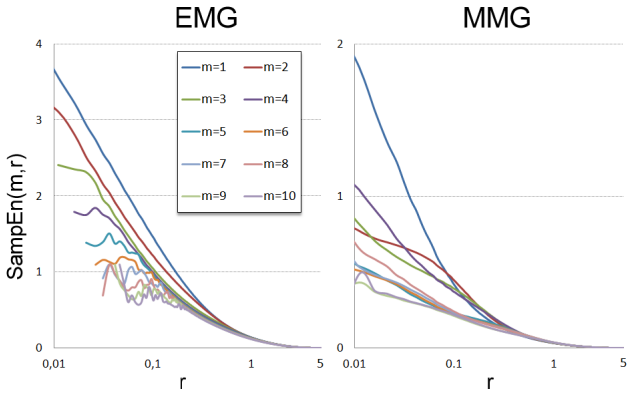

Figure 3. *SampEn* as function of embedding dimension *m* and threshold *r*: average over the group of 8 subjects.

Table I reports values of E_{MAX} and of the corresponding tolerance thresholds, r_{MAX}, over the whole group. For both signals E_{MAX} decreases and r_{MAX} increases with *m*. Moreover, also differences between these EMG and MMG parameters decrease with *m*. In particular, E_{MAX} and r_{MAX} of MMG are 28% and 55% the corresponding EMG values at *m*=1, 64% and 60% the EMG values at *m*=10.

No relative maxima appear in *SampEn* profiles averaged over the group (figure 3). EMG estimates look noisy when $m \geq 5$ and $r < 0.1$. The *SampEn* estimate of MMG at *m*=2 crosses estimates at contiguous embedding dimensions (*m*=1, 3 and 4) when $r \leq 0.2$, a behavior suggesting inconsistency of the estimator under certain conditions.

The *E(m,r)* profile naturally indicates in r_{MAX} a preferential *r* value. This approach univocally selects *r* taking into account its dependence on *m* and the intrinsic dynamics of the time series. *SampEn* calculated following this approach is shown in figure 4, as mean (SD) over the group. Values appear stable even at large *m*, probably because the choice $r=r_{MAX}$ avoids *r* thresholds associated to noisy estimates. This result is not granted if *r* is selected within the recommended 0.10-0.25 range, as figure 3 suggested. Interestingly, the profile of *SampEn* as function of *m* indicates higher unpredictability of EMG compared to MMG at the lower embedding dimensions only. This could be related to the dampening of faster mechanical responses to EMG potentials due to muscle inertia, and/or to resonance phenomena in the muscular fibers. Both these mechanisms might have affected the MMG irregularity as observed in low-dimensional embedding spaces only.

IV. Conclusion

The availability of a fast algorithm for calculating correlation sums allows to describe in details the structure of *SampEn* of physiological time series. As to *SampEn* of EMG and MMG, we found high sensitivity to *r*, possible instability with *m*, and that the criterion of "maximum entropy" proposed for selecting *r* in *ApEn* cannot be applied. However, the analysis of the profiles of correlation

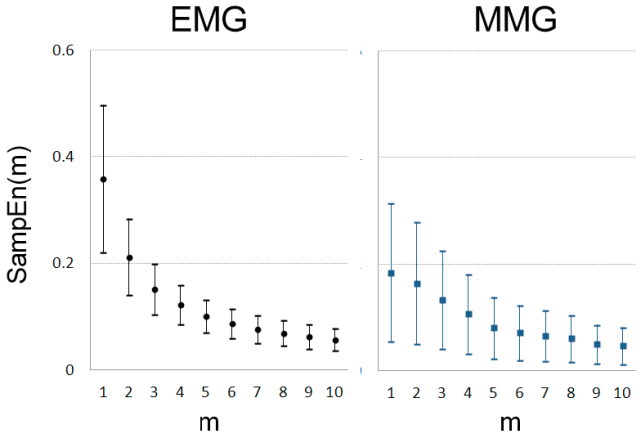

Figure 4. *SampEn* estimated at r_{MAX} for *m* between 1 and 10: average ±SD over the group of 8 subjects.

sums suggests an alternative criterion for choosing *r* at any embedding dimension *m*: the *r* value maximizing the number of "escaping vectors" *E*. This approach appears promising and deserves further studies to better understand its theoretical and physiological meanings.

Acknowledgment

S.Ż. is scholar within Sub-measure 8.2.2 Regional Innovation Strategies, Measure 8.2 Transfer of Knowledge, Priority VIII Regional Human Resources for the Economy, Human Capital Operational Programme co-financed by the European Social Fund and state budget.

References

[1] Pincus S.M, Goldberger A.L. Physiological time-series analysis: what does regularity quantify? *Am J Physiol Heart Circ Physiol* 1994; 266: H1643-56.

[2] Richman J.S, Moorman J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol* 2000;278:H2039–49.

[3] Pincus S.M, Huang W.M. Approximate entropy : statistical properties and application. *Commun Statist-Theory Meth* 1992;21:3061-3077.

[4] Pincus S.M. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci.* USA 1991;88:2297-2301.

[5] X. Chen, I. C. Solomon, and K. H. Chon. Comparison of the Use of Approximate Entropy and Sample Entropy: Applications to Neural Respiratory Signal. *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 2005: 4212-5

[6] K. Chon, Scully C., Lu S. "Approximate entropy for all signals," *Engineering in Medicine and Biology Magazine, IEEE*, 2009, 28:6, 18-23

[7] Lu S, Chen X, Kanters JK, Solomon IC, Chon KH. Automatic selection of the threshold value R for approximate entropy. *IEEE Trans Biomed Eng*. 2008;55:1966-72.

[8] Castiglioni P, Di Rienzo M. How the threshold "r" influences approximate entropy analysis of heart-rate variability. *Computers in Cardiology* 2008;35:561-4.

[9] Lake DE, Moorman JR. Accurate estimation of entropy in very short physiological time series: the problem of atrial fibrillation detection in implanted ventricular devices. *Am J Physiol Heart Circ Physiol.* 2011; 300: H319-25.

[10] Żurek S, Guzik P, Pawlak S, Kośmider M, Piskorski J. On the relation between correlation dimension, approximate entropy and sample entropy parameters, and a fast algorithm for their calculation, *Physica A*, 2012; 391: (24) 6601-10.