Bubble Entropy of fractional Gaussian noise and fractional Brownian motion

George Manis[1], Matteo Bodini[2], Massimo W Rivolta[2], Roberto Sassi[2]

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

Abstract

Aims: Bubble Entropy (bEn) is a metric which links the complexity of the series to the cost of sorting its samples, with limited dependence on parameters. Fractional Brownian motion (fBm) is a long-memory process, which has largely been used in modeling heart rate variability (HRV). fBm displays ephemeral regularities and periodicity at multiple time scales, which then vanish to reform differently. In here we tested if the continuously growing or decaying trends in fBm, which hint a broad range of swaps necessary for sorting, lead to maximal values of bEn.

Methods: We synthetically generated realizations of fBm (10^6 samples), along with its increments, the fractional Gaussian noise (fGn), a time-discrete process. The Hurst exponent H, on which fBm and fGn are parameterized, was varied in the entire range $(0, 1)$. bEn was computed with m ranging up to 200 (typically beyond the scope of other entropy metrics).

Results: For fGn, a stationary process, bEn showed a very small, if minimal, dependence on m. Empirically, it scaled as $H/2 + 3/4$. At low values of m, the dependence was more significant for fBm, a non-stationary process. When m grew, bEn approached a constant value.

Conclusions: bEn behaves like a scaling estimator for stationary Gaussian long-memory processes, but less so when non-stationarity becomes relevant (as it is for HRV).

1. Introduction

Bubble Entropy [1] is a metrics introduced to quantify the complexity of a time series by measuring the increase in the entropy of the series of sorting steps (swaps), necessary to order its portions of length m, when adding an extra element. It does not quantify directly the (differential) entropy rate of the series, like Approximate Entropy ($ApEn$) [2] and Sample Entropy ($SampEn$) [3] do. A clear advantage is its limited dependence on parameters, which are often critical to set. First, the time series $x_1, \ldots, x_N$ is embedded into an m dimensional space and then, for each of the $N - m + 1$ embedded vectors, the number of swaps Bubble Sort requires to sort them is assessed (in ascending order, but the result is invariant to the ordering selected). The second-order Rényi entropy of the series of swaps is computed as

$$H^m_{swaps} = -\log \sum_{i=0}^{\binom{m}{2}} p_i^2,$$

where p_i is the probability mass function (pmf) of having i swaps, estimated from the histogram of the counts. The maximum values of H^m_{swaps} appears when p_i is a uniform distribution, that is when all the possible ordering of the samples are equally likely. While there are $m!$ permutations for a sequence of m samples, we need a maximum of $m(m-1)/2$ swaps to generate any sequence from a given one. Thus, the maximum swap entropy is

$$U^m_{\text{swaps}} = \log \left[\frac{(m-1)m}{2} + 1 \right].$$

While other common entry measures, like Permutation Entropy [4], maximize when the input signal is generated by a white noise, this is not the case here. In fact, we recently derived [5] an exact formula for the swap entropy of a white Gaussian noise (WGN) and proved that when m is large

$$W^m_{\text{swaps}} \approx \frac{1}{2} \log \left[\pi \frac{m(m-1)(2m+5)}{18} \right] < U^m_{\text{swaps}}.$$

This was clearly exemplified in [6], where signal produced by autoregressive (AR) models with a large and positive one-step autocorrelation required a broader range of swaps than WGN. This is not surprising, as with bEn, the *complexity* of a time series is measured is term of added diversity in the ordering of the samples across scales, and not as lack of similar patterns (which would favor WGN).

Finally, bEn is computed as the increase (entropy rate) in swap entropy, when an extra element is added to each of the vectors, normalized with respect to bEn of uniform pmfs

$$bEn = \frac{H^{m+1}_{swaps} - H^m_{swaps}}{U^{m+1}_{swaps} - U^m_{swaps}}.$$

As a (slightly) alternative definition, in [5] we proposed to change the normalization factor to

$$bEn^* = \frac{H_{swaps}^{m+1} - H_{swaps}^{m}}{W_{swaps}^{m+1} - W_{swaps}^{m}},$$

so that a value $bEn^* = 1$ corresponds always to a WGN.

Heart Rate Variability (HRV), as well as series generated from long-memory processes, displays ephemeral regularities and periodicity at multiple time scales, which then vanishes to reform differently. The persistence between subsequent values (*e.g.*, continuously growing or decaying trends) and the self-similarity across different scales suggest that a broad range of sorting swaps might be required to bubble sort the m dimensional vectors embedded from these series. As a consequence, fractal processes should display a value of bEn^* larger than one. In this paper, we investigated this hypothesis and assessed the values of bEn and bEn^* for synthetic series generated by Guassian process displaying long memory.

2. Methods

As shown in [5, 6], estimates of bubble entropy for series derived from an AR process grow while the process approaches a Gaussian random walk, which is the limiting case of $x[n] = -a_1 x[n-1] + w[n]$ for $a_1 \to 1$, with $w[n] \sim \mathcal{N}(0, \sigma^2)$. A random walk is a time-discrete process, which weakly converges to a (time-continuous) Brownian motion, when the length of the sequence tends to infinity (Donsker's theorem, see [7]). More generally, fractional Brownian motion (fBM) is a non-stationary time-continuous long-memory process, displaying self-similarity and a slope of the (generalized) spectral density in the low frequencies scaling as $1/f^{\alpha}$. fBm, along with its increments, *i.e.*, the fractional Gaussian noise (fGn), a time-discrete stationary Gaussian process [8], have largely been used in modelling heart rate variability series. Both fBm and fGn are parametrized by a scalar parameter, H, the Hurst exponent, which specifies the extent of the correlations. For $H = 1/2$, fGn becomes a WGN and fBm a Brownian motion. For $H > 1/2$ fGn displays long memory (autocorrelation decaying at a polynomial rate), as well as fBn for any value of H.

We generated synthetic fGn and (sampled) fBm series using the algorithms proposed by [9] and [10], respectively, by varying the value of the Hurst parameter between 0 and 1 (both excluded), in steps of 0.025. For each value of H, 1000 series of 10^6 samples were produced and for each we computed the values of bEn and bEn^* for several values of m, from 3 to 200.

Given the computational load, due to the length of the series, we implemented a very fast algorithm which was $O(N)$. In practice (please also check [11]), once ordered the first vector, composed of the samples $x_1 \cdots x_m$, the first element x_1 was removed from the ordered sequence (decreasing the number of swaps) and the new element x_{m+1} added in the sequence in its ordered position (incrementing accordingly the number of swaps). Then the process was repeated for any remaining sample x_i. While requiring a bit more of memory and bookkeeping, the computational time was significantly reduced, in particular for large value of m (at $m = 200$, $\approx 330\times$ speedup on an Intel Core i7-7500U CPU).

3. Results

The results are reported in fig. 1. For clarity the horizontal axis is given in terms of α, the slope of the power spectral density for $f \to 0$ (or generalized power spectral density for fBm, which is a non stationary process) in a log-log plot, linked to H by the two scaling relations $\alpha = 2H - 1$ (fGn, red axis) and $\alpha = 2H + 1$ (fBm, blue axis). Empirically, we verified, that while getting to the same values, a faster convergence was obtained averaging, over the Monte Carlo runs, the estimates of the pmf p_i instead of H_{swaps}^{m} (rare events more likely appear in the final pmf). These are the values reported in the figure.

For small value of m, bEn increased with growing values of α up to about $\alpha = 2$, which corresponds to a random walk. Then it decreased. For large m, and an increasingly larger range of α values around 2, it reached a plateau where bEn saturated to a value close to 1. In these circumstances, with the augmenting long-range correlation, H_{swaps}^{m} tends to U_{swaps}^{m} (the swaps pmf tends to uniform). Then, the pmf starts displaying peaks of probability at the two extremes (no swaps and $m(m-1)/2$ swaps) and bEn decays after growing over 1. These is due to the fact that when the memory of the process is large, growing and decaying trends of length m are likely in the signal.

These conclusions can be reached more clearly looking at bEn^* in the the top panel of the same figure. For fGn, which is a stationary process, except when $m = 200$ and $\alpha < -0.5$, the values of bEn^* are well described by the line $bEn^\star \approx H/2 + 3/4$ (also shown) and bubble entropy behaves like a scaling exponent estimator. When the process is non-stationary (fBm), for small values of H a linear growth with about the same slope can be still observed. Then, close to $\alpha = 2$,

$$\begin{aligned} bEn^* &\approx \lim_{m\to\infty} \frac{U_{\text{swaps}}^{m+1} - U_{\text{swaps}}^{m}}{W_{\text{swaps}}^{m+1} - W_{\text{swaps}}^{m}} \\ &= \lim_{m\to\infty} \frac{\log\left[\frac{m(m+1)+2}{(m-1)m+2}\right]}{\frac{1}{2}\log\left[\frac{(m+1)(2m+7)}{(m-1)(2m+5)}\right]} = \frac{4}{3}, \end{aligned}$$

which supports the idea that the pmf tend to uniform.

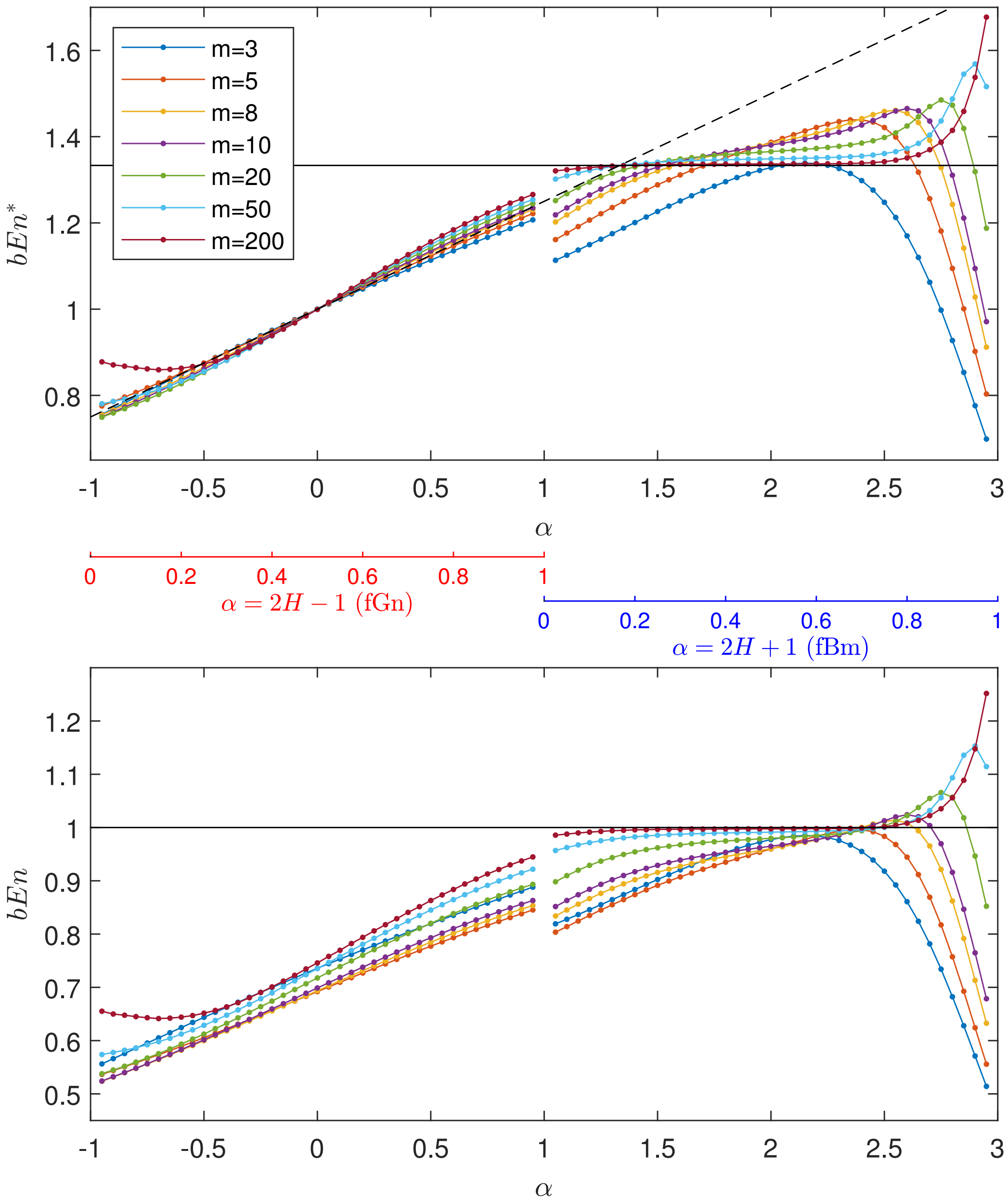

Figure 1. Values of bEn^* (top) and bEn (bottom) for synthetic series obtained from a fGn (left, red axis) and fBm (rigth, blue axis) as a function of α, the slope of the (generalized) power spectral density in the low frequencies. The slope is linked by different scaling algebraic relations to the Hurst exponents H, as specified in the axes. Each dot is the average of 1000 Monte Carlo simulations on sequences of 10^6 samples. Values of bEn^* for fGn are well approximated by the dashed line $bEn^* = H/2 + 3/4$. For large values of m, likely due to the non-stationary nature of fBm, the swaps' pmf tends to uniform for a large range of values around $H = 0.5$ or $\alpha = 2$ (which corresponds to the random walk) and the values of Bubble Entropy approach the limit set for a uniform distribution, the horizontal line at the level $bEn^* = 4/3$ or $bEn = 1$.

In the context of HRV series obtained from Holter 24h recordings, the slope α is usually found to be in the range $0.9 - 1.2$ for normal subjects, > 1.33 for congestive heart failure (CHF) patients and > 1.5 for subjects who underwent a myocardial infarction [12]. In this entire range, on the synthetic fGn and fBm signals, bEn and bEn^* displayed to grow linearly with α.

4. Discussion

In this work, we studied, experimentally, the behaviour of bubble entropy for long-memory Gaussian processes. We verified than when the extent of the correlations increases, the number of swaps necessary to sort sequences of length m tend to be uniformly distributed. As a consequence, bubble entropy is larger for fBm and fGn, than for uncorrelated white noises. This is coherent with the empirical understanding that in series displaying ephemeral regularities and periodicity at multiple time scales, as those produced by long-memory processes, always growing and decaying portions have a finite probability, which tend to be of the same magnitude of any other ordering.

For fGn and all the m values considered, bubble entropy scaled linearly with H (and α), thus behaving like a scaling exponent estimator. Similarly happened for fBm and some values of m. Many other estimators do exists, and one of the most common in the context of HRV is the long-term detrended fluctuation analysis DFAα_2 exponent [13]. In this respect, we can reconsider figure 6 of [5], which compared the discriminative capabilities of bubble entropy and DFA between long term HRV of normal subjects and CHF patients. bEn was always significantly different between the two groups for $m \geq 11$, but DFAα_2 (estimated for lags ≥ 11) was not. Thus, the differences in the ordering of the samples seem to tell something more than the aggregated scaling exponent.

Interestingly, the simulations we performed explored a much larger range of scales m, than what possible with techniques like permutation entropy (and even more for sample entropy). In fact, while there are $m!$ permutations for a sequence of m samples, we need a maximum of $m(m-1)/2$ swaps (a much smaller number) to generate any sequence from a given one. As a concluding remark, the number of swaps is related to the length of the shortest program which is needed to generate any sequence. Thus, the swaps pmf is also the pmf of the length of the codes required to produce any m sequence contained in the series. There might be interesting connections between bEn with the description length theory, which might be worthwhile of further analysis.

References

[1] Manis G, Aktaruzzaman M, Sassi R. Bubble entropy: An entropy almost free of parameters. IEEE Trans Biomed Eng 2017;64:2711–2718.

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

[3] 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:R789–R797.

[4] Bandt C, Pompe B. Permutation entropy: A natural complexity measure for time series. Phys Rev Lett 2002;88.

[5] Manis G, Bodini M, Rivolta MW, Sassi R. A two-steps-ahead estimator for bubble entropy. Entropy 2021;23(6).

[6] Bodini M, Rivolta MW, Manis G, Sassi R. Analytical formulation of bubble entropy for autoregressive processes. In 2020 11th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO). 2020; 1–2.

[7] Taqqu MS. Weak convergence to fractional Brownian motion and to the Rosenblatt process. Z Wahrscheinlichkeit 1975;31(4):287–302.

[8] Cerutti S, Esposti F, Ferrario M, Sassi R, Signorini MG. Long-term invariant parameters obtained from 24-h holter recordings: A comparison between different analysis techniques. Chaos 2007;17(1):015108.

[9] Paxson V. Fast approximation of self-similar network traffic. Technical report, LBL-36750/UC-405, 1995.

[10] Abry P, Sellan F. The wavelet-based synthesis for fractional brownian motion proposed by F. Sellan and Y. Meyer: Remarks and fast implementation. Appl Comput Harmon Anal 1996;3(4):377–383.

[11] Manis G, Sassi R. A Python library with fast algorithms for popular entropy definitions. In Proc. of Computing in Cardiology. 2021; (in press).

[12] Sassi R, Cerutti S, Lombardi F, Malik M, Huikuri HV, Peng CK, Schmidt G, Yamamoto Y, Reviewers: D, Gorenek B, Lip GY, Grassi G, Kudaiberdieva G, Fisher JP, Zabel M, Macfadyen R. Advances in heart rate variability signal analysis: joint position statement by the e-Cardiology ESC Working Group and the EHRA co-endorsed by the APHRS. EP Europace 2015;17(9):1341–1353.

[13] Peng CK, Havlin S, Stanley HE, Goldberger AL. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 1995;5(1):82–87.

Address for correspondence:

Roberto Sassi
Dipartimento di Informatica
Università degli Studi di Milano
via Celoria 18, 20133, Milano (MI) Italy
email: roberto.sassi@unimi.it