A New Heuristic Method for the Optimal Selection of Tolerance r for Entropy Indices

Dominique Makowski[1]

[1] School of Social Sciences, Nanyang Technological University, Singapore

Author Note

Correspondence concerning this article should be addressed to Dominique Makowski, HSS 04-18, 48 Nanyang Avenue, Singapore (dom.makowski@gmail.com).

Abstract

The tolerance threshold r is a key parameter of several entropy algorithms (e.g., *SampEn*). Unfortunately, the gold standard method to estimate its optimal value - i.e., the one that maximizes *ApEn* - is computationally costly, prompting users to rely on cargo-cult rules-of-thumb such as 0.2 * SD. This simulation study aims at validating a new heuristic, based on the embedding dimension m and the signal's length n (optimal r = SD * 0.281(m-1) + 0.005(log(n)) - 0.02(m-1 * log(n))), which was found to be superior to other existing heuristics. All the methods of optimal tolerance r estimation used in this study are available in the *NeuroKit2* Python software (Makowski et al., 2021).

Keywords: chaos, complexity, entropy, tolerance, r, physiology

Word count: 925

A New Heuristic Method for the Optimal Selection of Tolerance r for Entropy Indices

Introduction

Complexity analysis is an increasingly popular approach to physiological signals, including cardiac (e.g., Heart Rate Variability, [1]) and brain activity [2]. It is an umbrella term for the usage of various complexity indices that quantify concepts such as chaos, entropy, fractal dimension, randomness, predictability, and information. Importantly, some of the most popular indices of entropy (e.g., *ApEn*, *SampEn*, their fuzzy and multiscale variations) and recurrence quantification analysis (RQA), rely on a similar set of parameters. Namely, these are the delay τ, the embedding dimension m, and the tolerance r, which are critical to accurately capture the space in which complexity becomes quantifiable. Unfortunately, despite the existence of methods to estimate optimal values for these parameters depending on the signal at hand, their choice often relies on simple heuristics and cargo-cult conventions.

Such is the case of the tolerance threshold r, which typically corresponds to the minimal distance required to assign two points in a state-space as belonging to the same state. It directly impacts the amount of "recurrences" of a system and the measure of its tendency to revisit past states, which is the base metric for the calculation of the aforementioned entropy indices. Despite its importance, it is often selected as a function of the standard deviation (SD) of the signal, with (in)famous arbitrary values including 0.1 or $0.2 * SD$ [3]. One of the reason for the longevity of such an approach is 1) the past literature (as it is consistently used in the existing literature, the choice of the same values becomes the default and does not require much justification) and 2) the fact that other algorithms to estimate the optimal r are computationally costly.

The aim of the present study is to investigate the relationship between different methods for optimal tolerance r estimation. The ground-truth method used is the

tolerance value corresponding to a maximal value of Approximate Entropy - *ApEn* [4–6]. As this method is computationally expensive, the objective of this study is to assess whether fast heuristic proxies can be used to approximate $r_{maxApEn}$.

Methods

For $n = 5760$ combinations of different signal types and lengths, as well as noise types and intensities (the procedure used was the same as in [7], and the data generation code is available at **https://github.com/DominiqueMakowski/ComplexityTolerance**), we computed the Approximate Entropy (*ApEn*), which peak is used to estimate the optimal tolerance level for time-delay embedding spaces ranging from of 1 to 9 embedding dimensions m.

The aim of the analysis is to establish a new heuristic based only on the signal's SD and the embedding dimension m; and compare all of these approximations with other existing heuristics such as *0.2 SD*, *Chon* [6], and the *Schötzel* method $(1.3334 + 0.5627 * log(dimension))$ implemented in the package *nolds* [8].

Results

Maximum Approximate Entropy. **Figure 1** shows the normalized value of Approximate Entropy *ApEn* as a function of tolerance r and embedding dimension m. As expected, the value of *ApEn* peaks at certain values of r (hence its usage as an indicator of the optimal tolerance). The location of this peak seems strongly impacted by the embedding dimension m, getting more variable - and on average larger - as m increases.

New Heuristic. Selecting the tolerance based on the signal's SD alone does not appear as a good default, given the strong impact of embedding dimension. In order to validate a new heuristic to approximate the optimal r value based on the embedding dimension m and the signal's length, we compared different regression specifications using the BIC-based Bayes Factor test. The model which included the log-transformed signal

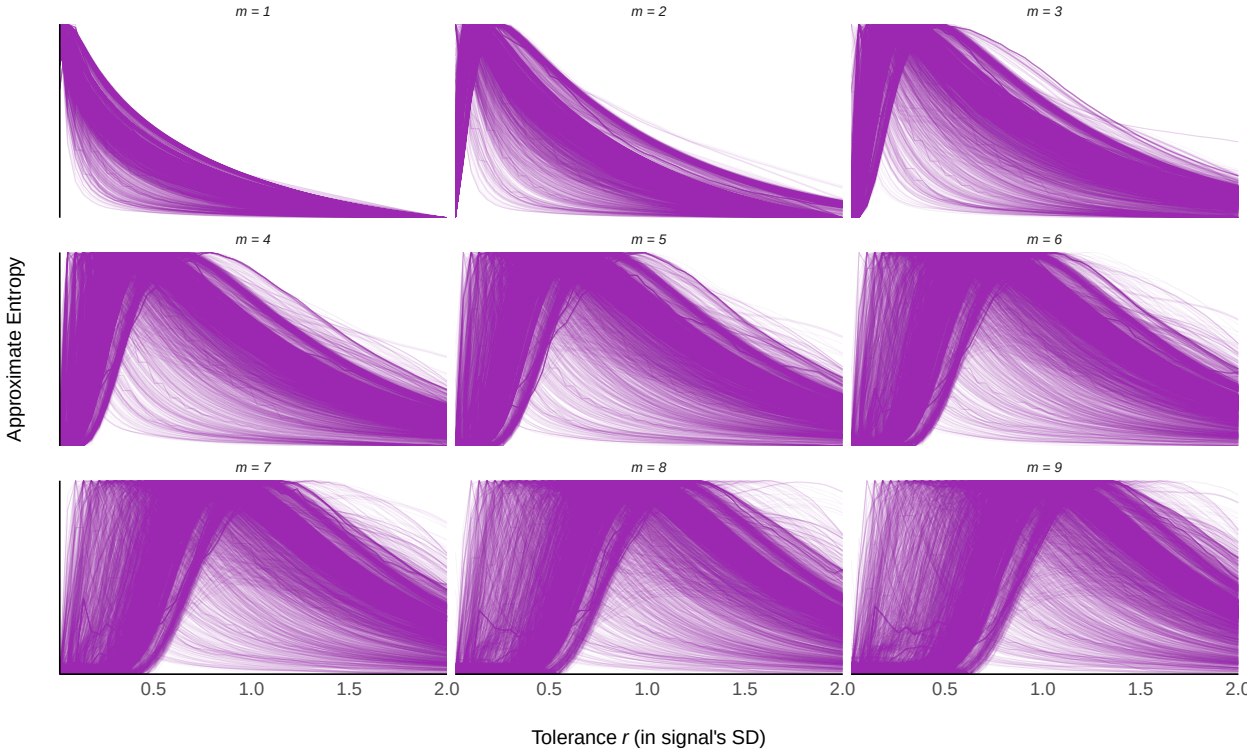

Figure 1. Approximate Entropy *ApEn* as a function of tolerance r (expressed in signal SD) and embedding dimension m.

length (in samples) and the embedding dimension m minus 1 (with no intercept) performed significantly better ($BF_{10} > 1000$) than any other model, with an explained variance of 92.64%. Based on this simple regression model, we can derive the following approximation (assuming a standardized signal with an SD of 1):

$$\widehat{r} = 0.2811(m - 1) + 0.0049(\log(n)) - 0.02(m - 1 \times \log(n)) \quad (1)$$

It should be noted that shorter signals require larger tolerance values, and the impact of length lowers as the signal gets longer. Also, for an embedding dimension of 2 (and short signal lengths), this equation returns values close to the *0.2 SD* heuristic, which is not entirely surprising as the latter was initially derived under such conditions (and that are common in some applications, such as heart-rate intervals).

Heuristics Comparison.

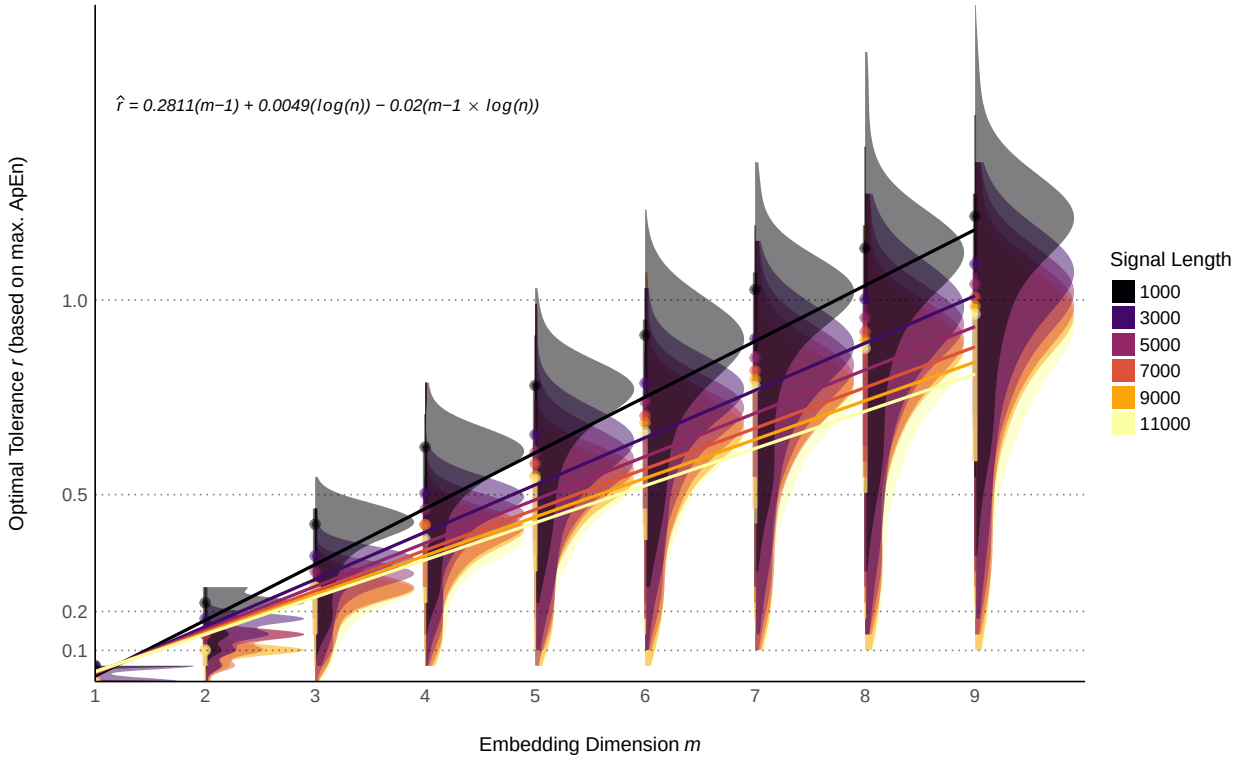

Figure 2. Optimal tolerance values approximated by a new heuristic model based on the embedding dimension m and the signal length n (in samples). The density plots show the true optimal tolernace values as based on max. ApEn.

Table 1

Comparison of Model Performance Indices

Model	BIC	AIC	R2
Makowski	-41375.73	-41402.30	0.77
Scholzel	-24637.03	-24663.60	0.68
Chon	24614.17	24587.60	0.18
SD	35119.90	35102.18	0.00

We compared together different methods to approximate $r_{maxApEn}$ (see **Table 1**) by comparing $r_{maxApEn}$ (our ground-truth) to the values estimated by different methods. The new heuristic method introduced in this study, based on the signal's SD, the embedding dimension and the log-transformed length, surpassed all other models ($BF_{10} > 1000$, $R^2 =$

0.77).

Discussion

The tolerance threshold r is a key parameter of several entropy algorithms, including widely popular ones like *SampEn*. The current gold standard method to estimate the optimal r is to compute Approximate Entropy (*ApEn*) over a range of different r values and to select the one corresponding to the maximum *ApEn* value. Unfortunately, this method is computationally costly.

In this study, we showed that a simple heuristic approximation based on the embedding dimension m and the log-transformed signal length is a valid approximation of $r_{maxApEn}$, showing superior performance to other heuristic methods. We recommend the use of this method as a default alternative to the *0.2 SD* rule of thumb.

While we believe that our data generation procedure was able to generate a wide variety of signals, and that our results are to some extent generalizable, future studies could attempt at refining the estimation procedures for specific signals (for instance, EEG, or heart rate data). All the methods of optimal tolerance r estimation used in this study, including our new proposal, are available in the *NeuroKit2* open-source Python software, as an option in the `complexity_tolerance()` function [9].

References

1. Pham, T.; Lau, Z.J.; Chen, S.; Makowski, D. Heart Rate Variability in Psychology: A Review of HRV Indices and an Analysis Tutorial. *Sensors* **2021**, *21*, 3998.
2. Lau, Z.J.; Pham, T.; Annabel, S.; Makowski, D. Brain Entropy, Fractal Dimensions and Predictability: A Review of Complexity Measures for EEG in Healthy and Neuropsychiatric Populations. **2021**.
3. Pincus, S.M.; Viscarello, R.R. Approximate Entropy: A Regularity Measure for Fetal Heart Rate Analysis. *Obstetrics and gynecology* **1992**, *79*, 249–255.
4. Chen, X.; Solomon, I.; Chon, K. Parameter Selection Criteria in Approximate Entropy and Sample Entropy with Application to Neural Respiratory Signals. *Am. J. Physiol. Regul. Integr. Comp. Physiol., to be published* **2008**.
5. Lu, S.; Chen, X.; Kanters, J.K.; Solomon, I.C.; Chon, K.H. Automatic Selection of the Threshold Value r for Approximate Entropy. *IEEE Transactions on Biomedical Engineering* **2008**, *55*, 1966–1972.
6. Chon, K.H.; Scully, C.G.; Lu, S. Approximate Entropy for All Signals. *IEEE engineering in medicine and biology magazine* **2009**, *28*, 18–23.
7. Makowski, D.; Te, A.S.; Pham, T.; Lau, Z.J.; Chen, S.H.A. The Structure of Chaos: An Empirical Comparison of Fractal Physiology Complexity Indices Using NeuroKit2. *Entropy* **2022**, *24*, 1036, doi:10.3390/e24081036.
8. Schölzel, C. *Nonlinear Measures for Dynamical Systems*; Zenodo, 2019;
9. Makowski, D.; Pham, T.; Lau, Z.J.; Brammer, J.C.; Lespinasse, F.; Pham, H.; Schölzel, C.; Chen, S. NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing. *Behavior research methods* **2021**, *53*, 1689–1696.