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 - the one that maximizes *ApEn* - is computationally costly, prompting users to rely to cargo-cult heuristics such as 0.2 * SD. In this study, we first compared the relationship bewteen the amount of Nearest Neighbours (*NN*) and the Recurrence Rate (*RR*), and showed that these values cannot be used to approximate the optimal r value. Secondly, we established a new heuristic, based only on the signal's SD and the embedding dimension m (optimal r = -0.032 + 0.1497 * m), which was 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, fractal, physiology, tolerance

Word count: 1156

Optimal Selection of Tolerance r for Entropy Indices

Introduction

Complexity analysis is a growing 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 the same subset 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 or cargo-cult conventions.

Such is the case of the tolerance threshold r, which typically corresponds to the minimal distance required to assigning two points in a state-space as belonging to the same state. It directly impacts the amount of "recurrences" of a system and 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 magic values including 0.1 or $0.2 * SD$ [3]. One of the reason for the longevity of such approach is 1) past literature (as many past studies used it, it becomes easier to justify the choice of the same values) and 2) the fact that other approaches 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 is to see whether any heuristic proxies can be used to satisfyingly approximate the ground-truth value.

Methods

For $n = 2880$ combinations of different signal types and lengths, as well as noise types and intensities (the procedure used was the same as in ..., and the data generation code is available at **https://github.com/DominiqueMakowski/ComplexityTolerance**), we will compute 3 different scores as a function of difference tolerance values (expressed in SDs of the signal): Approximate Entropy (*ApEn*), which peak is used to estimate the optimal tolerance level; the average number of nearest neighbours *NN*, which is the underlying quantity used by several entropy algorithms; and the recurrence rate *RR*, one of the core index of recurrence quantification analysis (RQA). These 3 scores are computed based on time-delay embedding spaces that we will create ranging from of 1 to 9 embedding dimensions m.

The goal of the analysis is to 1) investigate the possibility of using alternative scores, namely *RR* and *NN*, to approximate the location of the *ApEn* peak; 2) establish a new heuristic based on signal's SD and Dimension; and 3) compare all of these approximations with other existing heuristics such as *0.2 SD*, *Chon* [6], and the *Schötzel* method implemented in the package *nolds* [7].

Results

Descriptive Results. **Figure 1** shows the normalized value of Approximate Entropy *ApEn*, the amount of nearest neighbours *NN* and the Recurrence Rate *RR* as a function of tolerance r. 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 as m increases. Does this peak consistently correspond to certain values of *NN* and *RR*?

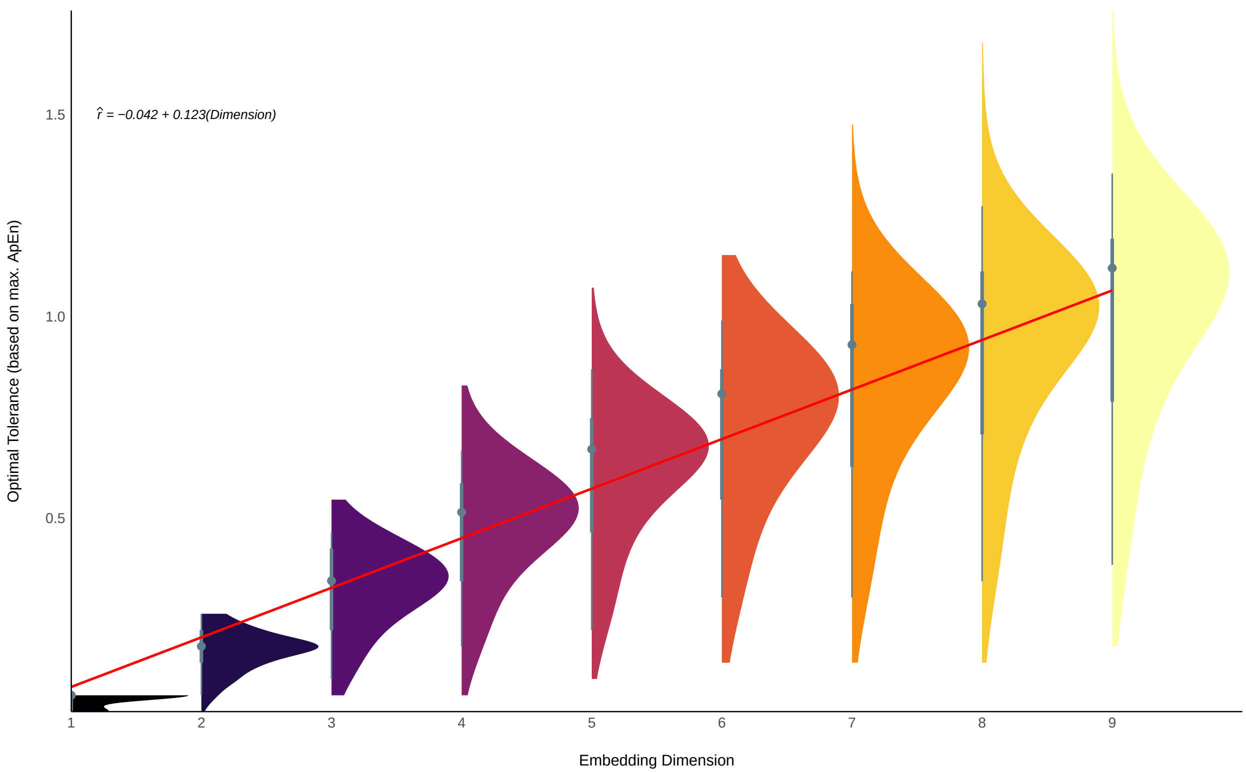
Using NN and RR. In order to assess whether the amount of nearest neighbours *NN* and the Recurrence Rate *RR* can be used to approximate the optimal tolerance threshold r (as estimated by *max. ApEn*), we fitted for each index 3 regression models to predict the index' value that corresponds to the location of *max. ApEn*: one without the embedding dimension m as predictor, one with it, and one with the dimension's logarithm. For both *NN* and *RR*, the model with the log-transform dimension as predictor was the best $BF_{10} > 1000$. However, *NN* did not share a strong relationship with the embedding dimension, as the explained variance of its (best-performing) model was low ($R^2 = 2.20\%$). It was higher for the model based on *RR* ($R^2 = 50.39\%$). The models were as follows:

$$\widehat{NN} = 0.0221 - 0.0107(\log(Dimension)) \tag{1}$$

$$\widehat{RR} = 0.0221 - 0.0107(\log(Dimension)) \tag{2}$$

According to these models, for an embedding dimension of 2, the target *NN* and *RR* values are 1.9% and 1.5%, respectively.

New Heuristic. Because computing *RR* or *NN* is also an expensive procedure, we also attempted at validating a new heuristic based only on the signal's *SD* and the embedding dimension m.



Selecting the tolerance based on the signal's SD alone makes hardly sense, as the embedding dimension has a strong impact on it. We fitted two models to predict the optimal tolerance (as estimated by *max. ApEn*), with the embedding dimension and its log-transformation as predictors, respectively. The model without the log-transformation performed significantly better ($BF_{10} > 1000$), with an explained variance of 77.38%. Based on this simple regression model, we can derive the following approximation (assuming a standardized signal with an SD of 1):

$$\widehat{maxApEn} = -0.0417 + 0.1228(Dimension) \tag{3}$$

Interestingly, for a dimension m of 2, this equation approximates the *0.2 SD* heuristic ($r = 0.204$), which actually was initially derived under this condition.

Heuristics Comparison. We compared together different approximations of $r_{maxApEn}$ (see **Table 1**). Our heuristic method presented in this study, based on the signal's SD and the embedding dimension, surpassed any other models ($BF_{10} > 1000$, $R^2 = 0.77$), followed by the *Schötzel* adjustment ($R^2 = 0.74$). The methods based on RR ($R^2 = 0.64$) and NN ($R^2 = 0.62$) were next, followed the *Chon* adjustment ($R^2 = 0.25$) and, finally, the fixed 0.2 SD value.

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 have shown that a simple heuristic approximation based on the signal's SD and the embedding dimension m was the best at approximating $r_{maxApEn}$, showing superior performance to procedures involving state-phase reconstruction related quantities, such as the amount of Nearest Neighbours (NN) and the Recurrence Rate (RR). We suggest 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 are available in the *NeuroKit2* open-source Python software [8].

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. Schölzel, C. *Nonlinear Measures for Dynamical Systems*; Zenodo, 2019;
8. 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.

Table 1

Comparison of Model Performance Indices

Model	BIC	R2	BF
Makowski	-17820.13	0.77	1.00
Scholzel	-14317.36	0.74	< 0.001
RR	-5691.58	0.64	< 0.001
NN	-4435.10	0.62	< 0.001
Chon	13318.91	0.25	< 0.001
SD	20694.14	0.00	< 0.001