An improved estimator of Shannon entropy with applications to systems with memory

Juan De Gregorio, David Sánchez, and Raúl Toral

IFISC (CSIC-UIB) Palma de Mallorca - Spain

juan@ifisc.uib-csic.es

AGENCIA ESTATAL DE INVESTIGACIÓN

UNIT OF EXCELLENCE MARÍA DE MAEZTU

Motivation: Given a sequence, how to determine its memory?

Memory m: minimum number of past states needed in order to faithfully determine the probability of a future state.

Random variable $X \longrightarrow L$ possible outcomes $z_1, \ldots, z_L$

$\downarrow$

Repeated N times $\longrightarrow$ sequence $S = X_1, \ldots, X_N$

S has memory m if:

$$P(X_{s+1} = z_j | X_1 = z_k, \ldots, X_s = z_l) = P(X_{s+1} = z_j | X_{s-m+1} = z_r, \ldots, X_s = z_l)$$

Method: study the block Shannon entropy

Group S in blocks of size $n \rightarrow L^n$ possible blocks $\{b_i^{(n)}\}_{1 \leq i \leq L^n}$

Block entropy: $H_n = -\sum_{i=1}^{L^n} p(b_i^{(n)}) \log(p(b_i^{(n)}))$

Theorem: H_n is a linear for $n \geq m \Leftrightarrow S$ has memory m

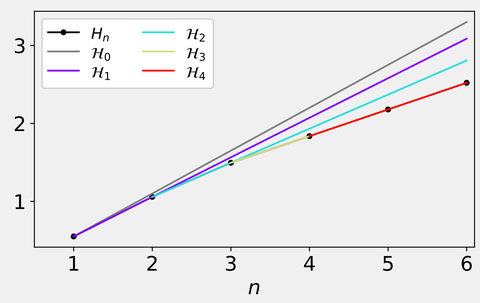
We can define:

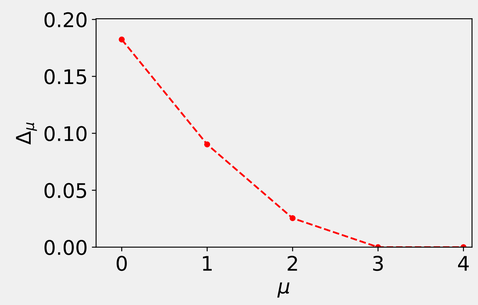
- μ: trial memory
- $\mathcal{H}_\mu(n)$: block entropy as if the system had memory μ

$$\begin{cases} \mathcal{H}_0(n) &= nH_1 \qquad n \geq 1 \\ \mathcal{H}_\mu(n) &= (H_{\mu+1} - H_\mu)(n - \mu) + H_\mu \quad n \geq \mu > 0. \end{cases}$$

- $\Delta_\mu = \dfrac{1}{n_{max} - \mu + 1} \sum_{n=\mu}^{n_{max}} (\mathcal{H}_\mu(n) - H_n)^2$

$$m = \min(\{\mu : \Delta_\mu = 0\})$$





Application of the method for a binary system with memory $m = 3$. $\Delta_\mu = 0$ if $\mu \geq 3$ which means that the system has memory 3 as we already knew.

A new entropy estimator

Problem: There is not known unbiased estimator of H

Our proposed estimator is an improvement of Chao-Shen's [1]:

- Horvitz-Thompson correction [2] to account for missing elements in S

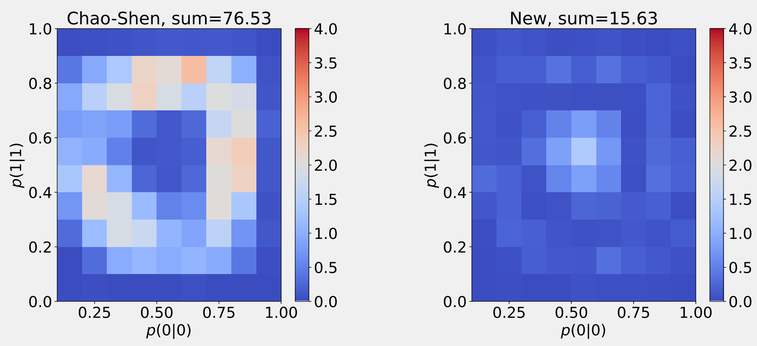
$$\hat{H}^{HT} = -\sum_{i \in S} \frac{p_i \log(p_i)}{P(i \in S)} \approx -\sum_{i \in S} \frac{p_i \log(p_i)}{1 - (1 - p_i)^N}$$

- Estimator for the p_i's$\rightarrow$ $\hat{p}_i^{MLE} = \dfrac{\#\text{ of occurrences of element } i}{N}$

- Since $\sum_{i \in S} p_i \leq 1$ but $\sum_{i \in S} \hat{p}_i^{MLE} = 1 \Rightarrow$ a correction to the estimated probabilities is needed: $\hat{p}_i = C\hat{p}_i^{MLE}$

- Correction that takes into account correlations:

$$C = 1 - \sum_{j=1}^{N/2} \frac{1}{N/2 + j} I(X_{N/2+j} \notin \{X_1, \ldots X_{N/2+j-1}\})$$

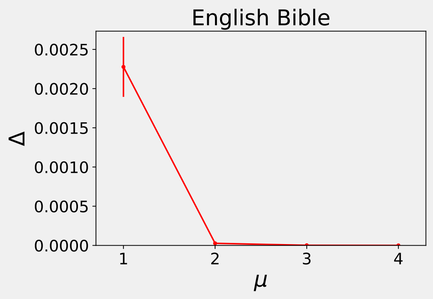


Comparison of estimators for binary systems with $m = 1$. The colors represent how far the quantity $\sum_{n=1}^{17} (H_n - \hat{H}_n)^2$ is from 0. The estimated entropies were calculated from numerically generated sequences of 10^4 realizations each.

Applications

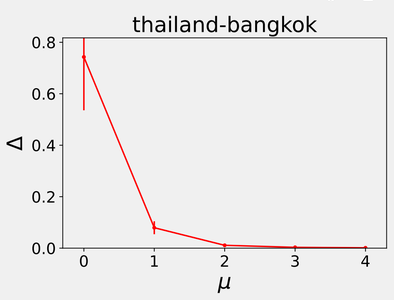
Lexical statistics of texts

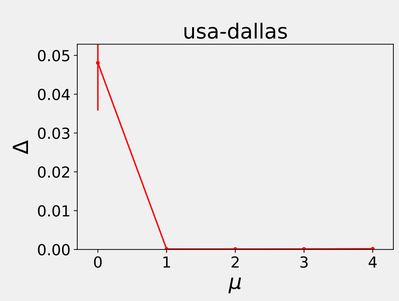
- Given a text, each word is replaced by its ranking r_i
- Pattern:
  - if $r_i < r_{i+1} \longrightarrow 1$
  - if $r_i > r_{i+1} \longrightarrow 2$



We analyzed 12 languages using our method and we found that all of them can be described with a model of memory $m = 2$.

Daily precipitations





We analyzed data of daily precipitations worldwide and we found that the choice of memory depends on the location. For example, Thailand can be described with a model of memory 2, whereas for the U.S. a memory 1 is better.

References

[1] A. Chao, T. J. Shen, *Nonparametric estimation of Shannon's index of diversity when there are unseen species in the sample*, Environ. Ecol. Stat. **10**, 429-443 (2003).

[2] D. G. Horvitz, D. J. Thompson, *A generalization of sampling without replacement from a finite universe*, Journal of the American Stat. Assoc. **47**, 663-85 (1952).