Temporal Entropy Measures

October 16, 2023

1 Background

Entropy is often used to represent an aggregated measure of the information flow between two sources. Often estimators assume a stationary entropy rate with respect to time. In settings that only consider a short time frame, this assumption may be reasonable, but over longer time periods, this may be invalid.

In some settings, entropy estimators are used to detect changes in entropy over time.

Consider a directed information flow between a source, S, and a target, T.

We can consider the information flow between these two nodes $I(S,T)$ as a time-varying quantity. At any instant, this can be represented by the entropy rate $h(S,T)$ or as an overall measure of entropy, $H(S,T)$.

In practice, estimating the entropy rate from a continuous data source can be problematic when measurements cannot be repeated. To make a Shannon estimate of the entropy rate, the size of the typical set or the discrete probability space is used. When a single data stream exists, this results in an estimate over many parameters using a small sample size. This results in an estimate with high variance, however the size of this variance is often not included in final estimate values.

However, understanding how the entropy rate between two sources varies over time can reveal dynamics between two actors indicative of interesting or abnormal behaviours which would not be detected using an aggregated estimator (Figure 1).

In this work, we will present two temporal entropy estimators. These two estimators capture two different aspects (Figure 2, Figure 3). The first measures the entropy rate across time, using properties of existing estimators to impose smoothness constraints on the resulting estimate, reducing variance.

The second approach is designed to capture temporal patterns created by delayed relationships in information flows between a source and a target.

2 Mathematical Background

2.1 Shannon entropy estimators

To motivate the two approaches for representing temporal relationships in the information flow between a source and a target, we will consider information flows through the lens of Shannon entropy estimators.

In particular, we will consider estimators which aim to estimate the size of the typical set, using the asymptotic equipartition property (AEP) to estimate the entropy of a sequence.

Definition 2.1 (Asymptotic Equipartition Property)

$$\frac{1}{n}\log_2\frac{1}{p(X_1,...,X_n)} \to H$$

for i.i.d X_i, $p(X_1,...,X_n) \to 2^{-nH}$

The typical set contains all sequences with an entropy close to H.

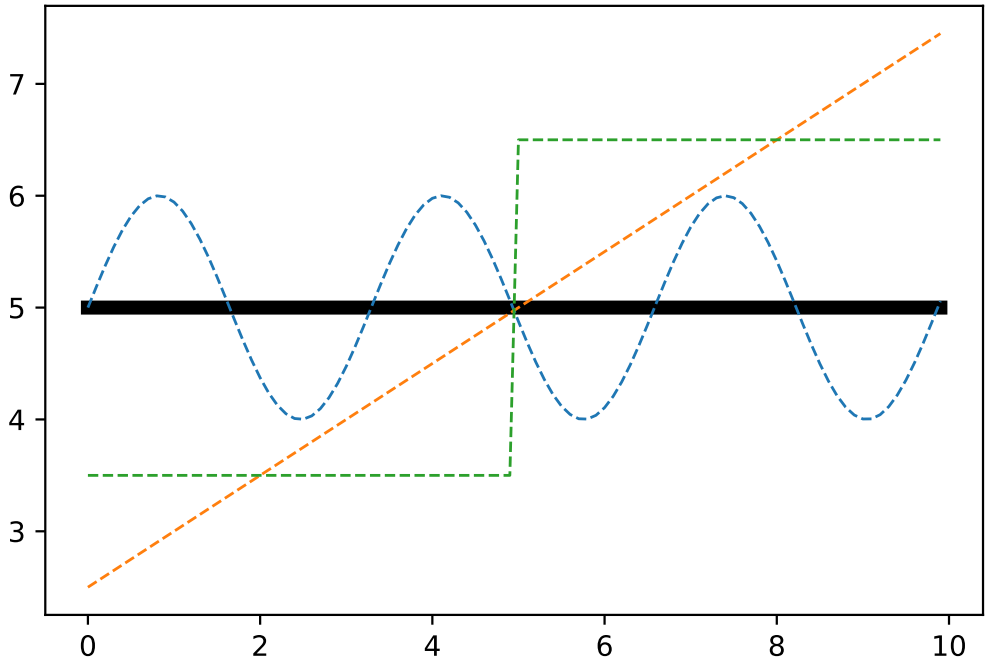


Figure 1: Example entropy rates shown with dashed lines which all have the same average entropy rate.

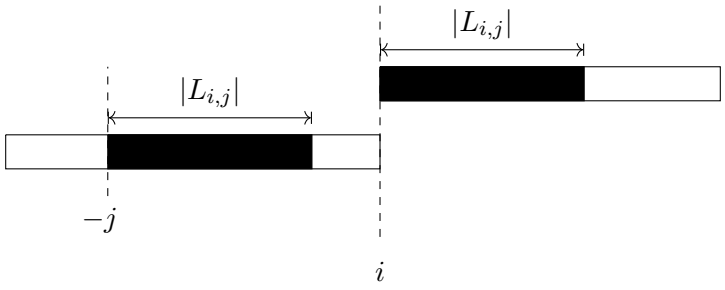


Figure 2: Smaller eg

3 Notation

Let the string from the source and target processes be denoted, S and T respectively.

The realisations from each process are drawn from a random process with a finite sample space containing values from the set $\mathcal{V}$.

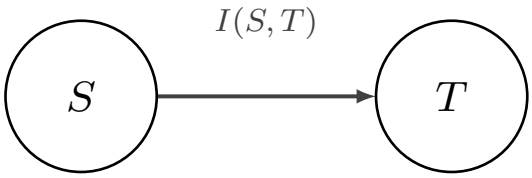
When applied to language, $\mathcal{V}$ represents the vocabulary with each value representing a distinct word and $|\mathcal{V}|$ representing the true vocabulary size.

For the sequence S, we let $S_i^j, i \leq j$ represent the subsequence given by $(S_i, S_{i+1}, ..., S_j)$.

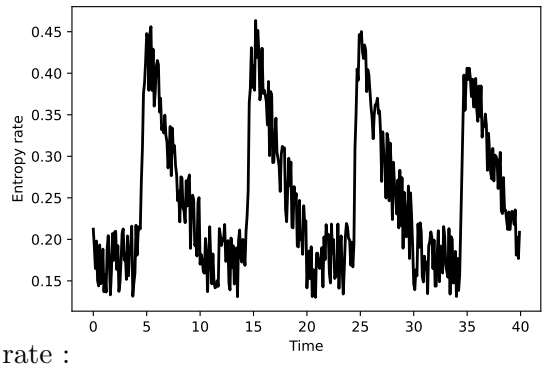
To estimate the Shannon cross entropy rate between S and T, at an instant i, we estimate the entropy rate between S_0^i and T_{i+1}^n. That is, we consider the cross entropy between $S_0, ..., S_i$ and $T_{i+1}, ..., T_n$.

Let the longest match length between the sequence starting at i and the match starting at j be denoted $L_{i,j}$.

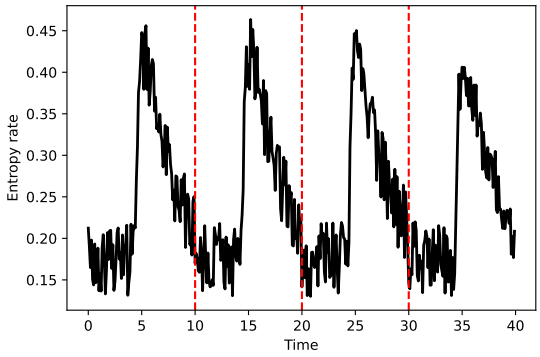
We can define Λ_i^n is the length of the shortest substring S_i^{i+l-1} (length l) starting at position i that does not appear as a contiguous substring of the previous n symbols S_{i-n}^{i-1}. This is a



(i) Entropy rate with time :



(ii) Partitioned entropy rate :



(iii) Entropy rate kernel :

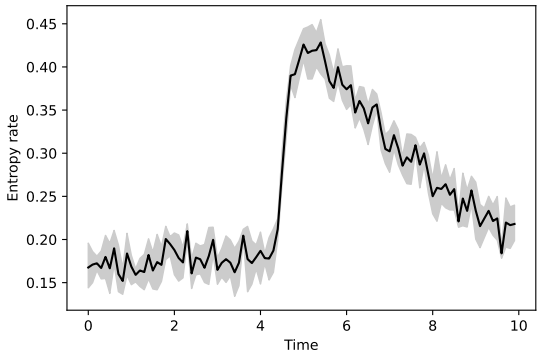


Figure 3: For an information flow between a source, S, and a target, T, we can consider the entropy rate with time (i). We can choose to represent this as a quantity which varies with time, or we can consider the entropy rate as a function of delay between the two sources. One approach to do this is to partition the continuous entropy rate into equally sized windows (ii) and aggregate the entropy rate within each window to estimate the entropy rate kernel (iii).

mis-quote from Wyner Ziv (seems to lose its boundary crossing abilities between WZ to OW and K).

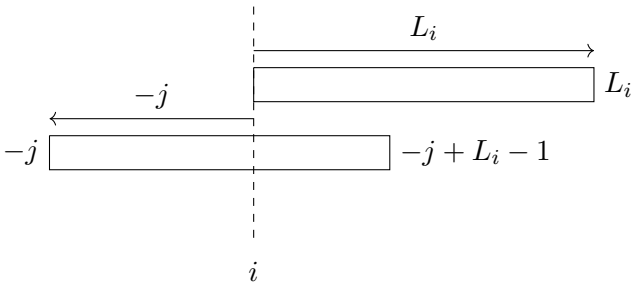


Figure 4: $\tilde{N}_l(x)$ as defined by Wyner Ziv **?**.

We can alternatively express Λ_i^n as

$$\max\{L_{i,j} | n \leq j < i\}$$

[[To Do: modify to show L not N]]

The quantity Λ_i^n is used to estimate the entropy in non-parametric estimators proposed by **?**Ornstein & Weiss (1993), Kontoyiannis et al. (1998). This quantity is used to estimate the size of the typical set, and can be used to make a point or averaged estimate of the entropy rate of a sequence.

To estimate the cross entropy between S and T. For each value of m, we calculate the match lengths l, such that $S_m, ..., S_{m+l} = T_i, ..., T_{i+l}$ for $m + l \leq i$. We then set, $\Lambda_i = \max(L_{i,\cdot}) + 1$, which represents the length of the longest unseen sequence beginning at instant i.

To estimate the entropy, we calculate each value of Λ_i for $i \leq n$, and use one of the following approximations proposed by Kontoyiannis:

(a)
$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\Lambda_i^n}{\log n} \to \frac{1}{H}$$

(b)
$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\Lambda_i^i}{\log i} \to \frac{1}{H}$$

(c)
$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\Lambda_i^i}{\log n} \to \frac{1}{H}$$

which converges almost surely **???**.

Here is an overview of methods I tried:

3.1 Estimating the size of the typical set

The AEP gives two sets:

1. typical set: the set of all sequences with an entropy close to H
2. non-typical set: opposite of the typical set

The proof is something like: If X_i (items in the sequence are i.i.d) then $f(X_i)$ are also i.i.d. Let $f(\cdot) = \log(\cdot)$.

By the weak law of large numbers, then

$$\begin{aligned}
&= -\frac{1}{n} f(X_1, ..., X_n) \\
&= -\frac{1}{n} f(X_1) ... f(X_n) \\
&= -\frac{1}{n} \sum_i \log p(X_i) \left(= \frac{S_n}{n} \right) \\
&= -\mathrm{E}[\log p(X_i)] \text{ in probability} \\
&= H(X)
\end{aligned}$$

Let us denote the typical set $A_\epsilon^{(n)}$.

Definition 3.1 (Typical set) *$A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences, $X_1, ..., X_n$, with the property:*

$$2^{-n(H(X)+\epsilon)} \leq p(X_1, ..., X_n) \leq 2^{-n(H(X)-\epsilon)}.$$

$A_\epsilon^{(n)}$ has the properties:

1. For $(X_1, ..., X_n) \in A_\epsilon^{(n)}$ then $H(X) - \epsilon \leq \frac{-1}{n} \log p(X_1, ..., X_n) \leq H(X) + \epsilon$.
2. $\Pr(\{A_\epsilon^{(n)}\}) > 1 - \epsilon$ for all sufficiently large n.
3. $|A_\epsilon^{(n)}| \leq (1 - \epsilon) 2^{n(H(X)+\epsilon)}$, where $|A|$ is the size of the set A.
4. $|A_\epsilon^{(n)}| \geq (1 - \epsilon) 2^{n(H(X)-\epsilon)}$ for sufficiently large n.

A sequence from the typical set has probability close to 1. The number of elements in the typical set is close to $2^{n(H(X)}$.

The first avenue of consideration was to use the number of elements in this typical set to construct a critical value. i.e. to identify $P(\text{observed sequence} > all in typical set)$ to determine size of typical set and thus the entropy. (For proof see Cover & Thomas (1991).

Note, $A_\epsilon^{(n)}$ does not necessarily contain all of the most likely sequences.

For example, for a Bernoulli with $p = 0.9$, $A_\epsilon^{(n)}$ will contain all sequences with a proportion of 1's equal to 0.9. But it will not contain a sequence of all ones.

Smallest probable set is

$$2^{nH(X)} = P(B_\delta^{(n)}) > 1 - \delta$$

for $delta < \frac{1}{2}$.

However this becomes tricky, as you need to have an idea of what the entopy is to be able to determine if a sequence is typical, and it falls apart for smaller samples.

3.1.1 Binomial Example

[[To Do: move notes from notebook over p(61-63)]]

Essentially we can say that the typical set has size:

$$1 \leq |T| \leq n^{|V|}$$

which isn't super helpful.

4 Entropy Landscape

Estimating entropy rate across time.

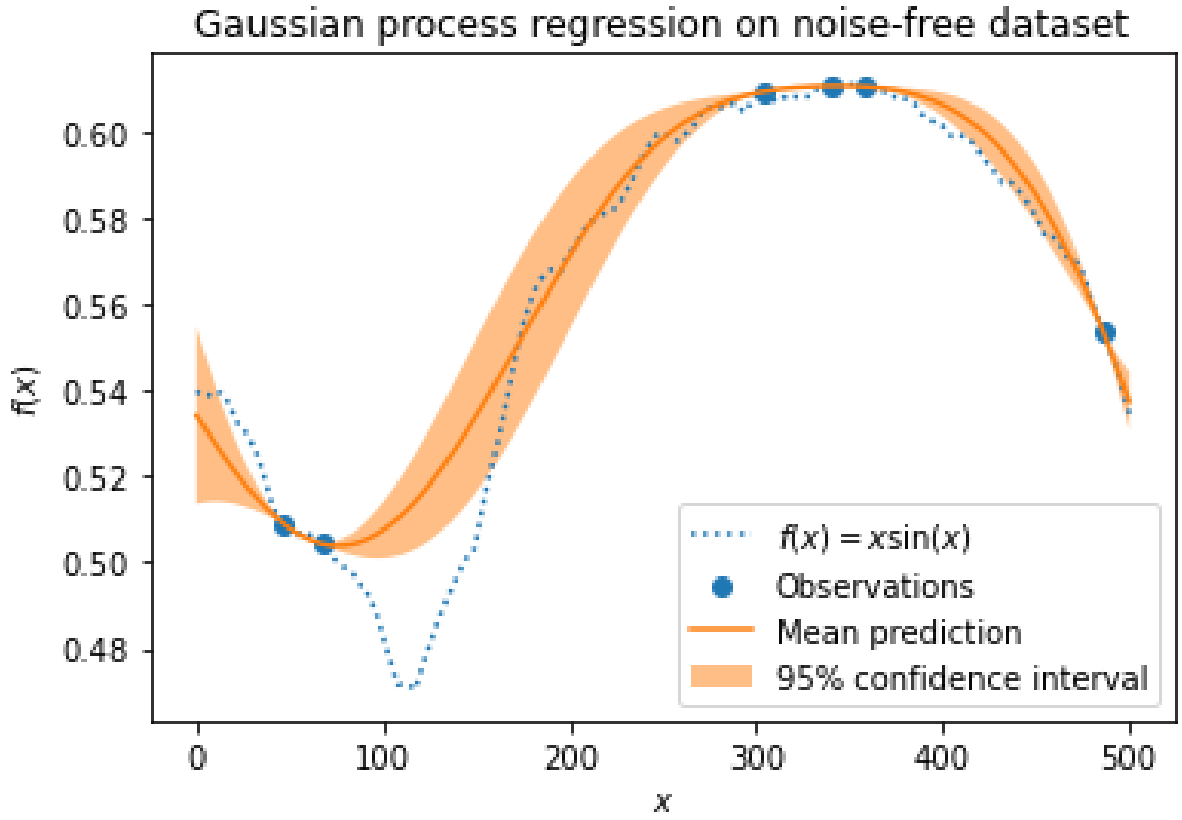


Figure 5: Gaussian Process Fit

4.1 Smoothness on temporal estimates

This estimator is used to estimate the entropy rate at all possible values of i across the sequence. By construction, for ascending i, each sequence considers matches to a target sequence starting 1 after the current starting index. This means that the size of the match at $i+1$ is constrained to be at least one smaller than the length of the match at i, given no constraints on j.

When we consider a point estimate on the entropy, we can consider the entropy estimate at $i+1$ as a function of the estimate at i.

$\Lambda_i = \max(L_{i,\cdot}) + 1 > \max(L_{i+1,\cdot}) + 1$

This means we can consider techniques to fit a smooth curve to the entropy estimates.

If we assume that $h(t)$ follows a continuous sample path, we can take three possible approaches: 1. Estimate E

$$h(t)$$

Two possible approaches include using a smoothing spline or Gaussian process model.

2. Estimate Var

$$h(t)$$

Estimate using the WLLN.

or, 3. Fit a model to the distribution of Λ^n, updating this distribution at each time step.

We ruled out (3) due to wanting to avoid imposing additional constraints, and wanting an approach which would work for a single data stream with small sample size. If you are willing to impose assumptions or an informative prior, a Bayesian type approach can be a great way to approach this.

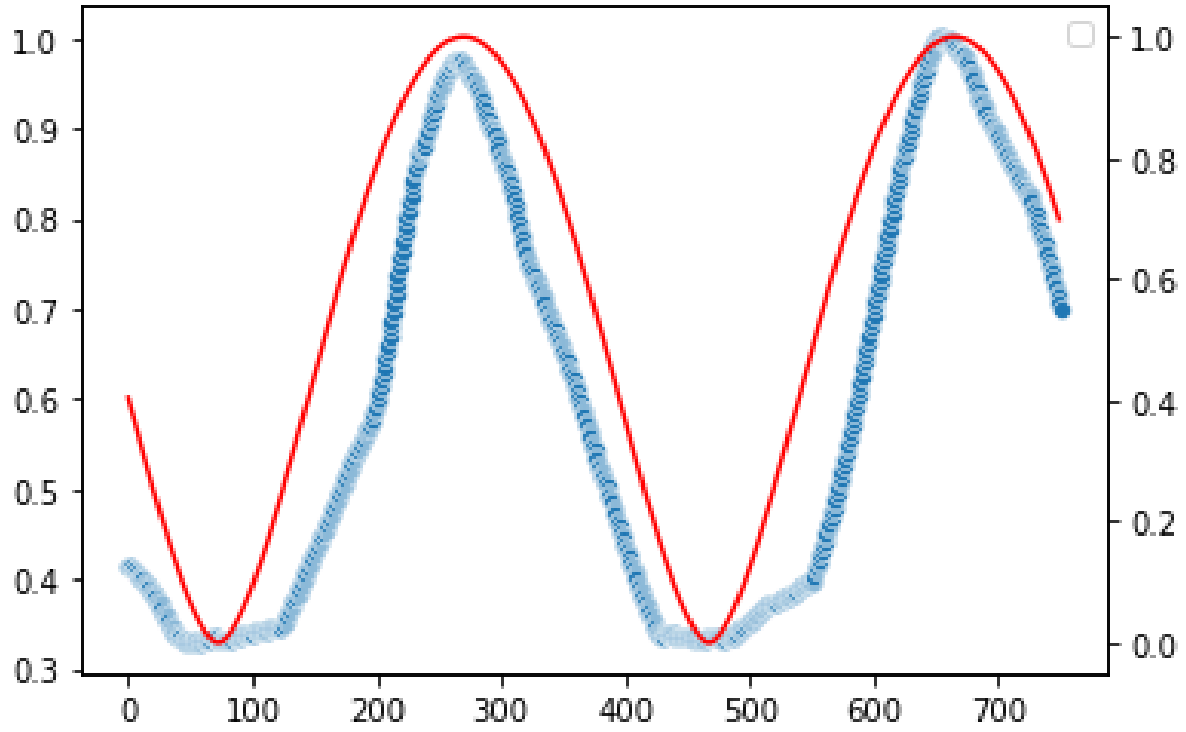


Figure 6: Smoothing Spline

4.2 Results

Maybe Gaussian process is useful if the kernels are interpretable, but for a function, smoothing spline doesn't require parameters to be chosen which is really good.

Smoothing splines worked better.

Process is: (1) fit moving average model then (2) fit smoothing spline using a grid search.

Seems like choice is resonably robust to choice of k and s (there are some rules which seem to work).

Even works for pretty wobbly functions.

4.3 Entropy Kernel

Idea: estimate the longest match length for a given time delay, and check if this is dependent on the size of the delay to determine if there is a meaningful temporal relationship between a source and a target.

So we find $L_{i,j}$ for all i and j across a pair of sequences. Using these to get a point entropy estimate on a sequence with a period of 20 demonstrates that we can see this effect:

There are artifacts in this for when j constrains the length of the match (overlaps/finite sequences). We do this by disregarding any estimates formed using a history length smaller than the 95th percentile of the observed lambdas.

We still have a strong spike.

4.3.1 Null model

Null model is that the match length is independent of the choice of j.

So we can estimate the empirical distribution of lambdas by permuting j with respect to L. This gives a null distribution which can be removed from the dataset. For our example,

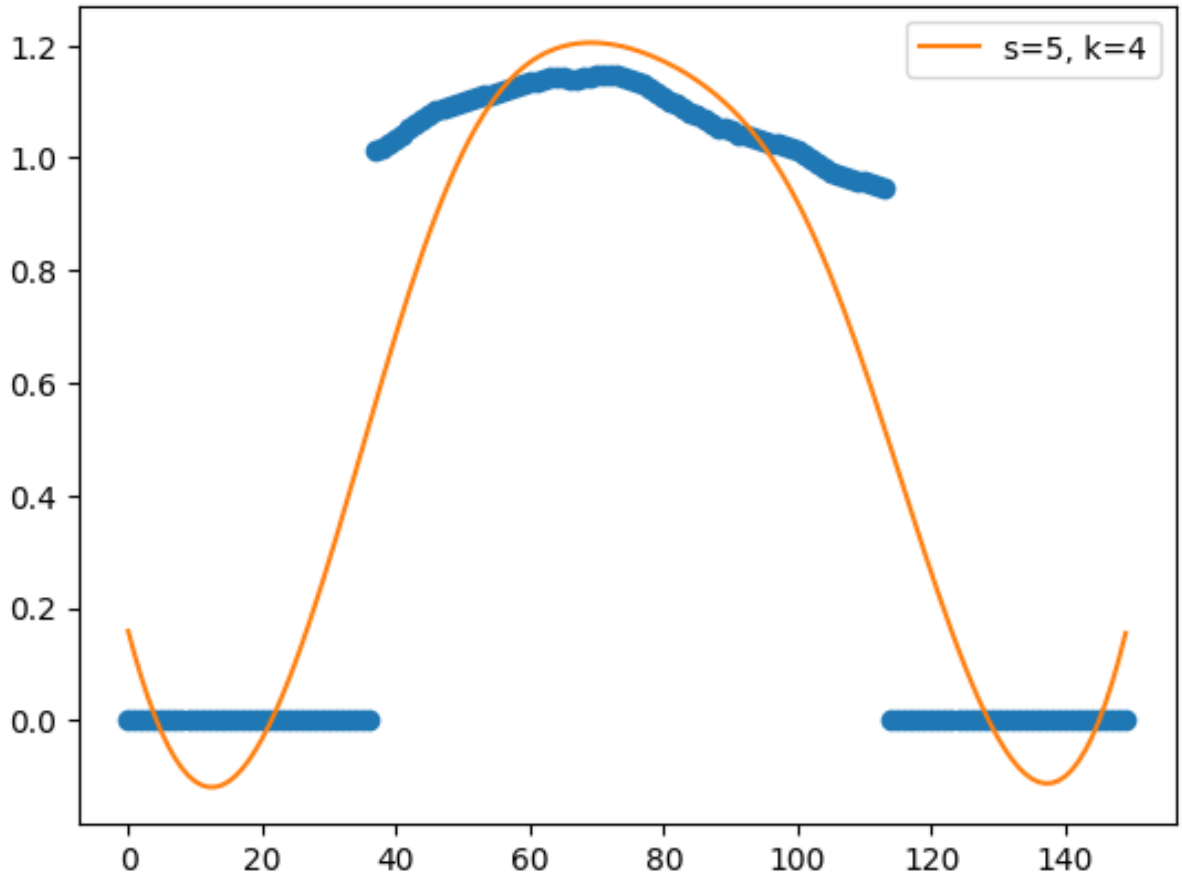


Figure 7: Smoothing Spline

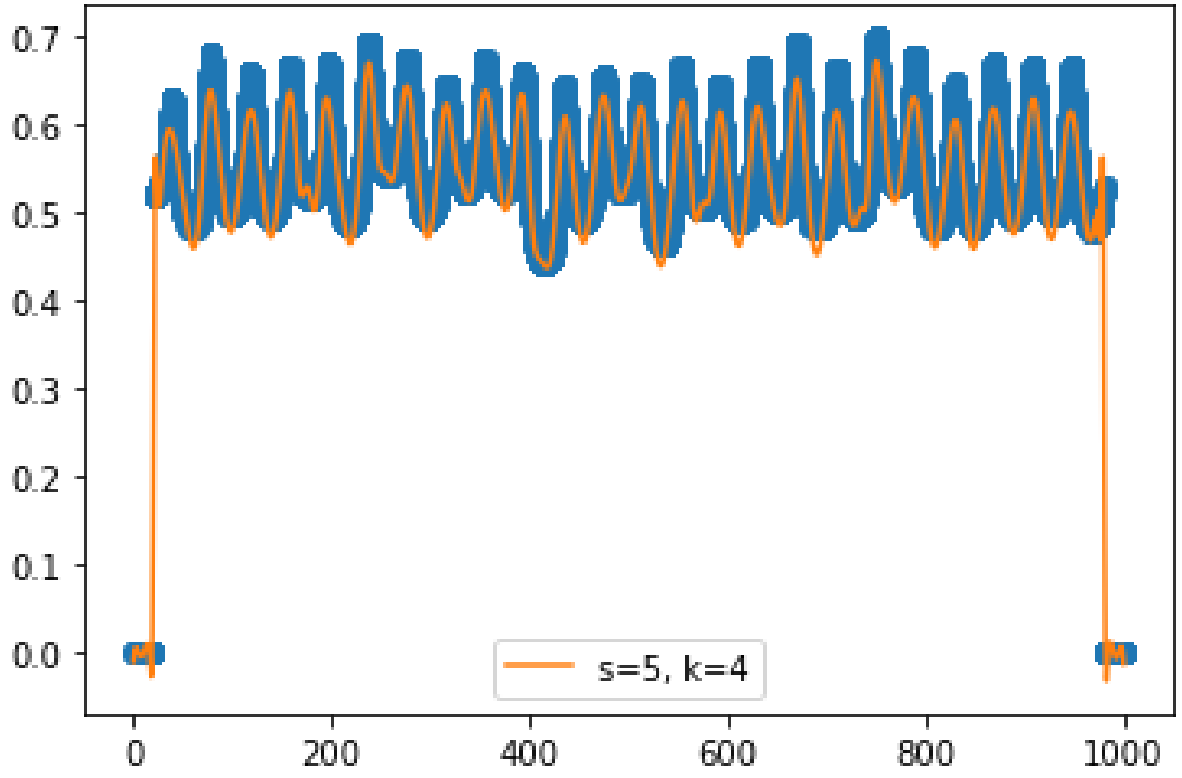


Figure 8: Smoothing Spline on wobbly function

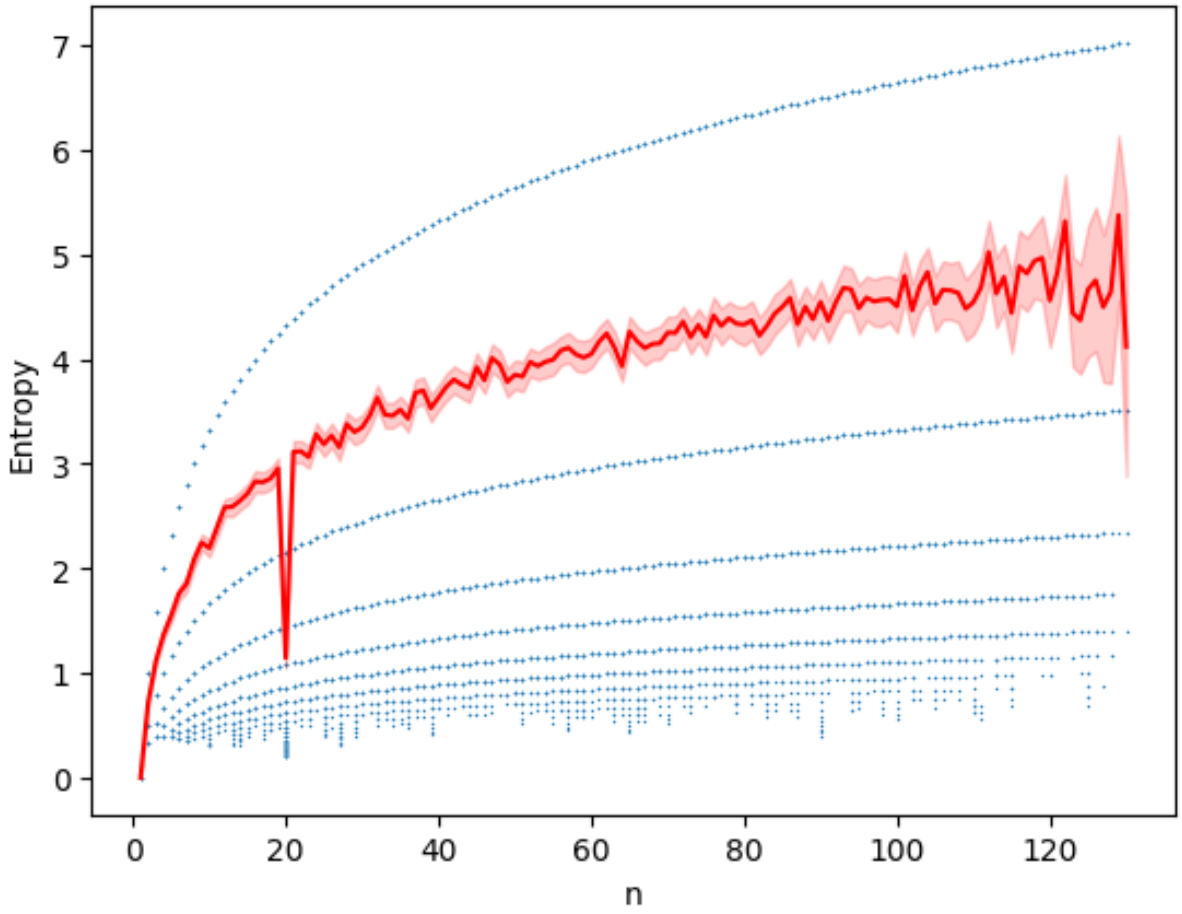


Figure 9: Point entropy estimate. n is the length of the history.

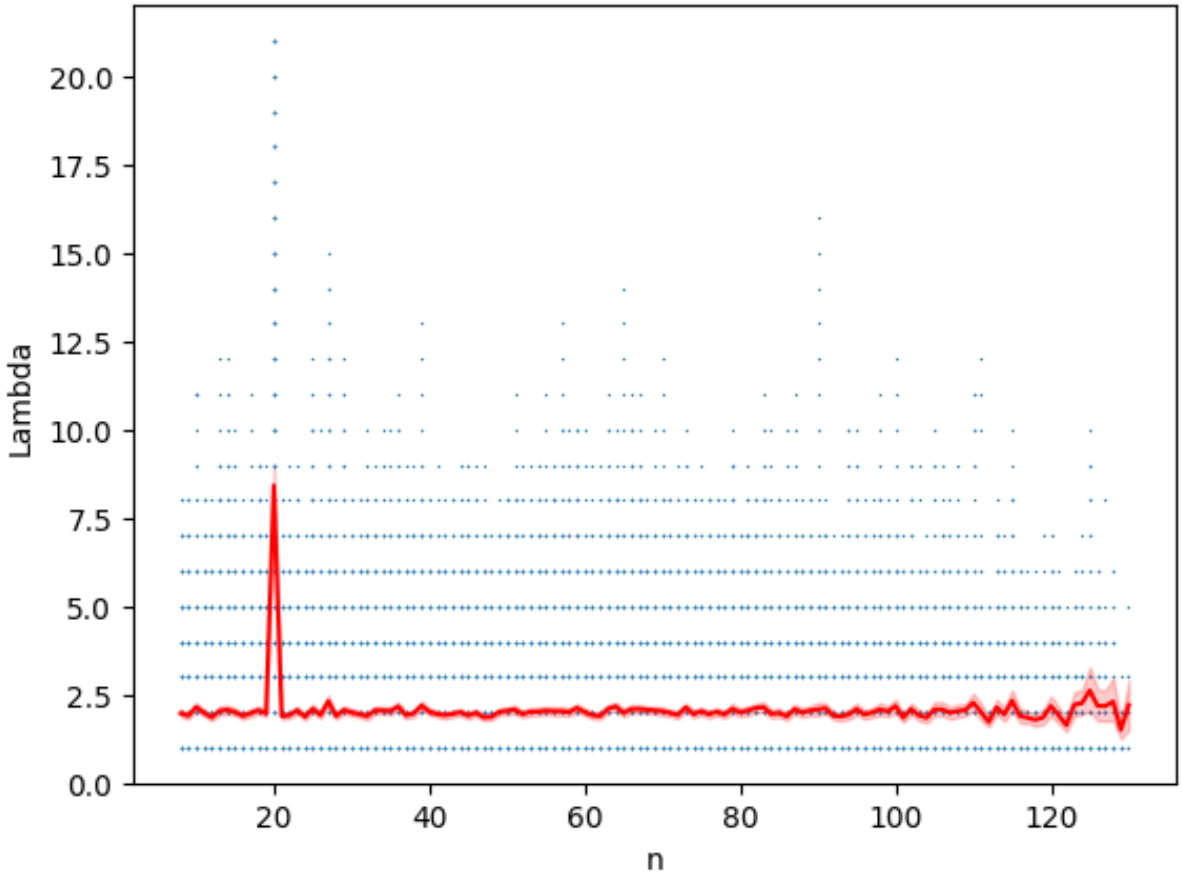


Figure 10: Filtered estimates

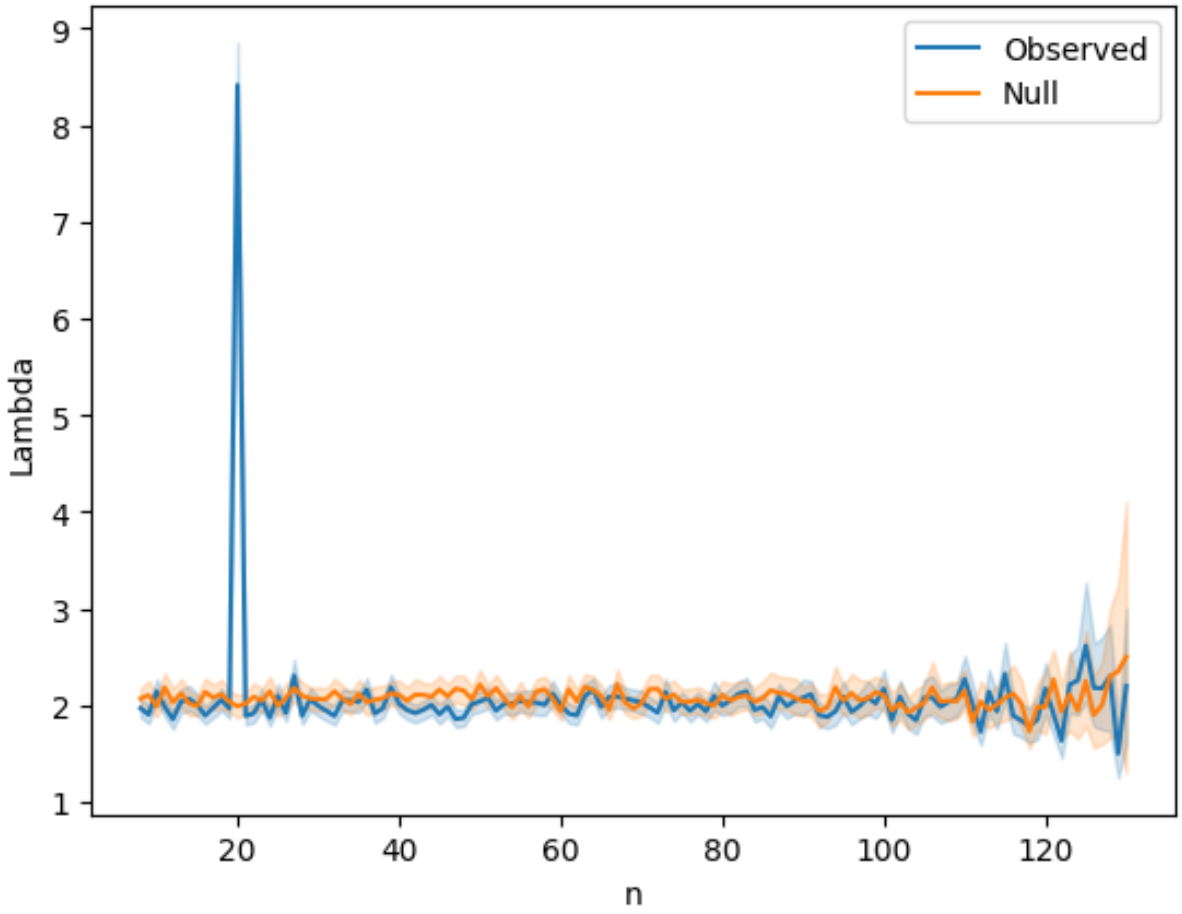


Figure 11: Null model

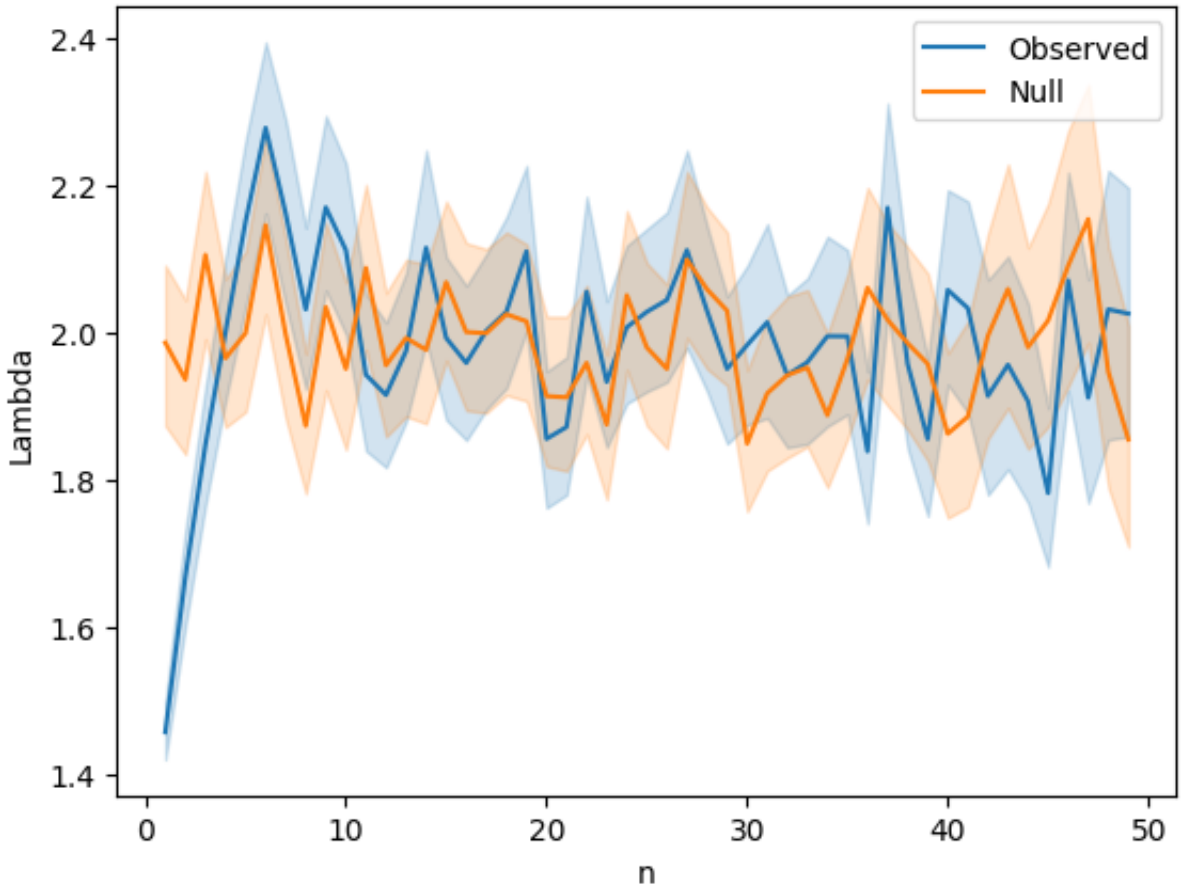


Figure 12: Sensitivity limitations

4.3.2 Sensitivity

Method needs this periodicity to occur **very** regularly. In the testing, we simulate sequences by copying the sequence element k steps back with probability p. If $p < 0.8$ it becomes really hard to detect the spike.

Q: What does this mean in a real data sense?

Can explicitly express this using a significance test

5 Entropy estimators in general

- don't want to impose window size

- consider distribution of match lengths as a function of the distance between the start of the sequence and the start of the match

- isolates ONLY very periodic patterns

References

Cover, T. M. & Thomas, J. A. (1991), *Elements of Information Theory*, Wiley Series in Telecommunications, Wiley, New York.

Kontoyiannis, I., Algoet, P., Suhov, Y. & Wyner, A. (1998), 'Nonparametric entropy estimation for stationary processes and random fields, with applications to English text', *IEEE Transactions on Information Theory* **44**(3), 1319–1327.

Ornstein, D. & Weiss, B. (1993), 'Entropy and data compression schemes', *IEEE Transactions on Information Theory* **39**(1), 78–83.