Honors Independent Study Paper

A Data-Analytic Exploration of Wikipedia Pageviews Data

Paper By: Matthew Kulec

Worked with: Professor Asohan Amarasingham and Anthony Alston

May 27, 2019

1 Cross Correlation

1.1 Definition and Speed

A useful way to compare how structurally similar two time series are relative to one another is by looking at the correlation, but over different time lags, which is known as cross-correlation. That is, by displacing one time series with respect to another and then computing the correlation as a function of that displacement. The correlation $\rho(\tau)$, at lag τ is defined by:

$$\rho(\tau) = \frac{\mathbb{E}[(X_{t+\tau} - \mu_X)(Y_t - \mu_Y)]}{\sigma_X \sigma_Y} \overset{P}{\leftarrow} \frac{1}{n\sigma_X\sigma_Y} \sum_{t=1}^{n-\tau} (X_{t+\tau} - \mu_X)(Y_t - \mu_Y), \qquad \tau \geq 0 \qquad (1)$$

$$\rho(\tau) = \frac{\mathbb{E}[(Y_{t-\tau} - \mu_Y)(X_t - \mu_X)]}{\sigma_X \sigma_Y} \overset{P}{\leftarrow} \frac{1}{n\sigma_X\sigma_Y} \sum_{t=1}^{n+\tau} (Y_{t-\tau} - \mu_Y)(X_t - \mu_X), \qquad \tau < 0 \qquad (2)$$

Which can be approximated by a sample mean, according to the law of large numbers, as the right hand side shows. n represents the length of the time series $\{X_t\}_{t=1}^{n}$, and expects that both time series $\{X_t\}_{t=1}^{n}$ and $\{Y_t\}_{t=1}^{n}$ share the same length. If not, an imputation of zeroes for the shorter time series is necessary.

Note that the cross correlation can be computed over $2n - 1$ lags, and for each lag there are $\mathcal{O}(n)$ operations. Therefore the total time complexity amounts to $\mathcal{O}(n^2)$. To help scale better for large n, notice that the cross correlation is essentially a discrete correlation, respecting a signal processing point of view. It also bears a resemblence to convolution by reversing one of the time series. This is transparent via the following relation:

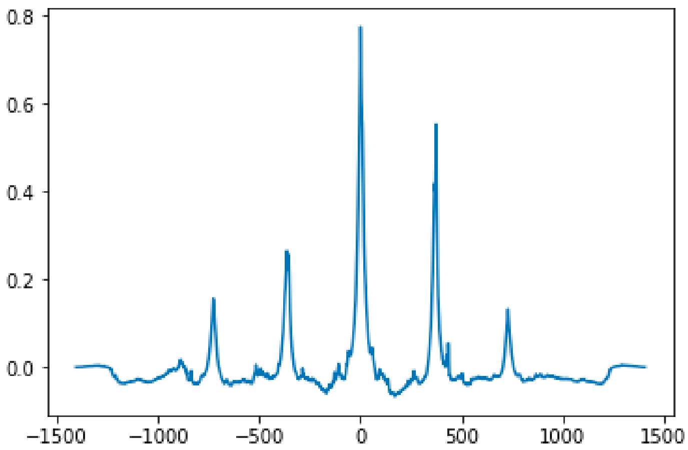
$$\mathcal{F}(g \star f)(s) = \mathcal{F}g(s)\mathcal{F}f(-s) \tag{3}$$

Which gives away what is known as the Convolution Theorem (proof in the appendix). Note the symbol $\star$ denotes a correlation, not a convolution. One can pair this result with the Fast Fourier Transform to compute at a $\mathcal{O}(n \log n)$ time complexity instead [3].

1.2 Interesting Properties

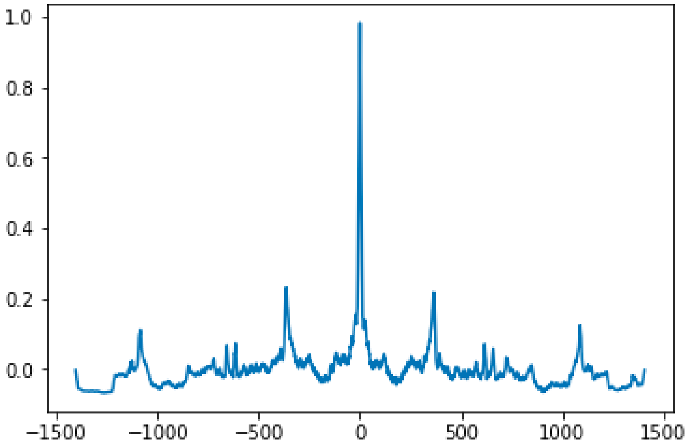
We obtained the time series of daily pageview counts over the time span from July 1, 2015 up to a couple of days in the past month (data was collected at different days during the past month, resulting in time series of lengths differing by a couple of days) for certain Wikipedia pages. This resulted in time series lengths n that ranged from 1350-1407 days of page views recording. We limited our analysis for this section to people, movies and characters associated with fan-favorite franchises. This is because we speculate that fans of a particular franchise have a high sensitivity to absorb new information (regarding their franchise) that is present in the news. They might feel the inclination to notice changes or quick look-ups in Wikipedia. Further motivation for why Wikipedia can be a reliable source stems from a study which demonstrated that changes in Wikipedia activity can help with detection of legitimate world-wide events in Twitter [6].

We observed several interesting properties resulting from the plots. The first of which is the ability to detect the presence of news-worthy events through sharp peaks besides the one at lag zero. For example, observe the graph of the cross correlation for Wikipedia pages "Princess Leia" and "Star Wars":

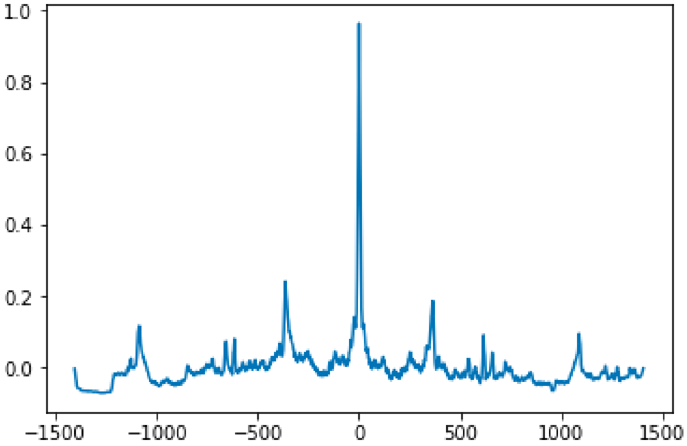


The second largest peak occurs in the neighborhood of 372-377th positive lag. Note that $\{X_t\}_{t=1}^{n}$ denotes the time series for "Princess Leia" and $\{Y_t\}_{t=1}^{n}$ denotes the time series for "Star Wars," so that second-highest peak happens when the "Princess Leia" time series is shifted to the right by the amount of lags specfied with respect to the "Star Wars" time series. One useful way to interpret a high correlation is that, on average, one (or both) time series is larger than its respective mean. This encourages us to examine what happened in the news 372 to 377 days ago relating to Star Wars. A custom Google news search from May 13, 2018 to May 18, 2018 (377 and 372 days ago, respectively) shows multiple news reviews about *Solo: A Star Wars Story*, which first aired May 10, 2018. This explosion of information diffusion could have been a reason for that spike.

Another discovery is the existence of a striking similarity in the shape of the ridges between the plots of two or more cross correlations, given that they each have a time series in common. A plausible explanation for this phenomenon is that there are similar reasons for viewing page A because of page B as there for viewing page C because of page B (assuming the cross correlation of A and B are similar in shape with the cross correlation of C and B). The following is a plot of the cross correlation of Wikipedia pages "Robert Downey Jr." and "Robert Downey Sr" (his father):



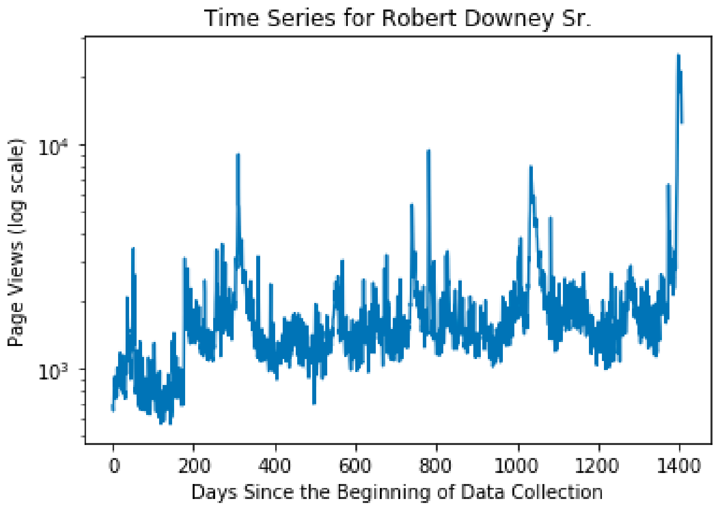
and this is a plot of the cross correlation of "Robert Downey Jr" and "Susan Downey" (his wife):

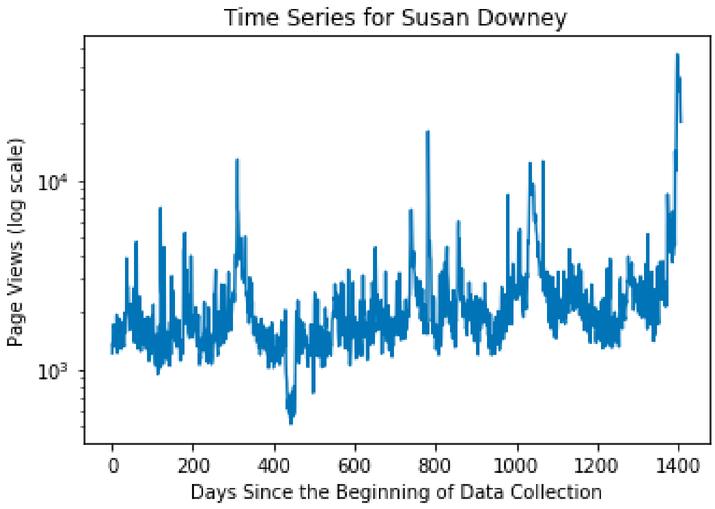


The cross correlations are nearly identical. Intuitively, we could then hypothesize that people visit the "Robert Downey Sr" Wikipedia webpage for the same reason they visit the "Susan Downey" Wikipedia webpage — because they are both related to "Robert Downey

Jr.” and because of virtually no other reason.

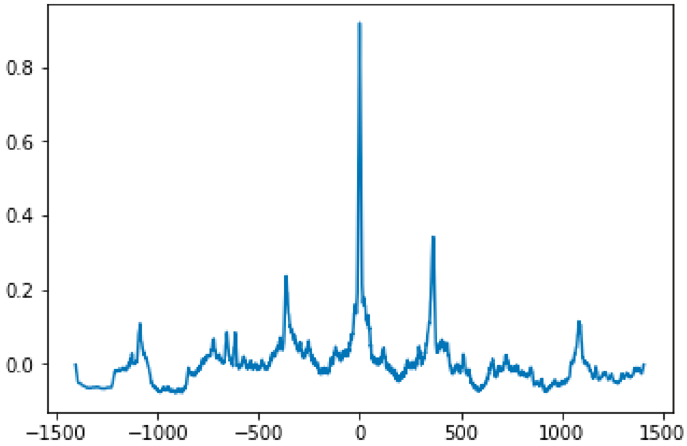
The only way this could happen is, if the time series for “Robert Downey Sr” and the time series for “Susan Downey” experience a proportional growth and decline in pageviews per day. In other words, the shapes of the ridges of those time series should be equal too. Indeed, they are:





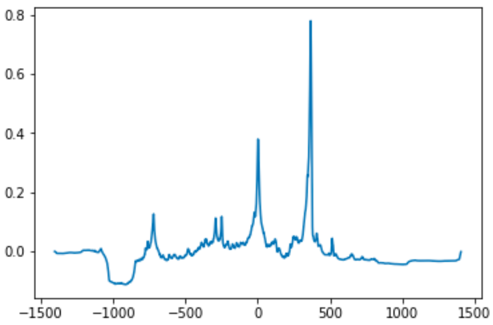
Conversely, we compared “Robert Downey Jr.” with a webpage that preserves much of the fine timescale correlations that was seen in the past two examples, but is noticeably different in some areas. A perfect example of the other webpage would be a movie in which

Robert Downey Jr played the only main protagonist Iron Man, but the movie would contain other characters and events (but not as important as Iron Man, relatively speaking) and thus opens multiple reasons from viewing that webpage. This webpage was "Iron Man (2008 film)":



Reasons for the discrepancy could be from the fact that there were a majority of people who viewed "Robert Downey Jr." who also viewed "Iron Man (2008 film)," but there are some people who viewed "Iron Man (2008 film)" without looking at the "Robert Downey Jr." webpage, and vice versa.

When we inspect the cross correlation for "Robert Downey Jr." and "Avengers: Infinity War," we should expect the shape to remember nothing like the previous plots because we can't guarantee that the people who viewed the "Avengers: Infinity War" webpage to have also seen the "Robert Downey Jr." webpage, since those viewers could be strictly Thor fans or Spiderman fans:



2 Hawkes Processes

2.1 Definition

In a Poisson Process, events occur independently of one another, but governed by their intensity function, or the expected number of events per unit time (assuming the point process is a temporal one). On the other hand, a Hawkes Process is a type of point process (specifically, a self-exciting process) where events need not be independent, and that the occurence of an event increases the likelihood (or intensity) for future events [7]. Formally, the event intensity function is defined by:

$$\lambda(t) = \mu_0(t) + \sum_{i:t>T_i} \varphi(t - T_i)$$

Where φ is typically taken to be a monotonically increasing kernel function. Some kernel functions used in practice are the exponential kernel $\varphi(t) = \alpha \exp(-\beta t)$ (with $\alpha \geq 0$, $\beta > 0$, and $\alpha < \beta$) and the power law kernel $\varphi(t) = \dfrac{\alpha}{(t+\beta)^{\eta+1}}$ (with $\alpha \geq 0$, $\eta, \beta > 0$, and $\alpha < \eta\beta^{\eta}$). $\mu_0(t)$ is known as the background, or base intensity function, and is an exogenous factor. Hyperameters α and β (and η if power-law kernel is chosen) govern the degree of endogenous self-excitation. An exogenous contribution can be the effect of new media spreading information to the public, while an endogenous contribution can be effect of participants linked by a social network spreading information (i.e. an information cascade), for instance. It is also important to note that the Hawkes intensity function depends on previous history of event times, i.e. $\{T_i \mid t > T_i, \forall i\}$.

2.2 Parameter Estimation

In order to estimate the parameters, one can perform maximum likelihood estimation. The maximum likelihood function, or the product of the probability densities for each of the event times $T_1, T_2, \ldots T_n$ over the period $[0, T]$ is given by (derivation in the appendix):

$$L(\Omega) = \exp\left(-\int_0^T \lambda(s)\mathrm{d}s\right)\prod_{i=1}^{n} \lambda(T_i)$$

Where Ω denotes the parameter space. For the sake of floating point accuracy and numerical complexity, one may wish to estimate the parameters based on the log-likelihood. However, the log-likelihood is not optimization friendly due to the fact that it is not globally convex, resulting in estimates that are local maximums, instead of than global ones [7]. Another strategy for parameter estimation can be a Bayesian one, by maximizing the evidence function [7].

3 Main Explorations

3.1 Non-Parametric Bootstrap and Parametric Bootstrap

The central idea of bootstrapping is to “recreate the the relation between the ‘population’ and the ‘sample’ by considering the sample as an epitome of the underlying population” [5]. Suppose you have a sample from some population: $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and you compute some statistic $\hat{\theta} = s(\mathbf{X})$ based on that original sample. One can treat each element in $\mathbf{X}$ as equally likely to have occurred, and sample with replacement from that original sample to obtain n bootstrap observations $\mathbf{X}_i^* = (X_{1,i}^*, X_{2,i}^*, \ldots, X_{n,i}^*)$. Using this artifical sample, or surrogate, we can compute a surrogate-based test statistic $\hat{\theta}_i^* = s(\mathbf{X}_i^*)$. We can continue this process to obtain a large amount of surrogate-based, or bootstrap test statistics B such that $1 \leq i \leq B$. The key observation is that there is an empirical distribution $\hat{F}$ that gives rise to these bootstrap samples [4], and that it tries to mimic the true distribution F. This is known as the non-parametric bootstrap.

On the other hand, parametric bootstrap makes some model-based assumptions based on the original sample, such as guessing what distribution it resembles and simulating random samples according to that hypothesized distribution for every step i. One can use maximum-likelihood estimation to estimate the parameters of the hypothesized distribution. Note that this variation of bootstrap does not sample from replacement like the previous type. As a result, there may be some specification error resulting from a wrong guess about the underlying distribution generating the original sample [8]. The resulting empirical distribution may be biased. On the bright side, since simulated samples are more diverse than resampled ones, there is a faster convergence in distribution [8].

3.2 Conditional Inference and Interval Jitter

Both parametric and non-parametric bootstrap are based on unconditional inference — there is no conditioning on events. Conditional inference arises when you condition on some event. Interval jitter is a type of conditional inference. The interval-jitter procedure was initially meant for point processes, but for time series, the analog would be to split the time series into continguous blocks of size Δ, and permute the observations within each of those blocks while preserving the sum of the observations (call this counts) for each of those blocks. The null hypothesis of interval jitter states that all possible time series realizations, given the condition just described and having the vector of counts stay the same across all realizations, are equally likely to occur. Specifically, the null hypothesis would be saying that [1]

$$\Pr\{\boldsymbol{X} = \{X_t\}_{t=1}^{n} | N(\boldsymbol{X}) = \boldsymbol{n}\} = \frac{\mathbb{I}\{N(\{X_t\}_{t=1}^{n}) = \boldsymbol{n}\}}{\sum_{\boldsymbol{Y} \in \mathcal{A}} \mathbb{I}\{N(\boldsymbol{Y}) = \boldsymbol{n}\}}$$

Where $\boldsymbol{X}$ denotes a time series, $\boldsymbol{n}$ denotes the vector of counts for each of the $\frac{n}{\Delta}$ bins, and $\boldsymbol{Y}$ denotes a random time series drawn from the space of all $n!$ possible time series $\mathcal{A}$.

Because these time series are equally likely, the null synonymously states that there is no temporal structure finer than Δ, and that there could be temporal structure at a scale coarser than Δ. A rejection of the null would be indicative in a low p-value, which is given by the fraction:

$$\frac{\#_{i=1}^{B}\{\rho_i^*(0) \geq \rho(0)\}}{B}$$

That is, the proportion of times the cross correlation evaluated at lag 0 of each of the surrogate time series is greater than the cross correlation at lag 0 for the original time series. In other words, a small p-value would mean that most correlations by chance are not greater than that of the original correlation, alluding to a finer temporal structure.

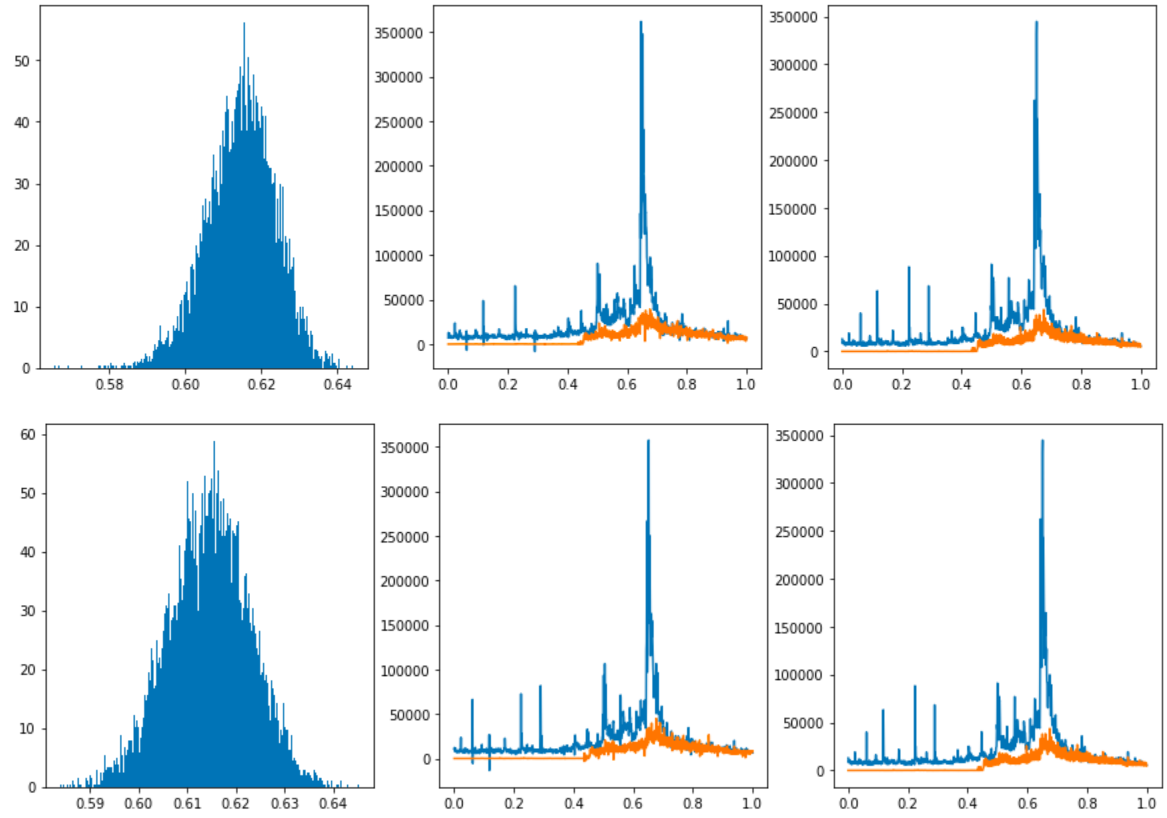
A last remark about the difference between unconditional inference, like bootstrap, and conditional inference, like interval jitter, is that the size of the null hypothesis for the latter is more constrained. In other words, conditional distributions have smaller variances than their unconditional counterparts, and this is important because a larger null translates to a greater difficulty in rejection [1][2].

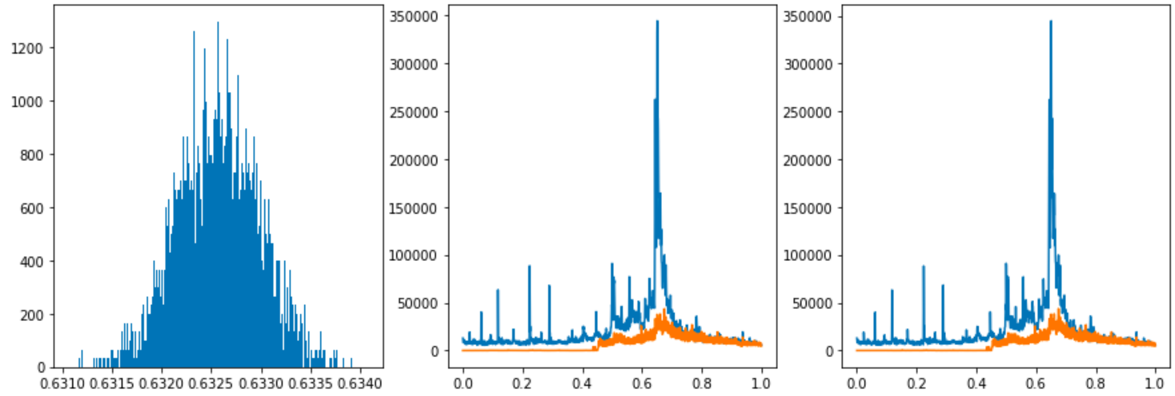
3.3 The Role of Occam's Razor in Hypothesis Testing

In hypothesis testing, if the p-value, which is the value of the original test statistic provided that H_0 is true, falls in the critical region, we reject the null hypothesis. This is because we have two explanations to consider from: either the p-value was a lucky occurence, or H_0 is most likely false. Occam's Razor tells us to choose the more likely explanation, which is the latter, especially if we're performing bootstrap with a large number of surrogate-based test statistics. The premise is attempting to answer the question: in a large pool of test statistics, where does the p-value settle among that pool?

It is important to mention that the the choice of the critical region need not be the tails of the null distribution. The choice of the critical region depends on the hypothesized position of the class of alternative distributions. For example, if the alternative distributions are along the domain, then a natural choice would be the tails. However, an alternative distribution may be thinner in comparison to the null, but be centered about the same mean just as the null. Then the critical region could be selected to be about the mean, rather than the tails of the null. Although these types of critical regions are uncanny, they are typically used when testing to check if the results of a scientific paper are too good to be true (Amarasingham, personal communication). If the p-value finds itself away from the tails of the null distribution (if we choose them as the critical region) and somewhere near the center, we can say that it is plausible that H_0 is true because we fail to reject H_0.

To illustrate this concept of alternative distributions, I performed 3 parametric bootstrap experiments. For each experiment, I divided the time series of “Bitcoin” and “Blockchain” into blocks of length Δ and for each block, computed the maximum likelihood estimators to estimate distribution parameters. Then I sampled observations from the distributions of each block and concatenated everything to form surrogate time series of the same length as the original. I repeated this for $B = 10000$ bootstrap trials. The null for the first experiment is that the blocks are each Poisson distributed with their λ’s being their respective sample means. The null for the second experiment was that the blocks are each normally distributed with their μ’s being their respective sample means and σ’s their respective sample unbiased standard deviations. Lastly, the null for the third experiment was that the blocks are each Laplace distributed with their μ’s (location parameter) being their sample medians, and b’s (scale parameter) being their respective mean absolute deviations from their medians. The first row corresponds to the first experiment, the second row corresponds to the second experiment, and the third row corresponds to the third experiment. The cross correlation at lag-0 was the test statistic and the empirical distributions of this test statistic are shown in the first column. Notice that each of these empirical distributions have most of their mass concentrated at various parts of the domain. This is why it seems fitting to use the tails as the critical region. The second column shows two example surrogate time series with "bitcoin" in blue and "blockchain" in orange. The plots in the third column show the actual, original time series for "bitcoin" and "blockchain."





I found interesting, nonzero results for the p-value only for $\Delta = 1, 2$, and 2 (with respect to each experiment). The original value of the lag-0 cross correlation is approximately 0.633. These p-values were $0.451, 0.0132$, and 0.017 (with respect to each experiment). Indeed, the maximum likelihood estimates for each of the blocks (of these sizes) are questionable. In time series data, nearby observations may be dependent and so the distribution of each datapoint may vary, which would explain why Δ needs to be small. However, judging from these p-values, we may reject the null hypotheses of experiments 2 and 3, but say it is plausible for the null of experiment 1 to be true.

4 Appendix

4.1 Proof of Convolution Theorem for Correlation

First note that:

$$(g \star f)(s) = \int_{-\infty}^{\infty} g(s+x) f(x) \mathrm{d}x$$

The forward fourier transform $\mathcal{F}m(s)$ for any function m is defined by:

$$\mathcal{F}m(s) = \int_{-\infty}^{\infty} e^{-2\pi ist} m(t)\mathrm{d}t$$

. Therefore,

$$\mathcal{F}(g\star f)(s) = \int_{-\infty}^{\infty} e^{-2\pi isu} \left(\int_{-\infty}^{\infty} g(u+x)f(x)\mathrm{d}x \right) \mathrm{d}u = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} e^{-2\pi isu} g(u+x)\mathrm{d}u \right) f(x)\mathrm{d}x$$

Using the substitution $u + x = t$, we continue with

$$\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} e^{-2\pi is(t-x)} g(t)\mathrm{d}t \right) f(x)\mathrm{d}x = \int_{-\infty}^{\infty} e^{-2\pi ist} g(t)\mathrm{d}t \int_{-\infty}^{\infty} e^{-2\pi i(-s)x} f(x)\mathrm{d}x = \mathcal{F}g(s)\mathcal{F}f(-s)$$

4.2 Heuristic derivation of MLE for Hawkes Process

An important remark is that the intensity function of a point process is formally defined by [7]:

$$\lambda(t) = \lim_{h\to 0} \frac{\mathbb{P}(N(t+h) - N(t) = 1)}{h}$$

where $N(t)$ denotes the number of arrivals by time t. Therefore, the intensity function can act as a probability density. Furthermore, the likelihood L can be described by multiplying the intensities for all events in the span $[0, T]$ and then multiplying by the probability of no arrivals over each of the interarrival intervals:

$$
\begin{aligned}
L(\Omega) &= \exp\left(-\int_0^{T_1}\lambda(s)\mathrm{d}s\right)\lambda(T_1)\exp\left(-\int_{T_1}^{T_2}\lambda(s)\mathrm{d}s\right)\lambda(T_2)\cdots\lambda(T_n)\exp\left(-\int_{T_n}^{T}\lambda(s)\mathrm{d}s\right)\\
&= \exp\left(-\int_0^{T}\lambda(s)\mathrm{d}s\right)\prod_{i=1}^{n}\lambda(T_i)
\end{aligned}
$$

Note the use of the Poisson distribution, which is useful whenever one wishes to describe the probability of having a certain count of events n over some continuous interval. Here $n = 0$ for each interarrival interval. Also note that by performing the integration of the intensity function over each interarrival intervals, we are essentially yeilding the average number of arrivals within each interarrival interval. Which is exactly what y in $\frac{y^n e^{-y}}{n!}$ dictates.

5 Bibliography

1. **Amarasingham A, Harrison MT, Hatsopoulos NG, Geman S**. Conditional Modeling and the Jitter Method of Spike Resampling, *J Neurophysiol* 107:517-531, 2012.

2. **Amarasingham A, Harrison MT, Hatsopoulos NG, Geman S**. Conditional Modeling and the Jitter Method of Spike Resampling: supplement, arXiv: 1111.4296 [stat.ME], 2011

3. **Cormen TH, Leiserson CE, Rivest RL, Stein C.** Introduction to Algorithms, Third Edition, MIT Press, 2009.

4. **Efron B, Tibshirani RJ.** An Introduction to the Bootstrap, Springer: Monographs on Statistics and Applied Probability 57, 1993.

5. **Lahiri SN**. Resampling Methods for Dependent Data, Springer, 2003 6. textbfOsborne

M, Petrovic S, McCreadie R, Macdonald C, Ounis I. Bieber No More: First Story Detection Using Twitter and Wikipedia, *ACM*: 2012.

7. **Rizoiu MA, Lee Y, Mishra S, Xie L.** A Tutorial on Hawkes Processes for Events in Social Media, arXiv: 1708.06401 [stat.ML], 2017.

8. **Shalizi CR**. Advanced Data Analysis from an Elementary Point of View, n.d.