Time series

Tom Rochette <tom.rochette@coreteks.org>

October 11, 2017 — e6ce246f

0.1 Context

0.2 Learned in this study

0.3 Things to explore

- How to match "aperiodic" periods? That is, period that may be separated by some data, or noisy periodic data?
- How can we make use of discontinuity in time (2300 -> 0200) to help detect periodic behavior?
- How should we deal with data that is not recorded at the same time, but close? What about data points that are far?
- Is there a difference between X = 1 to X = 10 with Y = 5 and X = 1, ..., 10 with Y = 5 (2 pts vs 10 pts)?

1 Overview

In this article, we study the properties that are of interest when trying to determine the correlation between two time series of discrete values.

- Time series with different length
- Time series with mismatching x-axis values
- Time series with variable x-axis separation

Condition	1	2	3	4	5	6
Different length	o	o	o	o	x	x
Mismatching x-axis values	o	o	o	o	x	x
Variable x-axis separation	o	o	o	o	o	x

2 1. Time series with similar lengths, matching and constant spaced x-axis values

This is most likely the simplest case. In this case, you can drop the x-axis values as they will not bring in any value.

As we move one series along the x-axis, the cross-correlation will inevitably drop as less as and less data points match one another. In order to prevent that from happening, we would need for the series to be stationary, that is, we can take the series and extend it to the infinite in both directions. If our time series contains n values, then our period is n, which means that we will have to try up to n positions in order to find the one with the cross-correlation closest to one, that is, the correlation is computed as $\rho_{X,Y}(\tau) = \frac{E[(X_n - \mu_X)(Y_{n+\tau} - \mu_Y)]}{\sigma_X \sigma_Y}$ and we want to maximize the

$$\arg\max_{\tau} |\rho_{X,Y}(\tau)|$$

However, in the case our time series is not stationary, we will have to repeat the procedure by testing decreasing time series length down to a given threshold.

3 2. Time series with similar lengths, mismatching and constant spaced x-axis values

Once again, we can probably drop the x-axis values. In this case, the mismatch in x-axis values means that our analysis will be done with a given lag.

4 3. Time series with similar lengths, matching and variable spaced x-axis values

Here our data points match on the x-axis, but it does not mean that the cross-relation between two series is without lag. Furthermore, the addition of the variable spaced x-axis values adds an additional problem to our analysis: since time points a separated by different amount of space, we have to decide what to do:

- Ignore all points that do not match in both series
- Create interpolation points for all missing values in the opposite series

In all cases, we will drop the data point outside of the interval common to both series as it would constitute extrapolation.

As we have to deal with variable spaced x-axis values, if we decide to ignore all points that do not match both series or we create interpolation points for all missing values in the opposite series, what we end up is case 1.

5 4. Time series with similar lengths, mismatching and variable spaced x-axis values

6 5. Time series with different lengths, mismatching and constant spaced x-axis values

7 6. Time series with different lengths, mismatching and constant spaced x-axis values

8 Other ideas

Start from the biggest period possible and reduce it (or the opposite, smallest to biggest)
Generate a diagram of the highest correlation per period size
Stop algorithm as soon as a specific correlation is achieved

Most (if not all data) needs to be normalized [x, y] => [0, 1]

Go for easy to calculate, generic algorithm, to more specific, time consuming algorithms.
If an algorithm can verify a premise, say for instance, that the data has a similar distribution, then it is likely to have a similar period. Likewise, if they have a different distribution, then it is likely that they will not share a similar period.

9 Possible analysis

- Minimum
- Average

- Maximum
- Median
- Distribution analysis
- Seconds (0-59), Minutes (0-59), Hours (0-23), Days (1-31), Week days (1-7), Weeks (1-52), Months (1-12), Years * analysis
- Movement analysis (up/down)

Look into FFT (fast fourier transform) and DTFT (discrete time fourier transform)

10 From theory

Time series are decomposed into three components:

- Trend
- Seasonal
- Random

11 See also

12 References

- https://en.wikipedia.org/wiki/Cross-correlation
- http://www.jasonbailey.net/stuff/wp-content/uploads/2013/04/Time_series_and_fft_big_data_brighton.pdf
- http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics