Associating heterogeneous time series

Master Thesis

Alexandre GUINAUDEAU

# Association between time series

# Defining association for heterogeneous time series

### Properties

In order to find time series which can help us troubleshoot failures, we must first define association in a clear way. To do so, we will start by defining the properties our metric should abide by, and then study metrics which fulfill these requirements.

Given two time series x and y, we note A(x, y) the association between them. Without loss of generality, we can assume that our metric gives a score from 0 to 1.

A good metric to measure association between time series should follow the following properties:

1. A(x, y) = 0 iff x and y are independent
2. A(x, y) > 0 iff L(x | y) < L(x)
3. If L(x) > 0, A(x, y) = 1 iff L(x|y) = 0
4. A(x, y) = A(y, x): A should be associative

## Metric comparison

We will start by studying usual metrics that capture the association between time series and then see how they can be generalized to apply to any time series

**Pearson Correlation**

The most common measure of correlation is the one defined by Pearson. To capture both correlated and anti-correlated time series we actually consider its square value \rho^2. It verifies all implications (=>) of the above properties: a high Pearson correlation indicates that the knowledge of a series increases our knowledge of the other. However, it only captures linear correlation, so we could miss association between series with non-linear correlation.

**Spearman and Kendall Correlations**

Spearman and Kendall both defined correlation metrics based on the ranking of elements rather than their exact value. While they capture non-linear correlations, they are less impacted by outliers which are important to associate time series – time series having outliers simultaneously are likely to be associated.

**Mutual Information**

If we consider categorical data, measures of association usually involve the information entropy. The metric that measures shared entropy is the mutual information. Its definition can be generalized to numerical data using approximations, for instance based on the k nearest neighbors.

***Bayesian definition of proba correlation: 0=random / 1=best split – EM algorithm***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Measure | P1 => | P1 <= | P2 => | P2 <= | P3 => | P3 <= | P4 |
| Pearson | T | T | T | F | T | F | T |
| Kendall | T | T |  |  | T |  | T |
| MI | T | T | T | T | T | T | T |
|  |  |  |  |  |  |  |  |

# Application to examples

### Categorical and event data

Our first example compares an event time series and a categorical one. This could for instance be useful to determine whether an installed part tends to fail more often than another. We will start with the case of 2 categories and then generalize to any number of categories.

Let x be a time series of events, that occur with probability p1 during t1 and then at frequency p2 during t2. Let t=t1+t2 the total time and p=p1\*n1+p2\*n2/n1+n2 the average probability of failure. Let y be a categorical time series with value C1 during t1 and C2 during t2,

**Pearson**

In order to compute the pearson correlation, we need the categorical data to have numerical values, we set C1=1 and C2 = 0, the pearson factor is equivalent to the $\chi^2$ test in this case:

\rho = …. =

**Null hypothesis**

We can formulate this as a null hypothesis problem. The null hypothesis is  
H0: “All the events have been drawn with the same probability p”

The z-test associated with this hypothesis is

Z = …

We reject the null hypothesis if Z > $\alpha$. We note that Z is proportional to pearson factor as $Z = \sqrt(n)\*\rho$. However, the null hypothesis gives us a good way to interpret the association score, when $Z > alpha\_{0.99}$ we can reject the null hypothesis with 99% confidence, in other words the probability of occurrence if very likely to have changed at some point.

**Spearman**

In this case, Spearman and Kendall correlations are equivalent to Pearson correlation because both series can only take 2 different values, so sorting them gives the same result as taking their absolute value

**Mutual information**

We note n\_{tot} = and p\_{tot} =

The mutual information is H\_{1} + H\_{2} – H\_{tot} where H\_{i} =

**Ranking candidate time series**

***Theory*** Let’s consider other candidate time series yi with value C1 during t1’ and C2 during t2’=t-t1’. A good association metric should rank yi time series such that y\_{t\_1} has the highest association with x. In this example, all metrics studied here find the same optimal associated time series in theory, because p1’-p2’ decreases for values of t1’ close to t2’, and 1/t1’ or 1/t2’ is very large for extreme values of t1’, so z’ < z

***Practice*** In practice, we used the Monte-Carlo method to compute the error between the expected best candidate and the actual one. We note that results are pretty good, even for small differences between p1 and p2 where it’s even hard for a human eye to detect the change in probability. The main bias is towards extreme values of t1’, when the random samples had an unusually high (or low) frequency at the beginning or at the end. However, using the Null Hypothesis method we can ignore the best candidate when the hypothesis is rejected, in which increases the accuracy of the model.

**Generalization**

We can generalize the results above to any categorical data, and to events with any number of change of frequency. Using the null hypothesis, we can recursively split the event time series into phases where the frequency of occurrence is statistically constant. This can be used to preprocess event time series.

The mutual information formula remains valid for any number of categories. It biases towards many categories, as the mutual information is always strictly positive. To balance this, we could penalize the mutual information such as defined in [A penalized mutual information criterion for blind separation of convolutive mixtures Mohammed El Rhabi, Guillaume Gelle, Hassan Fenniri, Georges Delaunay]. To keep our algorithm efficient enough to compare thousands of sub-second time series, we decided not to penalize the mutual information.

### Event and numerical data

For numerical data, we cannot expect to have a metric that computes the association in the right way on any raw dataset, because the association depends on what the data represents. Events can be triggered by extreme values or by sudden changes, they can depend on the original value, or on the de-trended or de-seasoned value. Events could also have an impact in the future, or for a specific duration. Therefore, we assume the data is preprocessed – potentially de-trended, de-seasoned, shifted, smoothed, derived, normalized – such that large values have a higher risk of triggering an event. This is not a real loss of generality, because one can derive multiple preprocessed series out of the original one (for instance using all the methods described above), and the algorithm will only find candidates with high association with the event series.

Let’s define a simple example once again, where we can define the expected behavior of the association metric.

Let x be a time series of events, and y be a series of Gaussian white noise, which could be the residual of a de-seasoned and de-trended series. When an even occurs, y has a higher value in average:

{y = N(0, 1) if x=0 else N(0, 1)+1}

In this case, we expect x and y to be highly associated. Furthermore, if we define other numerical series that are impacted by a fraction of the events, or by additional events, they should have a lower association with x.

**Pearson**

Pearson correlation can be applied in this case, as all the data is numerical. However, the correlation score will be quite low even when there is a high association because the correlation is not linear.

**Null hypothesis**

A hypothesis we could refute here is that both groups (the values of y when x occurs and its values when x does not occur) were drawn from the same distribution. This should give good results but requires a prior kind of distribution to be defined, so it could only be applied to normal distributions.

**Spearman**

In this case, the Spearman correlation can give better results than the Pearson correlation because it captures non-linear correlations. However, it will bias towards series that are impacted by few events, as it will consider the association to be perfect if the events $x$ correspond exactly to the largest values of y.

**Mutual information**

The previous definition of mutual information was defined for categorical data. For numerical data, the joint probability of two variables is also well defined if the underlying distribution is known:

[insert formula here]

However, it is possible to approximate this underlying distribution by considering the k-nearest neighbors of each point. Given k, for each point P, we compute the distance \varepsilon to its k-nearest neighbor. If the underlying distribution \mu was uniform, then the probability of being in the ball of center P and of diameter \varepsilon would be \mu \times V, where V is the volume of this ball. By deriving this expression, we get an approximation of the entropy:

[insert here]

And by extension of the mutual information between two numerical series. This expression is only valid for numerical data where all values are different (or at least where there are never more than k-1 elements with the same value) – which can be circumvented by adding a small random noise.

**Ranking candidate time series**

While all the above methods