Cross-Correlation Functions

The aim of a correlation function is to determine how similar two time series are and also to determine how much you would have to shift one timeseries in time to make it line up with another timeseries. This shift in time is known as a lag. We need a quantitative way of measuring how similar two timeseries are.

There are two main types of correlation function: the Interpolation Correlation function and the Discrete Correlation Function (Edelson and Krolik 1988, ApJ 333, 646).

The Interpolation Cross-Correlation Function

Look at my diagrams on the Teams chat.

1. Measure the mean and standard deviation for both time series, X and Y.
2. For each datapoint, measure its distance from the mean, $x - < x >$.
3. Shift Y by variable time τ.
4. Time t on X now lines up with $t - \tau$ on Y.
5. Multiply $x(t) - < x >$ by $y(t - \tau) - < y >$
6. The interpolation part comes in when sampling is not entirely even (ie, very often) and so, after shifting, there is no value of Y at the exact time $t - \tau$. In that case you have to interpolate between the values of Y on either side of $t - \tau$.
7. Repeat for all values of t; sum the result and divide by N-1 where N is number of data pairs.
8. Divide the answer by the product of the two standard deviations. That normalises the answer so that it won't depend on units.
9. The above gives you a value for ICCF(τ). To calculate the full correlation function, repeat for all values of τ, both positive and negative.
10. For a perfect correlation, the value of the ICCF is $+1$.
11. An autocorrelation function (either interpolation or discrete) is provided when you input the same time series twice, ie X and X, not X and Y. If the code is correct, that should give a value of $+1$ at $\tau = 0$, ie zero lag.
12. The width of the correlation function gives an indication of the memory of the system, ie for how long a data point remembers what previous data points were. You can take any measurement but FWHM is commonly used.

The Discrete Cross-Correlation Function

1. The difference here is that we do not shift Y.
2. Go to the first X point. Measure $x_1 - < x >$ as above. Do the same for first Y point.
3. Measure the time difference between the first X and the first Y point, τ_{11}.

4. Calculate the same statistic as above for that one pair of points, ie

$$DCF(\tau_{11}) = \frac{(x_1 - < x >)(y_1 - < y >)}{\sigma_x \sigma_y}$$

5. Save the values of τ_{11} and DCF(τ_{11}) in a 2-column list.

6. Staying with the first X point, repeat the process for all of the other Y points. Save the resulting values of DCF and τ in your list.

7. Now move to the second X point and repeat the whole process. Then repeat for all other X-points.

8. You can now plot your many points of DCF(τ) vs τ. That is the unbinned DCF. You will note that there is a large scatter of points, many exceeding the values of $+1$ or -1. That is normal.

9. However now sort your large list of DCF points into τ order. Select suitable sized bins for τ. Calculate the mean and standard deviation of the DCF in each bin and plot the results.

Points to Note

..which I shall probably continue to update

1. There are pros and cons to the two methods. The DCF is probably easier to code as it doesn't involve the interpolation step. It also introduces no additional information so is technically more correct. However it is not as sensitive as the interpolation method which, unsurprisingly, produces smoother results with less noisy looking correlation functions.

2. For our purposes use the directly measured standard deviation, straight from the data. You will see, in the Edelson and Krolik paper, a discussion of the 'error-subtracted variance', and the standard deviation, of course, is the square root on the variance.**Do not use error-subtraction here.**

   The error-subtracted variance takes away the observational errorbars. This provides a useful number if you want to know what the true underlying variance is. For example, if you measure a constant source, eg the X-ray emission from a faint cluster of galaxies which ought to be rock steady, many times, you will have statistical measurement error based solely on the number of photons detected. Thus you will measure a positive variance which is not a true measurement of the source variability. In this case you can also determine the size of the observational measurement errors from the number of photons. This error is exactly the same as the total error. So if you subtract the 'average observational error squared' from the total observed variance, you get the true source variance which, in this case, would be zero.

   Although not used here, we often divide the square root of the error-subtracted variance by the mean. That is the 'fractional variability', F$_{var}$.

   The correlation functions we want here should represent how closely two observationally determined time series are. If we divided by the square root of the error-subtracted variance, which could be very small, then the peak of the correlation function might be much larger than 1 and would be hard to interpret.

3. **The significance of the correlation** is not simply given by the peak value of the CCF, although in general high CCF peaks tend to be more significant. The peak value just tells you how much of the two time series are in common. You might have two time series, each comprising the sum of more than one process, but where one process is in common with both time series and the others

are not. That might give a very low peak CCF. However if you have enough good quality data, you might still be able to obtain a highly significant measurement of the correlation of that one component. The only way to properly calculate significances is with simulations, which we won't do here.

4. **Error on the lag:** It is often hard to measure precisely what the best lag is, and what the error on that lag is. Again, we require simulations and we won't do it here.