Cointegration Kernel with Tom Lee and Paul Lintilhac

In this project, we define a “Cointegration Kernel” that can be used to analyze time-series that are “causally related”. The overall motivation for this research is to give an alternative approach for analyzing large numbers of time series. The basic setting of such a problem might be like this: you are working at a company with a large amount of private data in the form of time-series, and you want to know if any of your time series might be predictive of important external economic indicators, financial indexes, or any other potentially valuable signals. The classic approach to this would be to look for high leading and lagging correlations between any of your internal time series and any external timeseries.

While this problem in general suffers from extremely high computational complexity when there are a large number of time-series involved, certain methods such as the Discrete Time Warp have been used that can leverage FFT in order to efficiently handle the added complexity of calculating cross correlations by viewing this calculation as a convolution, which can be represented as an inner product in Fourier space. In recently released white-papers, Google has demonstrated the use of multi-level clustering techniques in their Google Correlate product in order to pre-process the time series so that querying for nearest neighbors is much more efficient. Recently, a paper by Cuturi in 2007 suggests how to define a PDS Global Alignment Kernel based on the DTW algorithm, which allows us to implement the DTW as the kernel of an SVM.

However, even with all of these efficiency gains, we note that if we are under the assumption of a large number of time-series, then no matter how efficient our algorithms are, there is a very significant chance that any highly correlated time-series could be spurious correlations. Therefore we may want to impose an additional constraint that the time-series be co-integrated. Often time series that are not only correlated but also cointegrated are considered to have a stronger causal relationship than two time series with the same correlation but no significant correlation. This is because their relationship appears to have some “memory”, in that a certain linear combination of the time-series is error-correcting.

The classical test to see whether two time-series X and Y are cointegrated is the Engle-Granger, which consists of two parts:

1. Regress X on Y using least-squares, and then take the residual. We can define this residual explicitly using the normal equations:
2. Next, we run a unit-root test on in order to determine if it is stationary. If it is, then we say the two series X and Y are cointegrated. This usually takes the form of an Augmented Dickey-Fuller Test. In our case, we opt for a straightforward calculation of the autocorrelation with lag

Left to do:

* this process as an inner product of certain features of X and Y which are symmetric, i.e. such that our kernel
* prove that this kernel is PDS
* apply to economic data (e.g. moody’s or bloomberg) are compare results to a standard cross-correlation or DTW