Clustering of Big Data Time Series

The current revolution in Artificial Intelligence, Machine Learning, and Big Data presents new possibilities and challenges for researchers and analysts alike. This is particularly true in the case of time series, as for many domains, analysts now have access to collections of very long time series in many areas of interest, such as astronomy, geophysics, medicine, social media, economics and finance. These researchers and analysts are challenged when they are required to compare and cluster such long and diverse time series. On the whole, it is not usually possible to use traditional methods of analysis for these tasks, such as estimating models and comparing features, as these methods imply lengthy computations, including the inversion of extremely large matrices.

Caiado, Crato, and Poncela (2020) proposed a spectral method of synthesizing and comparing time series characteristics which is nonparametric and is focused on the data's cyclical features. Instead of using all the information available from the data, which is computationally very time-consuming, this procedure uses regularization rules to select and summarize the most relevant information for clustering purposes. This method does not imply the computation of the full periodograms, but only that of the periodogram components related to frequencies of interest, comparing and clustering the respective periodogram ordinates for the various time series using common clustering methods. They named this approach the 'fragmented-periodogram method'.

More recently, Albino, Caiado and Crato (2023) proposed a new approach for clustering big data time series which can be considered to be an alternative to the periodogram method in the case of time series: the 'fragmented-autocorrelation based method'. Essentially, these authors suggest using the autocorrelation function of time series which is only computed for the lags of greatest interest. In a large Monte Carlo simulation study, they explore whether this procedure is able to condense the relevant information of the time series. This method is illustrated in an empirical study of the evolution of various stock market indices.