Time Series Clustering and Classification Methods

Himanshu Jha and Radhakrishnan Ravi Vignesh

Introduction

The main objective for using clustering analysis is to group or cluster data that are more similar (based on a user defined similarity matrix) than those in the other groups for extracting features/patterns from complex datasets. With advent of the era of big data, clustering methods have been gaining tremendous popularity across multiple disciplines such as statistics, medicine, signal processing, engineering, and computer science and have become a key part of exploratory data analysis.

Multiple algorithms/approaches have been proposed for carrying out the task of clustering analysis, and the choice of algorithm/approach depends on the general task to be solved. However, at their core, all clustering analysis methods rely on utilizing a distance matric/measure of dissimilarity which is used to classify/divide datasets into different clusters. The choice of this distance matric/ measure of dissimilarity depends on the task at hand and the ultimate objective.

One of the most common choice of dissimilarity used in the clustering analysis is some sort of distance measure (Euclidean, Manhattan, etc.) However, using the distance between the i^{th} point on time series 1 and i^{th} point of time series 2 will often produce poor measure of dissimilarity (think identical time series shifted by some value in time or in magnitude, etc.) A wide variety of alternative measures of dissimilarity have been proposed for clustering time series data. In first half of this report, we focus on some of the popular methods for clustering/classifying complex time series data in the frequency domain and in the second part of the report, we will look at a popular machine learning algorithm for doing the same. The referenced approaches do not constitute a comprehensive list by any measure. They are merely our attempt at discussing various approaches that anyone may find while analyzing the time series data.

1. Time series clustering and classification via frequency domain

Time series clustering and classification methods via frequency domain use the covariance structure or equivalently the spectral density function and its sample version, the periodogram, for defining various multivariate measures of disparity between any two-time series. We focus primarily on understanding the clustering of linear stationary time series for this project.

1.1 Conditions for a valid measure of dissimilarity

a) Nonnegativity ($D(x, y) \geq 0$)
b) Symmetry ($D(x, y) = D(y, x)$)
c) Reflexivity ($D(x, x) = 0$)
d) Identity ($D(x, y) = 0\ if\ and\ only\ if\ x = y$)
e) Triangle identity ($D(x, z) \leq D(x, y) + D(y, z)$)

A quasi-distance matric satisfies all but the triangle inequality.

1.2 Estimation of spectral density from periodogram

Spectral density can be estimated from the rough sample estimate of a population function called periodogram. The periodogram is a “rough” estimation of the spectral density because it is only calculated at discrete fundamental frequencies, whereas spectral density is defined over a continuum of frequencies.

Two of the most common methods of estimating spectral density from the periodogram are

a) Centered moving average – Daniell kernel with parameter m is a centered moving average which creates a smoothed value at time t by averaging all values between $t - m$ and $t + m$. Modified Daniell kernel uses for a weighting coefficient scheme for calculating the centered values. It is a non-parametric method as it does not use any parametric model for the underlying time series.

b) Parametric estimation of spectral density (requires detrending of the time series data first) – parametric estimation of the spectral density from the periodogram by finding the best fitting AR model for the series and plotting the spectral density of that model. This model is supported by the theorem which says that the spectral density of any time series can be approximated by spectral density of an AR process.

1.3 Measure of dissimilarity for linear stationary time series

1.3.1 Kullback-Leibler (KL) discrimination information – The Kullback-Leibler (KL) discrimination information (Kullback and Leibler 1951; Kullback 1978) is given by

$$I(p;q) = E_p\left\{\log\frac{p(X)}{q(X)}\right\}$$

where E_p denotes the expectation under the density $p(.)$. The KL discrimination information takes the form

$$I(p;q) = \frac{1}{2}\left(tr\{R_p R_q^{-1}\} - \log\frac{|R_p|}{|R_q|} - mT\right)$$

when $p(X)$ and $q(X)$ correspond to competing zero-mean multivariate normal distributions. A symmetric measure of disparity ($quasi - distance$), the $J\ divergence$, is defined as

$$J(p;q) = I(p;q) + I(q;p)$$

1.3.2 Chernoff information – Chernoff information (Chernoff 1952; Renyi 1961) measure proposed by Perzen (1990) for measuring the difference between the two densities is given by

$$B_\alpha(p;q) = -\log E_p\left\{\left(\frac{q(X)}{p(X)}\right)^\alpha\right\}$$

For two normal random vectors differing only in the covariance structure, the foregoing measures takes the value

$$B_\alpha(p;q) = \frac{1}{2}\left(\log\frac{|\alpha p(X)| + |(1-\alpha)q(X)|}{|q(X)|} - \alpha\log\frac{|p(X)|}{|q(X)|}\right)$$

where the measure is indexed by $\alpha, 0 < \alpha < 1$. Chernoff information measure tends to behave like the two Kullback-Leibler measures near the two extreme values of α. $B_\alpha(p,q)$ scaled by $\alpha(1-\alpha)$ converges to $I(p;q)$ for $\alpha \to 0$ and $I(q;p)$ for $\alpha \to 1$. A quasi distance may be defined by letting

$$B_\alpha(p;q) = B_\alpha(p;q) + B_\alpha(q;p)$$

The α value is chosen such that it maximizes the dissimilarity between group populations by searching $B_\alpha(p;q)$ over the range of values of α.

1.3.3 Ravishankar et al. (2010) proposed a quasi-distance from the likelihood ratio and profile likelihood ratio test statistics for testing pairwise equality of L spectral matrices via $H_0\colon p(X) = q(X)$. The results were derived because the limiting distribution of the smoothed periodogram $\widehat{p(X)}$ is related to a complex Wishart distribution with $(2M + 1)$ degrees of freedom (M being the raw periodogram ordinates in the neighborhood of ω_n) and scale parameter $p(X)$ and $q(X)$ to construct a test statistic Q^* which is an average of the S smallest values of

$$Q(\omega) = 2^{2p(2M+1)} \frac{\left|\widehat{p(X)}\right|^{2M+1}\left|\widehat{q(X)}\right|^{2M+1}}{\left|\widehat{p(X)} + \widehat{q(X)}\right|^{2(2M+1)}}$$

where S should be chosen to large enough to smooth the local variations but small enough to characterize the overall spectrum (usually in the range of $\left[\sqrt{T}\right]$to $[T/10]$ where $[a]$ denotes the floor of a). The associated pair-wise quasi-distance is calculated by –

$$D\big(p(X), q(X)\big) = 1 - [Q^*]^{1/(2M+1)}$$

1.3.4 Hierarchical Spectral Merger Algorithm

The Hierarchical Spectral Merger Algorithm given by Euan et. al. (2018) [1] uses a dissimilarity measure called Total Variation Distance as the dissimilarity measure between two spectral densities (p and q). The TV Distance is calculated by 1 - Area under the curve formed by minimum of both the spectral densities at each frequency

$$TV\ Distance\ (p, q) = 1 - \int \min\big(p(\omega), q(\omega)\big)\, d\omega$$

which is then used for the hierarchical clustering of the time series. The algorithm starts with initializing each time series as a cluster. Then, two clusters with the least dissimilarity are combined to form a single cluster. Then, the spectral estimate of the cluster is obtained by the weighted average of individual time series. The process continues until the dataset is reduced to the required number of clusters.

The required number of clusters is obtained by an empirical method and Bootstrapping method in Euan et al. (2018) [1]. The empirical method works by calculating the elbow point of the inverse relationship curve obtained between minimum TV Distance and the number of clusters. The Bootstrapping method approximates the TV Distance distribution between the spectral densities and helps us choose a threshold for the TV Distance. A hypothesis testing is performed against adding another cluster with the help of Bootstrap quantiles to calculate the threshold.

The example discussed in the paper analyzes the height of the ocean waves. The wave height was non-stationary because of storm in the region. So, the time series was split into 30 minutes segments. Clustering is performed on these 30-minute chunks using the Bootstrap procedure

Euan et. al (2019) [2] have clustered ocean waves based on the method discussed above with a slight modification. A function representing the directional distribution of energy at different frequencies is multiplied with the existing spectral density to include the effects of direction.

2 Time series clustering and classification using dynamic time warping

Dynamic time warping (DTW) is a very popular algorithm for measuring the similarity between two time series by using a non-linear mapping of one signal to another by minimizing the distance between the two under following restrictions (Berndt and Clifford, 1994):

a) Every index from the first time series must be matched with one or more indices from the other sequence and vice versa.
b) Boundary condition – the start and the end points of the warped path must be the first and the last points of the aligned time series.
c) Monotonicity condition – The time order of the points in two time series must be monotonically increasing.

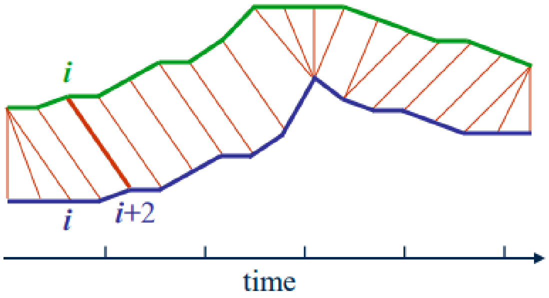


Figure 1: Non-linear mapping between two time series

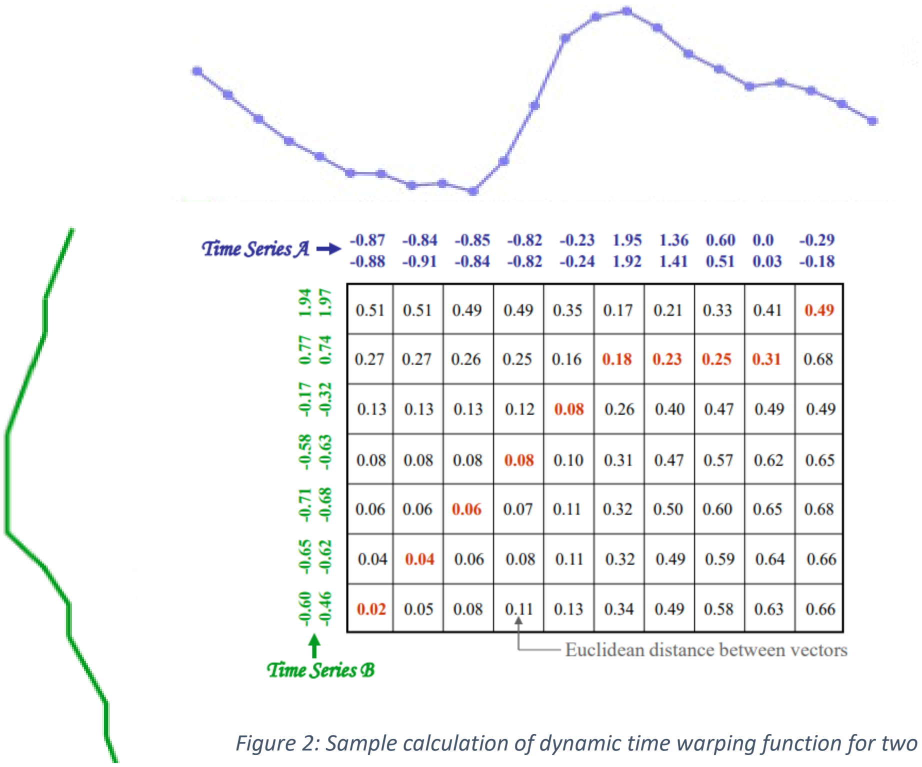
The optimal path alignment between the two-time series is the path that minimizes the warping cost

$$DTW(p, q) = \min \left\{ \sqrt{\sum_{k=1}^{K} w_k} \right.$$

Where w_k the matrix element $(i, j)_k$ that belongs to the k^{th} element of a warping path W, a continuous set of matrix elements that represent the mapping between p and q. The warping path can be found using the dynamic programming to evaluate the following recurrence

$$\gamma(i, j) = d(p_i, q_i) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\}$$

Where $d(i, j)$ the distance found in the current cell, and $\gamma(i, j)$ is the cumulative distance of $d(i, j)$ and the minimum cumulative distances from the three adjacent cells.



Time Series A →	-0.87 -0.88	-0.84 -0.91	-0.85 -0.84	-0.82 -0.82	-0.23 -0.24	1.95 1.92	1.36 1.41	0.60 0.51	0.0 0.03	-0.29 -0.18
1.94 1.97	0.51	0.51	0.49	0.49	0.35	0.17	0.21	0.33	0.41	**0.49**
0.77 0.74	0.27	0.27	0.26	0.25	0.16	**0.18**	**0.23**	**0.25**	**0.31**	0.68
-0.17 -0.32	0.13	0.13	0.13	0.12	**0.08**	0.26	0.40	0.47	0.49	0.49
-0.58 -0.63	0.08	0.08	0.08	**0.08**	0.10	0.31	0.47	0.57	0.62	0.65
-0.71 -0.68	0.06	0.06	**0.06**	0.07	0.11	0.32	0.50	0.60	0.65	0.68
-0.65 -0.62	0.04	**0.04**	0.06	0.08	0.11	0.32	0.49	0.59	0.64	0.66
-0.60 -0.46	**0.02**	0.05	0.08	0.11	0.13	0.34	0.49	0.58	0.63	0.66

Figure 2: Sample calculation of dynamic time warping function for two time series

3 Experiment

3.1 Results from simulation study

For comparing the clustering algorithm using spectral density method (HSM) and DTW, we simulated 18 complex AR(2) processes given by $X_t = 2 \times 0.95 \cos \omega_s X_{t-1} - 0.95^2 X_{t-2} + \epsilon_t$ for $\omega_s = 0, \frac{2\pi}{17}, \frac{3\pi}{17}, \dots, \pi$.

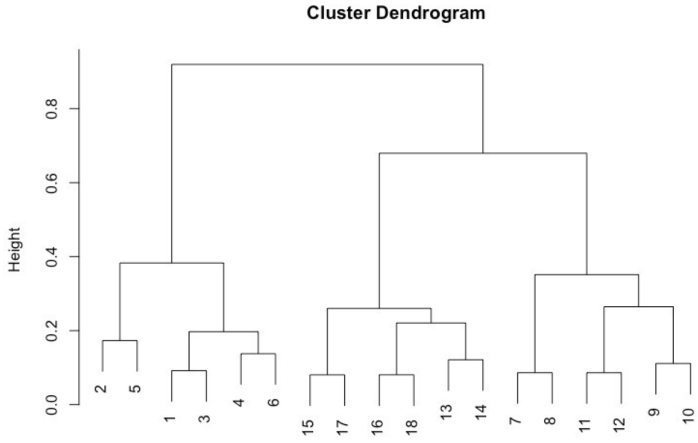


Figure 3: Result of HSM clustering on simulated AR(2) process

We can see that HSM clustering algorithm successfully clusters time series which are closer to each other in terms of ω_s. This helps us conclude that HSM clustering is successfully able to cluster with respect to frequency.

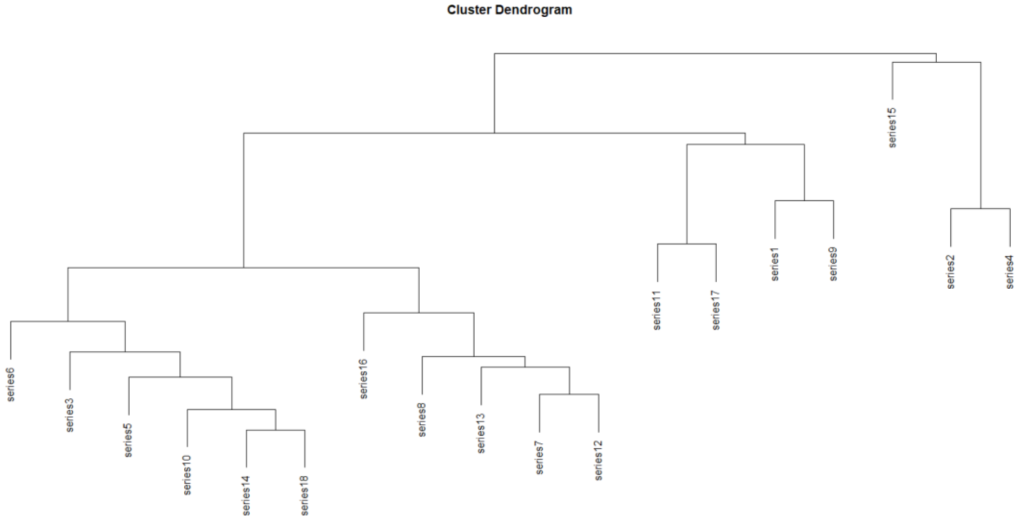


Figure 4: Result of DTW clustering on simulated AR(2) process

Since DTW clustering algorithm clusters based on the shape of the time series, we can see that it is not able to place AR(2) processes with closer frequencies together.

3.2 The stock prices of different companies listed in NYSE and NASDAQ are clustered using the HSM function in HMClust package in R.

It is an implementation of Euan et al., 2018[1] in which a Parzen window is used to smoothen the spectral density. As the HSM clustering is suitable for only linear stationary time series data, the first difference of the logarithm of the stock prices i.e., Log returns is calculated to obtain a stationary time series. We also carry out clustering using the Dynamic Time Warping procedure using the raw time series data as we wish to use the similarity in the shapes of the times series as a measure for clustering. The companies are represented by their Ticker symbols as listed on the two exchanges in the following two Dendrograms.

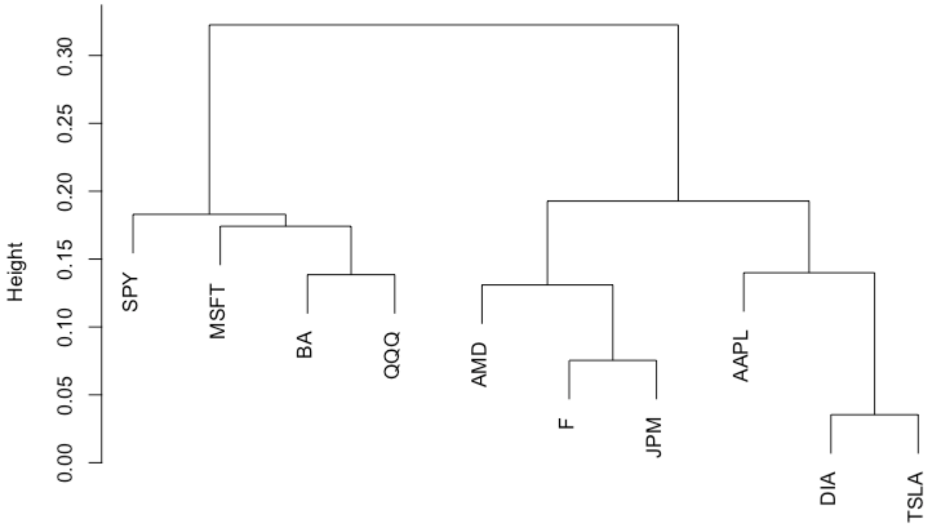


Figure 5: Result of HSM clustering on stock market data

HSM clustering clusters the assets based on their trading frequency and volatility.

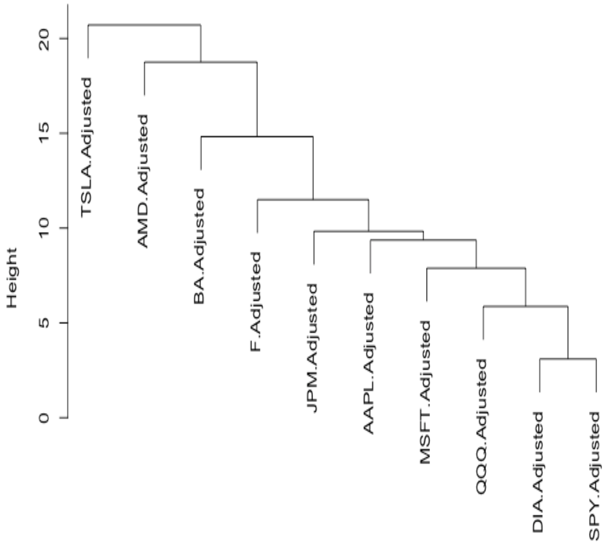


Figure 6:Result of DTW clustering on stock market data

From the dendrogram generated using DTW clustering, we can see that all ETFs are placed closest to each other due to the similarity in their shapes.

4. Conclusion

In this study, we have explored different measures of dissimilarity in frequency domain and one measure of dissimilarity with respect to shape of the time series. The choice of measure of dissimilarity depends on the general task to be solved and problem objective. If the objective is to cluster the time series data based on the underlying data generating process, dissimilarity measures in the spectral density function could be a good choice for the clustering. However, if the objective is to find similarity in general shape of the time series, DTW may be a better choice.

Reference

Kullback S., and Leibler, R. A. (1951)," On Information and Sufficiency," Annals of Mathematical Statistics, 22, 79-86.

Kullback, S. (1978), Information Theory and Statistics, Gloucester, MA: Peter Smith

Chernoff, H. (1952), "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of the Observations," Annals of Mathematical Statistics, 25, 573-578.

Renyi, A. (1961), "On Measure of Entropy and Information," in Proceeding of 4th Berkeley Symposium on Mathematical Statistics and Probability, 1, Berkeley: University of California Press, pp. 547-561.

Parzen, E. (1990), "Time Series, Statistics and Information," IMA Preprint Series 663, Institute for Mathematics and Its Applications, University of Minnesota.

Ravishankar, N., Hosking, J. R. M., & Mukhopadhyay, J. (2010). Spectrum based comparison of multivariate time series. Methodology and Computing in Applied Probability, 12, 749-762.

Euán, C., Ombao, H., & Ortega, J. (2018). The hierarchical spectral merger algorithm: a new time series clustering procedure. Journal of Classification, 35(1), 71-99.

Euán, C., & Sun, Y. (2019). Directional spectra-based clustering for visualizing patterns of ocean waves and winds. Journal of Computational and Graphical Statistics, 28(3), 659-670.

Berndt, D., & Clifford, J. (1994). Using Dynamic Time Warping to Find Patterns in Time Series. KDD Workshop.

R packages in used this study

https://rdrr.io/github/CarolinaEuan/HMClust/src/R/HMClust.R

https://rdrr.io/github/CarolinaEuan/HMClust/man/spec.parzen.html

http://www.rdatamining.com/examples/time-series-clustering-classification