Time Dependent Clustering of Time Series

Pinto da Costa, Joaquim F.
Faculdade de Ciências da Universidade do Porto, Departamento de Matemática Aplicada
Rua do Campo Alegre,687
4169-007 Porto, Portugal
E-mail: jpcosta@fc.up.pt

Silva, Isabel
Faculdade de Engenharia da Universidade do Porto, Departamento de Engenharia Civil
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
E-mail: ims@fe.up.pt

Silva, M. Eduarda
Faculdade de Ciências da Universidade do Porto, Departamento de Matemática Aplicada
Rua do Campo Alegre,687
4169-007 Porto, Portugal
E-mail: mesilva@fc.up.pt

In this work we consider the problem of clustering time series. Contrary to other works on this topic, our main concern is to let the most important observations, for instance the most recent, have a larger weight on the analysis. This is done by defining a similarity measure between two time series, based on Pearson's correlation coefficient, which uses the notion of weighted mean and weighted covariance, where the weights increase monotonically with the time.

As pointed out by Caiado et al. (2006), a fundamental problem in the clustering of time series is the choice of a relevant metric. For us, two time series are similar to each other, and should therefore fall into the same cluster, if their evolution over time shows similar characteristics. Consider for instance the example in Beringer and Hüllermeier (2006), where two stocks both of which continuously increase between 9:00 AM and 10:30 AM but then started to decrease until 11:30AM are considered similar, no matter what their absolute values are. That is to say that what interest us is not the distance between two time series but the distance between their "profiles", which in our case consist in the standardization of the two time series. We will start thus by deriving an expression for this distance.

Let $E = \{X_1, X_2, \ldots, X_n\}$ be a set of n time series each one with t observations and $X_i = (x_{i1}, x_{i2}, \ldots, x_{it})'$ and $X_l = (x_{l1}, x_{l2}, \ldots, x_{lt})'$ represent the values of two time series which, without loss of generality, started in time 1 and are currently in time t. Our dataset is represented by a nxt matrix $\mathbf{X}_{nxt}$ of real numbers whose lines represent the n time series and the columns the observation times; thus, X_i and X_l are two lines of this matrix.

Let us start by standardizing the data:

$$(1) \quad x_{ij} \leftarrow \frac{x_{ij} - \overline{x}_{i\bullet}}{s_{i\bullet}},$$

where $\overline{x}_{i\bullet} = \frac{1}{t}\sum_{j=1}^{t} x_{ij}$ is the usual average of the values of time series X_i; that is, the average of the values inside the line of the data matrix which corresponds to time series X_i. $s^2_{i\bullet} = \frac{1}{t}\sum_{j=1}^{t}(x_{ij} - \overline{x}_{i\bullet})^2$ is the variance of time series X_i. The usual squared Euclidean distance between the normalized values of the time series X_i and X_l is

$$\sum_{j=1}^{t}\left(\frac{x_{ij} - \overline{x}_{i\bullet}}{s_{i\bullet}} - \frac{x_{lj} - \overline{x}_{l\bullet}}{s_{l\bullet}}\right)^2 = \sum_{j=1}^{t}\left\{\frac{(x_{ij} - \overline{x}_{i\bullet})^2}{s^2_{i\bullet}} + \frac{(x_{lj} - \overline{x}_{l\bullet})^2}{s^2_{l\bullet}} - 2\frac{(x_{ij} - \overline{x}_{i\bullet})(x_{lj} - \overline{x}_{l\bullet})}{s_{i\bullet}s_{l\bullet}}\right\}.$$

This equation gives $2t(1-r)$, where r is the Pearson's correlation coefficient between the two series X_i and X_l and so we conclude that the squared Euclidean distance between two standardized time series is proportional to $1-r$.

As it is clear from the above expressions, the sample mean and variance give the same importance (weight) to all the values of the time series, namely $\frac{1}{t}$. However there are situations where this should not be the case; particularly with time series data. It is frequent that with this kind of data the most recent values should be given higher weight, as they are most important for the analysis. Consider for instance again the situation of the two stocks mentioned above. It is common that investors want to know which stocks are correlated but the recent behavior of the stocks is certainly more important for them then what happened one year ago, let's say. In order to take this into account, we will define now a weighted measure of correlation between two time series. Let us start by defining the weighted moments of mean and variance of time series X_i by

$$(2) \quad \overline{x}_{Pi\bullet} = \sum_{j=1}^{t} w_j x_{ij}, \;\; s^2_{Pi\bullet} = \sum_{j=1}^{t} w_j (x_{ij} - \overline{x}_{i\bullet})^2,$$

where the weights w_j are such that $w_j \geq 0$ and $\sum_{j=1}^{t} w_j = 1$. If now we use a weighted Euclidean distance between the weighted standardizations of the time series X_i and X_l we get

$$\sum_{j=1}^{t} w_j \left(\frac{x_{ij} - \overline{x}_{i\bullet}}{s_{Pi\bullet}} - \frac{x_{lj} - \overline{x}_{l\bullet}}{s_{Pl\bullet}} \right)^2 = 2(1 - r_P), \text{ where}$$

$$(3) \quad r_P = \frac{\sum_{j=1}^{t} w_j (x_{ij} - \overline{x}_{i\bullet})(x_{lj} - \overline{x}_{l\bullet})}{\sqrt{\sum_{j=1}^{t} w_j (x_{ij} - \overline{x}_{i\bullet})^2} \sqrt{\sum_{j=1}^{t} w_j (x_{lj} - \overline{x}_{l\bullet})^2}}$$

is a weighted measure of correlation between the two time series X_i and X_l. Now, instead of using the dissimilarity $d = 1 - r$ we can use $d_1 = 1 - r_P$ to define the distance between the time series. On the other hand, if instead of transformation (1) we start by doing the data transformation,

$$(4) \quad x_{ij} \leftarrow \frac{\sqrt{w_j}(x_{ij} - \overline{x}_{i\bullet})}{s_{Pi\bullet}},$$

(similarly for time series X_l) and then we use the usual squared Euclidean distance, the result will be the same.

As in this work we want to give higher importance to the most recent observations, we will use weights like $w_j = j$, $w_j = j^2$ and $w_j = \alpha^j$, for a suitable choice of α (in this work $\alpha = 1.3$). In this work we want to give higher importance to the most recent values but in other situations or for other types of data we might want to privilege other observations. All that is needed is to choose an appropriate weight function.

Weighted Clustering of Time Series

The aim of cluster analysis is to find a structure, if it exists, in a dataset, which means to group similar elements in the same cluster and dissimilar elements in different clusters. In our case we want to cluster the n time series in homogeneous clusters. One of the fundamental aspects of cluster analysis is the definition of a proper similarity or dissimilarity index between the elements to be clustered. In our case we choose the indices described in the previous section, which are metrics between time series. There are essentially two types of clustering methods: hierarchical and partitional. We will use a very well known partitional method, namely the K-means with some adaptations to make it able to choose the number of clusters (see Lerman et al., 2002).

Applications

In this section we will apply our weighted clustering in order to analyse a time series dataset that consists of 20 time series with 309 data points, about the Industrial Production (by Market Group) indices in the United States, from January 1977 to September 2002 (source: http://www.economagic.com). This dataset has originated another dataset, the one used in Caiado et al. (2006) which contains a specific transformation ($y_{ij} = \log(x_{ij}) - \log(x_{i(j-1)})$) in order to turn on the time series into stationary series.

In Table 1, we present the data transformation (DT), the weight function, the number of clusters (K) and number of series for cluster for the two datasets. In Figure 1, we present the diagram for the partitions obtained with the original Industrial Production indices series, to illustrate our procedure. The first observation regarding these results is that the number of clusters has a tendency to reduce when we go from the non-weighted situation (a) to the linear, quadratic and then exponential weighted cases. It seems that the larger the weight the less clusters we have. However this conclusion is for this dataset only and we can not extrapolate. Other experiments are needed and we believe that it is possible that an opposite behavior can be observed with other datasets; it all depends on the structure of the most recent observations of the time series compared to the first. Secondly, the homogeneity between the initial values of the time series inside each cluster seems to decrease as again we go from the non-weighted situation to the extreme exponential weight case. This behavior was expected as these initial values have smaller importance in the weighted cases. As for the last values of the time series in each cluster, it is difficult to take a conclusion, because on the one hand the time series should be more and more homogeneous in the last values in the weighted cases compared to the non-weighted; on the other hand, as in the weighted cases we have fewer and fewer clusters, the homogeneity inside each cluster decreases. We plan to analyse in the future other time series datasets, more complex, and use other weighted distances between time series to reach more solid conclusions about this novel method of clustering of time series.

Table 1: Weighted clustering results for the two datasets

Dataset: Ind. Prod.	DT	Weight function	K	Series by cluster
original	(1)	–	11	2; 6; 2; 2; 1; 1; 2; 1; 1; 1; 1
”	(4)	linear	3	16; 3; 1
”	(4)	quadratic	3	17; 2; 1
”	(4)	exponential	2	18; 2
stationary	(1)	–	20	1×20
”	(4)	linear or quadratic	2	13; 7
”	(4)	exponential	2	13; 7

Acknowledge

For the second author, this work reports research developed under financial support provided by “FCT - Fundação para a Ciência e Tecnologia”, Portugal.

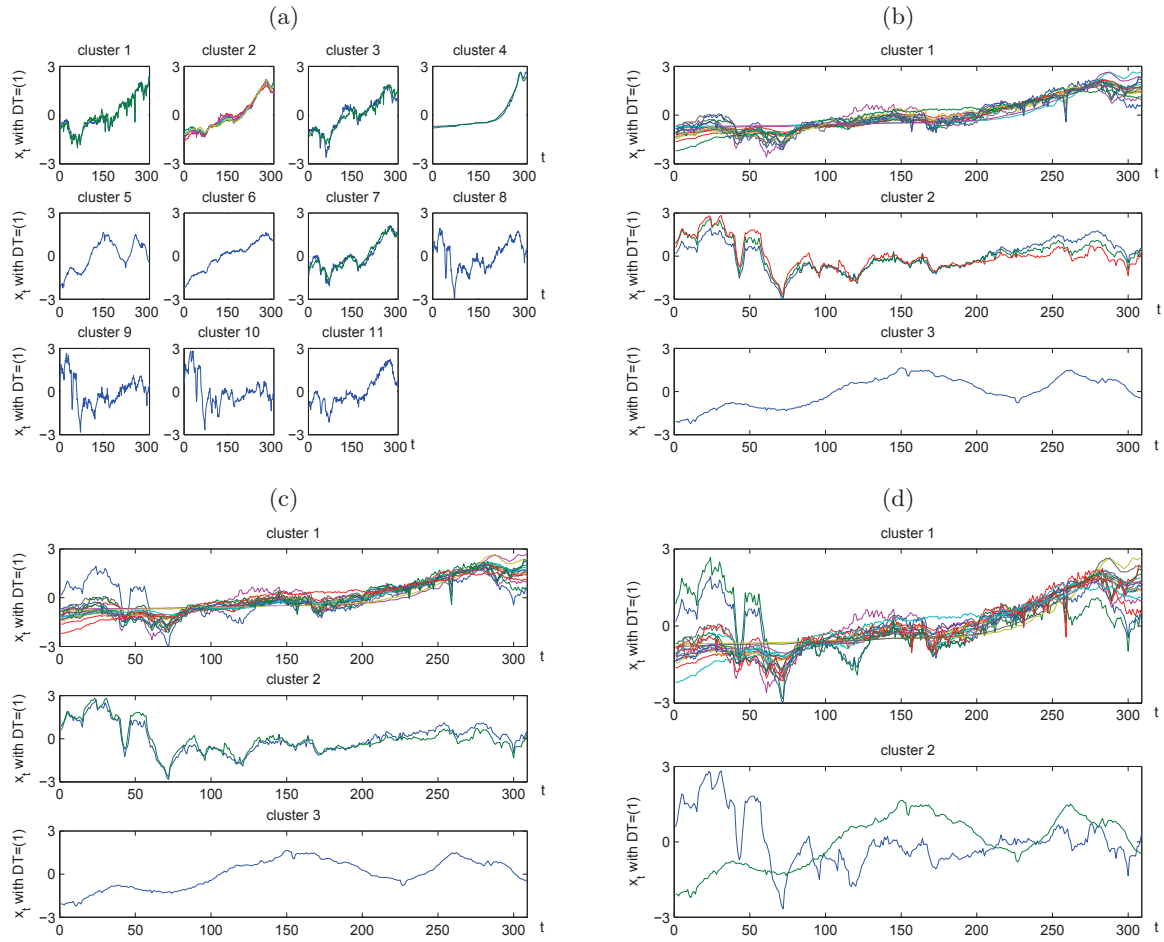

Figure 1: Cronograms of the original Industrial Production indices series (after the data transformation (1)) for each weighted cluster: (a) for DT=(1) and (b), (c) and (d) for DT=(4) with linear, quadratic and exponential weights, respectively.

REFERENCES

Beringer, J. and Hüllermeier, E. (2006). Online clustering of parallel data streams. *Data Knowl. Eng.*, vol. 58 (2), pp. 107-204.

Caiado, J. and Crato, N. and Peña, D. (2006). A Periodogram-Based Metric for Time Series Classification. *Comput. Stat. Data Anal.*, vol. 50 (10), pp. 2668-2684.

Lerman, I.C. and Pinto da Costa, J. and Silva, H. (2002). Validation of Very Large Data Sets Clustering by Means of a Nonparametric Linear Criterion. In *Classification, Clustering and Data Analysis. Proceedings of the 8th Conference of the International Federation of Classification Societies (IFCS 2002)*, vol. Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, pp. 147-157