

Searching for the Optimal Smoothing before applying Dynamic Time Warping

Selezione Automatica del Parametro di Smoothing per il Calcolo della Dissimilarità tra Curve attraverso il Dynamic Time Warping

Aldo Corbellini and Isabella Morlini

Dipartimento di Economia, Università di Parma, Via J.F. Kennedy, 6 – 43.100 Parma
e-mails: aldo.corbellini@unipr.it, isabella.morlini@unipr.it

Riassunto: Il lavoro estende i risultati del calcolo delle dissimilarità tra curve proposto in Morlini (2003) e considera la validità dell'utilizzo di stime *smooth* al posto dei dati originari nel calcolo del *dynamic time warping cost* (DTWC). La scelta del parametro λ nelle splines utilizzate per conseguire una stima *smooth* dei punti di ogni serie viene effettuata in via automatica attraverso la *cross validation*. Si propone un'applicazione sulle quantità di biossido di azoto rilevate dalle 85 stazioni per il monitoraggio della qualità dell'aria in Emilia Romagna. Il programma in Matlab per l'adattamento delle splines alle serie e il calcolo del DTWC può essere richiesto via e-mail al primo autore.

Keywords: Cluster Analysis, Cross Validation, Functional Data, Dynamic Time Warping, Spline, Smoothing Parameter.

1. Introduction

This work deals with the problem of transforming the arguments of curves before clustering functional data (see, among others, Ramsay & Li, 1998 and Ramsay & Silverman, 1997). By means of the Dynamic Time Warping (DTW) algorithm this problem is solved in a non parametric way, since a smooth monotone transformation of time t for aligning curves $x_i(t_j)$ is not explicitly defined. The couple of points $x_i(t_j)$ and $x_k(t_l)$ for sequences i and k ($i \neq k$) to be aligned are directly determined, instead. DTW is an easy to implement technique which can be straightforwardly applied to vector valued series, as would be the case if they indicate positions in two or three dimensional space. For a description of the algorithm we refer to Rabiner & Juang, 1993 and Morlini, 2003 while for the improvements that can be reached by smoothing the series before warping we refer to Morlini, 2003. Here we remark that DTW applied to row data may produce incorrect alignments, from a geometrical perspective, when pairing two values x_1 and x_2 which are identical but the first is a part of a rising interval in series 1 and the second is a part of a falling interval in series 2. An additional problem is that the algorithm may fail to align a salient feature (i.e. a peak, a valley) simply because in one series this is slightly higher or lower than in the other one. In this paper we extend previous results by automatically selecting the smoothing parameter λ of the interpolating splines. In Section 2 we briefly introduce the algorithm implemented in Matlab for the choice of λ (the program can be requested to the first author). In Section 3 we present an application regarding the series of nitrogen dioxide concentration measured by 85 air quality monitoring stations in the Emilia Romagna region and give some final remarks

2. Automatic selection of the spline smoothing parameter λ

The process of splines smoothing employs a roughness penalty approach to establish a trade-off between regularity of the spline and goodness of fit. Let $S(g)$ be the penalized sum of squares:

$$S(g) = \sum_{i=1}^n \{x_i - g(t_i)\}^2 + \lambda \int \{g''(t)\}^2 dt$$

with λ being a positive smoothing parameter. The penalized least square estimator $\hat{g}$ is the minimizer of the functional $S(g)$ over the class of all twice differentiable functions g . For estimating λ , we have implemented in Matlab the General Cross Validation (GCV) criterion, which automatically selects it. Given the following GCV score function:

$$GCV(\lambda) = \frac{1}{T} \frac{\sum_{i=1}^T \{x_i - \hat{g}(t_i)\}^2}{\left\{1 - T^{-1} \text{tr} A(\lambda)\right\}^2}$$

where T is the length of the series and $A(\lambda) = (I + \lambda QR^{-1}Q^T)^{-1}$, Q and R being two band matrices, the GCV choice of λ is carried out by minimizing the function $GCV(\lambda)$ over λ . The advantage of employing GCV over Cross Validation (see Green and Silverman, 1994, for a complete description of these criteria) is of computational matter, since it is possible to find the trace of $A(\lambda)$ without finding all its diagonal elements.

3. Air Quality Data Application

As an example to illustrate the technique proposed in Section 2 for smoothing each series before applying DTW and computing the DTW cost as a dissimilarity between curves, we consider measurements from the ARPA Agency of the nitrogen dioxide (NO_2) concentration in different sites in Emilia Romagna. Note that the DTW cost is a dissimilarity built according to the geometrical features of the trajectories like the instantaneous position, the slope of inter-temporal segments (velocity), the concavity and convexity of inter-temporal segments (acceleration), the polygonal line (shape) and the portion of area between each pair of trajectories. Following the definition proposed in D'Urso (2000), it belongs to the class of 'geometrical' dissimilarities between trajectories, which is alternative to the 'structural' and the 'correlative' classes. In the latter the dissimilarities consider the structural aspect of the series, like the trend and the seasonality, or the autocorrelation functions. Since the quality of the alignment and then the validity of the dissimilarities between each pair of sequences cannot be objectively determined in this real data set, we indirectly compare the DTW cost computed with row series and with smoothed series by analyzing the successive partition in groups. In the data set we 75 daily registrations (from the 1st October 2003 to the 14th December 2003) for each of the 85 air quality stations. Missing data are replaced with moving averages of length 7. Fig. 1 shows the dendrogram obtained with

DTW applied to smoothed series. Fig. 2 reports the dendrogram obtained with no previous smoothing. In both graphic the agglomerative method is the average linkage and labels indicate the province in which the measurements have been recorded. In the first dendrogram groups are well defined and well separated even in the lower levels of the hierarchy. In the second case (few) clusters are characterized only in the upper levels. For example, for a 0.2 distance, 11 and 84 groups are formed respectively in the first and in the second dendrogram. For a 0.25 level, groups are 9 and 67, respectively.

Figure 1. *Dendrogram obtained with cross validated smoothed series.*

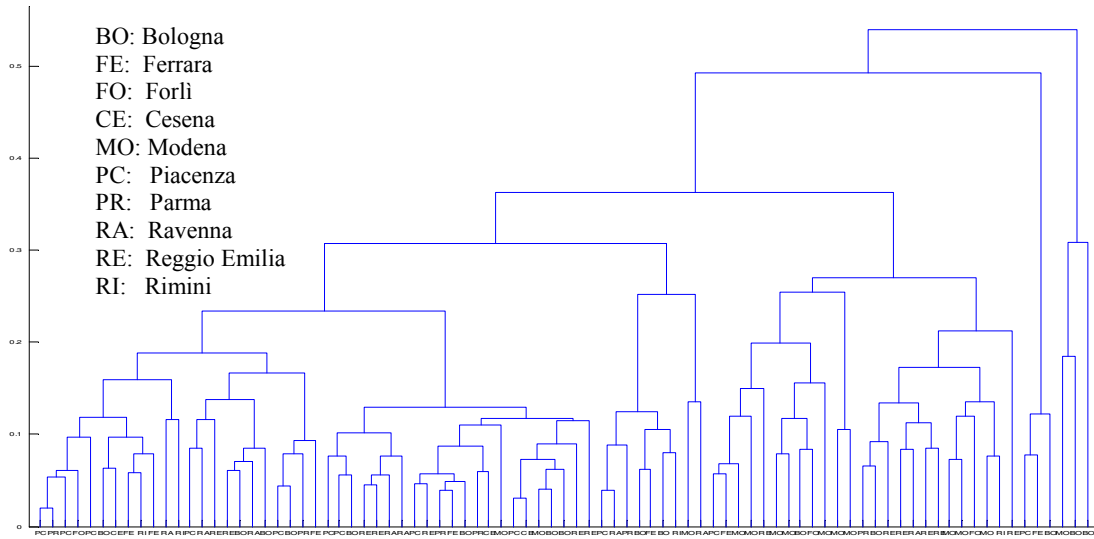
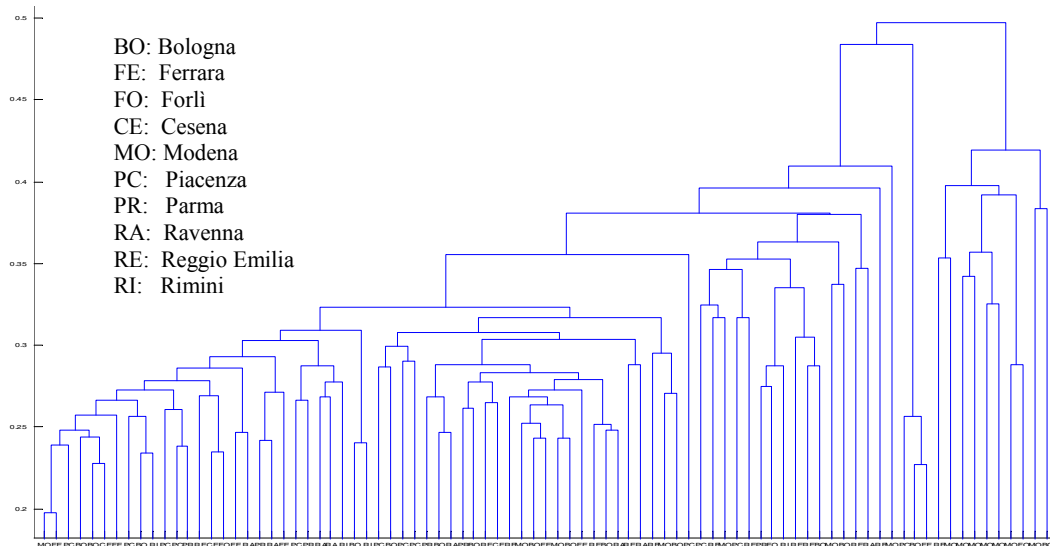


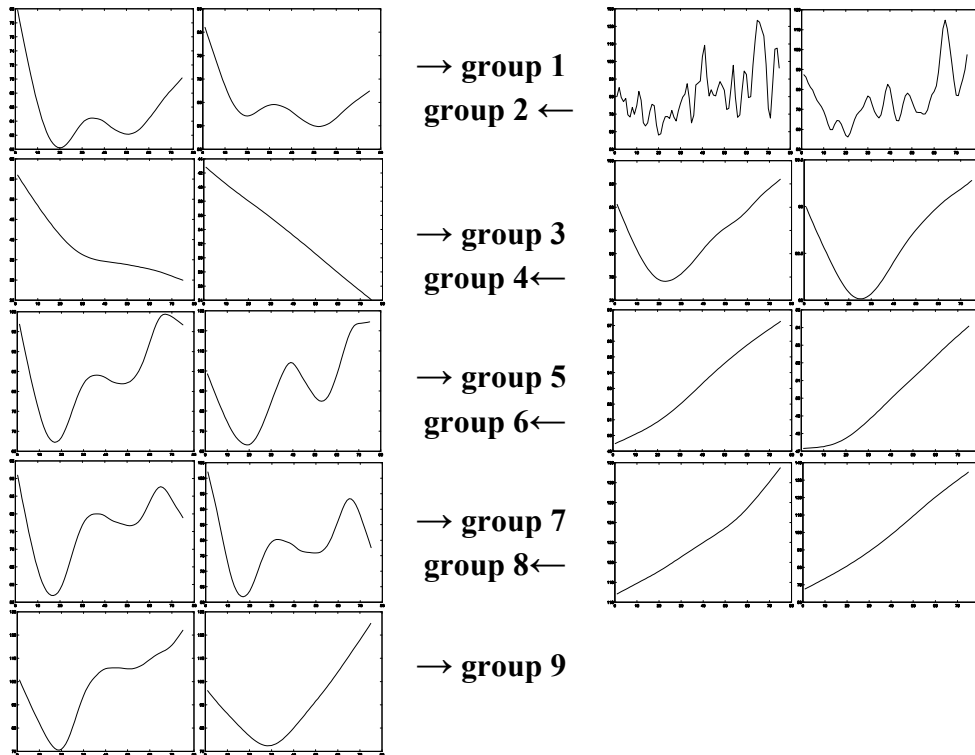
Figure 2. *Dendrogram (group average cluster analysis) obtained with row data.*



A deeper inspection of the first dendrogram shows that the 9 clusters group series very similar in shape (Fig. 3). Curves with many local maxima and minima (group 2) are separated from those with no local maxima or minima (clusters 3, 6 and 8) and from those with one local minimum (groups 4 and 9) or two local minima (groups 1, 5 and 7). The number of local maxima is also a distinctive characteristic of the clusters

Considering the labels of each series we see that the monitoring stations are clustered on the base of the place in which they are located (city center, countryside, industrial estate, ...). rather than on the province.

Figure 3. Examples of curves belonging to each of the 9 clusters (first dendrogram).



In conclusion, the main objectives of this work was to improve the DTW algorithm in order to achieve a dissimilarity measure based on the geometrical aspects of the trajectories to be classified, like the slope, the concavity or convexity of inter temporal segments and the number of local maxima and minima. Unlike the others geometrical dissimilarities reviewed in D'Urso (2000), the DTW cost also include the possibility of curve registration (i.e. a different alignment on the time axis). In this paper we have made automatic and data dependent the preliminary smoothing process of the series and we have indirectly measured the validity of the smoothing step before applying DTW on the base of the successive clustering results.

Main References

- D'Urso, P. (2000) Dissimilarity measures for time trajectories, *JISS*, 1-3, pp. 53-83.
- Green P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall, London
- Morlini, I (2003) On the criterion of dynamic time warping for computing the dissimilarity between time series, *Cladag 2003*, Bologna, pp. 285-288.
- Rabiner, L., Juang, B. (1993) *Fundamentals of Speech Recognition*, Prentice Hall, N.J.
- Ramsay J. O., Li, X. (1998) Curve registration, *J.R.S.S., Series B*, 60, 351-363.
- Ramsay J. O., Silverman, B.W. (1997) *Functional Data Analysis*, Springer, New York.