Curves Clustering and Outliers Detection: New Proposals and Open Problems

Classificazione di curve ed individuazione di outlier: nuove proposte e problemi aperti

Elvira Romano
Facoltà di Studi Politici per l’Alta Formazione Europea e Mediterranea
Seconda Università di Napoli
e-mail: elvroman@unina.it

Riassunto: Negli ultimi anni sono state proposte differenti metodologie per la classificazione di dati funzionali. Tuttavia, la maggior parte di esse risulta essere estremamente dipendente dalla tipologia di trasformazione utilizzata per l’approssimazione dei dati (forma funzionale vera). In questo lavoro viene proposta una strategia per la classificazione di dati funzionali che contemporaneamente stima e classifica le curve attraverso una generalizzazione dell’algoritmo di classificazione dinamica. La procedura utilizza spline con nodi variabili allo scopo di determinare dinamicamente la migliore forma funzionale sottostante i dati senza specificazioni a priori. Si propone, inoltre, un miglioramento della strategia per individuare eventuali outlier e per gestire problemi di sfasamento.

Keywords: functional data, curve prototype, free-knot spline, dynamic clustering, outliers

1. Introduction

Clustering methods in functional data framework, are often useful to detect groups of similar curve shapes (e.g. Abraham et al., 2003; James and Sugar, 2003; Tarpey and Kinateder, 2003). The problem is that the clustering results can depend on how the curves fit the data (Tarpey, 2007). A clustering method should provide a set of clusters with a reduced dependence from the chosen functional form.

Classical clustering approaches deal with data fitting using different basis functions and classifying the obtained model coefficients using a k-means method afterwards. B-spline basis functions are usually chosen for their numerical proprieties and their flexibility. However universal B-spline basis functions could not be appropriate for a set of curves with different shape, because the number and the location of knots influence the curve fitting.

Here we introduce a clustering strategy which includes the choice of an appropriate data functional form in the clustering process, that is the Dynamical Curves Clustering Algorithm with Free-knot Spline Estimation (DCC&FSE) (Romano, 2006).

The strategy uses B-spline basis with free knots to estimate a representative regression model for each cluster by optimizing a non linear minimization problem for the knots computation. In this way the procedure finding an optimal set of knots for each cluster allows to estimate and classify curves at the same time.
In many cases, functional data come from sensor systems where low fidelity and frequent failures can lead to outliers presence. Sensors in a network are often loosely synchronized because of different delays in detecting environmental changes. In order to investigate this problem, firstly, we have suggested a criterion to detect an optimal time shifting in the DCC\&FSE strategy and, secondly, we have implemented a curve similarity measure in the clustering process in order to monitor and catching outliers (Balzanella et al., 2007).

The remainder of the paper is organized as follows: in section 2, we introduce the DCC\&FSE strategy; in the section 3, we propose a generalized version of DCC\&FSE strategy that deals with outliers and phase mismatching problems; in section 4, an application on real data is presented and finally in section 5 a discussion on perspectives and research directions is provided.

2. Dynamical curves clustering with free-knot spline estimation

The DCC\&FSE strategy is an extension of the Dynamical Clustering Method (DCM) (Diday, 1971), proposed on functional data structures. Let $E$ be a set of $n$ functions such that each $i^{th}$ function, for $i = 1, \ldots, n$, is given by the $(t_j^i, y_j^i)_{1 \leq j \leq J}$ list of $J$ pairs, with $t_j^i \in T$, a compact subset of $\mathbb{R}^J$, and $y_j^i \in \mathbb{R}$.

The DCM looks for the partition $P = \{P_1, \ldots, P_C\} \in \mathcal{P}_c$ (where $\mathcal{P}_c$ is the family of all the partition $\{P_1, \ldots, P_c\} \in \mathcal{P}_C$ in $C$ clusters) and a set $G = \{g_1, \ldots, g_C\} \in \mathcal{G}_c$ (where $\mathcal{G}_c$ is the family of all admissible representation of $C$ clusters prototypes) such that the criterion of best fitting between $G$ and $P$ is minimized. We generalize this criterion to curves as follows:

$$\Delta(P, G) = \sum_{c=1}^{C} \sum_{i \in P_c} \delta^2(y^i, g_c) \quad P_c \in P, \ g_c \in G \tag{1}$$

where $\delta(y^i, g_c) = \|y^i - g_c\|$ is an $L_2$ distance. The set $G = \{g_1, \ldots, g_C\}$ of prototypes are computed by optimizing an adequacy criterion $\phi(g) = \sum_{i \in P_c} \delta^2(y^i, g_c)$ that leads to a free-knot spline estimation of a prototype for each cluster.

Let $\xi^c \in \mathbb{R}^M$ be a vector of knots, we look for a good approximation of the prototype $g_c(\xi^c) \in \mathbb{R}^J$, that is the estimate of a representative function of a set of curves in a cluster, solution of a Least Squares problem. Formally we obtain that each prototype can be written as:

$$g_c(\alpha^c_1, \xi^c) = B(\xi^c)\alpha^c_1 \in \mathbb{R}^J \tag{2}$$

where the vector $\xi^c$ is the Jupp transformation of $\xi^c$ (Jupp, 1978), $B(\xi^c)$ is the matrix of $J \times (H + M)$ whose the generic row is the vector $B_j^j(t_j, \xi^c) \in R^{H+M}$ of the B-spline basis function and $\alpha^c_1$ is the vector of the estimated coefficients. According to the optimized criterion, initializing the procedure with random equispaced knots we obtain, for each cluster, a local model prototype identified from the best sets of knots.

The algorithm works like classical DCA by alternating representation and allocation step. In the following section we will describe these two steps.
2.1. The optimization problem and the allocation step

In order to solve the optimization problem we propose a new approach inspired by Gervini (2006) to estimate the mean of a set of curves. The mentioned approach uses the Jupp transformation of the knots that ameliorates the so called lethargy propriety and removes the redundancy in the knots vectors. Let \( \mathbf{\xi}^c = \mathcal{J}(\mathbf{\xi}^c) \) be the Jupp’s transformation of the vector of knots \( \mathbf{\xi}^c \in [a, b] \) where the generic element is defined as: \( \xi^c_m = \log \frac{\xi_m^{c+1} - \xi_m^c}{\xi_m^{c+1} - \xi_0^c} \) for \( m = \{1, \ldots, M\} \) \( \xi_0^c = a \) and \( \xi_m^c = b \), and let \( \mathbf{B}(\mathbf{\xi}^c) \) be the matrix of \( J \times (H + M) \) whose generic row is the vector \( B^T \in R^{H+M} \) of the B-spline basis function. Using this transformation the adequacy criteria can be written as

\[
\phi(g) = \sum_{i \in P_c} \delta^2(y^i, g_{c\mathbf{\xi}^c})_{\mathbf{\alpha}^c, \mathbf{\xi}^c} = \sum_{i \in P_c} \| y^i - \mathbf{B}(\mathbf{\xi}^c)\mathbf{\alpha}^c \|^2 
\]

(3)

and it’s possible to find the best approximation of \( g_c \) by the least squares, such that:

\[
(\hat{\mathbf{\alpha}}^c, \hat{\mathbf{\xi}}^c) = \arg\min_{(\mathbf{\alpha}^c, \mathbf{\xi}^c)} \sum_{i \in P_c} \| y^i - \mathbf{B}(\mathbf{\xi}^c)\mathbf{\alpha}^c \|^2 
\]

(4)

and \( \hat{g}_c = \mathbf{B}(\hat{\mathbf{\xi}}^c)\hat{\mathbf{\alpha}}^c \).

For given \( \mathbf{\xi}^c \), the optimal \( \hat{\mathbf{\alpha}}^c(\mathbf{\xi}^c) \) is \( \hat{\mathbf{\alpha}}^c = (\mathbf{B}(\mathbf{\xi}^c)^TB(\mathbf{\xi}^c))^T \mathbf{B}(\mathbf{\xi}^c)^T \bar{y} \) in this way \( \hat{\mathbf{\xi}}^c \) solves the following minimization problem

\[
\arg\min_{(\mathbf{\xi}^c \in R^H)} \sum_{i \in P_c} \| y^i - \mathbf{B}(\mathbf{\xi}^c) \left( \mathbf{B}(\mathbf{\xi}^c)^T \mathbf{B}(\mathbf{\xi}^c) \right)^{-1} \mathbf{B}(\mathbf{\xi}^c)^T \bar{y} \|^2 
\]

(5)

Since the dependence of the basis functions from the knots is nonlinear, it is a nonlinear minimization problem. The solution to this problem, as stated before, is given from a little modification of the algorithm proposed by Gervini (2006). The algorithm produces a sequence of knots vectors \( \mathbf{\xi}^c \) and \( \mathbf{\xi}^c \) of increasing dimensions for each cluster, and the relative estimated functional prototypes. The optimal number of knots will be chosen on the basis of intermediate knot sequences. Since for regression splines, the number of knots determines the bias/variance trade-off, among the produced sequence of knots, for each cluster, it is chosen the function such that the relative set of knots minimizes the Generalized Cross Validation Criteria (GCV). Contrary to the Gervini method we do not start from defined equispaced knots but with random equispaced ones.

The algorithm uses two-steps to produce several sequences of knots vectors, different in cardinality and location and the related set of prototypes.

In the first step the Jupp transformed values of an initial random sequence of knots are used to find a minimiser of (5). This phase may be thought as providing a fast but potentially approximate solution as the starting point for the second phase. The second phase called Forward Addition repeats the procedure for the \( m = 1 \ldots M \) knots until several sequences of vectors are identified.

The obtained function \( g_c \) represents a local function prototype of each cluster, that is able to characterize the whole cluster structure. It is summarized by several sets of knots vectors obtained from the stepwise algorithm. Since the choice of the knots influences the resulting smoothing spline prototype, a set of optimized knots leads to finely control the
shape and the structure of the cluster. Moreover each $c-th$ cluster can be described as a $j-$dimensional vector of curves that follows the model:

$$y^i = B(\hat{\zeta}^c)\alpha_l + \epsilon_i \quad \forall i \in C$$  \hspace{1cm} (6)

where $\epsilon^i$ is an observation noise with mean zero, constant variance and covariance zero for distinct arguments values. In this sense, the procedure allows to discover patterns of curves and their underlying functions.

With the aim to allocate the functions to the set of the cluster prototypes we utilize a criteria based on the MSE distance between the all functions to the cluster prototype.

A formal method to define good clustering is to use measurements for the quality of clustering: structural indexes. In our framework, in order to measure the closeness of the partition, we propose to use the classical quality of partition index

$$Q(P) = 1 - \frac{W(P, \hat{g}_c(\hat{\zeta}^c))}{V}$$  \hspace{1cm} (7)

where $V$ is the total variability and $W$ is the variability inside a cluster $P_c$ relates to its prototype $\hat{g}_c$. This index ranching between 0 and 1, where the values 0 indicates that there is no structure, on the contrary if this value is near to 1 it means that there is a good clustering structure.

### 3. DCC\&FSE strategy for handling phases and outliers problems

The presence of time shifted curves can lead to compromise the classification results, bringing the method to be unfeasible with respect to real anomalies. In order to overcome this disadvantage and for detecting outliers from a dataset, we improved the DCC\&FSE algorithm in two important directions.

The reason of the first modification is to support time shifting aspects. In our algorithm, a re-phasing step between the representation and the allocation step is introduced. In this phase the optimal time shifting $\tau^*$ is the shifting value which allows to minimize the Euclidean distance between the curve $y^i$ and the prototypes $\hat{g}_c$. Formally it is such to minimize:

$$\varphi(\tau) = \sum_{c=1}^{C} \sum_{i=1}^{n} \delta^2 (y^i_{j+\tau}, \hat{g}_c)$$  \hspace{1cm} (8)

The introduction of the re-phasing step brings out a second sensitive aspect: the presence of curves having anomalous behaviors which impact on the clustering process, contributing to bad prototypes computation and to artificial groups. This has led to define an additional strategy based on Dynamic Time Warping (DTW) dissimilarity measure to compare curves.

#### 3.1. Outlier selection strategy

Starting from a set of $C$ clusters obtained from the DCC\&FSE with the re-phasing step, we compute the dissimilarity among the curves and the obtained prototypes by DTW. It is worth to notice that the DTW gives more importance to the overall shape of the curves,
when compared to the Euclidean distance. Furthermore it is invariant to the time shifting. It defines the similarities between curves, regardless of their individual time courses, by finding the warping in the time dimension of one curve, which minimizes their difference, and then it determines the area between them. Specifically, in order to attenuate the noise impact on the dissimilarity computing, we use DTW on smoothed estimates of the values of the curves (Morlini, 2005).

Given two smoothed curves \( \hat{y}^i \) and \( \hat{y}^j \) the DTW algorithm implies a construction of a matrix \( D \) in which the generic element \( d_k \) is the distance between the value of the curve \( \hat{y}^i \) at a time \( t \) and the value of the curve \( \hat{y}^j \) at time \( t' \). The dissimilarity between \( \hat{y}^i \) and \( \hat{y}^j \) also called Dynamic Time Warping Cost (DTWC) is defined as follows:

\[
DTWC = \min \sqrt{\frac{1}{K} \sum_{k=1}^{K} d_k / K}
\]

(9)

where \( K \) in the denominator is used to compensate for the fact that the warping path (a contiguous set of matrix elements that defines a mapping between \( \hat{y}^i \) and \( \hat{y}^j \)) may have different lengths. In our case, starting from a grid of two different sampling value \( T_1 \) and \( T_2 \) for respectively \( \hat{y}^i \) and \( \hat{y}^j \), \( K \) is such that: \( \max(T_1, T_2) \leq K \leq T_1 + T_2 - 1 \).

Moreover the element of \( D \) are subject to the following condition:

- Boundary conditions for \( k = 1 \) and \( k = K \): \( d_1 = d(t_1, t_1) \) and \( d_K = d(T_1, T_2) \). This requires the warping path to start and finish in diagonally opposite corner cells of the matrix.
- Given \( d_k = d(\hat{y}^i(t_r), \hat{y}^j(t_c)) \) then \( d_{k-1} = d(\hat{y}^i(t_a), \hat{y}^j(t_b)) \) where \( r - a \leq 1 \) and \( c - b \leq 1 \). This restricts two successive elements \( d_e \) in the summation to be adjacent (including diagonally adjacent cells).
- Given \( d_k = d(\hat{y}^i(t_r), \hat{y}^j(t_c)) \) then \( d_{k-1} = d(\hat{y}^i(t_a), \hat{y}^j(t_b)) \) where \( r - a \geq 0 \) and \( c - b \geq 0 \). This forces the points in \( D \) to be monotonically spaced in time.

In order to discriminate the outliers, we define a cut off parameter \( \alpha \) as the dissimilarity value corresponding to the percentile \( \gamma \) of the empirical distribution of the dissimilarity function \( DTWC \) in the whole clusters. It is such that at most 100\( \gamma \)% of the dissimilarity measuring are less than this value and at most 100 (1 - \( \alpha \)) % are greater.

Let \( \hat{y}^i \) \( i = 1 \ldots n \) be the set of true functional forms coming from the smoothing process according to the best set of knots \( \hat{\zeta} \) for the cluster \( c = 1 \ldots C \) which the curve belongs to. An outlier is a curve \( y^i \) such that for every prototype \( \hat{g}_c DTWC(\hat{y}^i, \hat{g}_c) \) is larger than \( \alpha \).

From an algorithmic point of view the re-phased version of the DCC&FSE with the addition of DTW criterion can be summarized as follows:

- The optimal set of prototypes \( \hat{g}_c, c = 1 \ldots C \) is computed in the re-phased version of DCC&FSE algorithm;

For each curve \( y^i, i = 1 \ldots n \)

The curve \( y^i \) is smoothed according to the best vector of knots \( \hat{\zeta}_c \) for the cluster which the curve belongs to;

For each cluster \( c \)
The DTWC, via DTW procedure is determined between the current smoothed curve $\hat{y}_i$ and the curve prototype $\hat{g}_c$.

End for each

The minimum value for DTWC is chosen

The empirical distribution $\text{Distr}$ of dissimilarity measures is constructed incrementally

End for

For $i = 1$ to $n$

if $\text{Distr}[i] > \alpha$ then $y^i$ is selected as outlier.

End for

An index based on the optimized criterion is proposed to compare the DCC&FSE strategy with the re-phased version of the DCC&FSE and the addition of DTW criterion. It is defined as:

$$\Delta_s = \frac{\Delta (P, G) - \Delta_r (P, G)}{\Delta (P, G)}$$  \hspace{1cm} (10)$$

where $\Delta$ is the value of the optimized criterion for DCC&FSE and $\Delta_r$ is the value of the optimized criterion with the introduction of the DTW.

4. Application

To demonstrate the effectiveness of the proposed approach, an analysis on real data has been performed.

The data come from one of the several experiments conducted by the Department of Hydraulic and Environmental Engineering "Girolamo Ippolito" at the University of Naples (Panisi et al., 2006), on the consequences of the introduction of submerged breakwaters on sea waves.

Various kinds of breakwaters, in terms of geometry and permeability, have been tested in a protected environment with the aim to examine their role in determining the wave profiles. To perform the analysis, a detection system made by 90 sensors has been placed to detect the rupture process of each sea wave on the breakwaters. During the experiments, the sensors were arranged on a wide area and the monitoring activity has been performed without human assistance. These two circumstances rise both time delay problems, as a consequence of the different time needed to signals to reach the storage unit, and outliers problems like transducers failures, noise or curves not belonging to the studied phenomenon. In such a context, our strategy appears to be suitable to discover the clusters structure summarizing the sea waves behaviors. Furthermore it controls the impact of phase mismatching on the prototypes computation detecting outliers presence.

The object of the analysis consists in 90 curves, each of them recorded from different sensors.

For measuring the performance of our strategy, we have compared the results obtained by the DCC&FSE strategy with the proposed one.

For DCC&FSE strategy we need, at first, to choose the order of B-spline functions, the minimum and maximum values of the knots, and then the number of the clusters. We
have chosen B-spline of order 4, since these are the most flexible and numerical stable basis functions. The number of knots has been fixed to be in the range (3,12).

In a situation where non informative phase displacements are present, although DCC\&FSE strategy is able to discover the basic clustering structure, some of the data are not correctly allocated. This makes the outliers detection problematic.

Figure 1 illustrates our outliers detection procedure on not re-phased data. As we can see, even though DTW catches the most of outlying curves, some that are not anomalous are identified as outliers.

**Figure 1:** Sea waves clusters by DCC\&FSE strategy without the re-phasing criteria: for each cluster, the thick dashed curves represent the wrong selected anomalies

![Figure 1](image1)

By introducing the re-phasing step, the outliers detection is more effective (Fig.2), since it is able to find out the real anomalies.

**Figure 2:** Sea waves clusters by DCC\&FSE with the re-phasing criteria: for each cluster, the thick dashed curves represent the outliers

![Figure 2](image2)

The improvements in terms of clusters homogeneity are evaluable both from a graphical point of view and by comparing the quality partition index defined, on the basis of the variability decomposition, as ratio of the between variability and the total variability.

By repeating the experiment 20 times, so to mitigate the impact of the initial random partitioning, we have observed an improving of the quality partition index quantifiable in about 15%.
5. Conclusion and future work

This paper proposes a clustering strategy for functional data and its improved version for dealing with phase mismatching and outliers problems. We have reported some of the obtained results, more details can be found in Romano (2006).

As a next step, we are going to study the possibility of incorporating a criterion for the outliers selection in the optimization process, such to improve the strategy from a computational point of view. Moreover, since numerous applicative fields, such as data coming from sensors systems, require the management of functional data related to the space, we are working on the possibility to introduce the spatial component in the clustering process.

References


