

# Sequential Clustering for Functional Data

Ana Justel\* and Marcela Svarc\*\*<sup>1</sup>

\**Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain*

\*\**Departamento de Matemática y Ciencias, Universidad de San Andrés and CONICET, Argentina*

## Abstract

This paper presents SeqClusFD, a top-down sequential clustering method for functional data. The clustering algorithm extracts the splitting information either from trajectories, first or second derivatives. Initial partition is based on gap statistic that provides local information to identify the instant with more clustering evidence in trajectories or derivatives. Then functional boxplots allow reconsidering overall allocation and each observation is finally assigned to the cluster where it spends most of the time within whiskers. These local and global searches are repeated recursively until there is no evidence of clustering at any time on trajectories or first and second derivatives. SeqClusFD simultaneously estimates the number of groups and provides data allocation. It also provides valuable information about the most important features that determine cluster structure. Computational aspects have been analyzed and the new method is tested on synthetic and real data sets.

**Keywords:** Hierarchical Clustering, Functional Boxplot, Gap Statistics.

## 1 Introduction

Functional data analysis (FDA) is a very active area of research nowadays, mainly since it has become very easy to collect and store data in “continuous time” (see [11]). Although generally each data is recorded only in a finite number of moments, it is more common to analyze them as functions rather than as vectors. Instead of points, the observations in this context are each of the curves. Most existing procedures for functional data require a large amount of consecutive observations on which smoothing techniques are applied, see [26]. The classical assumption is that functions belong to a Hilbert space (for example, square integrable functions on a finite real interval  $[a, b]$ ) and can be

---

<sup>1</sup>Corresponding author: Marcela Svarc, Departamento de Matemática y Ciencias, Universidad de San Andrés.... Email: msvarc@udesa.edu.ar

represented with a convenient functional basis. Depending on the characteristics of the functions, the most common basis are Fourier, Haar or wavelets. A finite truncation of the series facilitates the analysis with conventional multivariate methods applied to a finite set of retained coefficients.

The aim of cluster analysis, or unsupervised classification, is to assign observations into subsets, such that similar objects are in the same group and dissimilar ones are in different groups. This is a complex problem since usually there is no previous information about the data structure. There is no “one size fits all” method for analyzing any data set and the nature of the data should determine the procedure to be used. In fact, there are clustering methods that have outstanding performance in certain configurations of data and a terrible behavior under different conditions. For instance, it is well known that the popular  $k$ -means is suitable in  $\mathbb{R}^d$  for round, Gaussian and well separated clusters but it is not able to find cluster structures with nested groups. Some multivariate clustering techniques have been successfully adapted to functional data after considering specific geometrical notions and can be classified in one of the next three categories.

*Centroid based clustering methods* use optimization algorithms for centroid finding of each group (as  $k$ -means). Several authors studied and adapted these methods to functional data, see for instance [1], [31], [9], [10], [34], [14], [27], [28] and [16].

*Model based clustering methods* assume different population model for each group, typically all from the same distribution family (as Gaussian mixture models). In FDA there are several procedures following this strategy, such as those proposed in [19],[5] and [17].

*Hierarchical clustering methods* are based on the idea of divisive and/or agglomerative sequential grouping to provide the best partition on each possible number of groups. Divisive methods begin allocating all observations into one group and sequentially separate into different groups the observations that are more distant from the rest. This idea is repeated until there are as many groups as observations. Agglomerative methods initially assign each observation to a different group and then sequentially join the closest observations until all observations are in a single group. For functional data Giraldo *et al.* [15] developed an algorithm for spatially correlated data and Boullè *et al.* [6] introduced a bayesian proposal of nonparametric hierarchical clustering for functional data. At the best of our knowledge, the particular case of hierarchical clustering based on decision trees with sequential binary partitions, have never been extended to FDA from multivariate data approaches where several extensions can be found, see for instance [20],[3] and [12].

In general, in centroid or model based methods the number of clusters is known while in hierarchical clustering is unknown. Recently, Jacques and Preda [18] described

thoroughly the state of the art of this field.

As it is well known, not all relevant information on a function is visible in the trajectory. Derivatives are usually very helpful to highlight differences. Although all clustering methods for functional data mentioned above can be applied to the set of trajectories, or the set of first derivatives, second derivatives, and so on, it is interesting to have a statistical procedure able to extract relevant information from any or all of these sets. In supervised classification of functional data, Alonso et al. [2] considered a distance base on using derivatives. Exploration of functions and derivatives is a key point since group structure must be associated to the trajectories for some groups and to derivatives in some others. To illustrate this idea, lets analyze a very simple example with the three groups of 25 functions shown in Figure 1. Functions in groups 1 and 2 are constant but with different levels, while functions in group 3 have constant positive slope and levels are similar to functions in group 1. Considering original functions, only two groups are identified by looking in any instant, a cluster that contains lines of groups 1 and 3 and the other with lines from group 2. Switching to apply clustering methods with first derivative functions, we note that all lines in groups 1 and 2 have zero constant derivative function, different than derivatives from group 3 that are also constant but nonzero. Any clustering method applied to derivatives also identify a maximum of two clusters, one with the lines of groups 1 and 2 and the other with lines from group 3. In summary, by applying clustering methods to trajectories, functions are separated by level and when applied to derivatives are separated by shape. We will only be able to succeed in identifying the three groups if we run a sequential clustering scheme taking into account trajectories and derivatives.

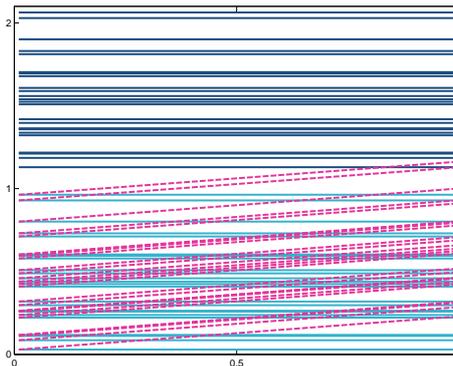


Figure 1: Light solid lines are in group 1; dark solid lines are in group 2; and dashed lines are in group 3.

The aim of this paper is to introduce **SeqClusFD**, a new hierarchical clustering method designed only for FDA. The idea is to develop a divisive top-down clustering based on sequential exploration of the functions and their derivatives. This means that

we consider local level and global shape properties that are available in functions, and not in finite dimensional data. Output from the algorithm includes number of groups, allocations and some guidelines for cluster interpretation.

The remainder of this paper is organized as follows. In Section 2, SeqClusFD method is presented and some details on practical implementation are analyzed. In Section 3, a simulation study with synthetic data is performed. In Section 4, well known real data are analyzed and a further study is conducted, including result interpretation. Conclusions are given in last Section.

## 2 Divisive top-down SeqClusFD method

We observe  $X_1(t), \dots, X_n(t)$  functions grouped into an unknown number  $k$  of populations,  $C_1, \dots, C_k$ . We assume all the functions are smooth and defined in the same compact real interval  $[a, b]$ . Each data  $X(t)$  is a realization of a stochastic process defined on the probability space  $(\Omega, F, P)$ , hence  $X(t)$  is a random variable for each  $t \in [a, b]$ .

Two facts are motivating the need of a new sequential procedure for cluster identification that takes into account the shape of the curves. The first is that functions that belong to different groups must differ by at least one of the following characteristics: trajectories, velocities or accelerations. The second is that it may happen that some groups differ by the shapes of the trajectories and some others by the shapes on any of their derivatives. We use the notation  $X_1^l(t), \dots, X_n^l(t)$ , where  $l = 0, 1, 2$  means the order of derivation. For the sake of simplicity, if  $l = 0$  we omit the superscript.

Hence, our idea is to start with a single group and, on each splitting stage, sequentially apply the following local and global clustering steps. We stop the divisions when there is no evidence of new cluster structure inside any group.

*Local clustering step:*

Considering the sets  $\{X_1(t), \dots, X_n(t)\}$ ,  $\{X_1^1(t), \dots, X_n^1(t)\}$  and  $\{X_1^2(t), \dots, X_n^2(t)\}$  separately, we find the instant  $t \in T$  in which the cluster structure is more evident in any of the three sets. Local clustering evidence is assessed by computing for each set and instant the gap statistic (GS) introduced by Tibshirani *et al.* [33] to estimate the number of clusters in a data set with any cluster method. The idea is to compare the change in the within cluster dispersion when one additional group is considered with the expected change under an appropriate reference distribution.

In our case  $X_1^l(t), \dots, X_n^l(t)$  is a sample of one dimensional observations and for a

$k$ -cluster structure  $C_1, \dots, C_k$ , the within cluster dispersions are estimated by

$$D_r = \sum_{i,i' \in C_r} d_{ii'}, \quad r = 1, \dots, k$$

where  $d_{ii'}$  is the euclidean distance between two observations of the same cluster. The gain of considering  $k + 1$  groups instead of  $k$  is

$$L_{k,k+1} = \ln(W_k) - \ln(W_{k+1}), \quad (2.1)$$

where  $W_k = \sum_{r=1}^k D_r / 2n_r$  is the total sum of pairwise distances within  $C_1, \dots, C_k$  and  $n_r$  is the cardinal of  $C_r$ .

As large  $L_{k,k+1}$  indicates evidence of  $k + 1$  groups, this value is compared with that obtained from a sample generated from an appropriate reference distribution. Tibshirani *et al.* [33] proved that, for the one-dimensional case, the reference distribution should be uniform. Hence we generate  $b = 1, \dots, B$  random samples of size  $n$  from the uniform distribution on the interval  $[\min X_i, \max X_i]$ . For each sample we compute  $\ln(W_{k,b})$  and  $\ln(W_{k+1,b})$ . We denote the empirical version of (2.1) by  $L^*(k, k + 1) = E^*(\ln(W_{k,b})) - E^*(\ln(W_{k+1,b}))$ , where  $E^*(\ln(W_{k,b}))$  and  $E^*(\ln(W_{k+1,b}))$  are empirical means. Empirical standard deviation of  $\ln(W_{k,b})$  is denoted by  $s^*(k)$ .

The gap statistic for the set  $X_1^l(t), \dots, X_n^l(t)$  is the minimum  $k$  that satisfies

$$L^*(k, k + 1) + n_{sd} \sqrt{1 + 1/B} s^*(k + 1) \geq L(k, k + 1), \quad (2.2)$$

where  $n_{sd}$  is the number of standard deviations to be fixed. In [33], the authors suggest  $n_{sd} = 1$ , but stronger rules could be considered.

We use the greatest gap statistic to decided the instant, feature (trajectory, first or second derivative) and number of groups for sample splitting. Then for the local tentatively classification we apply the one-dimensional  $k$ -means allocation on these instant to all the curves of this feature.

*Global revision clustering step:*

The previous functional data classification is only based on local information, then we need a global criteria to integrate the information from all the complete curves and correct possible local effects.

Reallocation of misleading data is done by identifying potential outliers using the functional boxplot (FB) introduced by Sun and Genton [30]. For each set of the proper feature curves inside the same cluster identified in the local clustering step we compute a FB. Construction is based on ordering functions from the center outward using band-depth definition of López-Pintado and Romo [21]. The appearance is determined by deepest curve (*median*), envelope of 50% central functions (*box borders*) and maximum

non-outlying envelope (*whiskers*). In analogy with classical boxplots, 1.5 times the 50% central region is typically considered the whiskers for the non-outlying envelope. In [21] the authors suggest that this value can be adjusted on practice.

We consider potential outlier to any curve that is outside the whiskers band at least for one instant. For these curves we compute the proportion of time that are inside each FB. In the final cluster allocation, they are moved to the group where they spend most of the time inside whiskers.

## 2.1 SeqClusFD algorithm

STEP 0.

- *Define* a grid on the interval  $[a, b]$ :  $a = t_0 < \dots < t_N = b$ .
- *Evaluate* trajectories, first and second derivatives in the grid:  $X_i^l(t_j)$  for  $i = 1, \dots, n$ ,  $l = 0, 1, 2$  and  $j = 0, \dots, N - l$ .
- *Start* by considering a single group.

STEP 1. *Local clustering.*

- *Calculate* gap statistics  $\tilde{k}(A_{j,l})$  in combination with  $k$ -means clustering for all possible data sets  $A_{j,l} = \{X_1^l(t_j), \dots, X_n^l(t_j)\}$ , where  $l = 0, 1, 2$  and  $j = 0, \dots, N - l$ .
- *Estimate* the number of groups in the functional data set by  $\hat{k} = \max_{j,l} \tilde{k}(A_{j,l})$ .
- *Set up*  $\hat{C}_1, \dots, \hat{C}_{\hat{k}}$  with the complete proper feature functions associated to the  $k$ -means clusters of  $A_{\hat{j},\hat{l}} = \arg \max_{j,l} \tilde{k}(A_{j,l})$ . In case of ties, choose  $A_{\hat{j},\hat{l}}$  that provides more evidence of  $\hat{k}$  clusters, this means that has larger value of  $L_{\hat{k},\hat{k}+1}(j, l)$ .

STEP 2. *Global cluster revision.*

- *Compute*  $\hat{k}$  functional boxplots with the curves inside  $\hat{C}_1, \dots, \hat{C}_{\hat{k}}$ . Consider 3 times point wise interquartil range for maximum length of whiskers, which is the usual rule in one dimensional boxplot to identify extreme outliers. For  $r = 1, \dots, \hat{k}$ , call  $(LB_r(t), UB_r(t))$  to the lower and upper extreme outlier whisker bands of functional boxplots.

- *Identify* potential outliers as functions outside  $(LB_r(t), UB_r(t))$ , at least for one instant, for  $r = 1, \dots, \hat{k}$ .
- *Reallocate* each potential outlier into the cluster  $\hat{C}_1, \dots, \hat{C}_{\hat{k}}$  where spends more time inside  $(LB_r(t), UB_r(t))$ , for  $r = 1, \dots, \hat{k}$ .

STEP 3.

If one group is divided, repeat step 1 and 2 for this group to find possible partitions. Stop division when gap statistic in step 1 determines that there is only one group for every instant  $t_0 < \dots < t_N$  and feature of the data set (trajectories or derivatives).

## 2.2 Practical implementation

Considering again the example shown in Figure 1, we find that SeqClusFD separates in the first step the third group using derivative functions. In second step, SeqClusFD separates the first two groups considering the trajectories. In addition, SeqClusFD determines that there are no further groups. Then, three groups are found and there are not misclassified observations. Ieva et al [16], perform a  $k$  means clustering procedure for functional data, but instead of using the classical  $L^2$  norm they work in the Sobolev  $H^1$  space with the ordinary norm, this means that they consider simultaneously the information of the function and the derivative. If we use that approach in this example, and estimate the number of clusters either by the Gap Statistics or by Calinski-Harabasz approach, which are among the best well known procedures to estimate the number of clusters, neither of them detect the correct number of clusters.

In more realistic problems computation of the first two derivatives is not a simple task. Typically, functional data are characterized by a high amount of consecutive observations defined in an interval of finite length, but frequency may defer from one individual to another and observations are not necessarily equidistant. When measures do not have sampling errors, function values can be interpolated and then derivatives are computed by differentiating. Otherwise it is necessary to use smoothing techniques in order to transform observations into functions that can be evaluated for any time  $t$ . Different smoothing techniques should be considered since each data set has different properties, as for instance functions could be monotone, or periodic or non-negative. Ramsay *et al.*, [25] discuss the main options that are consider in FDA, most of them implemented in Matlab and R routines that can be downloaded from <http://www.functionaldata.org>.

For every  $t_j$  and feature of the trajectories, uniform distribution boundaries for the bootstrap samples, that are necessary for gap statistic computations, will probably be

different. Three bootstrap procedures should be executed for each  $t_j$ . Then the thinner the grid  $t_0 < \dots < t_N$  defined in step 0 of SeqClusFD algorithm more computationally expensive is step 1. To speed up SeqClusFD we suggest pointwise rescaling to the interval  $[0, 1]$  for observations in subsets  $A_{j,l} = \{X_1^l(t_j), \dots, X_n^l(t_j)\}$  ( $l = 0, 1, 2$  and  $j = 0, \dots, N - l$ ) as follows,

$$Y_i^l(t_j) = \frac{X_i^l(t_j) - \min\{A_{j,l}\}}{\max\{A_{j,l}\} - \min\{A_{j,l}\}}, \text{ for } i = 1, \dots, n.$$

Now all bootstrap samples should be generated according to a uniform distribution on the interval  $[0, 1]$ . This means that bootstrap procedure can be done only once for each visit of the algorithm to step 1. However, when step 1 is repeated it is because a group is divided and then sample size is different. We can not avoid bootstrap should be redone on different visits.

### 3 Simulation Study

To show the performance of SeqClusFD we conducted a simulation study using different artificial data sets that have been previously proposed in the literature for clustering of functional data. In all cases we report the number of times that SeqClusFD selects the correct number of groups. For these successful examples, the correct classification rate (CCR) is estimated.

As SeqClusFD is based on gap statistics and functional boxplot, before executing the algorithm it is necessary to fix the value of some parameters related to these tools. For gap statistic we follow suggestions in the original paper, except for  $n_{sd} = 3$  on equation (2.2) since we are applying a stricter rule for identifying clusters with strong evidence at a certain time period. More than 500 bootstrap samples are not necessary. For functional boxplot we follow the classical rule to identify severe outliers, considering the maximum length of the whiskers as three times the interquartil range.

To reduce computational effort in the local clustering step we limit the maximum number of clusters to five. It is important to remark that, even though this parameter fixes an upper bound for the number of clusters on each partitioning step, there is not an upper bound for the general procedure. In the global cluster revision step, A minimum of 10 functions are required to compute a boxplot. The minimum cluster size to continue with splitting is also fixed on 10 functions

We propose the same values for all the simulations and real data examples.

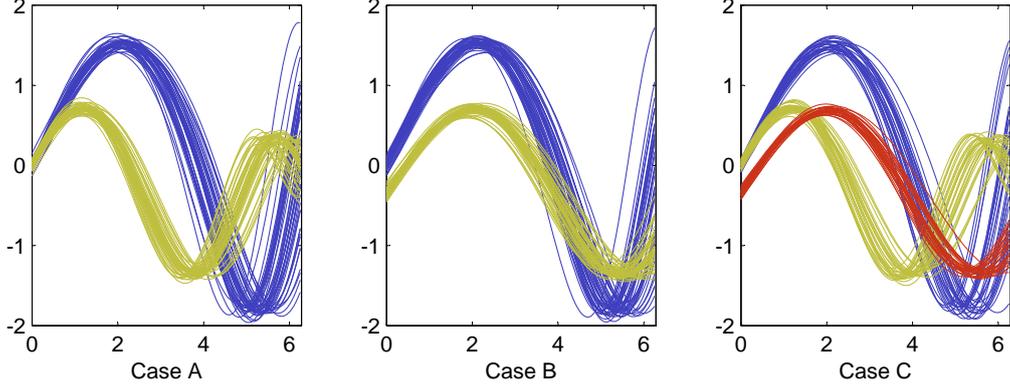


Figure 2: Examples of data set of functions simulated with models A, B and C.

### 3.1 Sampling errors associated with the entire curve

Sangalli *et al.* [27] introduced three models for curve generation. We follow their models for data generation with the idea of exploring the SeqClusFD performance in clusters structures with different amplitudes and curve registration problems.

*Model A:* Two clusters with  $n/2$  functions generated as follows,

$$\begin{aligned}
 X_i(t) &= (1 + \epsilon_{1i}) \sin(\epsilon_{3i} + \epsilon_{4i}t) + \\
 &\quad (1 + \epsilon_{2i}) \sin\left(\frac{(\epsilon_{3i} + \epsilon_{4i}t)^2}{2\pi}\right) \\
 &\quad \text{for } i = 1, \dots, n/2,
 \end{aligned} \tag{3.3}$$

$$\begin{aligned}
 X_i(t) &= (1 + \epsilon_{1i}) \sin(\epsilon_{3i} + \epsilon_{4i}t) - \\
 &\quad (1 + \epsilon_{2i}) \sin\left(\frac{(\epsilon_{3i} + \epsilon_{4i}t)^2}{2\pi}\right) \\
 &\quad \text{for } i = n/2 + 1, \dots, n.
 \end{aligned} \tag{3.4}$$

*Model B:* Two clusters with  $n/2$  functions generated in the first group following (3.3) and in the second group as follows,

$$\begin{aligned}
 X_i(t) &= (1 + \epsilon_{1i}) \sin\left(\epsilon_{3i} + \epsilon_{4i} \left(-\frac{1}{3} + \frac{3}{4}t\right)\right) - \\
 &\quad (1 + \epsilon_{2i}) \sin\left(\frac{(\epsilon_{3i} + \epsilon_{4i} \left(-\frac{1}{3} + \frac{3}{4}t\right))^2}{2\pi}\right) \\
 &\quad \text{for } i = n/2 + 1, \dots, n.
 \end{aligned} \tag{3.5}$$

*Model C:* Three clusters with  $n/3$  functions generated in the first group following (3.3), in the second group following (3.4) and in the third group following (3.5).

Table 1: Distribution of the Number of Groups found by SeqClusFD and Mean CCR.

Number of clusters	Model A	Model B	Model C
2	88	99.5	0
3	12	0.05	75.5
4	0	0	23.5
5	0	0	1
<b>Mean CCR</b>	99.67	99.96	99.45

We simulate 200 data sets of size  $n = 90$  for each model. All errors  $\epsilon_{1i}, \dots, \epsilon_{4i}$  are independent and normally distributed with mean 0 and standard deviation 0.05. Figure 2 shows three examples of 90 curves generated according with models A, B and C.

Table 1 reports for each model, percentage of data sets in which SeqClusFD finds the number of clusters indicated on first column. In bottom line are displayed the mean of correct classification rates (CCR) calculated only with data in which the number of clusters is correctly identified. In almost all cases it find the correct number of clusters. The result is poorer in the model C, which is also the most challenging problem, with a 75.5% of successful identifications. When the number of clusters is not identified correctly, generally it is determined that there is an additional cluster. We observed that in these cases one of the original cluster is divided into two and the other clusters are identified without error. Table 1 also shows the mean of correct classification rates (CCR), calculated only with data sets in which the number of clusters is correctly identified. It is clear that SeqClusFD has an outstanding performance in these cases.

Finally, we compare our procedure with other clustering techniques available in R for functional data, namely, *funclust* ([17]), *funHDDC* ([7]) and *kmeans.fd* which is a k-means procedure for functional included in the R package *fda.usc* ([13]). In all the cases the number of clusters must be given beforehand. Table 2 exhibits the mean CCR for the three procedures proposed. It is clear that SeqClusFD outperforms the three clustering strategies since it has higher CCR for all the models considered.

### 3.2 Sampling errors associated to each instant

We simulate 200 data sets with four clusters that are generated with a similar model to that introduced by Serban and Wasserman [29]. Each cluster contains 150 functions

Table 2: Mean CCR for funclust, kmeans.fd and funHDDC.

	Model A	Model B	Model C
funclust	75.62	73.57	48.48
kmeans.fd	67	76.77	68.45
funHDDC	85.81	47.39	37.86

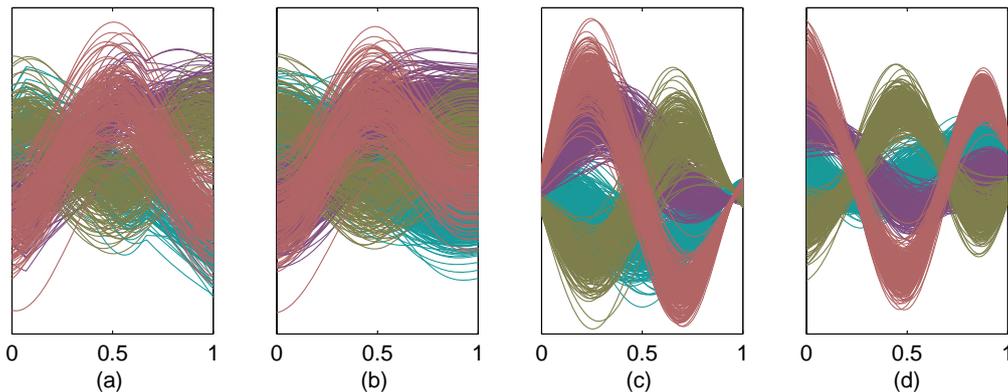


Figure 3: (a) Simulated data set with sampling errors associated to each instant; (b) smoothed data set; (c) first derivatives and (d) second derivatives.

that are generated as follows

$$X_{ij}(t) = f_j(t) + \epsilon_i(t),$$

$$\text{for } t \in [0, 1], i = 1, \dots, 150 \text{ and } j = 1, \dots, 4,$$

where

$$f_1(t) = \min \left( \frac{2 - 5t}{2}, \left( \frac{2 - 5t^2}{2} \sin \left( \frac{5\pi t}{2} \right) \right) \right),$$

$$f_2(t) = -f_1(t), \quad f_3(t) = \cos(2\pi t) \quad \text{and} \quad f_4(t) = -f_4(t).$$

In [29] the authors consider independent errors, while we consider correlated errors from a Gaussian process. For all the functions  $\epsilon(t)$  has normal distribution with mean 0.4, standard deviation 0.9 and covariance structure given by,

$$\rho(s, t) = 0.3 \exp \left( -\frac{(s - t)^2}{0.3} \right), \quad \text{for } s, t \in [0, 1].$$

Figures 3(a) and 3(b) show one of the generated data set and the corresponding smoothed data set with  $B$ -splines, respectively. Each color represents one theoretical cluster. In Figures 3(c) and 3(d) are displayed the firsts two derivatives. To prevent

Table 3: Mean CCR for *funclust*, *kmeans.fd*, *funHDDC* and *SeqClusFD*.

Clustering Procedure	Mean CCR
<i>funclust</i>	40.05
<i>kmeans.fd</i>	60.12
<i>funHDDC</i>	63.21
<i>SeqClusFD</i>	99.85

boundary effects we reflect one third of the observations at the beginning and at the end of each measurement. We also challenged our method with other clustering procedures for functional data: *funclust*, *funHDDC* and *kmeans.fd*, in these case the number of clusters was given as an input. Table 3 exhibits the mean CCR for the 200. It is important to highlight that *SeqClusFD* always identifies four groups and that has a much more higher mean correct classification rate than the rest of the procedures, achieving perfect classification in most of the replicates.

## 4 Real Data Examples

### 4.1 Berkeley Growth Study data

The Berkeley Growth Study is one of the best known long-term development research ever conducted, and the height growth data set introduced by Tuddenham and Snyder [35] is a reference to illustrate different methods for FDA. In particular, the heights of 54 girls and 39 boys measured between 1 and 18 years in 31 unequally spaced moments (see Figure 4a) are considered one of the most challenging data sets for clustering purposes. More measurements were taken when growth was more rapid, as childhood and adolescence, and least in the early years, when growth was more stable. As a consequence of that, for the computation of the first two derivatives, we need to transform observations into functions that can be evaluated for any time. We consider a monotonic cubic regression spline smoothing as suggested by Ramsay *et al.* [25].

Our first objective is to test the effectiveness of *SeqClusFD* in determining the number of groups in this data set without regard the gender information. Second, we use the output of *SeqClusFD* for classifying children. The information from step 1 gives us the key for understanding how the groups differ.

Using the same parameters as in simulation study for step 1 and 2 of *SeqClusFD*, final output identifies two clusters. The algorithm finds in the local step the highest

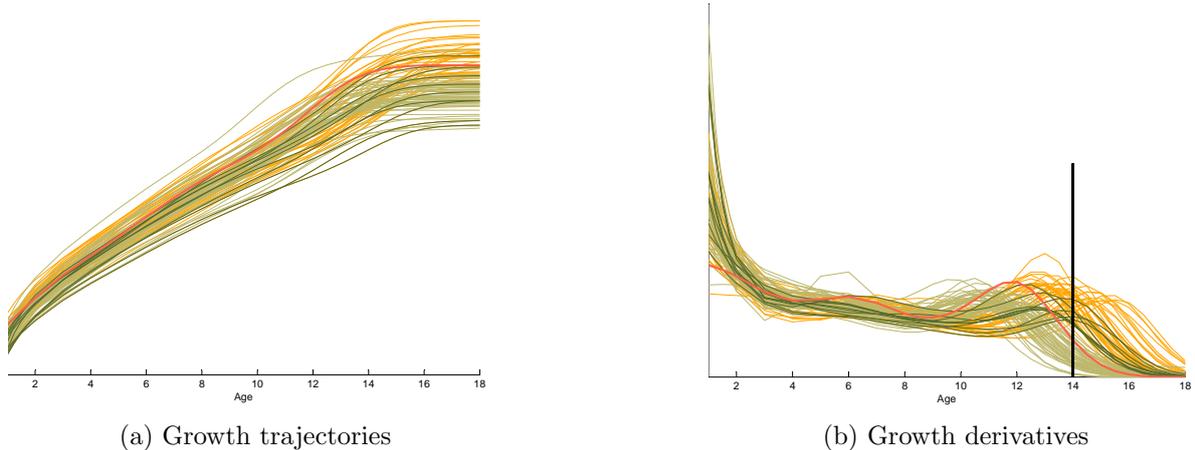


Figure 4: Heights of 54 girls (green curves) and 39 boys (orange curves) measured between 1 and 18 years. Misclassified curves by SeqClusFD are highlighted with thicker lines.

evidence of clustering in the first derivative (growth speed) at the age of 14, which is coincident with the end of puberty in girls but not in boys yet. In the final classification of the data, one cluster is coincident with boys and the other with girls. Only 10 curves are misclassified and are highlighted with thicker lines in Figures 4a and 4b. Misclassifications correspond to a boy with early puberty and 9 girls with late maturity. The correct classification rate is  $CCR = 89.25\%$ .

This data set has been used recently by Jacques and Preda [17] to compare several clustering methods for functional data. All analyzed methods make use of the information that the number of clusters is two, which could suggest that *a priori* SeqClusFD is less competitive. However the CCR for alternative methods are not always superior:  $\text{funclust-CCR} = 69.89\%$  (Jacques and Preda, [17]);  $\text{FunHDDC-CCR} = 96.77\%$  (Bouveyon and Jacques, [7]);  $\text{fclust-CCR} = 69.89\%$  (James and Sugar, [19]); and  $\text{kCFC-CCR} = 93.55\%$  (Chiou and Li, [9]). SeqClusFD is closer to the most successful (FunHDDC and KCFC) than to funclust and fclust. In addition, the output of SeqClusFD also includes an estimate of the number of groups and provides assistance for the interpretation of the clusters.

Sangalli *et al.* [28] also analyzed this data set. Although they find more evidence for the existence of a single cluster, they analyze the case of two clusters and the best CCR obtained is  $88.17\%$ .

## 4.2 EGC200 data

The 200 electrocardiograms of EGC200 data set can be found in the UCR Time Series Classification and Clustering website [8]. Data set has two groups, with 133 and 67 electrocardiograms each one, all of them recorded at 96 equally spaced instants. Left side of Figure 5 shows electrocardiograms ( $f$ ) and their first ( $f'$ ) and second ( $f''$ ) derivatives for the two clusters (orange and blue curves). These data has been analyzed by Jacques and Preda [17], among others, using the same clustering procedures than in the previous example, except kCFC. The results are: funclust-CCR = 84%; FunHDDC-CCR = 75%; and fclust-CCR = 74.5%.

Tree structure in central part of Figure 5 shows the results on the three iterations needed by SeqClusFD to complete the clustering procedure. We use the same parameters as in simulation study for step 1 and 2. In the first iteration, SeqClusFD algorithm detects two groups using the information provided by the electrocardiograms at  $t = 43$ . When the algorithm visits again step 1 for the first group, gap statistic determines that there is only one cluster and SeqClusFD stops division. The majority of the curves in this cluster are blue, 35, and 19 curves are orange. In the second iteration, the other group is divided in two clusters at  $t = 41$  with the information of the first derivatives. One cluster is a terminal node with 99 orange curves and only 7 blue that could be considered as misclassifications. In the third iteration, the remaining observations are separated in two groups at  $t = 24$  considering the information of the second derivatives. The green curves are a terminal cluster with 21 blue curves and 15 oranges. Finally, last four curves can not be separated because of the small size of the group (under 10). When exploring which are these curves we find that they can be considered as outliers. Some authors, see for instance [17], eliminate these electrocardiograms from the beginning of the analysis.

Right side of Figure 5 shows the final three cluster allocation in  $f$ ,  $f'$  and  $f''$ . The four outliers appear in dark green. Even though this data set only contains two clusters and we have found three, seen in the results that the group found in second iteration can be considered the same as the original cluster of orange curves, with 7 curves that are misclassified. Assuming that the original cluster of blue electrocardiograms is the union of the other two (blue and green), the CCR is 79.08%. For this calculation we have excluded the four outliers in order to compare SeqClusFD results with those of Jacques and Preda [17].

The result of SeqClusFD classification is very good in comparison with the other methods. Although SeqClusFD-CCR is not the highest CCR, note that the other methods start from an advantageous position by assuming the number of clusters is known. SeqClusFD is only overcome by funclust.

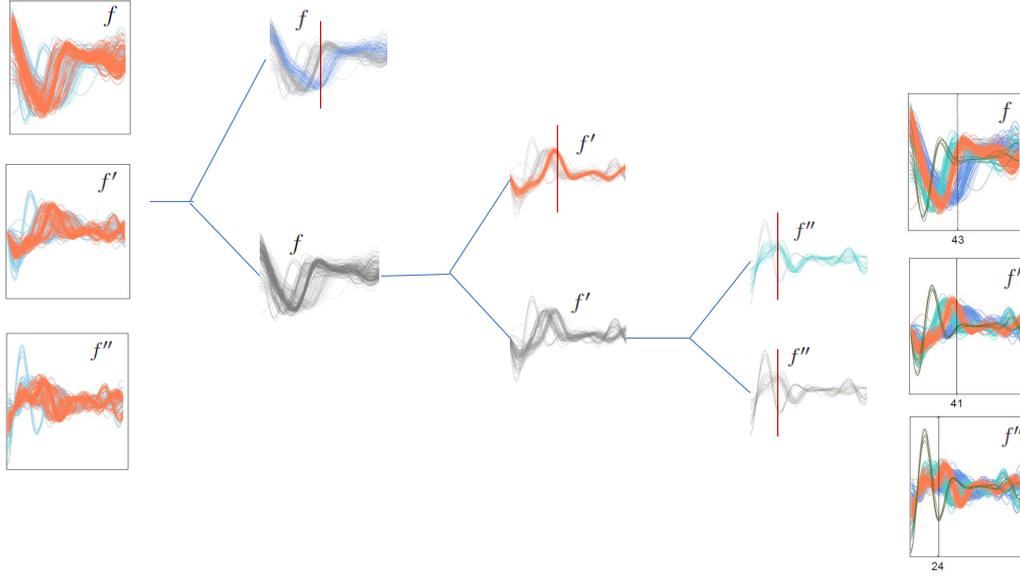


Figure 5: On the left side, electrocardiograms ( $f$ ), first ( $f'$ ) and second ( $f''$ ) derivatives for the two clusters (blue and orange) of ECG200 data set. On the central part, results of the three iterations executed by SeqClusFD. On the right side, final classification with SeqClusFD.

### 4.3 Italy Power Demand data

Italy Power Demand data set can also be found in the UCR Time Series Classification and Clustering website [8]. Data set contains two clusters: cluster 1 with 513 curves (see Figure 6(a)); and cluster 2 with 516 curves (see Figure 6(b)). Note that inside each cluster there are two different pattern of consumers. All curves were measured at 24 equally spaced moments along a day. To increase the grid size on step 0 of SeqClusFD algorithm, data are smoothed with six equally spaced knot splines (choosing more knots lead to similar results). To prevent boundary effects we reflect one third of the records at the beginning and end of each record. SeqClusFD is executed with the same parameters as in previous examples.

SeqclusFD finds 4 clusters in the two iterations shown in the tree of Figure 7, identifying the original clusters and the two different pattern of consumers within them. The partitions are always based on information provided by first derivatives, at  $t = 7$  in the first iteration and  $t = 23$  and  $t = 9$  in the second. The final classification is shown on the right side of Figure 7. Orange curves are cluster 1, while green curves are cluster 2. The darker green and orange curves are misclassified power demand curves. Collapsing the four clusters in only two we can calculate CCR, which is 93.49%.

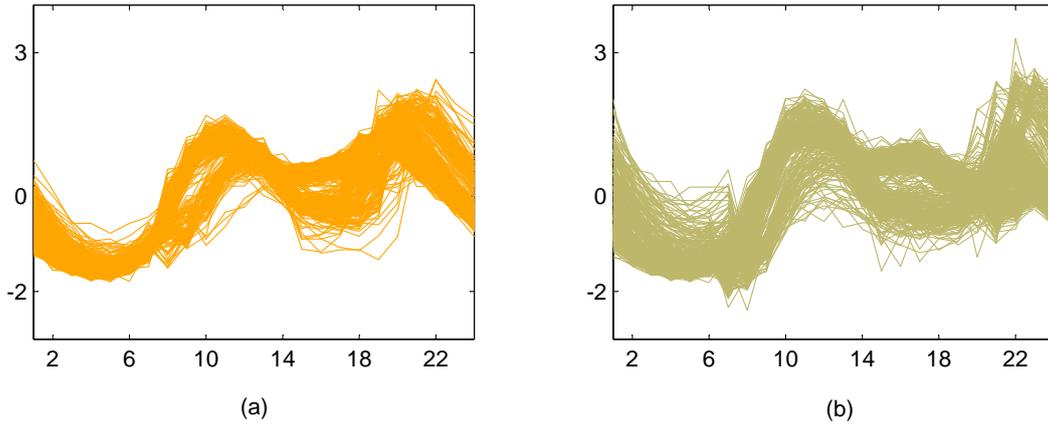


Figure 6: Power demand curves along a day in Italy: (a) 513 curves in cluster 1; and (b) 516 curves in cluster 2.

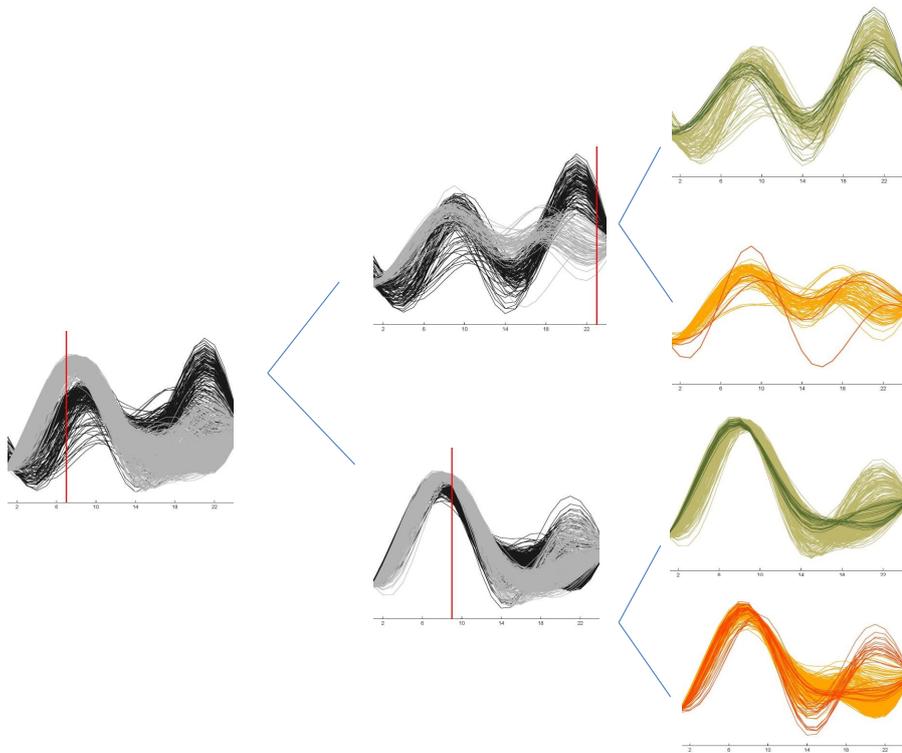


Figure 7: Results of the two iterations executed by SeqClusFD to classify curves of power demand in Italy. First derivatives are shown in all iterations.

## 5 Conclusion

We introduce SeqClusFD, a hierarchical clustering algorithm for functional data that simultaneously determines the number of groups and the clustering conformation. As we are interested in cluster structures that takes into account the shape of the curves,

we also consider the speed and acceleration functions linked to the original curves. It is a sequential procedure, that successively runs two steps, it searches the instant of time, along the functions and its derivatives, with most clustering ability, according to the gap statistics criterion and one dimensional k-means algorithm. Once this initial groups are conformed it builds up a functional boxplot for each group, with the objective of reallocating possibly misclassified data. Moreover, it provides helpful information towards the grouping structure.

Although the SeqClusFD is based on the one dimensional gap statistic, alternative methods to estimate the number of groups could be considered. Similarly, the functional boxplot could be adapted to consider different depth definitions for functional data. Several proposals can be found in [24].

The algorithm could be easily extended to the multivariate case, where each data is a vector of functions (see [4]). In the local step, any clustering method could be considered. On the global revision step, the boxplot could be designed using band depth proposed by Lopez-Pintado *et al.* [22] for multivariate functional data. The efficiency of the method should decrease with the dimensionality.

Finally, the output of the algorithm exhibits number of groups and clustering allocation. In addition, it gives information about the *key* moments  $t \in [a, b]$  and features, i.e.  $X(t)$ ,  $X^1(t)$  or  $X^2(t)$ . This yields to a better understanding of the clustering structure. This topic is closely related to the variable selection problem that has been widely studied recently in the supervised classification framework, see for instance Tian and James [32] and Martin-Barragan *et al.* [23]. However, this problem has not been analyzed for unsupervised classification problems, at the best of our knowledge.

## References

- [1] C. Abraham, P.A. Cornillon, E. Matzner-Løber and N. Molinari, “Unsupervised curves clustering using B-Splines”, *Scandinavian Journal of Statistics*, 30, 581–595, 1993. DOI: 10.1111/1467-9469.00350
- [2] A.M. Alonso, D. Casado and J. Romo, “Supervised classification for functional data: a weighted distance approach”, *Computational Statistics and Data Analysis*, 56, 2334–2346, 2012. DOI: 10.1016/j.csda.2012.01.013
- [3] J. Basak and R. Krishnapuram, “Interpretable hierarchical clustering by construction an unsupervised decision tree”. *IEEE Transactions on Knowledge and Data Engineering*, 17, 121–132, 2005. DOI:10.1109/TKDE.2005.11

- [4] J.R. Berrendero, A. Justel, and M. Svarc, “Principal components for multivariate functional data”. *Computational Statistics and Data Analysis*, 55, 2619–2634, 2011. DOI:10.1016/j.csda.2011.03.011
- [5] J.G. Booth, G. Casella and J.P. Hobert, “Clustering using objective functions and stochastic search”, *Journal of the Royal Statistical Society, B*, 70, 119–139, 2008. DOI: 10.1111/j.1467-9868.2007.00629.x.
- [6] M. Boullé, R. Guigourès, and F. Rossi, “Non parametric hierarchical clustering of functional data”. *Advance in Knowledge Discovery and Management, Studies in Computational Intelligence*, 5, 15–35, 2014. DOI:10.1007/978-3-319-02999-32.
- [7] C. Bouveyron and J. Jacques, “Model-based clustering of time series in group-specific functional subspaces”. *Advances in Data Analysis and Classification*, 5, 281–300, 2011. DOI:10.1007/s11634-011-0095-6.
- [8] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen and G. Batista. The UCR Time Series Classification Archive, 2015. ([http://www.cs.ucr.edu/~eamonn/time\\_series\\_data/](http://www.cs.ucr.edu/~eamonn/time_series_data/))
- [9] J.M. Chiou, and P.L. Li, “Functional clustering and identifying substructures of longitudinal data”. *Journal of the Royal Statistical Society, B*, 69, 679–699, 2011. DOI: 10.1111/j.1467-9868.2007.00605.x.
- [10] J.A. Cuesta Albertos and R. Fraiman, “Impartial Trimmed  $k$ -means for Functional Data”. *Computational Statistics and Data Analysis*, 51, 4864–4877, 2007. DOI:10.1016/j.csda.2006.07.011
- [11] A. Cuevas, “A partial overview of the theory of statistics with functional data”. *Journal of Statistical Planning and Inference*, 147, 1-23, 2014. DOI: 10.1016/j.jspi.2013.04.002.
- [12] R. Fraiman, B. Ghattas and M. Svarc, “Interpretable clustering using unsupervised binary trees”. *Advances in Data Analysis and Classification*, 7, 125–145, 2013. DOI:10.1007/s11634-013-0129-3
- [13] M. Febrero-Bande, M. Oviedo de la Fuente, “Statistical Computing in Functional Data Analysis: The R Package *fda.usc*”. *Journal of Statistical Software*, 51(4), 1-28, 2012. DOI: 10.18637/jss.v051.i04
- [14] C. Genolini and B. Falisard, “KmL:  $k$ -means for longitudinal data”. *Computational Statistics*, 25, 317–328, 2010- DOI : 10.1007/s00180-009-0178-4.

- [15] R. Giraldo, P. Delicado and J. Mateu, “Hierarchical clustering of spatially correlated functional data”. *Statistica Neerlandica*, 66, 403–421, 2012. DOI: 10.1111/j.1467-9574.2012.00522.x.
- [16] F. Ieva, A. M. Paganoni, D. Pigoli and V. Vitelli, “Multivariate functional clustering for the morphological analysis of electrocardiograph curves”. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62, 401-418, 2013. DOI: 10.1111/j.1467-9876.2012.01062.x
- [17] J. Jacques and C. Preda, “Model-based clustering of functional data.”, *Computational Statistics and Data Analysis*, 71, 92–106, 2014. DOI:10.1016/j.csda.2012.12.004
- [18] J. Jacques and C. Preda, “Functional Data Clustering: A Survey.”, *Advances in Data Analysis and Classification*, 8 (3), 231–255, 2014. DOI : 10.1007/s11634-013-0158-y.
- [19] G. James and C. Sugar, “Clustering for sparsely sampled functional data”. *Journal of the American Statistical Association*, 98, 397-408, 2003. DOI: 10.1198/016214503000189.
- [20] B. Liu, Y. Xia and P.S. Yu, “Clustering through decision tree construction”. *CIKM 00. In: Proceedings of the ninth international conference on information and knowledge management*. ACM, New York, NY, USA, pp 20–29, 2000. DOI:10.1145/354756.354775
- [21] S. López-Pintado and J. Romo, “On the concept of depth for functional data.” *Journal of the American Statistical Association*, 104, 718–734, 2009. DOI:10.1198/jasa.2009.0108
- [22] S. López-Pintado, Y. Sun, J.K. Lin and M.G. Genton, “Simplicial band depth for multivariate functional data”. *Advances in Data Analysis and Classification*, 8, 321–338, 2014. DOI:10.1007/s11634-014-0166-6
- [23] B. Martin-Barragan, R. Lillo and J. Romo, “Interpretable support vector machines for functional data”. *European Journal of Operational Research*, 232, 146.–155, 2014. DOI:10.1016/j.ejor.2012.08.017
- [24] K. Mosler, “Depth Statistics”. *Robustness and Complex Data Structures. Editors: Becker, C., Fried, R. and Kuhnt, S.* Springer Berlin Heidelberg. 17–34, 2013.
- [25] J. Ramsay, G. Hooker and S. Graves, *Functional Data Analysis with R and Matlab*. Springer, New York, 2009.

- [26] J. Ramsay and B.W. Silverman, *Functional Data Analysis* (2nd ed). Springer, New York, 2005.
- [27] L.M. Sangalli, P. Secchi, S. Vantini and V. Vitelli, “ $k$ -means alignment for curve clustering.” *Computational Statistics and Data Analysis*, 54, 1219–1233, 2010. DOI:10.1016/j.csda.2009.12.008
- [28] L.M. Sangalli, P. Secchi, S. Vantini and V. Vitelli, “Functional clustering and alignment methods with applications”. *Communications in Applied and Industrial Mathematics*, 1, 205–224, 2010. DOI: 10.1685/2010CAIM486.
- [29] N. Serban and L. Wasserman, “CATS: Cluster Analysis by Transformation and Smoothing”. *Journal of the American Statistical Association*, 100, 990–999, 2005. DOI 10.1198/016214504000001574
- [30] Y. Sun and M.G. Genton, “Functional boxplots”. *Journal of Computational and Graphical Statistics*, 20, 316–334, 2011. DOI: 10.1198/jcgs.2011.09224
- [31] T. Tarpey and K.K.J. Kinader, “Clustering functional data”. *Journal of classification*, 20, 93–114, 2003. DOI : 10.1007/s11634-013-0158-y
- [32] T.S. Tian and G.M. James, “Interpretable dimension reduction for classifying functional data”. *Computational Statistics and Data Analysis*, 57, 282–296, 2013. DOI:10.1016/j.csda.2012.06.017
- [33] R. Tibshirani, G. Walther and T. Hastie, “Estimating the number of data clusters via the gap statistic”. *Journal of the Royal Statistical Society, B*, 63, 411–423, 2001. DOI: 10.1111/1467-9868.00293.
- [34] S. Tokushige, H. Yadohisa and K. Inada, “Crisp and Fuzzy  $k$ -means clustering algorithms for multivariate functional data”. *Computational Statistics*, 21, 1–16, 2008. DOI:10.1007/s00180-006-0013-0
- [35] R.D. Tuddenham and M.M. Snyder, M.M., Physical growth of California boys and girls from birth to eighteen years, *Tech. Rep. 1, University of California Publications in Child Development.*, 1954.

## Acknowledgments

Ana Justel is supported by MINECO (Spain), grants CTM2011-28736/ANT and CTM2013-47381-P