Functional Data Analysis in Sport Science: Example of Swimmers' Progression Curves Clustering

Arthur Leroy, Servane Gey

FUNCTIONAL DATA ANALYSIS IN SPORT SCIENCE: EXAMPLE OF SWIMMERS' PROGRESSION CURVES CLUSTERING

Arthur Leroy
MAP5 - Paris Descartes University,
IRMES - INSEP,
arthur.leroy@insep.fr

Servane Gey
MAP5 - Paris Descartes University,
servane.gey@parisdescartes.fr

August 27, 2018

Abstract

Many data collected in sport science come from time dependent phenomenon. This article focuses on Functional Data Analysis (FDA), which study longitudinal data by modeling them as continuous functions. After a brief review of several FDA methods, some useful practical tools such as Functional Principal Component Analysis (FPCA) or functional clustering algorithms are presented and compared on simulated data. Finally, the problem of the detection of promising young swimmers is addressed through a curve clustering procedure on a real data set of performance progression curves. This study reveals that the fastest improvement of young swimmers generally appears before 16 years old. Moreover, several patterns of improvement are identified and the functional clustering procedure provides a useful detection tool.

Keywords Curve clustering; Functional Data Analysis; Swimming; Sport ; Detection

1 Introduction

For a long time, sport science has been interested by time dependent phenomenons. If, at first, people only kept track of performance records, there is currently a massive amount of various data. Among them, one specific type is called *time series* or *longitudinal data*. Many recorded and studied data can be considered as time series depending on the context. From the heart rate during a sprint [17], to the number of injuries in a team over a season [5], to the evolution of performances during a whole career [2], the common ground remains the evolution of a characteristic regarding a time period. An interesting property of such data lies in the dependency between two observations at

two different instants, leading, in mathematical terms, to the fact that the independent and identically distributed (iid) hypotheses are not verified. However, most of the usual statistical tools classically used in Sport Science, such as the law of large number and central limit theorem, need these properties[1]. Thus, all the statistical methods based on these results (hypothesis testing, method of moments, ...) collapse, and one needs specific tools to study time series. There is a whole literature related to the subject [4]. These methods focus on the study of time dependent processes that generate discrete observations. For instance, since an important topic of this paper concerns clustering, and a really comprehensive review about clustering of time series can be found in [25]. Despite the usefulness of such an approach, some theoreticians proposed a new modeling of the problem [7]. In many cases, the studied phenomenon is actually changing continuously over the time. Thus, the object we want to know about is more of a function than a series of point. In their paper [5], the authors highlight that it may be damageable to discretize phenomenons that are intrinsically functional. Moreover, they claim that continuous methods perform better than discrete ones on the specific case of the relationship between training load and injury in sport.

In some particular cases, it thus seems natural to model a continuous phenomenon as a random function of time, formally a stochastic process, and consider our observations as just few records of an infinite dimensional object. This approach is called functional data analysis (FDA) and gives a new range of methods well suited to work on longitudinal data. There was substantial theoretical improvements in the area the last two decades, and this paper intends to present some topics that might be useful to the sport science field. To our knowledge, there is very few paper in the sport literature that use FDA. We can cite [9] in which curve clustering is used to analyse the foot-strike of runners, or [18] for the study of muscle fatigue through a whole FDA analysis. Another example is given in [11] that proposes a functional version of ANOVA using splines to overcome common issues that occur in sport medicine.

The purpose of this paper is twofold : at first, it aims at providing a brief review of several methods and references for the theoretical aspects. Secondly, examples of practical tools and useful packages (on the software R as it is currently the most convenient to perform FDA) of state of the art methods are presented. Then, we also detail a specific study on a real data set, coming from our collaboration with the French Swimming Federation. This work focuses on the clustering of performance progression curves of young male swimmers and uses several FDA tools. We emphasise on the fact that FDA provides some tools that give information we could not exhibit otherwise, like the study of derivatives for example.

As mentioned previously, FDA allows to take into account the intrinsic nature of functional data. Apart from this philosophical advantage in term of modeling, one may note important benefits. For example, if one records several time series with observations at different instants and/or in different numbers, how to compare them ? How to study the evolution of performances of

[1]Note that there exist several versions of these theorems with more or less flexible hypotheses, depending on the context. We talk here about the most common versions, classically used in applied science.

swimmers from their competition times at given ages ? Competitors may have different number of races during their careers, and their performances are done at different ages (if one wants to avoid age discretization that have been shown problematic in [26]). This example illustrates exactly what we try to deal with in the following study on the data set of french male swimmers. Another fundamental advantage of FDA is the possibility to work on the derivatives of the observed functions. Indeed, it is often interesting to study the dynamic of a time dependent process. Even the second derivative, often referred as the *acceleration*, or a superior order derivative might provide valuable information in practice. The specific nature of functional data allows to study such properties, and the sport scientist may easily imagine the wide range of situations on which the study of derivatives might be interesting. One could think for example of the GPS position tracking analysis, the progression phenomenon of young athletes, or the following of actions of some muscles over time.

The first and fundamental step of a functional data analysis generally consists in the reconstruction of the function from the discrete set of observations. There is two cases at this step. Whether the observations are being considered as error-less (in term of measurement) and one can proceed to a direct interpolation through one of the multiple existing methods (linear, polynomial, ...). Or, more frequently, the set $x_{i,t_1}, \ldots, x_{i,t_n}$ is considered as observations at time $t_1, \ldots, t_n$ of a realisation $x_i(t)$ of a stochastic process $X(t)$. In this case, one can proceed to a *smoothing* step. It consists in the approximation of a function defined to be *close* to the observed points. To deal with noisy data, one always has to face the over-fitting/under-fitting issue. In most cases, one has to determine a smoothing parameter that define how much one wants to allow the function to contain *peaks*. These topics are largely detailed in the first chapters of [21]. Even if defining a consistent value of the smoothing parameter is a first work, one can see as an advantage the fact to explicitly control the signal-on-noise ratio of the data. The most common way to reconstruct the function from the observations is to use a basis of functions. A basis of functions is a set of specific functions ϕ_i of a functional space $\mathcal{S}$, such as each element of $\mathcal{S}$ can be defined as a linear combination of the ϕ_i. Formally, we can define the basis expansion f as :

$$f(t) = \sum_{i=1}^{N} \alpha_i \ \phi_i(t) \tag{1}$$

where $\phi_1, \ldots, \phi_N$ are the basis functions of a given functional space and $\alpha_1, \ldots, \alpha_N$ are real valued coefficients. Intuitively, if one fixes a common basis to fit observations, the information on individuals is contained in the vector of coefficients $\{\alpha_1, \ldots, \alpha_N\}$. That is why a common approach is to perform classical multivariate methods on these coefficients. Among the most common basis used in practice, we can cite Fourier basis and wavelets, which are well suited for periodic data [21] [13]. For non periodic data, the classical choice is spline basis, particularly the cubic splines in practice [7]. They allow to approximate a wide range of shapes with a rather good smoothness [20]. From a computational point of view, one can use the *R* package *fda*, on which one can find methods to fit observations into functional data, and way more tools for FDA. An overview of the *fda* package can be found in [20].

Once the data set is approximated by functions, one may perform analysis on them, and some classical statistical tools have been extended in the functional context. One of the first and most important adapted method was the functional principal component analysis (FPCA). Although slightly different, FPCA provides analogous information as the finite dimensional version [21]. This method allows to describe data into a non correlated low dimensional space. That is why it provides an excellent explanatory tool to visualize main features of the data as well as a way to reduce the number of informative dimensions. This can be particularly useful when one wants to apply algorithms on the vector $\{\alpha_1, \ldots, \alpha_N\}$ of coefficients of the basis expansion, with N rather large. It may accelerate calculation while retain most of the information as well as avoid curse of dimension in a big data context. We may also cite several methods presented in [21] such as *functional canonical correlation*, *discriminant analysis* and *functional linear models*.

In this article, we emphasise on the *clustering* approach, often fundamental when exploring a new data set or beforehand to a forecast. This method consists in computing sub-groups of individuals on a data set that make sense in the context of the study. Given K the number of clusters, a clustering algorithm would apply one or several rules to gather individuals presenting common properties. This problem has been largely explored these past ten years in the functional context and we will give some elements to summarize the state of the art. According to the survey [13], functional data clustering algorithms can be sorted in three distinct families, detailed below. We do not develop on direct clustering on raw observational points that does not take into account functional nature of the data and may give poor results.

(*i*) *2-steps methods*. The first step consists in the fitting procedure we detailed previously, choosing a common basis for all data. Then, a clustering algorithm such as k-means [1], or hierarchical clustering methods for example, is performed on the basis coefficients. If this vector of coefficients is in high dimension, one can add a step of FPCA and perform the clustering on the scores coming from the first eigenfunctions of the FPCA.

(*ii*) *Non-parametric clustering*. An overview of non-parametric functional data analysis is provided by [8]. It details many aspects where one does not assume that functional observations can be defined by a finite number of parameters. The idea is to define a *distance* between the functional observations without assumptions on the form of the curves. A classical measure of proximity between functions x_i and x_j is defined as :

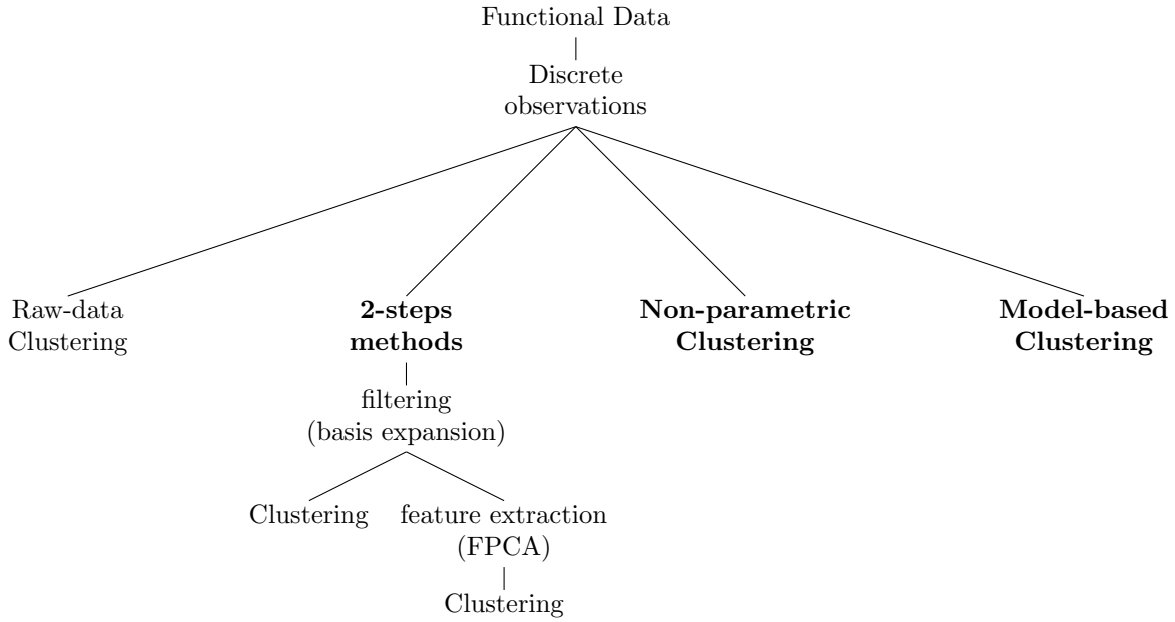
$$d_l(x_i, x_j) = \left(\int_{\mathcal{T}} x_i^{(l)}(t) - x_j^{(l)}(t) dt \right)^{\frac{1}{2}} \tag{2}$$

where $x_i^{(l)}$ is the l-th derivative of x. With such a measure of distance, one can run the heuristic of the k-means, for example, or any other distance-based clustering.

(*iii*) *Model-based clustering*. This approach has been widely developed in

the past years and gives good results. As the 2-step approach, it often uses basis expansion and/or FPCA to fit the data. However, rather than proceeding in two-step, the clustering is performed simultaneously. Many algorithms are based on Gaussian mixture models coupled with an EM-algorithm to compute the parameters [12] [23] [3]. We chose in this study to adapt the algorithm FunHDDC presented in [3] for several reasons that we develop in the following *Materials and Methods* section.

Note that the literrature does not give specific indications on which family of methods to use in a specific context and one might test several of them. Nevertheless, one should keep in mind that the right way to fit the data into functions strongly depends on the structure of the data. We also give some additional references in the next section, where we detail some algorithms that are easy to use in practice because of their implementation withon an unified R package. Below, the graph from [13] summarizes efficiently the different families and the process of clustering in a functional context :



2 Materials and Methods

Description of the real swimming data set

First of all, two types of data sets, on which we performed functional clustering algorithms, are described. The way we simulated data sets to test several methods will be described at the end of the current section. The real data have been collected by the French Swimming Federation. It gathers all the performances of french male swimmers, since 2002, for the 100m freestyle in a 50m pool. Because of confidentiality issues, athletes are identified by a number. The data set is composed of 46115 performances and ages of 1468 different swimmers, and is

available on the Github page of the corresponding author. All the algorithms were run on the R software and the corresponding packages will be named in the sequel.

Testing several algorithms on simulated data sets

To begin, a comparative study of several classical functional clustering algorithms has been performed on simulated data. Only few information are provided here on these methods, and we invite the reader to refer to the corresponding papers. For this work, the R package *funcy*, which compiles seven state of the art algorithms, was used. It gives a common syntax and format for the input and output data. The list below enumerates the algorithms, regrouped according to their family, that can be used with *funcy*.

(ii) distance-based:

- distclust, a distance based algorithm that allows irregular measurement. [19]

(iii) model-based :

- fitfclust, a mixed mixture model based algorithm that allows irregular measurement [14].
- iterSubspace, a model based algorithm, based on a subspace projection, that allows irregular measurements. Dimension between clusters can vary [6].
- funclust, a mixed mixture model based algorithm [12].
- funHDDC, a mixed mixture model based algorithm. Dimension between clusters can vary [3].
- fscm, a mixed mixture model based algorithm [15].
- waveclust, a mixed mixture model based algorithm that uses a fitting step with a Wavelet basis [10].

Unfortunuatly, the current version (1.0.0) of the *funcy* package has troubles with the funHDDC algorithm, which is not directly usable at the moment. All the remaining algorithms were applied on three simulated data sets, with $K = 4$ groups. The resulting clustering were compared to real group distributions using the Rand Index (RI)[22]. This measure, between 0 and 1, is computed by counting according pairs of individuals between two different partitions of a data set. The RI is provided as a result of the *funcit* function of the *funcy* package, and compares the ability of each procedure to retrieve the actual groups. Then, graphs of centers of each curve clusters were drawn to analyse consistency of our results according to the original data.

Clustering the real swimming data set

As mentioned above, the real data set is very irregular, with no accordance in time and in number of measurements between athletes. Thus, the first step

of the analysis was the definition of a common ground through a smoothing procedure. According to the non-periodic form of the data and the relatively low number of observational points (around 30) for each athlete, a B-spline basis was chosen. The study focus on the age period from 12 to 20 years old, which is crucial in the progression phenomenon that we aimed at studying. A basis of seven B-splines of order 4 was defined so that the derivatives remain smooth. Since we did not wish to focus on a specific time period, the knots were equally placed on ages 13 to 19. This fitting procedure was performed thanks to the *fda R* package. To analyse efficiently a real data set, one needs first to explore it, to figure out the more suited algorithm to use. To this purpose, a FPCA was performed on the progression curves and their derivatives, separately. We looked at the percentage of variance explained by each eigenfunction and the shapes of them, to understand the main features of the curves. The funHDDC algorithm was used as clustering procedure. One can find more details in the result section about the reasons of this choice. Although implemented in the *funcy* package, we chose to work with the original *funHDDC R* package, because of current problems of implementation on it. Several features of the package were used, as Bayesian Information Criterion (BIC), Integrated Classification Likelihood (ICL) and slope heuristic, to deal with problems of model selection and choice of the number K of clusters. The clustering was performed on the curves and their derivatives, separately at first. Then, the resulting clusters were compared thanks to the Adjusted Rand Index (ARI) [22], which is an extended version of the RI to partitions with different number of clusters. This measure allows to quantify the adequacy between individuals grouped whether by a clustering on progression curves or on derivatives. Noticing that athletes were clustered differently, providing two types of information, we decided to perform a third clustering procedure. This time, the multivariate clustering version on the funHDDC algorithm was used. The term multivariate clustering refers to a clustering algorithm that deals with multidimensional functions. The progression curves were defined as a first variable, while the derivatives as a second variable. For each clustering procedure, the resulting clusters centers and curves were plotted. Finally, the results were analysed and discussed with swimming experts to confront the found clusters to the sport logic.

2.1 Definition of the simulated data sets

We defined three simulated data sets to test the algorithms of the *funcy* package on different contexts. We used the function *sampleFuncy* of the *funcy* package that provides an easy way to simulate data sets suited to apply directly methods from *funcy* on them.

Sample 1:
Data are sampled from four different processes of the form $f(t) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 0.05)$. The four functions f are :

- $t \longmapsto t - 1$
- $t \longmapsto t^2$
- $t \longmapsto t^3$
- $t \longmapsto \sqrt{t}$

For each process, 25 curves are simulated and each is observed at 10 regular instants on the t-axis. This sample corresponds to a low variance situation with well separated processes.

Sample 2:
The data set is the same as Sample 1 with $\varepsilon \sim \mathcal{N}(0, 0.1)$. This sample corresponds to middle variance situation with well separated processes.

Sample 3:
Data are sampled from four different processes of the form $f(x) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 0.5)$. The four functions f are :

- $t \longmapsto t - 1$
- $t \longmapsto -t^2$
- $t \longmapsto t^3$
- $t \longmapsto sin(2\pi t)$

For each process, 25 curves are simulated and each is observed at *irregular* instants on the t-axis (less than 10 instants and/or not regularly placed). Since the observations are irregular, we had to proceed to a fitting step to use three of the six methods of the package, which are not implemented in this case. We used the function *regFuncy* of the *funcy* package for this purpose. This sample corresponds to a high variance situation with crossing processes observed irregularly.

3 Results

Results on simulated data

The Table 1 below provides results on the comparison between the six algorithms of the *funcy* package. These results are mainly illustrative and one should be aware that the quality of a clustering algorithm cannot be addressed through simulation. However, it can give some clues on the type of situations where algorithms seem to perform properly or not. The sample 1 was designed to be easy to cluster and most model based algorithms perform well. Nevertheless, they are outperformed by the only distance based method *distclust* gives almost perfect results. As Sample 2 is simply a noisier version of Sample 1, the problem becomes harder and results slightly decrease. One can note that, although the stochastic processes we sampled from are the same as in Sample 1, the "hierarchy" between methods changes. This might indicate differences at noise robustness between the methods. For example, performances of the *fcsm* algorithm decrease only slightly compared to *distclust*. Finally, as expected, the results fall on the fuzzy situation of Sample 3. Only three methods achieve moderate performances, and one can note that there is an algorithm of both families among them. Although Table 1 informs on the performances of these algorithms, it does not give information on the ability of the methods to retrieve the actual shape of the underlying functions. The following graphs will add some visual evidences to judge quality of the results.

Table 1: Mean Rand Index and (Standard Deviation) on 100 simulations of the tree samples.

Method	**Sample 1**	**Sample 2**	**Sample 3**
fitfclust	0.945 (0.14)	0.857 (0.01)	0.307 (0.06)
distclust	**0.996 (0.01)**	0.888 (0.05)	0.523 (0.07)
interSubspace	0.938 (0.14)	0.850 (0.12)	**0.527 (0.07)**
funclust	0.450 (0.17)	0.418 (0.16)	0.084 (0.07)
fscm	0.948 (0.12)	**0.902 (0.01)**	**0.527 (0.07)**
waveclust	0.920 (0.12)	0.810 (0.01)	0.324 (0.13)

Figure 1 gives one representation of the Sample 1 curves. In addition, the curves of each clusters centers of the best performing algorithm are drawn. One can see that Sample 1 is quite simple to deal with, since curves of different groups are well separated. Not surprisingly, the *distclust* clustering algorithm satisfyingly figures out the actual shape of each process.

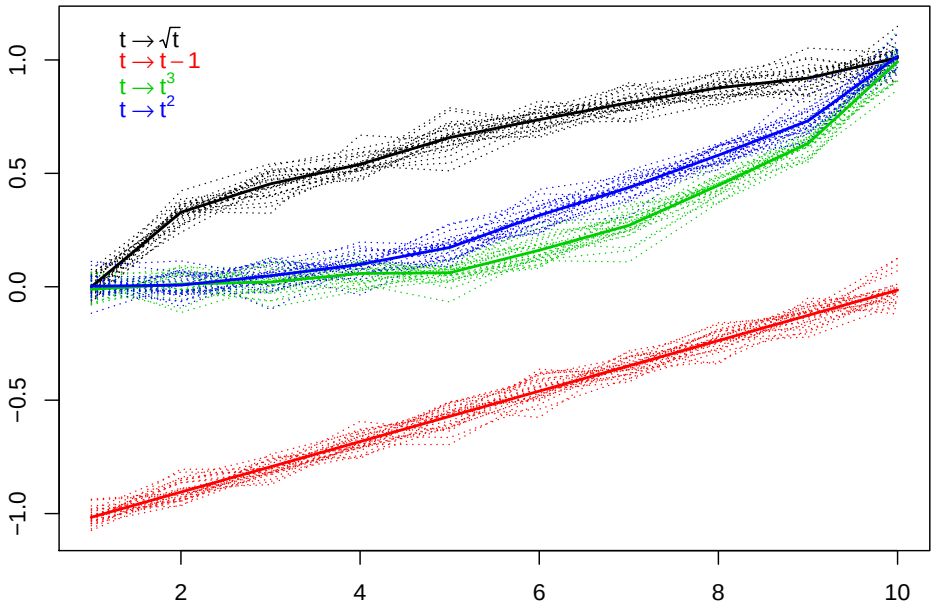


Figure 1: All curves (dotted lines) and cluster centers curves (plain lines) obtained with *distclust* algorithm for Sample 1

One can see on Figure 2 that, if the noisier situation of Sample 2 affects the good clustering rate, the shapes of the underlying functions remain correctly approximated by clusters centers of *fscm*.

Sample 3 was designed to be trickier since curves cross each other and the signal appears rather noisy. In this context, one can see on Figure 3 that, as expected, the algorithms retrieve approximately the true shapes of the underlying functions. While the *sinus* (in black) function seems correctly identified,

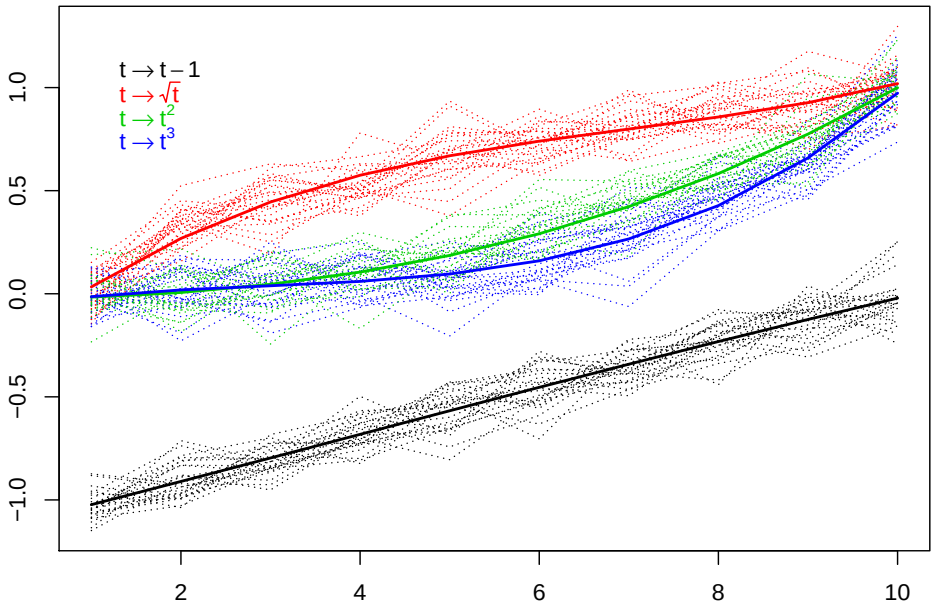


Figure 2: All curves (dotted lines) and cluster centers curves (plain lines) obtained with *fscm* algorithm for the simulated Sample 2

the *iterSubspace* algorithm struggles to separate the polynomial functions.

Data set of swimmers' progression curves

The choice of the funHDDC algorithm was motivated by two main arguments. First, this is a flexible method that has been shown efficient in various cases. Secondly, because of the results of the FPCA performed to explore the data set. Indeed, as presented on the top Figure A1 (Appendix), we notice that the underlying dimension of the data seems clearly lower than the original one: the entire variance of the data set can be expressed with only three scores. Additionally, the shapes of the first informative eigenfunctions are drawn (bottom Figure A1) and inquire on the main features of the data. One can see an analogous result of low underlying dimension for the derivatives (Appendix : Figure A2). Thus, it seems natural to work with a FPCA-based method. FunHDDC provides a flexible way to deal with the "extra-dimensions", proposing six models that represents six different ways to model covariance matrices. We tested each of them to figure out the more appropriate. As adviced by the authors in [23], the BIC is used for the model selection and the slope heuristic to choose the number K of clusters. According to these criteria, the best model, among the six, is composed of 5 clusters for the progression curves, and 4 clusters for the derivatives. Resulting clusters are represented on Figure A3 and Figure A4 (Appendix). At this stage, the Adjusted Rand Index (ARI) is used to compare the way athletes were grouped and give a value of 0.41. The value of ARI would be around 0.20 for a completely random clustering procedure. This result, far

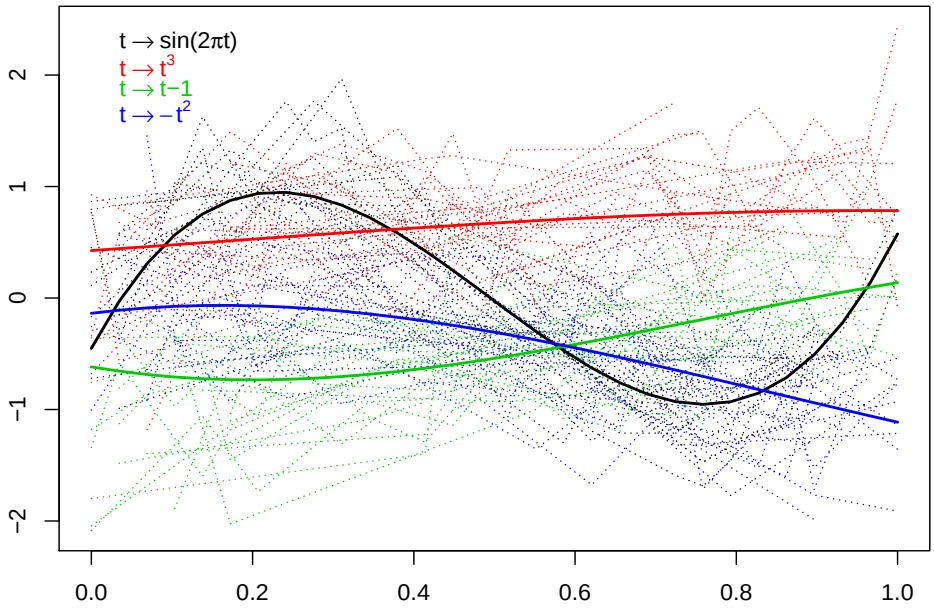


Figure 3: All curves (dotted lines) and cluster centers curves (plain lines) obtained with *iterSubspace* algorithm for the simulated Sample 3

from an ARI equals to 1 of complete adequacy, lets us think that different features of the data were used to group individuals in each context. A Discussion with swimming experts leads us to conclude that the clustering on progression curves mainly grouped athletes according to their level of performance, whereas the derivatives clustering seems to gather individuals presenting similar trends of progression (at a particular age, or with the same dynamic for example). These conclusions guided us to the multivariate clustering procedure, that gives results presented on Figure 2 and Figure 3. A close look at the groups on Figure 2 seems to indicate that multivariate clustering clusters combines information both on level of performance and trends of evolution. One can see that similar profiles are coloured the same way. We also verify this from a swimming expert point of view by checking samples of athletes in each groups. On Figure 3, one can see more clearly differences between each group thanks to the cluster center curves.

4 Discussion

As mentioned in the simulated data set context, we shall emphasise that no objective criterion might reflect correctly the quality of a clustering procedure. The authors of [24] recall that all clustering algorithms are some way subjective regarding how they gather individuals or which metric they use. Thus, the resulting clusters should be judged and analysed according to the context. Like many other statistical tools, a clustering procedure does not give any quantita-

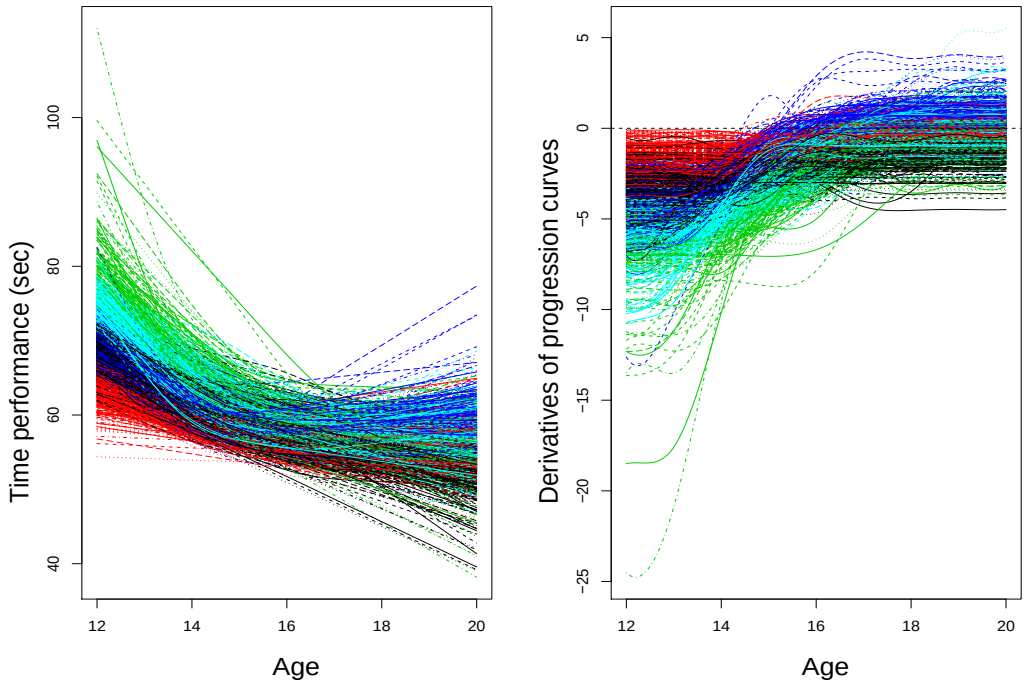


Figure 4: All progression curves of swimmers (left) and derivatives (right) coloured by clusters, obtained with the multivariate *funHDDC* algorithm.

tive certainty, but rather a new point of view on the data. One should consider as good results any useful perspective hidden in the raw data. Thus, we worked closely with sport experts, not only to analyse the results but throughout the entire analysis. All choices of parameters and/or methods were driven both by mathematical and sport considerations.

In this work, we enlighten some classical methods and useful practical packages as well as provide some clues on the particularities of the different algorithms. One can note that distance-based methods are generally easy to use and give rather good results for simple problems. In the other hand, model-based methods lie on more complicated design but often give good results in a wider range of problems. It explains why they are often recommended by experts of the field [13] and form most of the algorithms implemented in *funcy*. Algorithms using Gaussian mixtures are naturally more flexible than methods like k-means, since they might be considered as a generalisation with elliptic clusters rather than circular ones. However, one should also keep in mind that this flexibility often costs longer computational time. Indeed, even if the EM algorithm is really efficient to solve the mixture of Gaussian problem, the multiplicity of models and the number of clusters to test might take non negligible time to run. For our purpose, which is to help a swimming federation with the detection of young promising athletes, computational time was not an issue

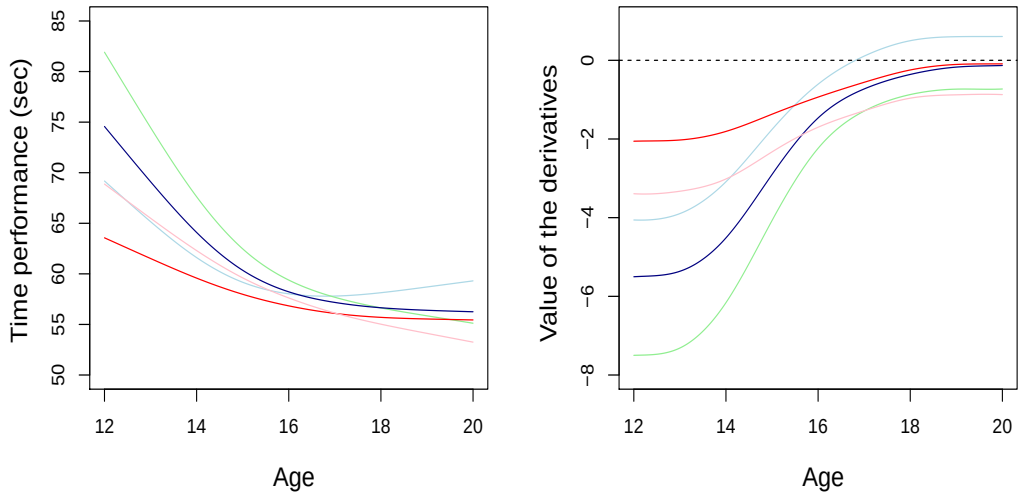


Figure 5: Cluster centers curves of swimmers (left) and derivatives (right) coloured by clusters, obtained with the multivariate clustering *funHDDC* algorithm.

since the aim was more of a long term decision making. Nevertheless, many sport-related problems need today to be solved quickly or even in live, and our methodological choices would have been different under such constraints.

About the results on the swimming data set, we observe consistent outcomes from both mathematical and sport point of views. If our work does not give any certainty about the progression phenomenon of young swimmers, it gives some enlightenments of its general pattern and provide a practical tool to gather similar profiles. These results might help the detection of promising young athletes with both a better understanding and graphical outcomes to support the decision process. Note that this work remains descriptive and thus preliminary, but one can think of it as a first step for a further predictive analysis. If we do not discuss here findings about any particular swimmers for confidentiality concerns, we can highlight some points that seem interesting to swimming experts. First, as mentioned in [2] [16], it seems difficult to precisely detect young talents before 16 years old, because of the fast evolution before this age. One can observe between 14 and 16 years old a huge decrease of the value of the derivatives and thus of the speed of progression. Moreover, athletes that seem to be better at 20 years old are often those who continue to progress, even slightly, after 16 years old. A classical pattern, confirmed with swimming experts, is the presence of a cluster of swimmers who are always among best performers. These athletes are typically often detected and can benefit of the best conditions to improve their performances. However, two clusters of athletes, often slightly slower than previous ones when young, present opposite behaviors. As one group stops rapidly to progress and performs rather modestly at 20 years old,

another cluster gathers swimmers with a fast improvement who often perform as good as best swimmers when older. One can think of these young athletes as the main target of a detection program, since they often remain away from top level structures at young ages. If these findings are promising, this work needs further developments to provide more quantitative and predictive outcomes. The FDA offers several methods of classification and regression, but as mentioned many times previously, it would be necessary to adapt them to our specific problem, or to develop new algorithms.

To conclude, we recall that the main purpose of this paper is to present a brief review of the functional data analysis and we emphasise one last time on the usefulness of such an approach. As supported by the example of curves clustering, FDA can offer new perspectives in the sport science field.

References

[1] C. Abraham, P. A. Cornillon, E. Matzner-Løber, and N. Molinari. Unsupervised Curve Clustering using B-Splines. 30(3):581–595.

[2] Gennaro Boccia, Paolo Moisè, Alberto Franceschi, Francesco Trova, Davide Panero, Antonio La Torre, Alberto Rainoldi, Federico Schena, and Marco Cardinale. Career Performance Trajectories in Track and Field Jumping Events from Youth to Senior Success: The Importance of Learning and Development. 12(1).

[3] Charles Bouveyron and Julien Jacques. Model-based Clustering of Time Series in Group-specific Functional Subspaces. 5(4):281–300.

[4] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer Science & Business Media.

[5] David L. Carey, Kay M. Crossley, Rod Whiteley, Andrea Mosler, Kok-Leong Ong, Justin Crow, and Meg E. Morris. Modelling Training Loads and Injuries: The Dangers of Discretization.

[6] Jeng-Min Chiou and Pai-Ling Li. Functional clustering and identifying substructures of longitudinal data. 69(4):679–699.

[7] Carl de Boor. On calculating with B-splines. 6(1):50–62.

[8] Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media.

[9] S. E. Forrester and J. Townend. The effect of running velocity on footstrike angle – A curve-clustering approach. 41(1):26–32.

[10] M. Giacofci, S. Lambert-Lacroix, G. Marot, and F. Picard. Wavelet-Based Clustering for Mixed-Effects Functional Models in High Dimension. 69(1):31–40.

[11] Nathaniel E. Helwig, K. Alex Shorter, Ping Ma, and Elizabeth T. Hsiao-Wecksler. Smoothing spline analysis of variance models: A new tool for the analysis of cyclic biomechanical data. 49(14):3216–3222, 10 03, 2016.

[12] Julien Jacques and Cristian Preda. Funclust: A curves clustering method using functional random variables density approximation. 112:164–171.

[13] Julien Jacques and Cristian Preda. Functional data clustering: A survey. 8(3):231–255.

[14] Gareth M James and Catherine A Sugar. Clustering for Sparsely Sampled Functional Data. 98(462):397–408.

[15] Huijing Jiang and Nicoleta Serban. Clustering Random Curves Under Spatial Interdependence With Application to Service Accessibility. 54(2):108–119.

[16] Philip E. Kearney and Philip R. Hayes. Excelling at youth level in competitive track and field athletics is not a prerequisite for later success. pages 1–8.

[17] Dayanne S. Lima-Borges, Paula F. Martinez, Luiz Carlos M. Vanderlei, Fernando S. S. Barbosa, and Silvio A. Oliveira-Junior. Autonomic modulations of heart rate variability are associated with sports injury incidence in sprint swimmers. pages 1–11.

[18] Fermin Mallor, Teresa Leon, Martin Gaston, and Mikel Izquierdo. Changes in power curve shapes as an indicator of fatigue during dynamic contractions. 43(8):1627–1631.

[19] Jie Peng and Hans-Georg Müller. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. 2(3):1056–1077.

[20] James O. Ramsay and Bernard W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*, volume 77. Springer New York.

[21] James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis — James Ramsay — Springer*.

[22] William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. 66(336):846–850.

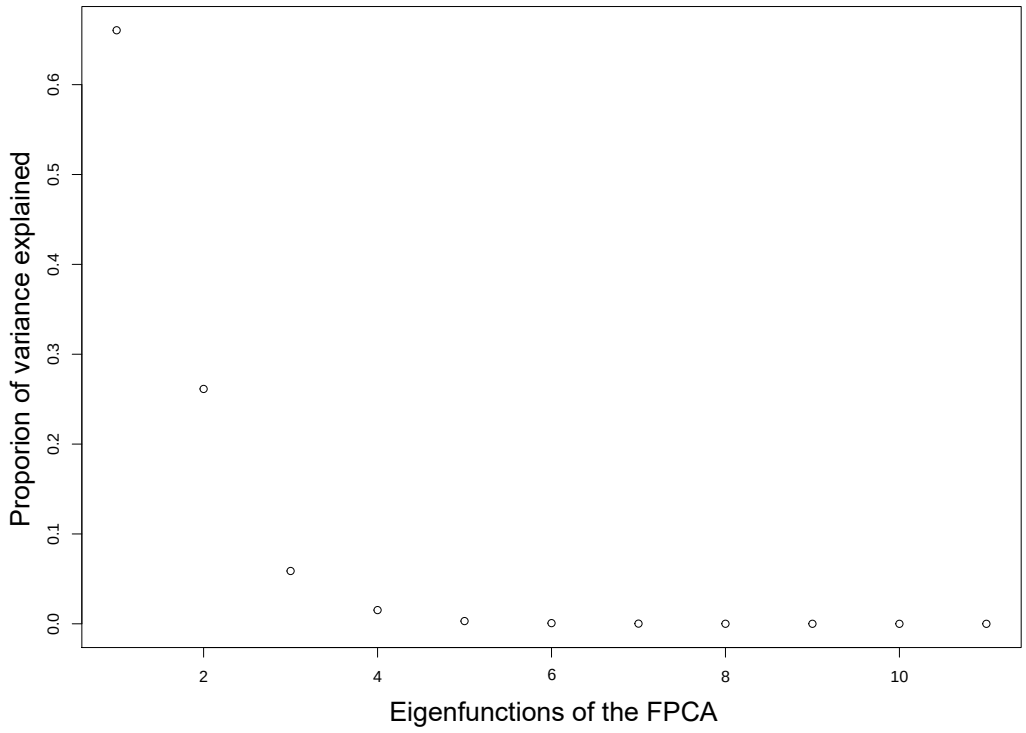
[23] Amandine Schmutz, Julien Jacques, Charles Bouveyron, Laurence Cheze, and Pauline Martin. Clustering multivariate functional data in group-specific functional subspaces.

[24] Ulrike von Luxburg, Robert C. Williamson, and Isabelle Guyon. Clustering: Science or Art? In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 65–79.

[25] T. Warren Liao. Clustering of time series data—a survey. 38(11):1857–1874.

[26] Nick Wattie, Jörg Schorer, and Joseph Baker. The relative age effect in sport: A developmental systems model. 45(1):83–94.

Appendix



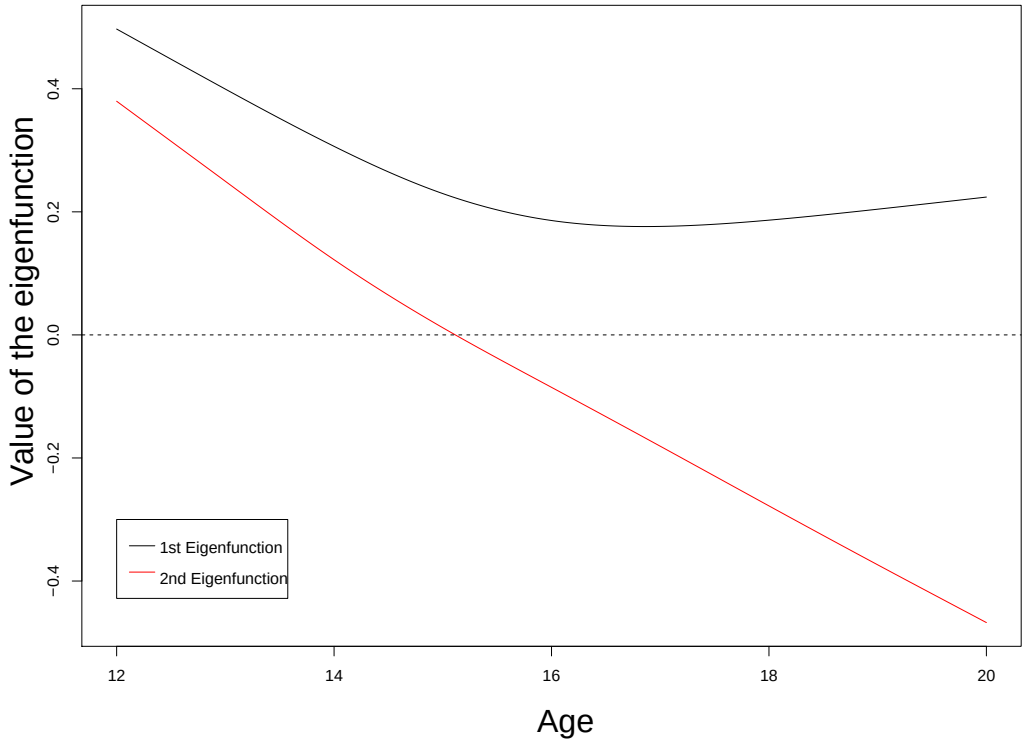
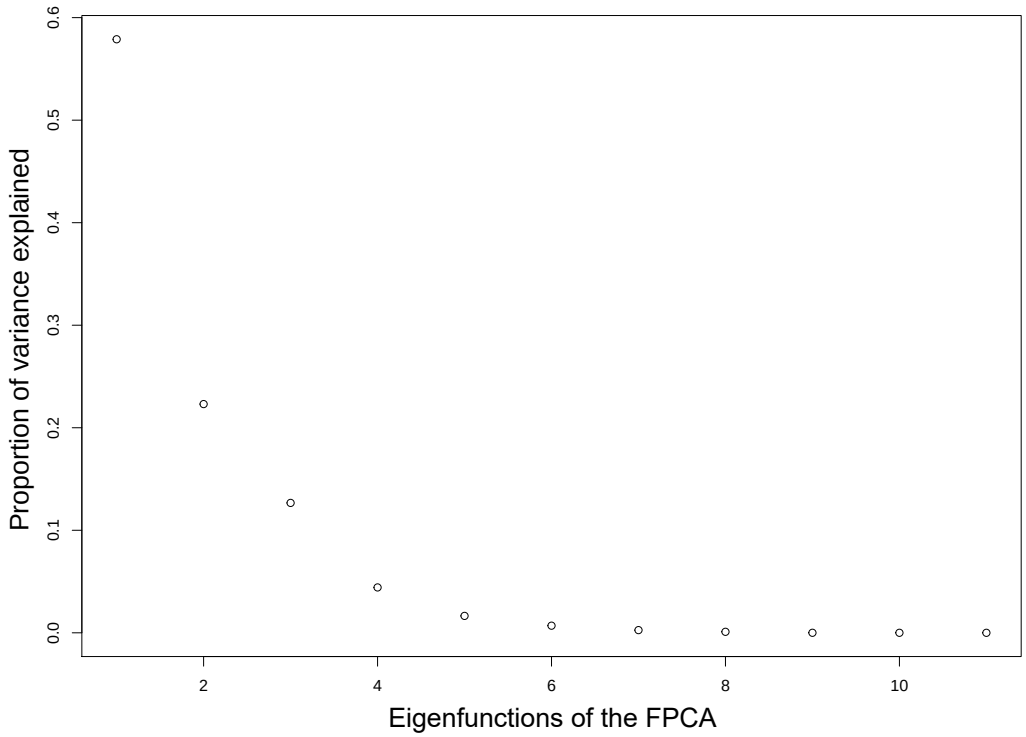


Figure 6: Result of the FPCA on the progression curves. Proportion of variance explained by each eigenfunctions (top). Values of the two first eigenfunctions (bottom)



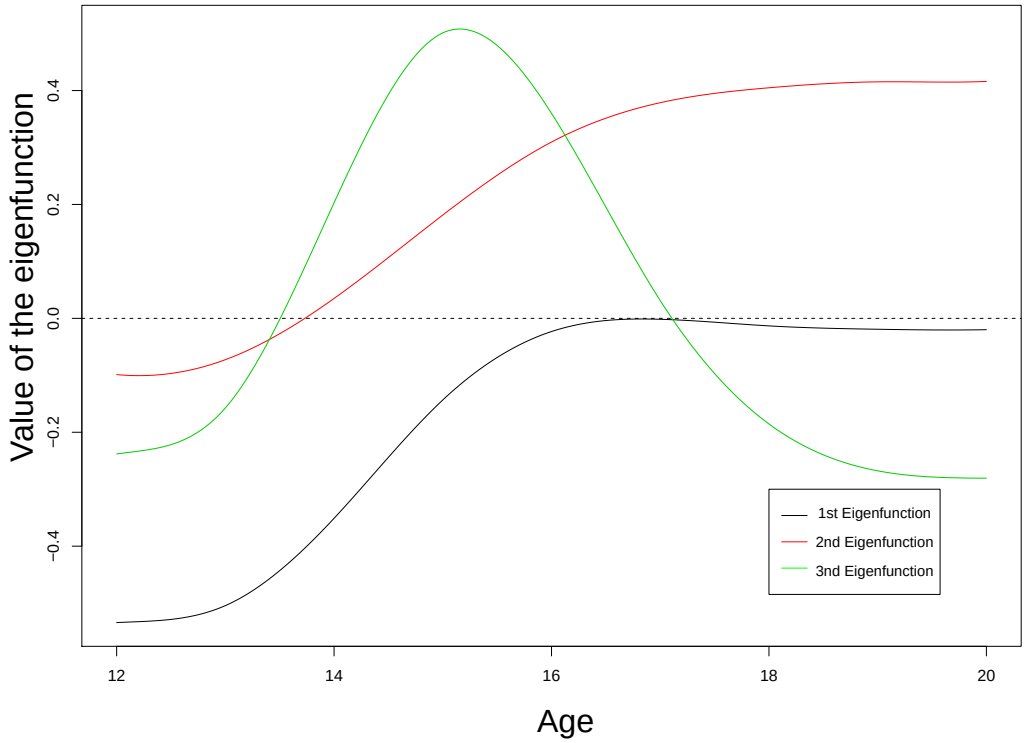


Figure 7: Result of the FPCA on the derivatives of the progression curves. Proportion of variance explained by each eigenfunctions (top). Values of the three first eigenfunctions (bottom)

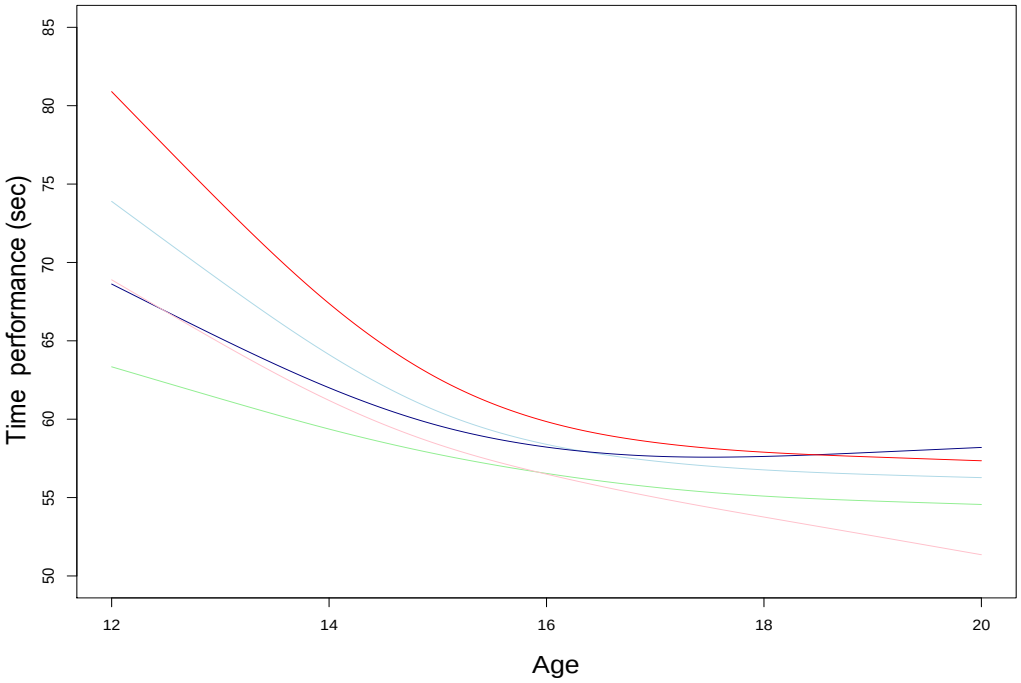


Figure 8: Clusters centres of the progressions curves. Computed with the univariate funHDDC algorithm.

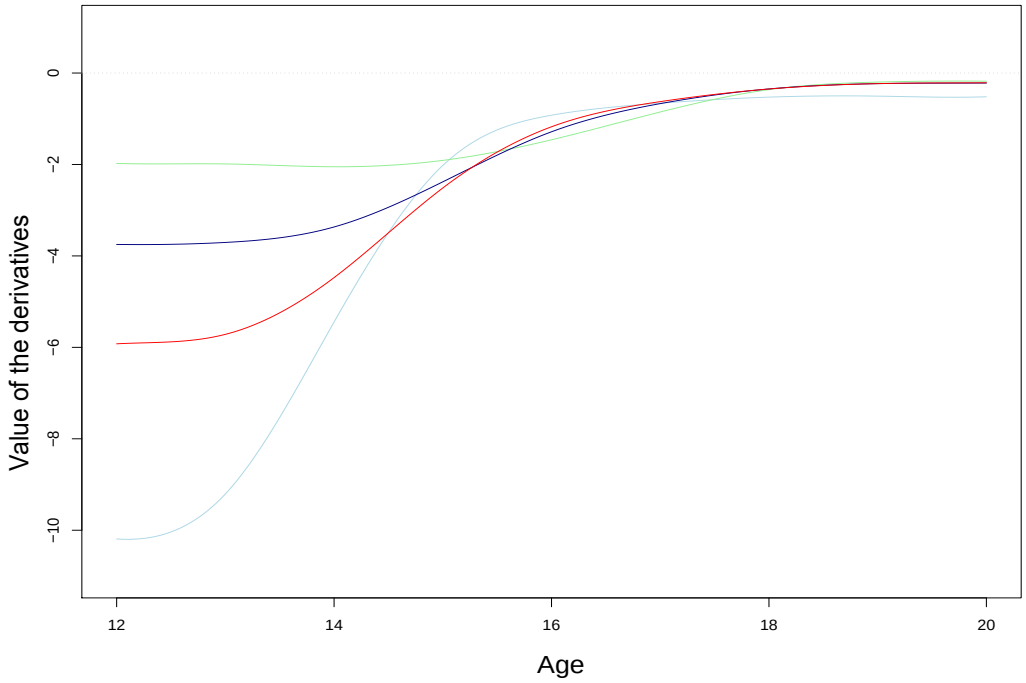


Figure 9: Clusters centres of the derivatives of the progressions curves. Computed with the univariate funHDDC algorithm.