



HAL
open science

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

Arthur Leroy, Andy Marc, Olivier Dupas, Jean Lionel Rey, Servane Gey

► To cite this version:

Arthur Leroy, Andy Marc, Olivier Dupas, Jean Lionel Rey, Servane Gey. Functional Data Analysis in Sport Science: Example of Swimmers' Progression Curves Clustering. Applied Sciences, 2018, 8 (10), pp.1766. 10.3390/app8101766 . hal-01862727v3

HAL Id: hal-01862727

<https://hal.science/hal-01862727v3>

Submitted on 13 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Article

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

Arthur Leroy ¹ , Andy Marc², Olivier Dupas³, Jean Lionel Rey⁴, Servane Gey⁵

¹ MAP5 - Paris Descartes University, IRMES - INSEP; arthur.leroy@insep.fr

² IRMES - INSEP; andy.marc@insep.fr

³ French Swimming Federation, olivier.dupas@ffnatation.fr

⁴ French Swimming Federation, jeanlionel.rey@free.fr

⁵ MAP5 - Paris Descartes University; servane.gey@parisdescartes.fr

* Correspondence: arthur.leroy@insep.fr; Tel.: +33 (0) 1 41 74 41 89

Academic Editor: name

Version October 13, 2018 submitted to Appl. Sci.

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

Longitudinal data in sport

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 [1], to the number of injuries in a team over a season [2], to the evolution of performances during a whole career [3], 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 properties¹. 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].

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.

25 These methods focus on the study of time dependent processes that generate discrete observations.
26 For instance, since an important topic of this paper concerns clustering, and a really comprehensive
27 review about clustering of time series can be found in [5].

28
29 Despite the usefulness of the time series approach, some theoreticians proposed a new modeling
30 of longitudinal data [6]. In many cases, the studied phenomenon is actually changing continuously
31 over the time. Thus, the object we want to know about is more of a function than a series of point. In
32 their paper [2], the authors highlight that it may be damageable to discretize phenomenons that are
33 intrinsically functional. Moreover, they claim that continuous methods perform better than discrete
34 ones on the specific case of the relationship between training load and injury in sport.

35
36 In some particular cases, it thus seems natural to model a continuous phenomenon as a random
37 function of time, formally a stochastic process, and consider our observations as just few records of
38 an infinite dimensional object. This approach is called functional data analysis (FDA) and gives a
39 new range of methods well suited to work on longitudinal data. There was substantial theoretical
40 improvements in the area the last two decades, and this paper intends to present some topics that
41 might be useful to the sport science field. To our knowledge, there is very few paper in the sport
42 literature that use FDA. We can cite [7] in which curve clustering is used to analyse the foot-strike of
43 runners, or [8] for the study of muscle fatigue through a whole FDA analysis. Another example is
44 given in [9] that proposes a functional version of ANOVA using splines to overcome common issues
45 that occur in sport medicine. Finally, the work presented in [10] uses curve clustering methods to
46 study different types of footfall in running. The methodology used in this paper is closely related
47 to our present article, and authors claim that this approach clearly improved analysis of footfall
48 compared to former empirical and observational ways to classify runners.

49
50 If such an approach remains marginally applied in the sport field, one would find many examples
51 in a wide range of other domains. We can cite for example meteorology, with the article [11] that
52 describes the study of temperature among Canadian weather stations, which has become a classic
53 data set over the years. Another famous data set is presented in [12] as an application to biology, by
54 studying the growth of children as a time continuous phenomenon. Those works and data sets are
55 today considered as benchmarks to test new methods, but many fields such as economy [13], energy
56 [14], medicine [15] or astronomy [16] have used FDA and contribute to this really active research topic.

57 *Detection of young athletes*

58 In the elite sport context, a classical problem lies in the detection of promising young athletes [17].
59 With professionalisation and evolution of training methods, differences in competition became more
60 and more tight in recent years [18]. Besides, it has been shown that the development of some specific
61 abilities during adolescence is a key component of improvement [19]. Hence, many sport federations
62 or structures have paid interest in the subject and tried to understand mechanisms behind what could
63 be called *talent* [20] and the evolution during young years of a career. A key feature to take into account
64 is morphology, since it obviously influences performance in many sports [21]. Morphology is also
65 known as a major noise factor in the detection issue, as the differences in physical maturity leads to
66 promote some young athletes over others [22] just because of their advantages in height or weight,
67 which will disappear when adult. Some problems raise when these maturity rhythms are ignored,
68 such as in training centers, with an over-representation of athletes born during the first months of
69 the year [23]. Moreover, it appeared in several studies that performance at young ages provides in
70 itself a poor predictor of the future competition results [3]. Only a small portion of elite athletes before
71 16 years old remains at a top level of performance later [24]. It thus seems clear that the classical
72 strategy, which consists in training intensively in specific structures only best performers of a young
73 age range, reaches its limits. If there are numerous elements that influence performance [25], several

74 works [26] seem indicate that the evolution over the time of a young athlete is more suited to predict
 75 future abilities than raw values at given ages. Different patterns of progression exist, and it might be
 76 important to take them into account if one wants to improve quality of talent detection strategies. Our
 77 work in this context aims to provide a more global vision of the progression phenomenon by saving its
 78 genuine continuous nature. Therefore, model data as functions and study them as such in the frame of
 79 FDA might offer a new perspective and provide insights to sport structures for their future decisions.

80 *Functional Data Analysis (FDA)*

81 As mentioned previously, FDA allows to take into account the intrinsic nature of functional data.
 82 Apart from this philosophical advantage in term of modeling, one may note important benefits. For
 83 example, if one records several time series with observations at different instants and/or in different
 84 numbers, how to compare them ? How to study the evolution of performances of swimmers from
 85 their competition times at given ages ? Competitors may have different number of races during their
 86 careers, and their performances are done at different ages (if one wants to avoid age discretization that
 87 have been shown problematic in [23]). This example illustrates exactly what we try to deal with in the
 88 subsequent curve clustering example. Another fundamental advantage of FDA is the possibility to
 89 work on the derivatives of the observed functions. Indeed, it is often interesting to study the dynamic
 90 of a time dependent process. Even the second derivative, often referred as the *acceleration*, or a superior
 91 order derivative might provide valuable information in practice. The specific nature of functional
 92 data allows to study such properties, and the sport scientist may easily imagine the wide range of
 93 situations on which the study of derivatives might be interesting. One could think for example of the
 94 GPS position tracking analysis, the progression phenomenon of young athletes, or the following of
 95 actions of some muscles over time.

96
 97 The first and fundamental step of a functional data analysis generally consists in the reconstruction
 98 of the function from the discrete set of observations. There is two cases at this step. Whether the
 99 observations are being considered as error-less (in term of measurement) and one can proceed to a
 100 direct interpolation through one of the multiple existing methods (linear, polynomial, ...). Or, more
 101 frequently, the set $x_{i,t_1}, \dots, x_{i,t_n}$ is considered as observations at time t_1, \dots, t_n of a realisation $x_i(t)$
 102 of a stochastic process $X(t)$. In this case, one can proceed to a *smoothing* step. It consists in the
 103 approximation of a function defined to be *close* to the observed points. To deal with noisy data, one
 104 always has to face the over-fitting/under-fitting issue. In most cases, one has to determine a smoothing
 105 parameter that define how much one wants to allow the function to contain *peaks*. These topics are
 106 largely detailed in the first chapters of [27]. Even if defining a consistent value of the smoothing
 107 parameter is a first work, one can see as an advantage the fact to explicitly control the signal-on-noise
 108 ratio of the data. The most common way to reconstruct the function from the observations is to use a
 109 basis of functions. A basis of functions is a set of specific functions ϕ_i of a functional space \mathcal{S} , such as
 110 each element of \mathcal{S} can be defined as a linear combination of the ϕ_i . Formally, we can define the basis
 111 expansion f as :

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

112 where ϕ_1, \dots, ϕ_N are the basis functions of a given functional space and $\alpha_1, \dots, \alpha_N$ are real valued
 113 coefficients. Intuitively, if one fixes a common basis to fit observations, the information on individuals
 114 is contained in the vector of coefficients $\{\alpha_1, \dots, \alpha_N\}$. That is why a common approach is to perform
 115 classical multivariate methods on these coefficients. Among the most common basis used in practice,
 116 we can cite Fourier basis and wavelets, which are well suited to model periodic data [27] [28]. Fourier
 117 basis is a widespread choice that works well when data present a repetitive pattern (such as day/night
 118 cycles for example) since the basis functions are sinusoids. However, their efficiency decrease when
 119 data are less regular, especially on the modelisation of derivatives. Wavelet basis are designed to settle

120 this sensibility to irregular signals. Coefficients are slightly longer to compute, but this basis has
121 really good mathematical property and progressively replace Fourier basis in several applications
122 [29] [30]. For non periodic data, the classical choice is spline basis, particularly the cubic splines in
123 practice [6]. B-splines are piecewise polynomial functions and require few coefficients to define a good
124 approximation, which make B-splines especially adequate when observations are sparse on the time
125 domain [31]. They allow to approximate a wide range of shapes with a rather good smoothness [30].
126 From a computational point of view, one can use the *R* package *fda*, on which one can find methods to
127 fit observations into functional data, and way more tools for FDA. An overview of the *fda* package can
128 be found in [30].

129
130 Once the data set is approximated by functions, one may perform analysis on them, and some
131 classical statistical tools have been extended in the functional context. One of the first and most
132 important adapted method was the functional principal component analysis (FPCA). Although slightly
133 different, FPCA provides analogous information as the finite dimensional version [27]. This method
134 allows to describe data into a non correlated low dimensional space. That is why it provides an
135 excellent explanatory tool to visualize main features of the data as well as a way to reduce the number
136 of informative dimensions. This can be particularly useful when one wants to apply algorithms on
137 the vector $\{\alpha_1, \dots, \alpha_N\}$ of coefficients of the basis expansion, with N rather large. It may accelerate
138 calculation while retain most of the information as well as avoid curse of dimension in a big data
139 context. We may also cite several methods presented in [27] such as *functional canonical correlation*,
140 *discriminant analysis* and *functional linear models*.

141
142 The purpose of this paper is twofold: at first, it aims at providing a brief review of several
143 methods and references for the theoretical aspects. Secondly, examples of practical tools and useful
144 packages (on the software *R* as it is currently the most convenient to perform FDA) of curve clustering
145 state of the art methods are presented. Then, we also detail a specific study on a real data set, coming
146 from our collaboration with the French Swimming Federation. This work focuses on the clustering of
147 performance progression curves of young male swimmers and uses several FDA tools. We emphasise
148 on the fact that FDA provides some tools that give information we could not exhibit otherwise, like the
149 study of derivatives for example.

151 *Clustering functional data*

152 In this article, we emphasise on the *clustering* approach, often fundamental when exploring a new
153 data set or beforehand to a forecast. This method consists in computing sub-groups of individuals on
154 a data set that make sense in the context of the study. Given K the number of clusters, a clustering
155 algorithm would apply one or several rules to gather individuals presenting common properties.
156 This problem has been largely explored these past ten years in the functional context and we will
157 give some elements to summarize the state of the art. According to the survey [28], functional data
158 clustering algorithms can be sorted in three distinct families, detailed below. We do not develop on
159 direct clustering on raw observational points that does not take into account functional nature of the
160 data and may give poor results.

161
162 (i) *2-steps methods*. The first step consists in the fitting procedure we detailed previously, choosing
163 a common basis for all data. Then, a clustering algorithm such as k-means [31], or hierarchical
164 clustering methods for example, is performed on the basis coefficients. If this vector of coefficients is in
165 high dimension, one can add a step of FPCA and perform the clustering on the scores coming from the
166 first eigenfunctions of the FPCA.

167

168 (ii) *Non-parametric clustering.* An overview of non-parametric functional data analysis is provided
 169 by [32]. It details many aspects where one does not assume that functional observations can be defined
 170 by a finite number of parameters. The idea is to define a *distance* between the functional observations
 171 without assumptions on the form of the curves. A classical measure of proximity between functions x_i
 172 and x_j is defined as :

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

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

175
 176 (iii) *Model-based clustering.* This approach has been widely developed in the past years and
 177 gives good results. As the 2-step approach, it often uses basis expansion and/or FPCA to fit the data.
 178 However, rather than proceeding in two-step, the clustering is performed simultaneously. Many
 179 algorithms are based on Gaussian mixture models coupled with an EM-algorithm to compute the
 180 parameters [33] [34] [35]. We chose in this study to adapt the algorithm FunHDDC presented in [35]
 181 for several reasons that we develop in the following *Materials and Methods* section.

182
 183 Note that the literature does not give specific indications about the family of methods that should
 184 be used in a specific context and one might test several of them. Nevertheless, one should keep in
 185 mind that the right way to fit the data into functions strongly depends on the structure of the data. We
 186 also give some additional references in the next section, where we detail some algorithms that are easy
 187 to use in practice because of their implementation within an unified *R* package. Below, Figure 1 from
 188 [28] summarizes the different families described above and the process of clustering in a functional
 189 context:

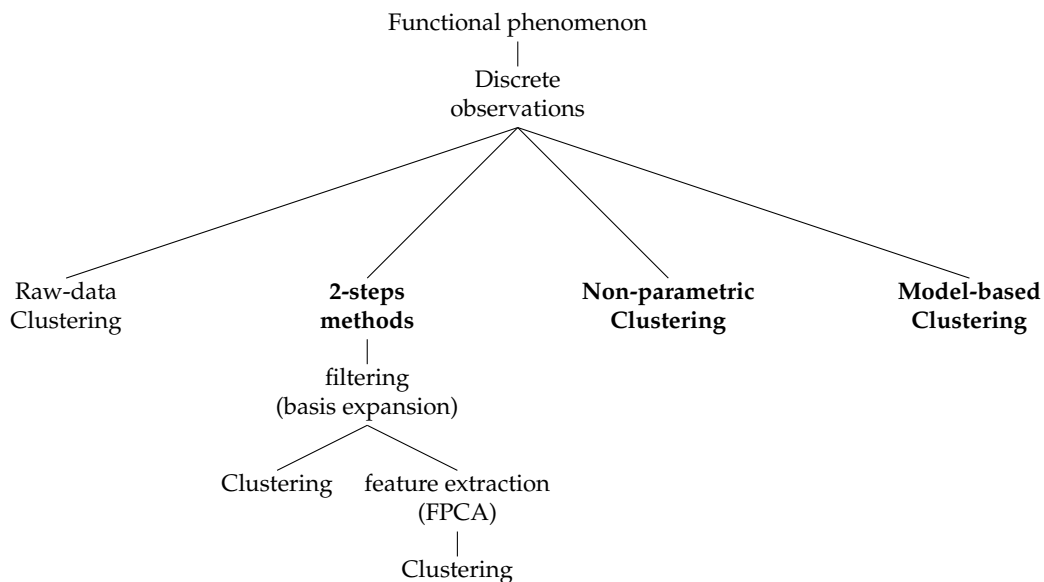


Figure 1. Summary of the different approaches to perform clustering on functional data, from raw data (top) to final clusters (bottom).

190 2. Materials and Methods

191 *Description of the real swimming data set*

192 First of all, two types of data sets, on which we performed functional clustering algorithms,
193 are described. The way we simulated data sets to test several methods will be described at the end
194 of the current section. The real data have been collected by the French Swimming Federation. It
195 gathers all the performances of french male swimmers, since 2002, for the 100m freestyle in a 50m
196 pool. Because of confidentiality issues, athletes are identified by a number. The data set is composed of
197 46115 performances and ages of 1468 different swimmers, and is available on the Github page of the
198 corresponding author. Raw data consists of time series of competition performances at different ages
199 for each swimmer. The number of competitions and the age at which swimmers participate differs
200 from one to another, leading to strongly uneven time series. This particularity of the data set (as well
201 as the ability to work on derivatives) led to model the observations as functions rather than time series.
202 Thus, a first step of fitting was performed to extract the functional nature of the data and deal with
203 the random fluctuations in the observations. All the algorithms were run on the *R* software and the
204 corresponding packages will be named in the sequel.

205 *Testing several algorithms on simulated data sets*

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

212

213 *(ii) distance-based:*

- 214 • *distclust*: An approximation of the L^2 distance between curves is defined, and a k-means heuristic
215 is used on individuals using this distance. This method is well designed in the context of sparsely
216 observed functions with irregular measurements. [36]

217 *(iii) model-based :*

- 218 • *fitfclust*: One of the first algorithm to use a Gaussian mixture model for univariate functions
219 that we briefly describe. This heuristic holds for all following algorithms described as Gaussian
220 mixture methods. Functions are represented using basis functions, and the associated coefficients
221 are supposed to come from Gaussian distributions. Given a number K of different means and
222 covariances parameters corresponding to the K clusters, an EM algorithm is used to estimate the
223 probability of each observational curves to belong to a cluster. When the iterations stop (various
224 stopping criteria exist), an individual is affected to its most likely cluster. A preliminar step of
225 FPCA can be added to work on lower dimensional vectors and thus speed up the calculations.
226 This method is well designed in the context of sparsely observed functions. [37]
- 227 • *iterSubspace*: A non-parametric model based algorithm. This method uses the Karhunen-Loeve
228 expansion of the curves, and perform a k-means algorithm on the scores of FPCA and the mean
229 process. This method can be useful when the Gaussian assumption does not hold but k-means
230 approach can lead to unstable results. [38]
- 231 • *funclust*: A Gaussian mixture model based algorithm. This method uses the Karhunen-Loeve
232 expansion of the curves and allows each cluster's Gaussian parameters to be of different sizes,
233 according to the quantity of variance expressed by the corresponding FPCA. The algorithm also
234 allows different covariance structures between clusters and thus generalizes some methods such
235 as *iterSubspace*. [33]

- 236 • funHDDC: A Gaussian mixture model based algorithm. This method presents lots of common
237 characteristics with funclust, but additionally allows clustering of multivariate functions. The
238 algorithm proposes six ways to model covariates structures, especially for the extra-dimension of
239 the FPCA. [35]
- 240 • fscm: A non-parametric model based algorithm. Each cluster is modeled by a Markov random
241 field, and functions are clustered by shape regardless to the scale. Observation curves are
242 considered as locally-dependent, and a K-nearest neighbors algorithm is used to define the
243 neighborhood structure. Then, an EM algorithm estimates parameters of the model. This method
244 is well designed when the assumption of independence between curves does not hold. [39].
- 245 • waveclust: A linear Gaussian mixed effect model algorithm. This approach uses a dimension
246 reduction step using wavelet decomposition (rather than classic FPCA). An EM algorithm is
247 used to compute parameters of the model and probabilities to belong to a cluster. This method is
248 well design for high-dimension curves, when variations such as peaks appears in data, and thus
249 wavelets performe better than splines. [29].

250 Unfortunately, the current version (1.0.0) of the *fancy* package has some troubles with the
251 funHDDC algorithm, which is not directly usable at the moment. All the remaining algorithms
252 were applied on three simulated data sets, with $K = 4$ groups. The resulting clustering were compared
253 to real group distributions using the Rand Index (RI)[40]. This measure, between 0 and 1, is computed
254 by counting according pairs of individuals between two different partitions of a data set. The RI
255 is provided as a result of the *funcit* function of the *fancy* package, and compares the ability of each
256 procedure to retrieve the actual groups. Then, graphs of centers of each curve clusters were drawn to
257 analyse consistency of our results according to the original data.

258 *Clustering the real swimming data set*

259 As mentioned above, the real data set is very irregular, with no accordance in time and in number
260 of measurements between athletes. Thus, the first step of the analysis was the definition of a common
261 ground through a smoothing procedure. According to the non-periodic form of the data and the
262 relatively low sampling of observational points (around 30) for each athlete, a B-spline basis was
263 chosen. The study focuses on the age period from 12 to 20 years old, which is crucial in the progression
264 phenomenon that we aimed at studying. A basis of seven B-splines of order 4 was defined so that the
265 derivatives remain smooth enough to work on derivatives. Since we did not wish to focus on a specific
266 time period, the knots were equally placed on ages 13 to 19. One should note that data are considered
267 as realisations of a stochastic process, and thus raw data are assumed to contain random fluctuations.
268 The function that is fitted using the B-spline basis has to represent properly the true signal and the well
269 known over/under-fitting issue appears in this case. In order to differentiate the true signal from the
270 noise, several methods can be used, knowing that there is always a trade-off between smoothness of the
271 function and closeness to observation points. A classical approach consists in the use of penalisation
272 in the least-square fitting calculation, and the signal-on-noise ratio would be controlled by a unique
273 hyper-parameter. In our case, we used a cross-validation criterion to compute an optimal value for this
274 hyper-parameter, and the resulting functional data were considered as coherent by swimming experts.
275 This whole fitting procedure was performed thanks to *R* (version 3.5.0) software, and especially the
276 *fda* (version 2.4.8) package. To analyse efficiently a real data set, one needs first to explore it, to figure
277 out the more suited algorithm to use. To this purpose, a FPCA was performed on the progression
278 curves and their derivatives, separately. We looked at the percentage of variance explained by each
279 eigenfunction and the shapes of them, to understand the main features of the curves. One can see on
280 Figure A1 of the appendix that main variations among level of performance appear at young ages
281 and a clustering procedure on progression curves tends to simply group individuals according to this
282 criterion. As displayed on Figure A2, first eigenfunctions of the derivatives represent three different
283 modes of variations localized at young, middle, and older ages. These characteristics of data would
284 be relevant to include to the clustering procedure beside the level of performance information. To

285 this purpose the funHDDC algorithm was used as clustering procedure, as this is one of the rare
286 implemented algorithm that works in a multivariate case and thus allow to consider both curves and
287 their derivatives simultaneously. One can find more details in the result section about the reasons of
288 this choice. Although implemented in the *funcy* package, we chose to work with the original *funHDDC*
289 *R* package, because of current problems of implementation on it. Several features of the package were
290 used, as Bayesian Information Criterion (BIC), Integrated Classification Likelihood (ICL) and slope
291 heuristic, to deal with problems of model selection and choice of the number K of clusters. Since no
292 particular assumptions were made on the covariance structure or the number of clusters from a sport
293 expert point of view, the hyper parameters of the model have been optimised from data. All possible
294 models were computed for different values of K and the best one (the sense of the term *best* is developed
295 in the Results section) was retained as our result clustering. In the funHDDC algorithm, each cluster
296 is considered to be fully characterised by a Gaussian vector, from which scores on eigenfunctions of
297 the FPCA are assumed to come. Thus, the clustering becomes a likelihood maximization problem
298 where one wants to find value of means and covariance matrices that fit the best to data, as well as
299 probabilities for each of data curves to belong to a cluster. All parameters influence values of each
300 others and this classical issue is addressed thanks to an Expectation-Maximization (EM) algorithm
301 that computes efficiently close approximations of optimal parameters. At the end of the procedure, a
302 data curve is considered to belong to the cluster within which it has the higher probability to come
303 from. The clustering was performed on the curves and their derivatives, separately at first. Then, the
304 resulting clusters were compared thanks to the Adjusted Rand Index (ARI) [40], which is an extended
305 version of the RI to partitions with different number of clusters. This measure allows to quantify
306 the adequacy between individuals grouped whether by a clustering on progression curves or on
307 derivatives. Note that many other indexes exist, such as Silhouette index or Jaccard index for example.
308 Although our results were quite comparable using one or another, the reader can find an extensive
309 comparative study of the different indexes in [41]. Noticing that athletes were clustered differently,
310 providing two types of information, we decided to perform a third clustering procedure. This time,
311 the multivariate clustering version on the funHDDC algorithm was used. The term multivariate
312 clustering refers to a clustering algorithm that deals with multidimensional functions. The progression
313 curves were defined as a first variable, while the derivatives as a second variable. For each clustering
314 procedure, the resulting clusters centers and curves were plotted. Finally, the results were analysed
315 and discussed with swimming experts to confront the found clusters to the sport logic.

316 2.1. Definition of the simulated data sets

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

320
321 Simulated data sets are sampled from four different processes of the form $f(t) + \epsilon$, with f
322 and ϵ detailed in Table 1 below. For each process 25 curves are simulated, thereby leading to 100
323 curves in each sample. The aim of the following clustering procedure is to gather themselves curves
324 that correspond to the same underlying process. An additional goal would be to retrieve, at least
325 approximately, the shapes of deterministic functions f that generated each data curves within a cluster.
326 Intuitively, Sample 1 depicts an easy situation with low noise and well separated processes, whereas
327 Sample 2 represents the same processes in a higher variance context. Finally, Sample 3 corresponds to
328 a high-noise and crossing processes context, which is designed to be trickier. Moreover, in the case of
329 Sample 3, observations of the curves are irregular on t -axis and thus, for three out of six algorithms of
330 the package that are not implemented in this case, we had to proceed to a previous fitting step. We
331 used the function *regFuncy* of the *funcy* package to this purpose.

Table 1. Details on the simulated samples. Processes are defined as $f(t) + \varepsilon$ with 4 different functions f in each sample and a varying noise parameter ε .

Data set	Functions	Noise	Observations on t-axis
Sample 1	$t \mapsto t - 1$ $t \mapsto t^2$ $t \mapsto t^3$ $t \mapsto \sqrt{t}$	$\varepsilon \sim \mathcal{N}(0, 0.05)$	10 points at <i>regular</i> instants
Sample 2	$t \mapsto t - 1$ $t \mapsto t^2$ $t \mapsto t^3$ $t \mapsto \sqrt{t}$	$\varepsilon \sim \mathcal{N}(0, 0.1)$	10 points at <i>regular</i> instants
Sample 3	$t \mapsto t - 1$ $t \mapsto -t^2$ $t \mapsto t^3$ $t \mapsto \sin(2\pi t)$	$\varepsilon \sim \mathcal{N}(0, 0.5)$	≤ 10 points at <i>irregular</i> instants

332 3. Results

333 Results on simulated data

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

Table 2. Mean Rand Index and (Standard Deviation) on 100 simulations of the tree samples. Each algorithm run in at most few seconds on our simulated data sets. Comparison in speed between algorithms is given as a multiple of the fastest which is set arbitrarily to 1.

Method	Sample 1	Sample 2	Sample 3	Running speed
fitfclust	0.945 (0.14)	0.857 (0.01)	0.307 (0.06)	2.8
distclust	0.996 (0.01)	0.888 (0.05)	0.523 (0.07)	19.2
iterSubspace	0.938 (0.14)	0.850 (0.12)	0.527 (0.07)	1
funclust	0.450 (0.17)	0.418 (0.16)	0.084 (0.07)	1
fscm	0.948 (0.12)	0.902 (0.01)	0.527 (0.07)	7
waveclust	0.920 (0.12)	0.810 (0.01)	0.324 (0.13)	34

349 Figure 2 gives one representation of the Sample 1 curves. In addition, the curves of each clusters
350 centers of the best performing algorithm are drawn. One can see that Sample 1 is quite simple to deal
351 with, since curves of different groups are well separated. Not surprisingly, the *distclust* clustering
352 algorithm satisfyingly figures out the actual shape of each process.

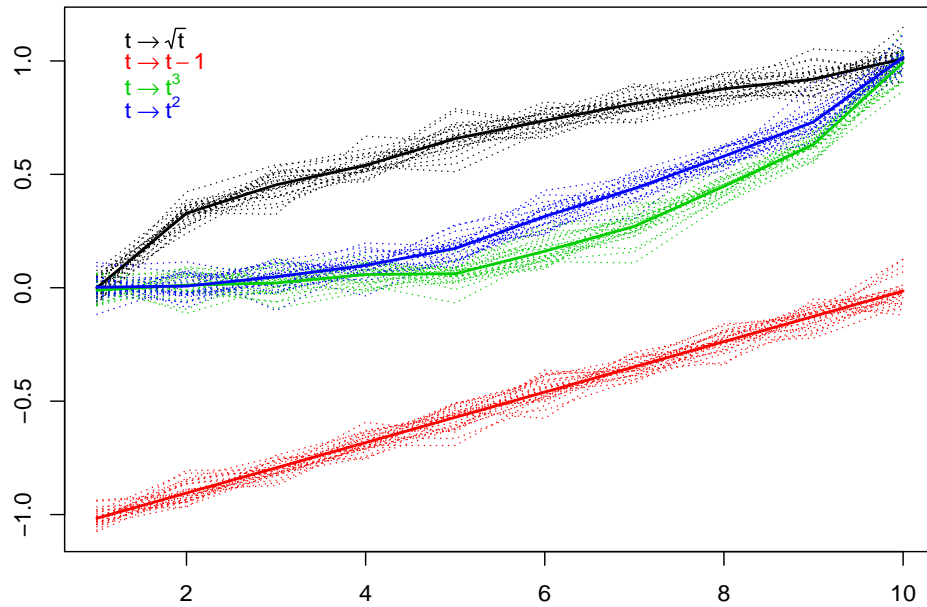


Figure 2. All curves (dotted lines) and cluster centers curves (plain lines) obtained with *distclust* algorithm for Sample 1. The algorithm correctly clusters curves and retrieves the underlying shapes of generating processes.

353 One can see on Figure 3 that, if the noisier situation of Sample 2 affects the good clustering rate,
 354 the shapes of the underlying functions remain correctly approximated by clusters centers of *fscm*.

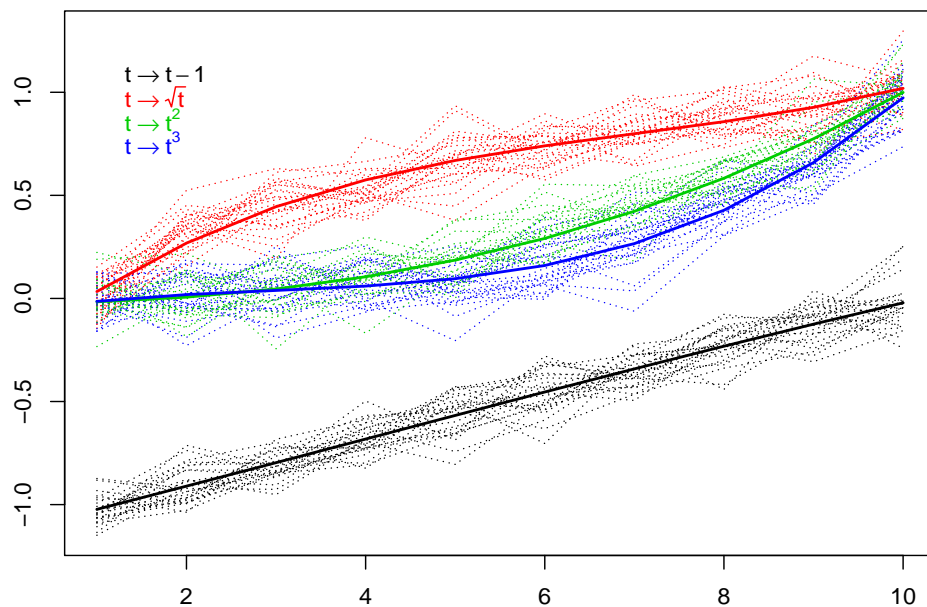


Figure 3. All curves (dotted lines) and cluster centers curves (plain lines) obtained with *fscm* algorithm for the simulated Sample 2. Clustering becomes more difficult between curves (e.g. blue and green curves) but the algorithm still performs well to figure out the underlying shapes.

355 Sample 3 was designed to be trickier since curves cross each other and the signal appears rather
 356 noisy. In this context, one can see on Figure 4 that, as expected, the algorithms retrieve approximately
 357 the true shapes of the underlying functions. While the *sinus* (in black) function seems correctly
 358 identified, the *iterSubspace* algorithm struggles to separate the polynomial functions.

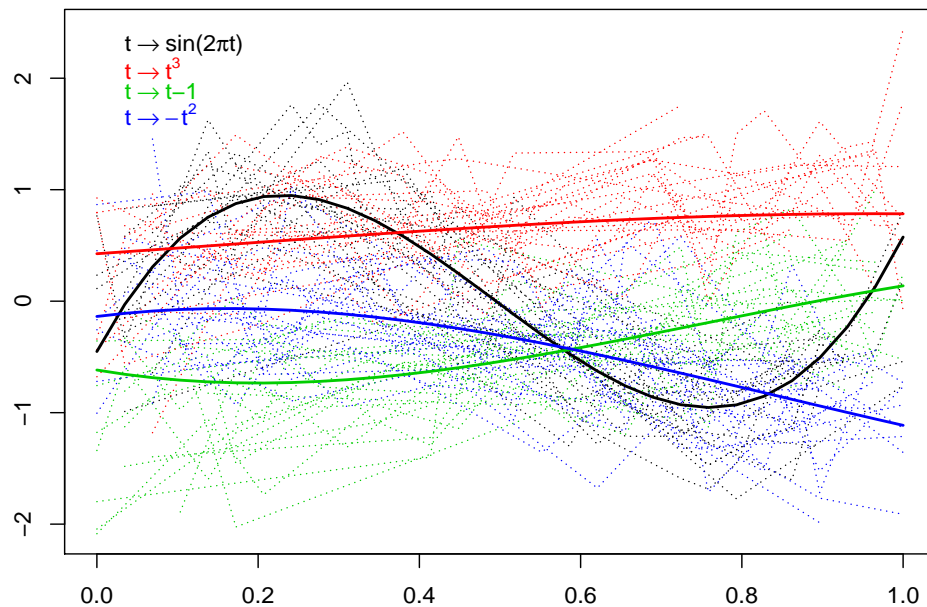


Figure 4. All curves (dotted lines) and cluster centers curves (plain lines) obtained with *iterSubspace* algorithm for the simulated Sample 3. Both clustering and detecting underlying shapes become difficult. The high noise makes the clustering fuzzy, which is affecting the cluster central curves.

359 *Data set of swimmers' progression curves*

360 The choice of the funHDDC algorithm was motivated by two main arguments. First, this is a
 361 flexible method that has been shown efficient in various cases. Secondly, because of the results of the
 362 FPCA performed to explore the data set. Indeed, as presented on the top Figure A1 (Appendix), we
 363 notice that the underlying dimension of the data seems clearly lower than the original one: the entire
 364 variance of the data set can be expressed with only three scores. Additionally, the shapes of the first
 365 informative eigenfunctions are drawn (bottom Figure A1) and inquire on the main features of the data.
 366 One can see an analogous result of low underlying dimension for the derivatives (Appendix : Figure
 367 A2). Thus, it seems natural to work with a FPCA-based method. FunHDDC provides a flexible way to
 368 deal with the "extra-dimensions", proposing six models that represents six different ways to model
 369 covariance matrices. We tested each of them to figure out the more appropriate. As advised by the
 370 authors in [34], the BIC is used for the model selection and the slope heuristic to choose the number K
 371 of clusters. According to these criteria, the best model, among the six, is composed of 5 clusters for the
 372 progression curves, and 4 clusters for the derivatives. Resulting clusters are represented on Figure
 373 A3 and Figure A4 (Appendix). At this stage, the Adjusted Rand Index (ARI) is used to compare the
 374 way athletes were grouped and give a value of 0.41. The value of ARI would be around 0.20 for a
 375 completely random clustering procedure. This result, far from an ARI equals to 1 of complete adequacy,
 376 lets us think that different features of the data were used to group individuals in each context. A
 377 Discussion with swimming experts leads us to conclude that the clustering on progression curves
 378 mainly grouped athletes according to their level of performance, whereas the derivatives clustering
 379 seems to gather individuals presenting similar trends of progression (at a particular age, or with the
 380 same dynamic for example). These conclusions guided us to the multivariate clustering procedure,

381 that gives results presented on Figure 2 and Figure 3. A close look at the groups on Figure 2 seems to
382 indicate that multivariate clustering clusters combines information both on level of performance and
383 trends of evolution. One can see that similar profiles are coloured the same way. We also verify this
384 from a swimming expert point of view by checking samples of athletes in each groups. On Figure 3,
385 one can see more clearly differences between each group thanks to the cluster center curves.

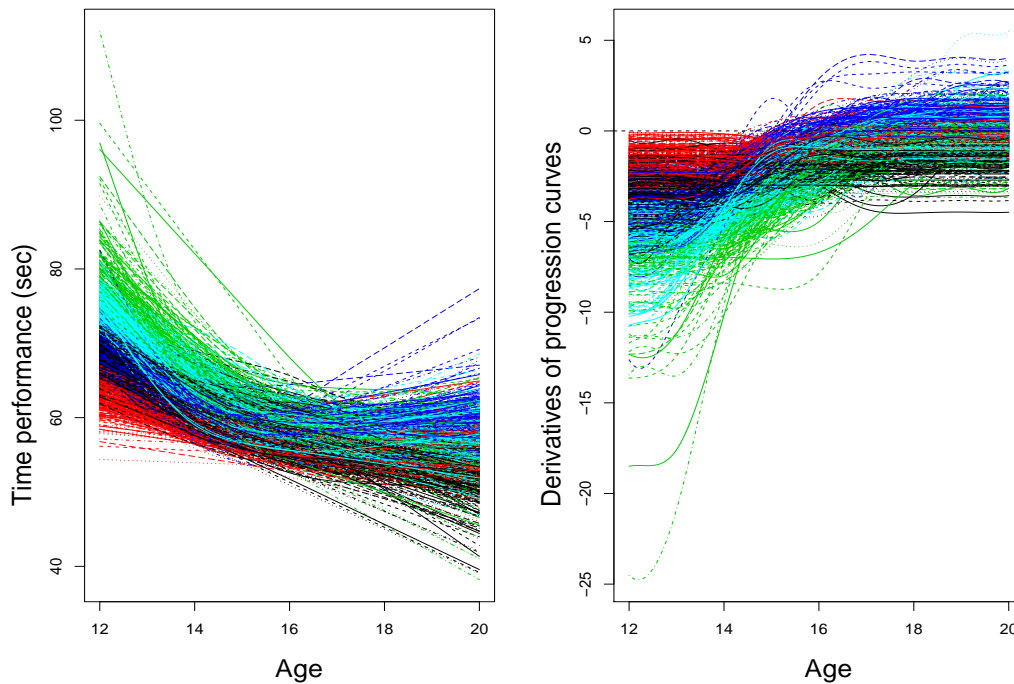


Figure 5. All progression curves of swimmers (left) and derivatives (right) coloured by clusters, obtained with the multivariate *funHDDC* algorithm. **Left:** Each curve represents the time performance for one swimmer between 12 and 20 years old. The clustering by level of performance can be observed particularly on this graph. **Right:** Each curve represents the derivative of the progression curve for one athlete between 12 and 20 years old. The clustering by speed of improvement and progression patterns is more clearly expressed on this graph.

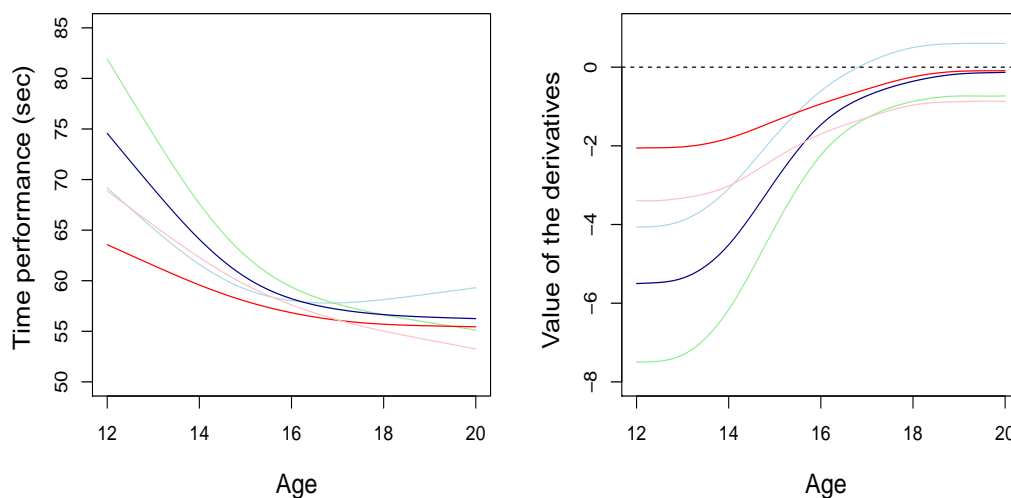


Figure 6. Cluster centers curves of swimmers (left) and derivatives (right) coloured by clusters, obtained with the multivariate clustering *funHDDC* algorithm. Similar information as on Figure 4 can be seen on this graph, in a clearer way with only center of clusters displayed.

386 4. Discussion

387 As mentioned in the simulated data set context, we shall emphasise that no objective criterion
 388 might reflect correctly the quality of a clustering procedure. The authors of [42] recall that all clustering
 389 algorithms are some way subjective regarding how they gather individuals or which metric they use.
 390 Thus, the resulting clusters should be judged and analysed according to the context. Like many other
 391 statistical tools, a clustering procedure does not give any quantitative certainty, but rather a new point
 392 of view on the data. One should consider as good results any useful perspective hidden in the raw
 393 data. Thus, we worked closely with sport experts, not only to analyse the results but throughout the
 394 entire analysis. All choices of parameters and/or methods were driven both by mathematical and
 395 sport considerations.

396
 397 In this work, we enlighten some classical methods and useful practical packages as well as
 398 provide some clues on the particularities of the different algorithms. One can note that distance-based
 399 methods are generally easy to use and give rather good results for simple problems. In the other
 400 hand, model-based methods lie on more complicated design but often give good results in a wider
 401 range of problems. It explains why they are often recommended by experts of the field [28] and form
 402 most of the algorithms implemented in *funcy*. Algorithms using Gaussian mixtures are naturally
 403 more flexible than methods like k-means, since they might be considered as a generalisation with
 404 elliptic clusters rather than circular ones. However, one should also keep in mind that this flexibility
 405 often costs longer computational time. Indeed, even if the EM algorithm is really efficient to solve the
 406 mixture of Gaussian problem, the multiplicity of models and the number of clusters to test might take
 407 non negligible time to run (few hours in our case). For our purpose, which is to help a swimming
 408 federation with the detection of young promising athletes, computational time was not an issue since
 409 the aim was more of a long term decision making. Nevertheless, many sport-related problems need
 410 today to be solved quickly or even in live, and our methodological choices would have been different
 411 under such constraints.

412
 413 About the results on the swimming data set, we observe consistent outcomes from both
 414 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. Moreover, using FDA, we were able to figure out information from uneven time series. Using smooth functions instead of raw data points provides a first understanding of the main trends and the continuous nature of the progression phenomenon. However, one should always pay attention to the random fluctuations of the data that serve to fit the studied functions. In order to improve the quality of the approximation and decrease the influence of the noise, we would like to collect more data on swimmers, with training performances for example. Nevertheless, 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 [3] [24], 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.

Author Contributions: Conceptualization, Arthur Leroy and Servane Gey; Methodology, Arthur Leroy; Software, Arthur Leroy; Validation, Arthur Leroy and Servane Gey; Formal Analysis, Arthur Leroy; Investigation, Arthur Leroy; Resources, Arthur Leroy; Data Curation, Arthur Leroy; Writing—Original Draft Preparation, Arthur Leroy; Writing—Review & Editing, Arthur Leroy and Servane Gey; Visualization, Arthur Leroy; Supervision, Arthur Leroy and Servane Gey; Project Administration, Arthur Leroy and Servane Gey; Funding Acquisition, Arthur Leroy

Funding: This research received no external funding

Acknowledgments: We thank French Swimming Federation for the data set they provided, their confidence and their continuous help. We also thank the two reviewers for their helpful comments and suggestions that improved, in our opinion, the quality of the paper.

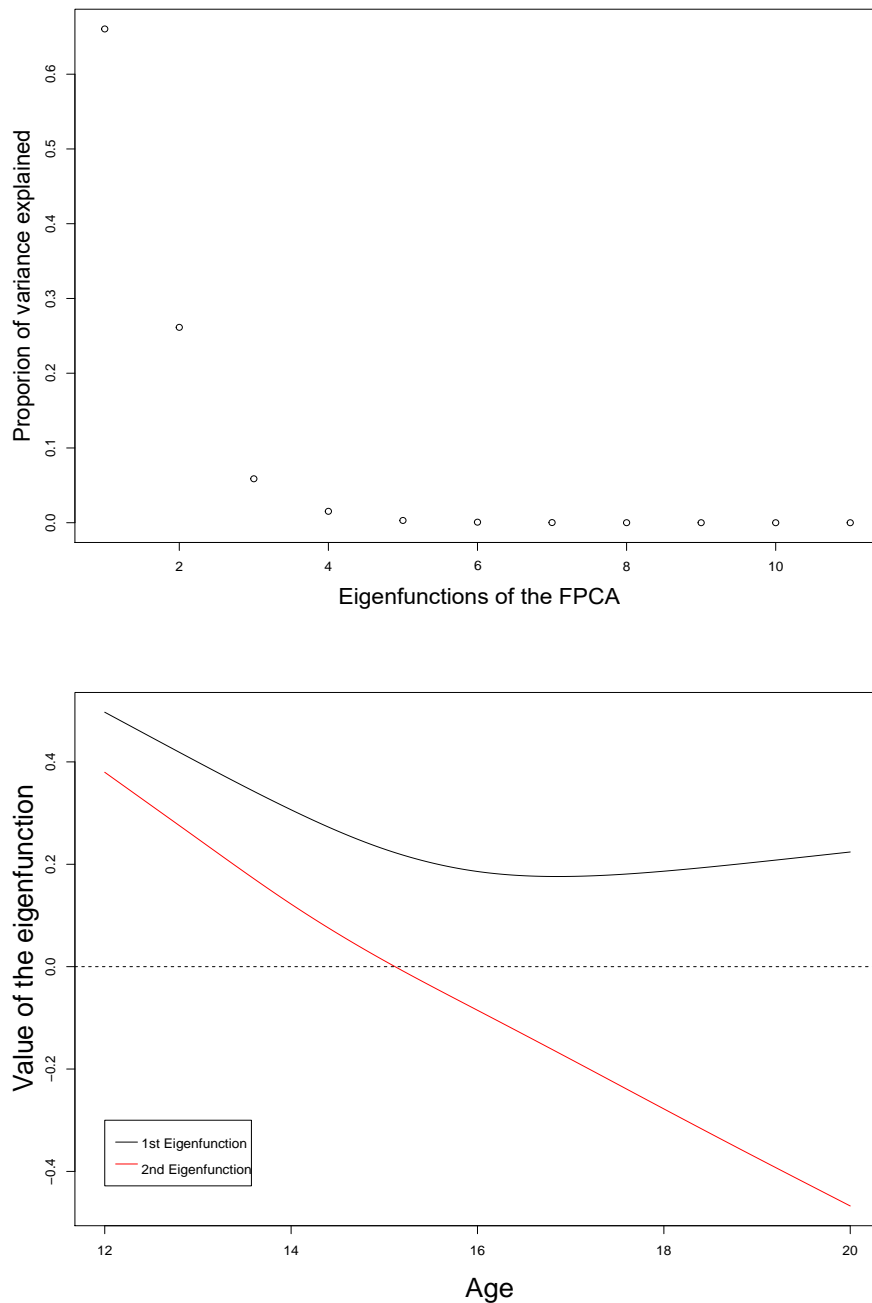
Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI Multidisciplinary Digital Publishing Institute
FDA Functional Data Analysis
FPCA Functional Principal Component Analysis
RI Rand Index
ARI Adjusted Rand Index
BIC Bayesian Information Criterion
ICL Integrated Classification Likelihood
EM Expectation-Maximization

461

462 **Appendix**

Bottom: Values of the two first eigenfunctions. Eigenfunctions are orthogonal each others and display the main modes of variation of the curves. The first eigenfunction mainly informs on differences at young ages, while the second focuses on the opposition between speeds at young and older ages.

Figure A1. Results of the FPCA on the progression curves. **Top:** Proportion of variance explained by each eigenfunction. With only 2 eigenfunctions, one can express about 90% of the total variance of the data set.

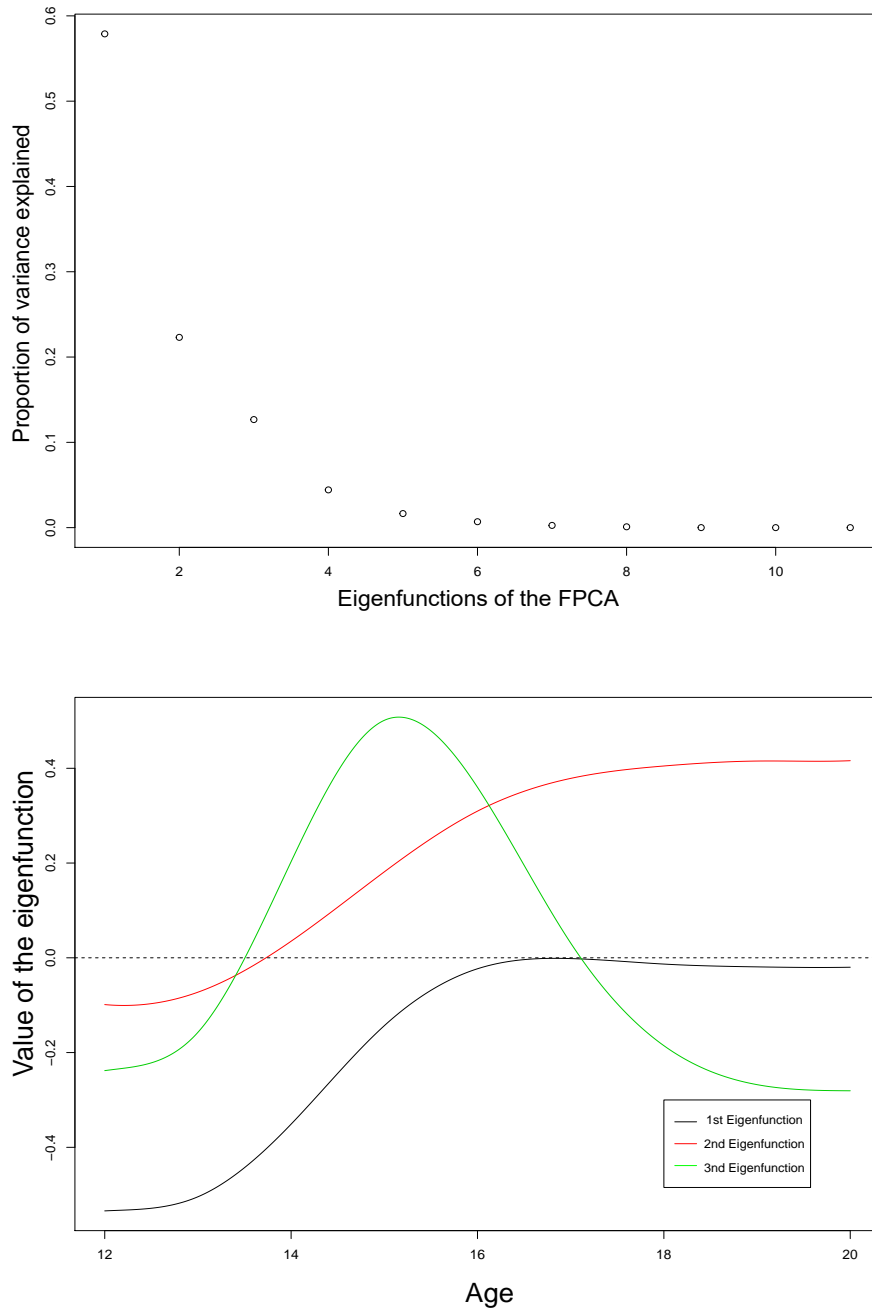


Figure A2. Results of the FPCA on the derivatives of the progression curves. **Top:** Proportion of variance explained by each eigenfunction. With only 3 eigenfunctions, one can express about 90% of the total variance of the data set. **Bottom:** Values of the three first eigenfunctions. Eigenfunctions are orthogonal each others and display the main modes of variation of the curves. The first eigenfunction mainly informs on derivative differences at young ages, while the second focuses on the behaviour between 16 and 18 years old. The third eigenfunction expresses the differences of swimmers improvement between the middle and the bounds of the time interval.

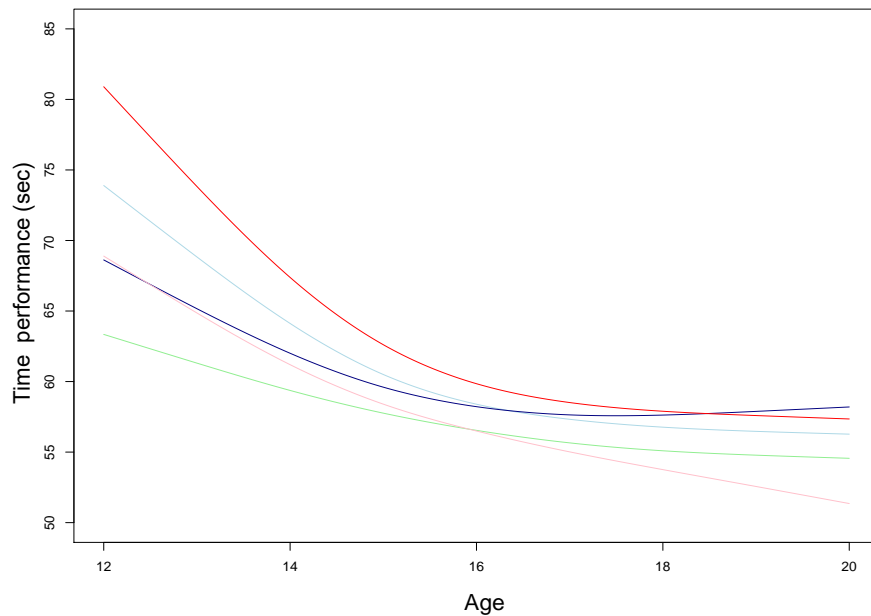


Figure A3. Clusters centres of the progressions curves. Computed with the univariate funHDDC algorithm. Clusters show different patterns of evolution, and some progression curves cross each others. The same level of performance at 12 years old (e.g. pink and blue curves) can lead to really different levels when older.

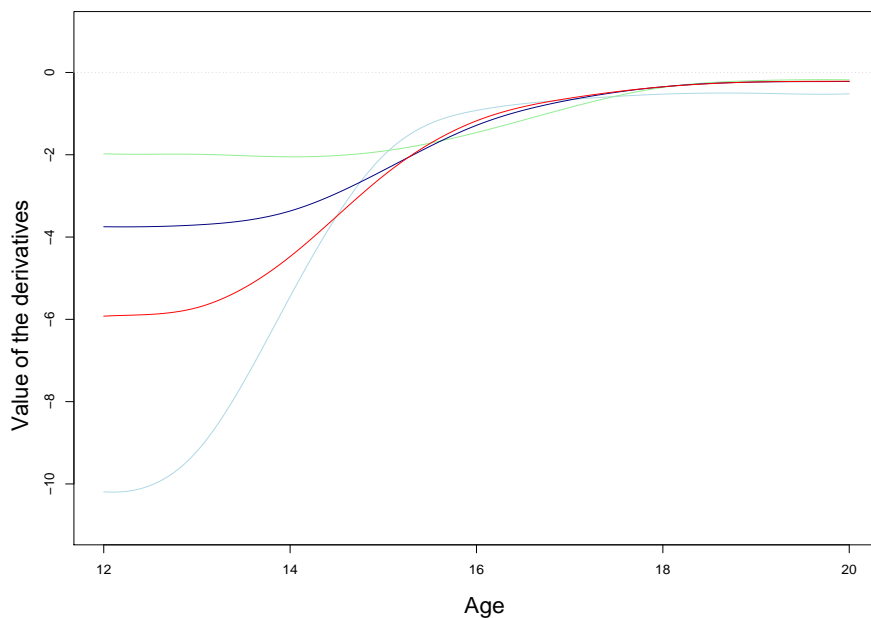


Figure A4. Clusters centres of the derivatives of the progressions curves. Computed with the univariate funHDDC algorithm. Clusters mostly differ on the value of the derivatives at young age and converge all to 0 at 20 years old.

463 **References**

- 464 1. Lima-Borges, D.S.; Martinez, P.F.; Vanderlei, L.C.M.; Barbosa, F.S.S.; Oliveira-Junior, S.A. Autonomic
465 Modulations of Heart Rate Variability Are Associated with Sports Injury Incidence in Sprint Swimmers.
466 *The Physician and Sportsmedicine* **2018**, pp. 1–11. doi:10.1080/00913847.2018.1450606.
- 467 2. Carey, D.L.; Crossley, K.M.; Whiteley, R.; Mosler, A.; Ong, K.L.; Crow, J.; Morris, M.E. Modelling Training
468 Loads and Injuries: The Dangers of Discretization. *Medicine and Science in Sports and Exercise* **2018**.
469 doi:10.1249/MSS.0000000000001685.
- 470 3. Boccia, G.; Moisè, P.; Franceschi, A.; Trova, F.; Panero, D.; La Torre, A.; Rainoldi, A.; Schena, F.; Cardinale,
471 M. Career Performance Trajectories in Track and Field Jumping Events from Youth to Senior Success: The
472 Importance of Learning and Development. *PLoS ONE* **2017**, *12*. doi:10.1371/journal.pone.0170744.
- 473 4. Brockwell, P.J.; Davis, R.A. *Time Series: Theory and Methods*; Springer Science & Business Media, 2013.
- 474 5. Warren Liao, T. Clustering of Time Series Data—a Survey. *Pattern Recognition* **2005**, *38*, 1857–1874.
475 doi:10.1016/j.patcog.2005.01.025.
- 476 6. de Boor, C. On Calculating with B-Splines. *Journal of Approximation Theory* **1972**, *6*, 50–62.
477 doi:10.1016/0021-9045(72)90080-9.
- 478 7. Forrester, S.E.; Townend, J. The Effect of Running Velocity on Footstrike Angle – A Curve-Clustering
479 Approach. *Gait & Posture* **2015**, *41*, 26–32. doi:10.1016/j.gaitpost.2014.08.004.
- 480 8. Mallor, F.; Leon, T.; Gaston, M.; Izquierdo, M. Changes in Power Curve Shapes as an
481 Indicator of Fatigue during Dynamic Contractions. *Journal of Biomechanics* **2010**, *43*, 1627–1631.
482 doi:10.1016/j.jbiomech.2010.01.038.
- 483 9. Helwig, N.E.; Shorter, K.A.; Ma, P.; Hsiao-Wecksler, E.T. Smoothing Spline Analysis of Variance Models: A
484 New Tool for the Analysis of Cyclic Biomechanical Data. *Journal of Biomechanics* **10 03**, **2016**, *49*, 3216–3222.
485 doi:10.1016/j.jbiomech.2016.07.035.
- 486 10. Liebl, D.; Willwacher, S.; Hamill, J.; Brüggemann, G.P. Ankle Plantarflexion Strength in Rearfoot and
487 Forefoot Runners: A Novel Clusteranalytic Approach. *Human Movement Science* **2014**, *35*, 104–120.
488 doi:10.1016/j.humov.2014.03.008.
- 489 11. Ramsay, J.O.; Dalzell, C.J. Some Tools for Functional Data Analysis. *Journal of the Royal Statistical Society.*
490 *Series B (Methodological)* **1991**, *53*, 539–572.
- 491 12. Gasser, T.; Muller, H.G.; Kohler, W.; Molinari, L.; Prader, A. Nonparametric Regression Analysis of Growth
492 Curves. *The Annals of Statistics* **1984**, *12*, 210–229.
- 493 13. Liebl, D. Modeling and Forecasting Electricity Spot Prices: A Functional Data Perspective. *The Annals of*
494 *Applied Statistics* **2013**, *7*, 1562–1592, [1310.1628]. doi:10.1214/13-AOAS652.
- 495 14. Bouveyron, C.; Bozzi, L.; Jacques, J.; Jollois, F.X. The Functional Latent Block Model for the Co-Clustering
496 of Electricity Consumption Curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **2018**,
497 *67*, 897–915. doi:10.1111/rssc.12260.
- 498 15. Shen, M.; Tan, H.; Zhou, S.; Smith, G.N.; Walker, M.C.; Wen, S.W. Trajectory of Blood Pressure Change
499 during Pregnancy and the Role of Pre-Gravid Blood Pressure: A Functional Data Analysis Approach.
500 *Scientific Reports* **2017**, *7*, 6227. doi:10.1038/s41598-017-06606-0.
- 501 16. Velasco Herrera, V.M.; Soon, W.; Velasco Herrera, G.; Traversi, R.; Horiuchi, K. Generalization of the
502 Cross-Wavelet Function. *New Astronomy* **2017**, *56*, 86–93. doi:10.1016/j.newast.2017.04.012.
- 503 17. Johnston, K.; Wattie, N.; Schorer, J.; Baker, J. Talent Identification in Sport: A Systematic Review. *Sports*
504 *Medicine* **2018**, *48*, 97–109. doi:10.1007/s40279-017-0803-2.
- 505 18. Berthelot, G.; Sedeaud, A.; Marck, A.; Antero-Jacquemin, J.; Schipman, J.; Saulière, G.; Marc, A.; Desgorges,
506 F.D.; Toussaint, J.F. Has Athletic Performance Reached Its Peak? *Sports Medicine (Auckland, N.Z.)* **2015**,
507 *45*, 1263–1271. doi:10.1007/s40279-015-0347-2.
- 508 19. Moesch, K.; Elbe, A.M.; Hauge, M.L.T.; Wikman, J.M. Late Specialization: The Key to Success in
509 Centimeters, Grams, or Seconds (Cgs) Sports. *Scandinavian Journal of Medicine & Science in Sports* **2011**,
510 *21*, e282–290. doi:10.1111/j.1600-0838.2010.01280.x.
- 511 20. Vaeyens, R.; Lenoir, M.; Williams, A.M.; Philippaerts, R.M. Talent Identification and Development
512 Programmes in Sport. *Sports Medicine* **2008**, *38*, 703–714. doi:10.2165/00007256-200838090-00001.

- 513 21. Mohamed, H.; Vaeyens, R.; Matthys, S.; Multael, M.; Lefevre, J.; Lenoir, M.; Philippaerts, R. Anthropometric
514 and Performance Measures for the Development of a Talent Detection and Identification Model in Youth
515 Handball. *Journal of Sports Sciences* **2009**, *27*, 257–266. doi:10.1080/02640410802482417.
- 516 22. Goto, H.; Morris, J.G.; Nevill, M.E. Influence of Biological Maturity on the Match Performance of
517 8 to 16 Year Old Elite Male Youth Soccer Players. *Journal of Strength and Conditioning Research* **2018**.
518 doi:10.1519/jsc.0000000000002510.
- 519 23. Wattie, N.; Schorer, J.; Baker, J. The Relative Age Effect in Sport: A Developmental Systems Model. *Sports*
520 *Medicine (Auckland, N.Z.)* **2015**, *45*, 83–94. doi:10.1007/s40279-014-0248-9.
- 521 24. Kearney, P.E.; Hayes, P.R. Excelling at Youth Level in Competitive Track and Field Athletics Is Not a
522 Prerequisite for Later Success. *Journal of Sports Sciences* **2018**, pp. 1–8. doi:10.1080/02640414.2018.1465724.
- 523 25. Vaeyens, R.; Güllich, A.; Warr, C.R.; Philippaerts, R. Talent Identification and Promotion Programmes of
524 Olympic Athletes. *Journal of Sports Sciences* **2009**, *27*, 1367–1380. doi:10.1080/02640410903110974.
- 525 26. Ericsson, K.A.; Hoffman, R.R.; Kozbelt, A.; Williams, A.M. *The Cambridge Handbook of Expertise and Expert*
526 *Performance*; Cambridge University Press, 2018.
- 527 27. Ramsay, J.O.; Silverman, B.W. *Functional Data Analysis*; Springer, 2005.
- 528 28. Jacques, J.; Preda, C. Functional Data Clustering: A Survey. *Advances in Data Analysis and Classification*
529 **2014**, *8*, 231–255. doi:10.1007/s11634-013-0158-y.
- 530 29. Giacomini, M.; Lambert-Lacroix, S.; Marot, G.; Picard, F. Wavelet-Based Clustering for Mixed-Effects
531 Functional Models in High Dimension. *Biometrics*, *69*, 31–40. doi:10.1111/j.1541-0420.2012.01828.x.
- 532 30. Ramsay, J.O.; Silverman, B.W. *Applied Functional Data Analysis: Methods and Case Studies*; Vol. 77, Springer
533 New York, 2002.
- 534 31. Abraham, C.; Cornillon, P.A.; Matzner-Løber, E.; Molinari, N. Unsupervised Curve Clustering Using
535 B-Splines. *Scandinavian Journal of Statistics* **2003**, *30*, 581–595. doi:10.1111/1467-9469.00350.
- 536 32. Ferraty, F.; Vieu, P. *Nonparametric Functional Data Analysis: Theory and Practice*; Springer Science & Business
537 Media, 2006.
- 538 33. Jacques, J.; Preda, C. Funclust: A Curves Clustering Method Using Functional Random Variables Density
539 Approximation. *Neurocomputing* **2013**, *112*, 164–171. doi:10.1016/j.neucom.2012.11.042.
- 540 34. Schmutz, A.; Jacques, J.; Bouveyron, C.; Cheze, L.; Martin, P. Clustering Multivariate Functional Data in
541 Group-Specific Functional Subspaces. *HAL* **2018**.
- 542 35. Bouveyron, C.; Jacques, J. Model-Based Clustering of Time Series in Group-Specific Functional Subspaces.
543 *Advances in Data Analysis and Classification* **2011**, *5*, 281–300.
- 544 36. Peng, J.; Müller, H.G. Distance-Based Clustering of Sparsely Observed Stochastic Processes, with
545 Applications to Online Auctions. *The Annals of Applied Statistics* **2008**, *2*, 1056–1077, [0805.0463].
546 doi:10.1214/08-AOAS172.
- 547 37. James, G.M.; Sugar, C.A. Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical*
548 *Association* **2003**, *98*, 397–408. doi:10.1198/016214503000189.
- 549 38. Chiou, J.M.; Li, P.L. Functional Clustering and Identifying Substructures of Longitudinal Data. *Journal of the*
550 *Royal Statistical Society: Series B (Statistical Methodology)*, *69*, 679–699. doi:10.1111/j.1467-9868.2007.00605.x.
- 551 39. Jiang, H.; Serban, N. Clustering Random Curves Under Spatial Interdependence With Application to
552 Service Accessibility. *Technometrics* **2012**, *54*, 108–119. doi:10.1080/00401706.2012.657106.
- 553 40. Rand, W.M. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical*
554 *Association* **1971**, *66*, 846–850. doi:10.1080/01621459.1971.10482356.
- 555 41. Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J.M.; Perona, I.n. An Extensive Comparative Study of
556 Cluster Validity Indices. *Pattern Recognition* **2013**, *46*, 243–256. doi:10.1016/j.patcog.2012.07.021.
- 557 42. von Luxburg, U.; Williamson, R.C.; Guyon, I. Clustering: Science or Art? *Proceedings of ICML Workshop on*
558 *Unsupervised and Transfer Learning* **2012**, pp. 65–79.