Noname manuscript No.
(will be inserted by the editor)

Archetypoid Analysis for Sports Analytics

Vinué, G. · Epifanio, I.

the date of receipt and acceptance should be inserted later

Abstract We intend to understand the growing amount of sports performance data by finding extreme data points, which makes human interpretation easier. In archetypoid analysis each datum is expressed as a mixture of actual observations (archetypoids). Therefore, it allows us to identify not only extreme athletes and teams, but also the composition of other athletes (or teams) according to the archetypoid athletes, and to establish a ranking. The utility of archetypoids in sports is illustrated with basketball and soccer data in three scenarios. Firstly, with multivariate data, where they are compared with other alternatives, showing their best results. Secondly, despite the fact that functional data are common in sports (time series or trajectories), functional data analysis has not been exploited until now, due to the sparseness of functions. In the second scenario, we extend archetypoid analysis for sparse functional data, furthermore showing the potential of functional data analysis in sports analytics. Finally, in the third scenario, features are not available, so we use proximities. We extend archetypoid analysis when asymmetric relations are present in data. This study provides information that will provide valuable knowledge about player/team/league performance so that we can analyze athlete's careers.

Keywords Archetype analysis · Sports data mining · Functional data analysis · Extreme point · Multidimensional scaling · Performance analysis

This work has been partially supported by Grant DPI2013- 47279-C2-1-R. The databases and R code (including the web application) to reproduce the results can be freely accessed at www.uv.es/vivigui/software.

Vinué, G.
Department of Statistics and O.R., University of Valencia, 46100 Burjassot, Spain
Tel.: +34-659363291
E-mail: guillermo.vinue@uv.es

Epifanio, I.
Dept. Matemàtiques and Institut de Matemàtiques i Aplicacions de Castelló. Campus del Riu Sec. Universitat Jaume I, 12071 Castelló, Spain

1 Introduction

A high level of professionalism, advances in technology and complex data sets containing detailed information about player and team performance have contributed to the development of sport science (Williams and Wragg, 2004). Sports performance analysis is a growing branch within sport science. It is concerned with the investigation of actual sports performance in training or competition (O'Donoghue, 2010). One of the most important issues in sport science is to identify outstanding athletes (or teams) based on their performance. In particular, the question regarding who the best players are in a competition is at the center of debates between sport managers and fans. There are lists and rankings, each with their own criteria and biases. A thorough analysis of the players' performance has direct consequences on the composition of the team and on transfer policies because this evaluation is used to decide whether the team should recruit or extend the player. To that end, managers and scouts assess players based on their knowledge and experience. However, this process is based on subjective criteria. The observer has developed notions of what a good player should look like based on his/her previous experience (Shea and Baker, 2013). Thus, the evaluation is subjective/biased, which may cause flawed or incomplete conclusions. Traditional means of evaluating players and teams are best used in conjunction with rigorous statistical methods. One interesting approach to provide objective evidence about how good (or bad) the players perform based on the statistics collected for them is described by Eugster (2012). The author uses archetype analysis (AA) to obtain outstanding athletes (both positively and negatively). These are the players who differ most from the rest in terms of their performance. It has been shown that extreme constituents (Davis and Love, 2010) facilitate human understanding and interpretation of data because of the principle of opposites (Thurau et al., 2012). In other words, extremes are better than central points for human interpretation.

AA was first proposed by Cutler and Breiman (1994). Its aim is to find pure types (the archetypes) in such a way that the other observations are a mixture of them. Archetypes are data-driven extreme points. As is rightly pointed out by Eugster (2012), in sports these extreme points correspond to positively or negatively prominent players. However, AA has an important drawback: archetypes are a convex combination of the sampled individuals, but they are not necessarily observed individuals. Furthermore, there are situations where archetypes are fictitious, see for example Seiler and Wohlrabe (2013). In sports, this situation can cause interpretation problems for analysts. In order to cope with this limitation, a new archetypal concept was introduced: the archetypoid, which is a real (observed) archetypal case (Vinué et al., 2015; Vinué, 2014). Archetypoids accommodate human cognition by focusing on extreme opposites (Thurau et al., 2012). Furthermore, they make an intuitive understanding of the results easier even for non-experts (Vinué et al., 2015; Thurau et al., 2012), since archetypoid analysis (ADA) represents the data as mixtures of extreme cases, and not as mixtures of mixtures, as AA does.

In this paper, we propose using ADA to find real outstanding (extreme) players and teams based on their performance information in three different scenarios. Firstly, in the multivariate case, where several classical sport variables (features) are available. Secondly, in combination with sparse functional data, for which archetypoids are defined for the first time in this work. Thirdly and finally, when only dissimilarities between observations are known (features are unavailable) and these dissimilarities are not metric, but asymmetric proximities.

Functional data analysis (FDA) is a modern branch of statistics that analyzes data that are drawn from continuous underlying processes, often time, i.e. a whole function is a datum. An excellent overview of FDA can be found in Ramsay and Silverman (2005). Even though functions are measured discretely at certain points, a continuous curve or function lies behind these data. The sampling time points do not have to be equally spaced and both the argument values and their cardinality can vary across cases, which makes the FDA framework highly flexible.

On the one hand, our approach is a natural extension and improvement of the methodology proposed by Eugster (2012) with regard to multivariate data. On the other hand, the methodology can also be used with other available information, such as asymmetric relations and sparse functional data. The main goal is to provide sport analysts with a statistical tool for objectively identifying extreme observations with certain noticeable features and to express the other observations as a mixture of them. Furthermore, a ranking of the observations based on their performance can also be obtained. The application of ADA focuses on two mass sports: basketball and soccer. However, it can be used with any other sports data.

The main novelties of this work consist of: 1. Introducing ADA to the sports analytics community, together with FDA; 2. Extending ADA to sparse functional data; 3. Proposing a methodology for computing archetypoids when asymmetric proximities are the only available information. The outline of the paper is as follows: Section 2 is dedicated to preliminaries. In Section 3 related work is reviewed. Section 4 reviews AA and ADA in the multivariate case, ADA is extended to deal with sparse functional data and an ADA extension is introduced when asymmetric relations are present in data. We also present how a performance-based ranking can be obtained. In Section 5, ADA is used in three scenarios. In the multivariate case, ADA is applied to the same 2-D basketball data used by Eugster (2012) and to another basic basketball player statistics data set, and compared with other alternative methodologies and previous approaches. In the second scenario, ADA and FDA are applied to longitudinal basketball data. In the third scenario, ADA is applied to asymmetric proximities derived from soccer data. Finally, Section 6 ends the paper with some conclusions.

2 Preliminaries

2.1 Functional Data Analysis (FDA)

Many multivariate statistical methods, such as simple linear models, ANOVA, generalized linear models, PCA, clustering and classification, among others, have been adapted to the functional framework and have their functional counterpart. ADA has also been defined for functions by Epifanio (2016), where it was shown that functional archetypoids can be computed as in the multivariate case if the functions are expressed in an orthonormal basis, by applying ADA to the coefficients in that basis. However, in Epifanio (2016) functions are measured over a densely sampled grid. When functions are measured over a relatively sparse set of points, we have sparse functional data. An excellent survey on sparsely sampled functions is provided by James (2010). In this case, alternative methodologies are required. Note that when functions are measured over a fine grid of time points, it is possible to fit a separate function for each case using any reasonable basis. However, in the sparse case, this approach fails and the information from all functions must be used to fit each function.

2.2 h-plot representation

Recently, a multidimensional scaling methodology for representing asymmetric data was proposed by Epifanio (2013, 2014) (it improved on other alternatives). The dissimilarity matrix $\mathbf{D}$ is viewed as a data matrix and their variables are displayed with an h-plot.

For computing the h-plot in two dimensions, the two largest eigenvalues (λ_1 and λ_2) of the variance- covariance matrix, $\mathbf{S}$, of $\mathbf{D}$, are calculated, together with their corresponding unit eigenvectors, $\mathbf{q_1}$ and $\mathbf{q_2}$. The representation is given by $\mathbf{H_2} = (\sqrt{\lambda_1}\mathbf{q_1}, \sqrt{\lambda_2}\mathbf{q_2})$. The goodness-of-fit is estimated by $(\lambda_1^2+\lambda_2^2)/\sum_j \lambda_j^2$, and the closer it is to 1, the better the fit. The Euclidean distance between rows h_i and h_j is approximately the sample standard deviation of the difference between variables j and i. Two profiles with a large (or small) Euclidean distance between them in the h-plot are different (or similar). Note that the goal of the h-plot is not to preserve the exact pairwise dissimilarities as in other multidimensional scaling methods, but to preserve the relationships between the profiles. This point of view is of particular interest when dealing with non-metric dissimilarities, since these dissimilarities cannot be exactly represented in a Euclidean space. To summarize, the original dissimilarity matrix $\mathbf{D}$ is mapped to a 2-D feature matrix.

3 Related work

The book by Shea and Baker (Shea and Baker (2013)) introduces original metrics for analyzing player performance and explores the question of who

the most valuable players are, among other matters. A second book was also written by Shea (Shea (2014)), which investigates player evaluation and types using SportVU data (SportVU is a camera system used to track player positions that collects data at a rate of 25 times per second). In Winston (2009) there is a description of mathematical methods that are used to assess players and team performance. Kubatko et al. (2007) published a paper with the aim of providing a common starting point for scientific research in basketball. In Bhandari et al. (1997) a data-mining application was developed to discover patterns in basketball data.

In recent years, AA has been used in different domains such as multi-document summarization (Canhasi and Kononenko, 2013, 2014), the evaluation of scientists (Seiler and Wohlrabe, 2013), developmental psychology (Ragozini et al., 2017), biology (D'Esposito et al., 2012), market research and benchmarking (Li et al., 2003; Porzio et al., 2008; Midgley and Venaik, 2013), industrial engineering (Epifanio et al., 2013), e-learning (Theodosiou et al., 2013), machine learning problems (Mørup and Hansen, 2012), image analysis (Bauckhage and Thurau, 2009) and astrophysics (Chan et al., 2003).

As regards FDA, it is increasingly being used in different fields, such as criminology, economics and archaeology (Ramsay and Silverman (2002)), biomedicine (Ullah and Finch (2013)) and psychology (Levitin et al. (2007)). In spite of the fact that time series data or movement trajectories are common in sports, we have only found applications in sport biomechanics or medicine (Epifanio et al., 2008; Harrison et al., 2007; Donoghue et al., 2008; Harrison, 2014) and player's ageing curves (Wakim and Jin, 2014). In Wakim and Jin (2014), k-means clustering of PCA scores computed as proposed by Yao and Müller (2005) is performed for Win Shares on a different database from those we use. The values of the mean curves for each cluster are between -2 and 6; no extreme trajectories are obtained.

4 Methodology

4.1 AA and ADA for multivariate numeric data

Let $\mathbf{X}$ be an $n \times m$ matrix of real numbers representing a multivariate data set with n observations and m variables. For a given k, the objective of AA is to find a $k \times m$ matrix $\mathbf{Z}$ that characterizes the archetypal patterns in the data. This method convexly approximates data points using archetypes that are themselves convex combinations of data points. More precisely, AA is aimed at obtaining an $n \times k$ coefficient matrix α and a $k \times n$ matrix β such that $min_{\alpha,\beta}\|\mathbf{X} - \alpha\beta\mathbf{X}\|$, where the elements of matrices α and β are not negative and the rows of α and columns of β add up to 1 ($\mathbf{Z} = \beta\mathbf{X}$). In other words, the objective of AA is to minimize the residual sum of squares (RSS) that arises from combining the equation that shows $\mathbf{x}_i$ as being approximated by

a convex combination of $\mathbf{z}_j$'s (archetypes), i.e. $\|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2$, and the equation that shows $\mathbf{z}_j$'s as convex combinations of the data ($\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$):

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2$$

under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ and $i = 1, \ldots, n$ and

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ and $j = 1, \ldots, k$

On the one hand, constraint 1) implies that the predictors of $\mathbf{x}_i$ are convex combinations of the collection of archetypes, $\hat{\mathbf{x}}_i = \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j$. Each α_{ij} is the weight of the archetype j for the individual i, i.e., the α coefficients represent how much each archetype contributes to the approximation of each observation. On the other hand, constraint 2) means that archetypes $\mathbf{z}_j$ are convex combinations of the data points. To solve AA, Cutler and Breiman (1994) proposed an algorithm using an alternating minimization algorithm, where each step involves solving several convex least squares.

According to the previous definition, archetypes are not necessarily real observed cases. The archetypes would correspond to specific cases when $\mathbf{z}_j$ is a data point of the sample, i.e., when only one β_{jl} is equal to 1 in constraint 2) for each j. As $\beta_{jl} \geq 0$ and the sum of constraint 2) is 1, this implies that β_{jl} should only take on the value 0 or 1. In ADA, the original optimization problem therefore becomes:

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2, \qquad (1)$$

under the constraints

1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ and $i = 1, \ldots, n$ and

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \in \{0, 1\}$ and $j = 1, \ldots, k$ i.e., $\beta_{jl} = 1$ for one and only one l and $\beta_{jl} = 0$ otherwise.

Before archetypoids appeared, the most widely used strategy to overcome the fact that archetypes are not sampled individuals was to compute the nearest individual to each archetype. This can be done in different ways, the three

most common of which are as follows. The first possibility consists of computing the Euclidean distance between the k archetypes and the cases and identifying the nearest ones, as mentioned by Epifanio et al. (2013) (this set is referred to as $cand_{ns}$). The second determines the cases with the maximum α value for each archetype, i.e. the cases with the largest relative share for the respective archetype (set $cand_{\alpha}$, as presented by Eugster (2012) and Seiler and Wohlrabe (2013)). The third possibility chooses the cases with the maximum β value for each archetype, i.e., the major contributors in the generation of the archetypes (set $cand_{\beta}$).

ADA can be solved by trying all the possible combinations (a combinatorial solution, which is certainly the optimal solution), but its computational cost is very high. Therefore, the archetypoid algorithm was proposed (see Vinué et al. (2015) for details). It has two phases: a BUILD phase and a SWAP phase. In the BUILD step, an initial set of archetypoids is determined. This initial set of archetypoids can be $cand_{ns}$, $cand_{\alpha}$ or $cand_{\beta}$. The aim of the SWAP phase is to improve the current set of archetypoids by exchanging selected cases for unselected cases and by checking whether or not these replacements reduce the objective function of Equation 1. In the SWAP phase, for each archetypoid a, for each non-archetypoid data point o, swap a and o and compute the RSS of the configuration, then select the configuration with the lowest RSS. This is done until there is no change in the archetypoids.

Note that neither archetypes nor archetypoids are necessarily nested. For instance, if three and four archetypoids are calculated, there is no reason for these four to include the first three computed, as the existing ones can change to better capture the shape of the data set.

The number of archetypes or archetypoids to compute is the user's decision (as is the number of clusters in a clustering problem). For guidance in this choice, the well-known elbow criterion can be used. The ADA (or AA) algorithm is run for different numbers of k and their RSS are plotted (ADA is run beginning from the three possible initializations, and the solution with the smallest RSS is considered). The point where this curve flattens suggests the correct value of k.

For $k = 1$, the archetype is the sample mean, whereas the archetypoid is the medoid (with one cluster) (Vinué et al., 2015). The medoid is the object in the cluster for which the average dissimilarity to all the objects in the cluster is minimal (Kaufman and Rousseeuw, 1990). As the number of points in $\mathbf{X}$ is finite, its convex hull is a convex polytope, which is the convex hull of its N vertices. For $1 < k < N$, archetypes belong to the boundary of the convex hull of $\mathbf{X}$, but archetypoids are not necessarily vertices, as shown in Vinué et al. (2015). For $k \geq N$, archetypoids (and archetypes) coincide with the vertices (Kreiman and Milman, 1940).

In Section 5, we will compare ADA with other unsupervised methods.These methods, together with their relations, are detailed in Vinué et al. (2015). We briefly reproduce them here.

The most closely related method is the Simplex Volume Maximization (SiVM) algorithm introduced by Thurau et al. (2012), where the same prob-

lem as ADA was formulated. SiVM sequentially chooses the $j + 1$ vertex that maximizes the simplex (polytope which is the convex hull of its vertices) volume given the first j vertices. Due to its efficiency (low running times), reasonable solutions are given by SiVM in the case of very large databases. However, SiVM assumes that archetypoids are vertices of the convex hull of $\mathbf{X}$, but this is not necessarily true as shown in Vinué et al. (2015) (archetypoids are not necessarily on the boundary of the convex hull of data like archetypes), which could prevent SiVM from finding the optimal solution. The other drawback of SiVM is that it is a greedy algorithm, which is fast and often returns good solutions, but a particular selection in a certain iteration could prevent a good solution being found because previous selections are not reconsidered. A stochastic version of SiVM was introduced by Kersting et al. (2012).

The Sparse Modeling Representative Selection method (SMRS), developed by Elhamifar et al. (2012), also addresses the problem of finding a subset of data points that efficiently describes the entire data set. It is assumed that each observation can be expressed as a linear combination of the representatives. Then, the problem of finding the representatives is formulated as a sparse multiple measurement vector problem. The representative points coincide with some of the actual data points, as is the case with archetypoids.

In the supplementary material of Vinué et al. (2015), computational costs are also analyzed for several methods. The speed of our algorithm depends on the efficiency of the convex least squares method, as was the case with the archetype algorithm implemented by Eugster and Leisch (2009). We use the penalized non-negative least squares method, that according to Cutler and Breiman (1994), is relatively slow but can be used if the number of variables is larger than the number of observations.

4.2 ADA for sparse time series data with FDA

Here, we extend functional archetypoid analysis (FADA) for sparse functional data. Let us assume that n smooth functions, $x_1(t)$, ..., $x_n(t)$, are observed, with the i-th function measured at t_{i1},..., t_{in_i} points, i.e. $x_{ij} = x(t_{ij})$. Based on the Karhunen-Loeve expansion, the functions are approximated by

$$\hat{x}_i(t) = \hat{\mu}(t) + \sum_{j=1}^{m} \hat{\xi}_{ij}\hat{\phi}_j(t), \tag{2}$$

where ξ_{ij} is the jth principal component score for case i, $\phi_j(t)$ represents the jth principal component function (eigenfunction), and m is the number of principal components used in the estimation. We use the geometric approach to obtain the maximum likelihood estimation of the functional principal components from sparse functional data proposed by Peng and Paul (2009), which outperforms other estimation procedures (James et al., 2000; Yao and Müller, 2005) and which also incorporates information from all the curves. In Peng and Paul (2009), a model selection procedure based on the minimization of

an approximate cross-validation (CV) score was also proposed for choosing both m and the number of basis functions M. These M functions are cubic B-splines with equally spaced knots, and are used in the model to represent the eigenfunctions. These procedures are implemented in the R package `fpca` (Peng and Paul, 2011). It also allows us to estimate the functional principal component scores using the best linear unbiased predictors (Yao and Müller, 2005) and to predict the trajectory (entire function) for each subject as in Equation 2. Note that the eigenfunctions are orthonormal; therefore, to obtain FADA we can apply ADA to the $n \times m$ matrix $\mathbf{X}$, with the scores (the coefficients in the Karhunen-Loeve basis).

4.3 ADA for dissimilarity data with h-plot

In sports, non-metric pairwise data with violations of symmetry (A may beat B at home, but may lose as a visitor) or triangle inequality (A may defeat B, B may defeat C, but C may beat A) are common. In Vinué et al. (2015), a methodology for computing archetypoids when features are not available was proposed. We extend that methodology for working with asymmetric proximities. The idea is to represent the dissimilarities in $\mathbb{R}^m$, while trying to preserve the information given by the pairwise dissimilarities. Then, ADA is computed in this representation. Note that if dissimilarities is the only information available, archetypes could also be computed in this new space, but a correspondence could not be established with the original objects, because they are not in a vector space. Therefore, the crux of the matter is to find an appropriate representation. We use the h-plot representation explained in Section 2.2 for mapping the dissimilarity matrix to a 2-D feature matrix in a Euclidean space, and ADA is applied to this feature matrix.

4.4 Ranking of the observations by ADA

D'Esposito and Ragozini (2008) proposed ranking multivariate performances by finding a "worst-best" direction, projecting the data on it and finally ranking the observations in the associated univariate space. We could assume that high values in the variables correspond to good performances. When $k = 2$, two opposite extremes are obtained as archetypoids. It is expected that one of archetypoids corresponds to a case with high values in many variables. This archetypoid could be the "best" case. Whereas, the other archetypoid would correspond to a case with low values in many variables, i.e the "worst" case. The ranks for each variable of the two archetypoids with respect to the other data values should be investigated. The 'worst' archetypoid should have low values for most of the ranks, and vice versa for the 'best'. This 'worst-best' direction was determined by D'Esposito and Ragozini (2008) using two archetypes instead of two archetypoids. Note that alpha values tell us the contribution of each archetypoid to each observation. Therefore, ordering observations (players or teams, for example) along the 'worst-best' direction can

be achieved by simply considering the ranks of alpha values corresponding to the 'worst' case (note that these alpha values are complementary to the alpha values corresponding to the 'best' case, as they add up to 1). In other words, the ranking procedure is equivalent to sorting by the best archetypoid's alphas descending.

When k is larger than 2, richer information could be extracted than simply reducing the problem to the 'worst-best' direction. In many situations, there are different kinds of 'good' players/teams, and these may not be fully extracted when only two archetypoids are considered. When a larger k is considered, alpha values can also be used for ranking with respect to the features highlighted for the corresponding archetypoid. However, the ranking is not unique, but rather several rankings corresponding to each archetypoid can be obtained, as mentioned by Eugster (2012).

5 Sports applications

Archetypes and archetypoids are computed by means of the `archetypes` R package (Eugster and Leisch, 2009; R Development Core Team, 2016) and the `Anthropometry` R package (Vinué et al., 2017; Vinué, 2017), respectively.

5.1 Player performance analysis with ADA

Two examples are discussed in this Section (more examples are analyzed in Section 5.1.1). We will demonstrate that archetypoids need not to be the same as $cand_{ns}$, $cand_{\alpha}$ (remember that this is the solution given in Eugster (2012)) or $cand_{\beta}$ individuals. In addition, we will see that ADA returns a more accurate solution.

First example The first example is used in Eugster (2012), where archetypes are calculated and real basketball players are analyzed. We focus on the NBA database that collects the total minutes played and field goals made by 441 players from the 2009/2010 season. As only two variables are considered, this allows us to illustrate the concepts by using bidimensional plots. Variables are standardized for both AA and ADA, as in Eugster (2012), since the range and meaning of variables are different.

The three archetypal players (this number was indicated by the elbow criterion) obtained by Eugster (2012) are Kevin Durant, Dwayne Jones and Jason Kidd. The first thing we do is compute the best possible set of archetypal players, the combinatorial solution for $k = 3$. This set is made up of Kevin Durant, Jason Kidd and Travis Diener and was obtained after 9 days of computation, using a forward sequential search procedure run on a single computer. When applying our archetypoid algorithm to the same database we did indeed obtain these three players as the final archetypoids in 25 seconds.

Fig. 1 shows the $cand_{\alpha}$ players, the archetypoid players and the players obtained with other unsupervised methods for $k = 3$: (i) SiVM and SMRS;

(ii) the Affinity Propagation algorithm (AP) proposed by Frey and Dueck (2007); (iii) a Bayesian partial membership model (BPM) (Mohamed et al., 2014); and (iv) PAM, k-means and fuzzy k-means clustering methods. SiVM and SMRS (with the regularization parameter equal to 20) obtain the same archetypal players as in Eugster (2012). With SMRS it is not possible to select exactly how many representatives have to be obtained. BPM seems to obtain separate athletes, but they are not as extreme as the archetypoids. AP and the clustering methods return representatives (AP does not allow us to set a specific number of representatives either) that are mainly in the middle of the data rather than at the boundaries, so they cannot be considered as outstanding players. Please also note that k-means and fuzzy k-means do not return observed individuals.

Next, a brief description of the main features of the archetypal players is introduced. In sports, a detailed analysis of the players' performance can help coaches to create individualized performance profiles.

Firstly, the archetypoids are described. Kevin Durant is a very good scorer because he scored a lot of shots in the time he was on court. According to this data, if he played an entire NBA game (48 minutes, without overtime periods), he would score almost 12 shots, which is a very good performance. Durant has won three NBA scoring titles to this day. Travis Diener has a similar profile (his ratio of field goals to minutes played is extremely low), as he only made 2 field goals in 50 minutes played. In addition, Jason Kidd might be considered an "ineffective scorer" because he played a large amount of minutes and he did not score many shots. However, it is well-known that Jason Kidd was a point guard whose main role was assisting instead of scoring. In fact, he is ranked second on the NBA's all-time assist list.

Regarding $cand_\alpha$ archetypes (solution in Eugster (2012)), Durant and Kidd was already described, so only Dwayne Jones remains to be described. Dwayne Jones was not able to score any points because he played very few minutes. This kind of players are called "benchwarmers" in basketball jargon. As pointed in Eugster (2012), Durant and Jones are the "natural" maximum and minimum in the 2D dataset.

Archetypoids, $cand_\alpha$ (solution in Eugster (2012)), SiVM and SMRS sets are quite similar in this simple example with two variables. The next examples will show the differences (and advantages) between archetypoids and the related approaches more clearly.

As an additional point, an interactive and easy-to-use web application to visualize and obtain this type of results is available [1]. The app can also be generated in R [2]. Please note that the R package `Anthropometry` and all its dependencies must be installed before launching the app in R.

Second example The second example consists of the basic statistics appearing in Hoopdata (2009) for NBA players from the 2010-2011 season who had

[1] http://bayes2.ucd.ie:3838/gvinue/AppBasketball

[2] **library(shiny) ; runUrl('http://www.uv.es/vivigui/softw/AppPlayers.zip')**

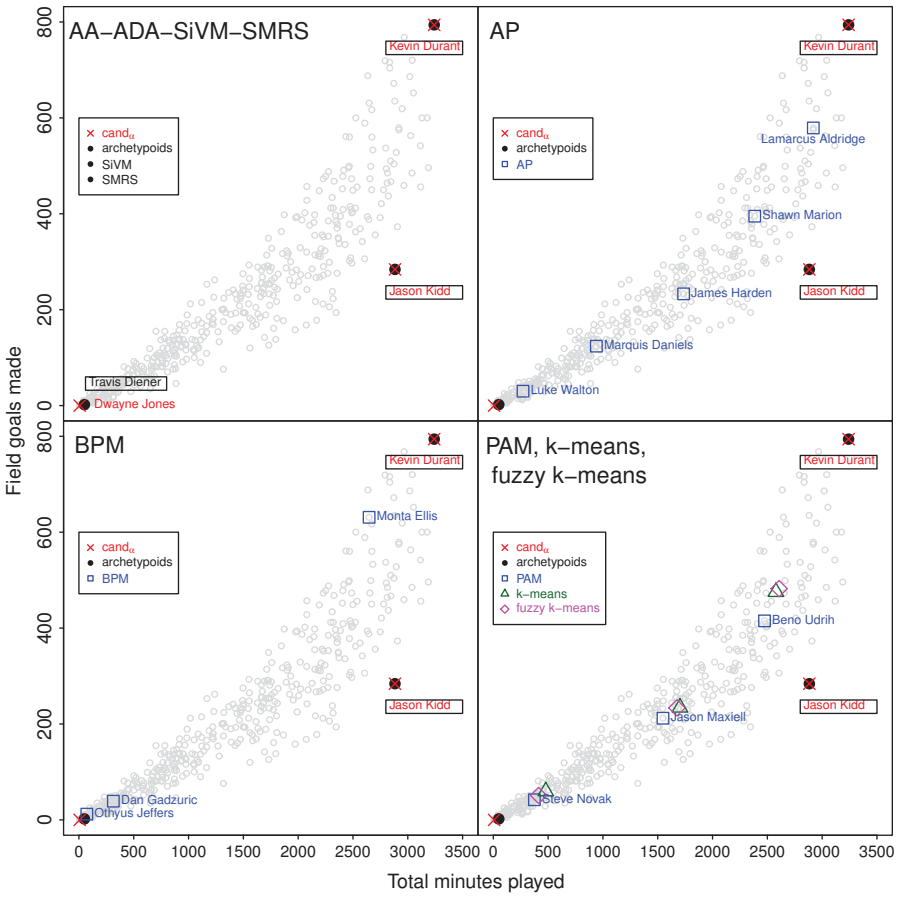

Fig. 1 $cand_\alpha$ players (with red crosses, obtained by Eugster (2012)) and archetypoid players (with solid black circles and frame box) for the total minutes played and field goals made by a set of NBA players from the 2009/2010 season, together with the representatives (with blue squares) obtained for the following methods, respectively: (a) SiVM and SMRS (not indicated because they match the $cand_\alpha$ players), (b) AP, (c) BPM and (d) PAM, k-means and fuzzy k-means (blue squares, green triangles and magenta diamonds, respectively). The RSS are: 0.00165 (ADA) and 0.00169 ($cand_\alpha$, SiVM and SMRS). The computational times are: AA 2 sec.; ADA for each initial candidate set, 25 sec.; SiVM $\ll$ 0.1 sec.; SMRS, 8 sec. (for regularization parameter 20: for others, for example 14 sec.).

played in at least 30 games and averaged at least 10 minutes per game (the same sample selection as made by Lutz (2012)), i.e. 332 players in total. However, the basic variables we use are different from the ones used by Lutz (2012). The variables used are (broken down per 40 minutes): GP (Games Played), GS (Games Started), Min (Minutes played), FG (Field Goals made), FGA (Field Goals Attempted), FG% (Field Goal Percentage: Field Goals made / Field Goals Attempted), 3P (Three Pointers made), 3PA (Three Pointers Attempted), 3P% (Three Point Percentage), FT (Free Throws made) FTA

(Free Throws Attempted), FT% (Free Throw Percentage), OR (Offensive Rebounds), DR (Defensive Rebounds), TR (Total Rebounds), AST (Assists), STL (Steals), TO (Turnovers), Blk (Blocks), PF (Personal Fouls), PTS (Points Scored). As before, variables are standardized. The elbow criterion suggests that 3 archetypoids should be chosen, as can be seen in Fig. 2 (there is also a less pronounced elbow at $k = 6$, but in the interests of brevity we examine the results of 3 archetypoids). Corresponding to Occams razor and following the same idea explained in Epifanio et al. (2013), three and six archetypoids can be considered as the best numbers of archetypoids (the law of parsimony is considered since a large numbers of representative cases may overwhelm the user and thus, be counterproductive, although if the user is interested in more archetypoids, they can be computed).

The ADA solution with $k = 3$, the solution given by Eugster (2012), $cand_\alpha$, and the first three SiVM representative players can be seen in Table 1, together with their corresponding RSS, runtimes and their percentiles in each variable. Note that the smallest RSS is obtained with archetypoids. It was not possible to select 3 representatives for any regularization parameter with SMRS. With the regularization parameter equal to 0.5, the smallest number of representative players are obtained, the following 7 players: Eddie House, Kobe Bryant, Joel Przybilla, Dwight Howard, Andris Biedrins, Eduardo Najera and Derek Fisher, given an RSS of 0.06982 and 20 sec. of computation for this regularization parameter (note that several regularization parameters were tested). The RSS for ADA with $k = 7$ is 0.05301, with archetypoids: Tim Duncan, Baron Davis, Matt Carroll, Kevin Durant, Aaron Gray, Daequan Cook and Derek Fisher.

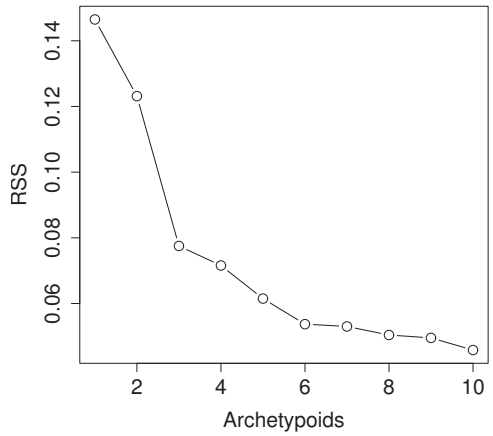

Fig. 2 Screeplot of the residual sum of squares for the 2010/2011 NBA database of basic statistics. The screeplot displays the RSS in descending order against the number of archetypoids.

Table 1 Percentiles for representative players with different methods for the 21 variables considered in the 2010/2011 NBA database of basic statistics. The top 3 features are highlighted in bold and the values above or equal 90 are in a frame box. The percentile for the mean is added under the column header. The RSS for $k = 3$ (chosen according to the elbow criterion) are: 0.07752 (ADA), 0.08455 ($cand_\alpha$) and 0.12709 (SiVM). The computational times are: AA (from 1 to 10) 1 minute; ADA with $k = 3$, 25 sec. (beginning from $cand_{ns}$), 20 sec. (beginning from $cand_\alpha$), 32 sec. (beginning from $cand_\beta$); SiVM $\ll$ 0.1 sec. .

Method	Players	GP	GS	Min	FG	FGA	FG%	3P	3PA	3P%	FT	FTA	FT%	OR	DR	TR	AST	STL	TO	Blk	PF	PTS
		44	55	51	52	51	58	52	53	36	59	55	42	60	56	58	68	52	54	64	59	52
ADA	**James**	78	90	**98**	**99**	97	84	58	59	52	97	98	49	38	78	64	92	86	97	54	8	**99**
	Dooling	**85**	47	43	25	37	10	75	**77**	61	27	19	78	3	5	3	**82**	60	54	13	27	30
	Favors	32	48	32	35	22	84	30	18	25	54	70	12	**98**	75	**88**	5	17	54	84	**98**	29
AA $cand_\alpha$	Foster	32	19	21	6	5	69	30	27	25	11	17	8	**100**	**93**	**99**	38	42	5	79	89	4
	Bryant	**100**	100	83	**100**	100	52	72	76	49	98	97	77	45	54	48	83	73	94	23	18	**100**
	Blake	**78**	9	33	3	8	3	75	72	**78**	3	1	**91**	14	30	20	77	42	38	13	18	4
SiVM	Howard	75	89	96	93	67	97	30	27	25	**99**	**100**	11	94	99	**99**	23	80	97	96	56	97
	Cardinal	32	20	3	3	4	30	**95**	83	**99**	8	4	**100**	20	20	19	53	86	5	47	88	7
	Nash	67	85	81	67	54	73	63	57	86	77	64	**98**	20	30	25	**100**	25	**99**	13	1	71

In this database, note that variable distributions are mostly positively skewed, because many players have low values in the variables and only a few players have high values. This fact should be taken into account when interpreting the percentile. As a consequence, the difference in values between two low percentiles (for example percentile 20 and 10) will generally be smaller than the difference in variable values for two high percentiles (for example 100 and 90).

According to the percentile information, the features of the players are as follows. We begin with the ADA solution. James has high values in nearly all variables (except in PF), representing a "very good" (star) player. However, Dooling and Favors have low-middle percentiles in many variables, and high percentiles only in some of them. In fact, Dooling and Favors complement each other: if Dooling has a high percentile in one variable, it implies that Favors has a low percentile in that variable, and vice versa. For example, Dooling has high percentiles, even beyond James' in GP, all Three Pointers variables and FT%, while Favors has high percentiles in FG%, all the Rebound variables, blocks and personal faults. Dooling and Favors do not stand out in many of the variables considered, but they do have some complementary strengths.

A correspondence between the archetypoid's profiles and those found in the $cand_\alpha$ set can be established. Bryant's profile would correspond with James' profile. Although, both have high percentiles in the majority of variables (except PF, where the interpretation is the opposite to the rest of the variables), on average James' percentiles are a little higher than Bryant's. On the other hand, Foster's profile would correspond with Favors', whereas Blake's would correspond with Dooling's.

As regards the SiVM solution, there is a direct correspondence between Howard's profile and James' profile. Both James and Howard were selected both in the All-NBA First Team and in the NBA All-Defensive First Team. Howard has high percentiles in many variables, except in Three Pointers and AST, which are aspects where most centers don't usually highlight (especially some years ago), and his major flaw: the percentage of free throws: FT%. Nash

was the leader of assists per game and he was also another star of the league. He has upper-middle percentiles in many variables, very high in FT% and AST, but low percentiles in all the Rebound variables and Blocks, also expected because he was a point guard, and also in Steals and Personal Fouls. The other representative player, Cardinal, has low percentiles in many variables except in Three Pointers variables, FT%, Steals and Personal Fouls. The correspondence of Nash and Cardinal with Dooling and Favors is not so clear.

The alpha coefficients of each case are of great interest because they provide us with information about the feature composition according to the archetypoids, and also a ranking. Next, we are going to present the archetypoids and their similar players saying also their achievements in the season 2010-2011. When there are no major achievements in this season, but there are for other seasons, we will also mention them.

The five cases (without considering the respective archetypoid whose α value is 100%) with the highest α for archetypoid LeBron James (All-NBA First Team, All-Defensive First Team and All-Star Game) are in this order (this is a ranking based on archetypoid James):

- Kobe Bryant (All-NBA First Team, All-Defensive First Team and MVP All-Star Game).
- Kevin Durant (Scoring Leader, All-NBA First Team and All-Star Game).
- Russell Westbrook (All-NBA Second Team and All-Star Game).
- Dwyane Wade (All-NBA Second Team and All-Star Game).
- Carmelo Anthony (All-Star Game).

Players with the five highest α for archetypoid Dooling (besides him) are (this is a ranking based on archetypoid Dooling):

- Steve Blake.
- Eddie House (NBA Champion in 2007-08 with Boston Celtics).
- DeShawn Stevenson (NBA Champion with Dallas Mavericks).
- Jason Kidd (NBA Champion with Dallas Mavericks. He was rewarded with lots of individual awards and honors between seasons 1994-1995 and 2006-2007).
- Mario Chalmers (All-Rookie Second Team in the 2008-2009. NBA Champion in 2011-12, 2012-13 with Miami Heat).

The third archetypoid player, Favors, won the All-Rookie Second Team honors in 2010-2011. The highest-ranking players based on archetypoid Favors (himself not included) are: Aaron Gray, Omer Asik, Jeff Foster, Joey Dorsey and Joel Przybilla.

The All-NBA first team in 2010-2011 was formed by:

- Derrick Rose (All-NBA First Team, Season MVP and All-Star Game).
- Kobe Bryant.
- LeBron James.
- Kevin Durant.

– Dwight Howard (Defensive Player of the Year, All-NBA First Team, All-Defensive First Team and All-Star Game).

LeBron James' profile is obviously 100% explained by archetypoid James'. He is the best player of the league with a huge consensus. Derrick Rose profile is 82% explained by James' and 18% by Dooling's. This means that Rose is similar to James but he also has similarities with Dooling, which could explain some of his flaws. Kobe Bryant's profile matches 96% of James' and 4% of Dooling's, so this is reflecting the fact that Bryant is another super star, but probably he is not an all-around player like LeBron is. The same can be said for Durant since his profile is 93% formed by James' profile and 7% by Dooling's. Howard's profile is a mixture between 64% of James' and 36% of Favors'. In this case, Howard is also in the list of very good players but with some more limitations, which might cause his higher similarity to other not so good players such as Favors.

If we compute ADA with $k = 2$, to establish a unique ranking, the two resulting archetypoids would be Patrick Patterson and Jameer Nelson. We do not have an archetypoid with low percentiles in all the variables because there is no such player in our data set (note that the players in our sample had played in at least 30 games and averaged at least 10 minutes per game). Therefore, in this example a real 'worst-best' direction is not obtained since 'bad' players do not exist in our data set, therefore ADA cannot find the 'worst' player.

Patterson has more low-middle percentiles than Nelson (they complement each other), so if we select Nelson as the 'best', the ranking would begin with (alphas for archetypoid Nelson for the following players are one or nearly one): Steve Nash, Stephen Curry, Derrick Rose, Jameer Nelson, Raymond Felton, Jason Kidd, Monta Ellis, Deron Williams, Russell Westbrook, Chris Paul, Manu Ginobili, Kobe Bryant, Chauncey Billups and Kevin Martin. But this ranking should be considered with a great deal of caution since a 'worst-best' direction was not obtained.

As regards ranking performance, the ranking obtained by ordering the alphas corresponding to archetypoid LeBron James with $k = 3$ would be more realistic, as in that case we obtain a 'very good' player as an extreme, so we can consider that having an alpha equal to zero for that archetypoid means that that player is not a star. The first 20 players would be: LeBron James, Kobe Bryant, Kevin Durant, Russell Westbrook, Dwyane Wade, Carmelo Anthony, Kevin Martin, Derrick Rose, Amare Stoudemire, Dirk Nowitzki, Blake Griffin, Monta Ellis, Deron Williams, Danny Granger, Dwight Howard, Kevin Love, Manu Ginobili, Tony Parker, Andrea Bargnani and Eric Gordon. If we do not filter out the data set and all the NBA players in the 2010-11 season had been used (even if they had played less than 30 games and averaged less than 10 minutes per game) with the same variables except GP, then a 'worst-best' direction is found and the ranking for the first 20 players of a total of 536 would be (the alpha value for the first 13 players is 1, so all of them would be in position 1): Kevin Martin, Peja Stojakovic (when he was playing for Toronto Raptors), Kobe Bryant, Kevin Durant, Monta Ellis, Derrick

Rose, Manu Ginobili, Eric Gordon, Chauncey Billups, Danny Granger, Ray Allen, LeBron James, Wesley Matthews, Russell Westbrook, Deron Williams, Dwyane Wade, D.J. Augustin, Tony Parker, Joe Johnson, Andrea Bargnani and Stephen Curry.

Note that variables were standardized and all variables have the same weight for ADA computation and the corresponding ranking obtained. If it is considered that some of the variables are more important than others for determining performance, then those variables could be appropriately weighted before ADA computation.

5.1.1 Comparison with previous approaches

The evaluation of players can lead to the definition of new basketball positions, as was analyzed in Lutz (2012)). In this paper, the author uses a multivariate cluster analysis to group NBA players and to look at how different types of players may affect winning. In order to see the differences between the results obtained by Lutz (2012) and those obtained with ADA, we have applied ADA to the same data. In Lutz (2012), 10 clusters were determined, so we apply ADA for $k = 10$, which gives the following archetypoids: Joey Dorsey, Will Bynum, Jason Smith, Tayshaun Prince, Mickael Pietrus, Jason Kidd, Greivis Vasquez, Monta Ellis, Dwight Howard and James Jones. Table 2 shows the z-scores and percentiles for the 10 archetypoids. Note that there is a large difference with respect to the average z-score of each cluster in Lutz (2012), where non-extreme (very big or small) z-scores are found, in contrast to the ADA solution.

Table 2 Z-scores and percentiles for each archetypoid. In percentiles, the top 3 features are highlighted in bold and the values above or equal 90 are in a frame box.

Archetypoids	Measure	GP	Min	%Ast	AR	TOR	ORR	DRR	Rim	3-9	10-15	16-23	3s	Stls	Blks
Joey Dorsey	Z-score	-1.49	-1.53	-0.16	-0.26	2.45	3.04	1.94	-0.30	-0.97	-1.10	-1.41	-1.05	-0.4 5	-0.21
	Percentile	11	7	37	52	**98**	**100**	**95**	47	11	4	2	22	40	55
Will Bynum	Z-score	-0.26	-0.75	-2.20	1.24	0.48	-1.00	-1.38	0.23	-0.36	-0.20	-0.24	-0.59	0.28	-0.90
	Percentile	40	28	2	**87**	**74**	15	4	**67**	48	60	48	42	67	11
Jason Smith	Z-score	0.84	-1.27	1.35	-0.79	0.06	1.10	0.39	-1.09	-0.97	-0.65	0.45	-1.05	-1.00	-0.17
	Percentile	**72**	12	**93**	17	61	**82**	68	13	11	36	70	22	14	57
Tayshaun Prince	Z-score	0.91	1.02	-0.50	-0.08	-1.51	-0.49	-0.36	0.29	0.77	2.20	1.69	-0.31	-0.90	0.06
	Percentile	75	**80**	27	60	2	46	45	69	86	**96**	**94**	49	20	66
Mickael Pietrus	Z-score	-1.83	-0.79	1.39	-0.91	-1.19	-1.02	-0.60	-1.16	-0.97	-0.80	-0.37	1.18	-0.60	-0.06
	Percentile	5	27	**94**	10	8	11	36	10	11	23	43	**85**	33	**61**
Jason Kidd	Z-score	1.04	1.07	0.53	4.66	2.14	-0.97	-0.22	-1.22	-0.77	-0.65	-0.58	1.75	2.37	-0.29
	Percentile	85	81	69	**100**	96	16	50	8	23	36	35	**96**	**98**	50
Greivis Vasquez	Z-score	0.36	-1.52	-1.59	2.07	2.04	-0.95	-1.05	-1.22	-0.46	-0.65	-1.06	-0.36	-1.07	-0.92
	Percentile	**54**	8	9	**97**	**95**	18	13	8	42	36	16	47	12	9
Monta Ellis	Z-score	1.04	1.96	-1.18	0.17	-0.25	-1.02	-1.07	1.80	1.08	1.00	3.14	1.64	3.40	-0.44
	Percentile	85	**100**	16	70	45	11	11	94	88	87	**100**	94	**99**	43
Dwight Howard	Z-score	0.91	1.61	-0.43	-1.02	0.69	1.84	2.77	2.99	3.44	0.70	-0.99	-0.99	1.57	3.93
	Percentile	75	**96**	29	6	80	95	**99**	99	99	82	20	30	92	**99**
James Jones	Z-score	1.11	-0.68	2.27	-0.67	-1.95	-0.92	-0.83	-1.62	-1.18	-1.10	-0.93	0.95	-0.92	-0.5
	Percentile	**91**	31	**100**	28	0	20	26	0	1	4	21	**79**	18	38

The same data as in Lutz (2012) were used by Gruhl and Erosheva (2014), except the GP variable. Some of the posterior means for the five pure type mean parameters that they obtained have values outside the range of observed

data (for example, negative values or percentages higher than 100%), which makes them quite difficult to interpret. However, the ADA solution with $k = 5$ for the same data is easy to interpret, as it is composed of real extreme players. The archetypoids are: Carlos Delfino, Xavier Henry, Jameer Nelson, LaMarcus Aldridge and Omer Asik.

Regarding the application of AA to sports analytics, there was only a previous reference (Eugster, 2012). As it is the most closely related work, we have also analyzed the same data as Eugster (2012), which consists of 19 variables relating to 441 players from the 2009/2010 season of the NBA. Variables are standardized. The elbow criterion suggests, as in Eugster (2012), that 4 archetypoids should be chosen. The archetypoids are: Gerald Henderson, Mike Bibby, Marc Gasol and Dwyane Wade. The solution given by Eugster (2012), $cand_\alpha$, is the set formed by: Dwayne Jones, Taj Gibson, Anthony Morrow and Kevin Durant. The representative players according to SiVM are (in this order): Dwight Howard, Dwayne Jones, LeBron James and Taj Gibson. It was not possible to select 4 representatives for any regularization parameter with SMRS. With the regularization parameter equal to 2, the following solution is obtained: Dominic McGuire, Dwight Howard, Aaron Brooks, Jason Collins and Taylor Griffin. The percentiles of these players can be seen in Table 3, together with the RSS for each method. Note that the smallest RSS is obtained with archetypoids, even when the SMRS solution has more representatives. As in Sect. 5.1, variable distributions are mostly positively skewed. The disqualification variable is a clear example of this, as percentile 59 is, in fact, the minimum value for that variable (0).

Table 3 Percentiles for representative players with different methods for the 19 variables considered in the 2009/2010 NBA database by Eugster (2012). The top 3 features are highlighted in bold and the values above or equal 90 are in a frame box. The RSS for $k = 4$ (chosen according to the elbow criterion) are: 0.04265 (ADA), 0.06604 ($cand_\alpha$), 0.08956 (SiVM) and 0.0745 (SMRS with $k = 5$).

		GamesPlayed	TotalMinutesPlayed	FieldGoalsMade	FieldGoalsAttempted	ThreesMade	ThreesAttempted	FreeThrowsMade	FreeThrowsAttempted	OffensiveRebounds	TotalRebounds	Assists	Steals	Turnovers	Blocks	PersonalFouls	Disqualifications	TotalPoints	Technicals	GamesStarted
Bad players																				
ADA	**Henderson**	28	20	18	22	41	**42**	33	30	24	19	18	23	17	38	15	**59**	21	**45**	21
AA $cand_\alpha$	Jones	1	0	0	0	**29**	14	3	2	2	2	3	3	1	5	1	**59**	0	**45**	21
SiVM																				
SMRS	McGuire	**35**	18	12	14	29	14	3	9	33	26	16	14	20	23	17	**59**	10	**45**	33
	Collins	14	9	7	7	**29**	21	3	3	10	7	10	10	8	15	14	**59**	5	**45**	21
	Griffin	5	2	3	3	**29**	24	6	5	2	2	5	3	2	15	2	**59**	3	**45**	21
Very good players																				
ADA	**Wade**	77	93	99	99	80	86	99	99	81	79	98	**99**	**99**	93	78	59	**100**	98	90
AA $cand_\alpha$	Durant	100	100	100	**100**	94	95	100	**100**	80	93	85	96	**100**	94	74	83	**100**	85	100
SiVM	James	76	98	100	**100**	95	96	100	**100**	70	91	99	98	99	91	52	59	**100**	93	89
SMRS	-																			
Point guards																				
ADA	**Bibby**	86	76	68	71	**93**	**93**	50	46	31	51	91	80	64	19	68	59	69	85	**93**
AA $cand_\alpha$	Morrow	59	71	77	76	**96**	**92**	61	56	66	66	59	79	63	54	69	59	78	**78**	66
SiVM	-																			
SMRS	Brooks	**100**	97	96	98	**100**	**100**	90	89	62	59	97	81	98	51	85	83	97	98	100
Centers																				
ADA	**M.Gasol**	59	86	81	72	29	21	90	93	96	95	76	81	84	**97**	**98**	**96**	83	78	83
AA $cand_\alpha$	Gibson	**100**	76	74	70	29	14	66	71	97	93	49	67	75	96	**99**	**100**	70	85	83
SiVM																				
SMRS	Howard	100	95	94	83	29	36	99	100	100	**100**	71	84	**100**	**100**	100	98	96	100	100

According to the percentile information, the features of the players are as follows. Firstly, the archetypoids are described. Gerald Henderson has low values in all statistics, representing a less skillful basketball player considering the variables used in this analysis. It is well known that Henderson is the type of player who works behind the scenes and he is very good in the so-called game intangibles. However, with the variables available in the traditional box score he is appearing in the negative side of the players spectrum. At the other extreme, Dwyane Wade has high values in all statistics (except in disqualifications), representing a "very good" (star) player. Mike Bibby and Marc Gasol are also "good" players but with some weak points. Mike Bibby's main weak point is free throws. Some of his other weak points are rebounds and blocks, but this is logical because Bibby was a 188 cm point guard with good shooting percentages, especially shooting from three-point range. The weak points of Marc Gasol (who represents the typical features of a good center who plays mainly in the paint and grabs a lot of rebounds) are threes and a high level of disqualifications.

A correspondence between the archetypoid's profiles and those found in the $cand_\alpha$ set of archetypes can be also established. Dwayne Jones's profile would correspond with Gerald Henderson's profile, and analogously Kevin Durant's with Dwyane Wade's, Anthony Morrow's with Mike Bibby's, and Taj Gibson's with Marc Gasol's. However, note that some of Anthony Morrow's percentiles are near 50 (offensive rebounds, assists, blocks), whereas Bibby's percentiles for these same variables are more extreme. As a consequence, Morrow's features are not as extreme as Bibby's. The same happens with the free throws made and assists percentiles of Gibson and Marc Gasol. This is consistent with reality: Bibby was one of the best guards of the league during his career and Marc Gasol has been one of the best centers of the league since making his debut in the NBA, according to the majority opinion.

Regarding the players returned by SiVM, we could also establish a correspondence with the archetypoids' profiles, except Mike Bibby. Dwayne Jones' and Taj Gibson's profiles have been already commented. LeBron James's profile would correspond with that of Dwyane Wade. Dwight Howard's profile would correspond again with that of Marc Gasol. Note that the SiVM solution returns two players, Gibson and Howard, with two very similar profiles, unlike the ADA solution, which has no repeated profiles, and gives new information. This could be due to the greediness of SiVM, with immobile previous selections.

Information from players obtained by SMRS is redundant. Three players (Dominic McGuire, Jason Collins and Taylor Griffin) have similar profiles to that of Dwayne Jones. The other two players are "good" but with some weak points: Dwight Howard (already discussed) and Aaron Brooks. Aaron Brooks' weak points are rebounds and blocks, in the sense that his percentiles for these variables are near 50. His profile can be considered similar to Bibby's. No "very good" player is returned by SMRS.

The alpha coefficients of each case are interesting because they provide us with information about the feature composition according to the archetypoids.

Again, we describe all players saying their achievements in the season 2009-2010 or in other seasons when relevant.

Let us look at this for the three archetypoids corresponding with "good" players. The five cases (without considering the respective archetypoid whose α value is 100%) with the highest α for archetypoid Dwayne Wade (All-NBA First Team, All-Defensive Second Team and All-Star Game) are in this order (this is a ranking based on archetypoid Wade):

- LeBron James (All-NBA First Team, All-Defensive First Team and Season MVP and All-Star Game).
- Kevin Durant (Scoring Leader, All-NBA First Team and All-Star Game).
- Kobe Bryant (All-NBA First Team, All-Defensive First Team, Finals MVP, NBA Champion and All-Star Game).
- Carmelo Anthony (All-NBA Second Team and All-Star Game).
- Stephen Jackson (NBA Champion in 2002-03 with San Antonio Spurs).

The players with the five highest α for archetypoid Mike Bibby (All-Rookie First Team in 1998-99) (besides him) are (this is a ranking based on archetypoid Bibby): Quentin Richardson, Carlos Delfino, Steve Blake, Rasual Butler and Rashard Lewis (NBA Champion in 2012-13 with Miami Heat and All-Star Game in 2004-2005 and 2008-2009).

For archetypoid Marc Gasol (All-Rookie Second Team in the 2008-2009 and several awards and honors from the season 2012-2013) (besides him) they are (this is a ranking based on archetypoid Marc Gasol):

- Andrew Bogut (All-NBA Third Team in the season 2009-2010. In addition, All-Rookie First Team in 2005-2006 and NBA champion in 2014-15 with Golden State Warriors among other awards).
- Taj Gibson (All-Rookie First Team).
- Samuel Dalembert.
- Jason Thompson.
- Nene Hilario (All-Rookie First Team in 2002-2003).

Marc Gasol got the NBA Defensive Player of the Year award in the season 2012/2013 and also belonged to the NBA-All-Defensive Team. In addition, Andrew Bogut belonged to the NBA-All-Defensive Team in the season 2014/2015. Therefore, the list of players similar to the archetypoid Marc Gasol can be considered as a set of defensive specialists.

The All-NBA first team in 2009-2010 was formed by Dwyane Wade, Kobe Bryant, LeBron James, Kevin Durant and Dwight Howard (Defensive Player of the Year, Rebounds Leader, Blocks Leader, All-NBA First Team, All-Defensive First Team and All-Star Game).

LeBron James (and obviously Dwayne Wade) are 100% explained by archetypoid Dwayne Wade. Both Wade and especially James have been NBA stars for the last seasons. The same can be said for Kevin Durant (his profile is 99% formed by Dwayne Wade's profile and 1% by Marc Gasol's). Kobe Bryant's profile is constituted 83% by Dwayne Wade's and 17% by Mike Bibby's, so Bryant is very similar to Wade as well with other features more related to

Bibby's. Dwight Howard profile is a mixture between 59% Marc Gasol's and 41% Dwayne Wade's. This means that he is especially similar to other of the outstanding centers, such as Marc Gasol.

Although the model with $k = 4$ is recommended (a higher k does not reduce the RSS very much), archetypoids can be computed for other k values. As previously discussed, archetypoids are not necessarily nested, but in this problem the archetypoids obtained for higher values of k give even more details to the previous ones obtained with $k = 4$, for example for $k = 5$ or $k = 7$.

With $k = 5$, the archetypoids are: Dan Gadzuric (with a similar profile to Gerald Henderson's), Mike Bibby (he appeared in the $k = 4$ solution), LeBron James (with a similar profile to Dwayne Wade's), Paul Millsap (with a similar profile to Marc Gasol's) and Pau Gasol (with a similar profile to Paul Millsap's, although with no disqualification). Note that Pau Gasol won the NBA title with the Lakers in that season and was a member of the All-NBA Third Team. Regarding, Paul Millsap, he was a member of the All-Rookie Second Team in 2006/2007, of the All-Defensive Second Team in 2015-1016, and four times All-Star Game between 2013-2014 and 2016-2017.

With $k = 7$, the archetypoids are: Mario West (with a similar profile to Gerald Henderson's), Dwyane Wade (he appeared in the $k = 4$ solution), Jose Calderon and Anthony Morrow, which give more details to Mike Bibby's archetypoid profile, and Taj Gibson, Elton Brand (Rookie of the Year in the season 1999-2000, some awards until the season 2005-2006) and Dwight Howard, which give more details to Marc Gasol's archetypoid profile.

5.2 Player career trajectory analysis with ADA+FDA

FADA is applied to the problem of finding archetypoid basketball players based on their Game Score (GmSc) over time ($x_i(t)$ represents GmSc of player i for a certain age t). GmSc is a measure of a player's productivity for a single game. The scale is similar to that of points scored, (40 is an outstanding performance, 10 is an average performance, etc.). The exact formula of GmSc can be found in the Glossary of basketball (2016).

Our database now contains the NBA players and their statistics for each game, including their age (year and day) when they played each game and their GmSc for that game, from the 2005-2006 season to the 2014-2015 season. There are 247577 rows in total [3]. In order to clean the data, we removed the entries where the players played for less than 5 minutes, and we also removed the entries where players were below 19 years old (due to the NBA's age restriction) or over 40 years old (we did not remove the players, it is simply that the age range considered is from 19 to 40 years old). Note that only 8 players in our database have played in their forties, which is a very small sample size. This may return biased results, since the players who play in their forties, are usually very good players, who are requested to extend their

[3] All the data was downloaded from:
www.basketball-reference.com/play-index/pgl_finder.cgi?lid=front_pi

careers. Therefore, they have a high GmSc, but this score is not representative of their age. Note that each player is measured at an irregular and sparse set of time points which differ widely across subjects. Players with only one measurement are also excluded. Finally, the data set contains 1071 players, with 231803 entries in total: the player with least entries has only 4 games recorded, whereas 1546 games are recorded for the player with most entries.

Following Peng and Paul (2009), we have chosen to use $m = 4$ eigenfunctions with $M = 5$ basis functions for representing the eigenfunctions, as it gives the smallest approximate CV score. The estimated eigenvalues are: $3.127180e+02$, $1.989077e+02$, $8.946427e+01$ and $3.68e-13$. Scores are also estimated. Note that the fourth eigenvalue, eigenfunction and their respective scores are almost negligible. Figures 3 and 4 show the mean function and the first four eigenfunctions. Although we do not present functional variance in the plots, it can be computed using Equation 6 in Peng and Paul (2009), which gives the projected covariance kernel and which is implemented in the accompanying software.

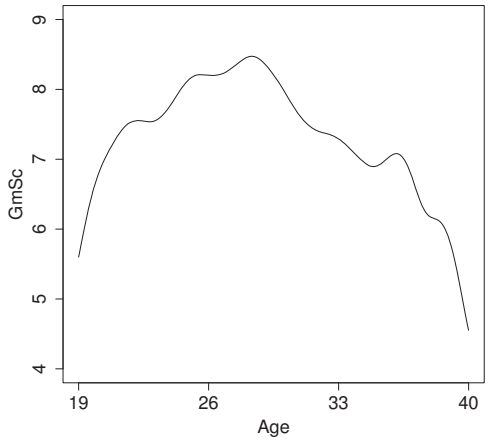

Fig. 3 Estimated mean function for GmSc data.

ADA is applied to the matrix of functional principal component scores. The elbow criterion suggests that 4 archetypoids should be chosen (the number of vertices of the convex hull is 26). The 4 archetypoids are: Lance Stephenson, LeBron James, Danny Granger and Stephen Jackson. Figure 5 displays the GmSc observed for each archetypoid together with a smooth curve using local fitting (the function *loess* from the R package `stats` (Cleveland et al., 1992)) only to aid interpretation.

The trajectories of each archetypoid are very different. Lance Stephenson is a replacement level player. He does not have high GmSc values; in fact, at

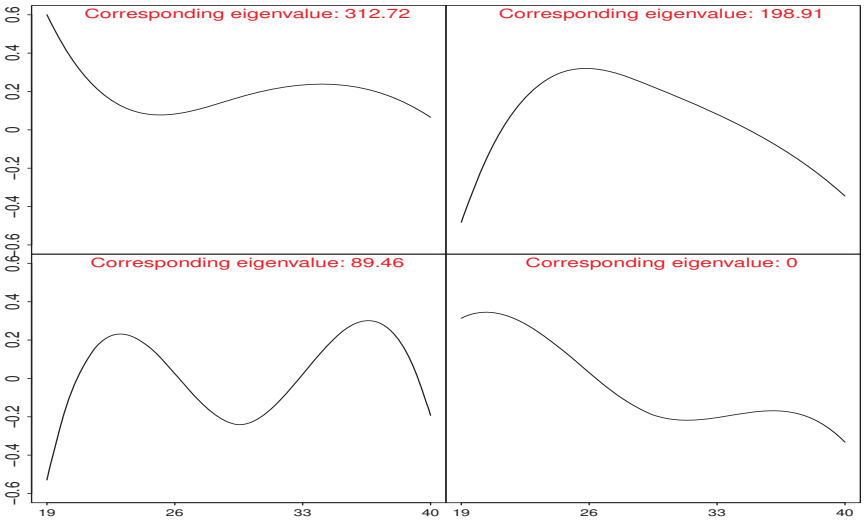

Fig. 4 Estimates of the first four principal components for GmSc data, from left to right.

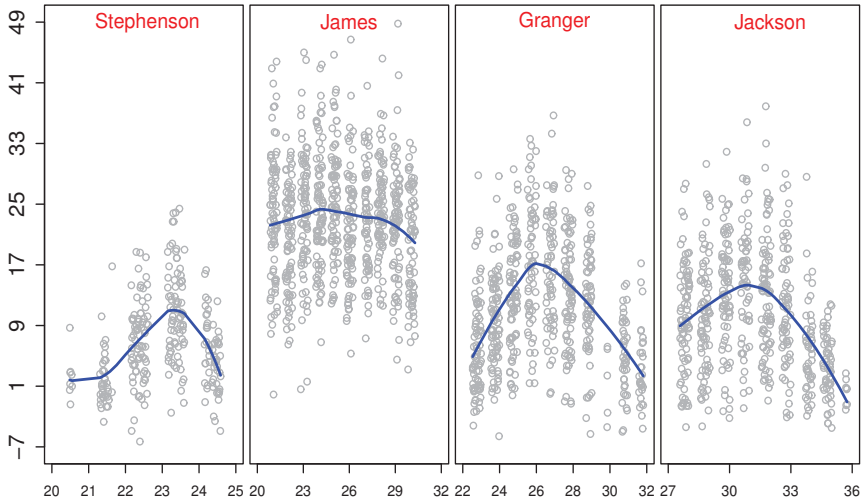

Fig. 5 GmSc observed (circles) for each functional archetypoid and a *loess* regression smoother (solid line).

age of 20 or 21 the values are below 5. These values increased to a peak (a little above 10) at the age of 23, then they decreased. LeBron James is the archetypoid who represents the NBA stars; in particular, he is a consistently strong performer. His GmSc is high (above 20) over the years. Danny Granger and Stephen Jackson's GmSc curves also have a mountain form like Stephenson, but with higher GmSc values, and the peak (around 15) is at other ages.

Granger only performed well early in his career (before major injuries) and Jackson is something of a late bloomer, starting his decline at around 32. The 10 players with the highest alphas corresponding with the archetypoid LeBron James, i.e. those with very high GmSc throughout their careers are, in this order (this is a ranking based on archetypoid James): LeBron James, Kevin Durant, Anthony Davis, Kobe Bryant, Chris Paul, Dirk Nowitzki, Carmelo Anthony, Blake Griffin, Dwight Howard and Chris Bosh. As a point of interest, Stephen Curry, who was named the 2015 NBA Most Valuable Player, appears in the 19th position. His GmSc curve is explained 58% by LeBron James archetypoid, 22% by Danny Granger archetypoid and 20% by Stephen Jackson archetypoid. A plot with his observed GmSc can be seen in Figure 6. Note that up to the age of 24, his GmSc values were below 15, then they began to increase to 20 at the age of 25.

The solution $cand_\alpha$ coincides with that of ADA. If the matrix of functional principal component scores is also used, the first four SiVM representative players are LeBron James, Greivis Vasquez, Stephen Jackson and Brandon Roy. Note that James and Jackson also appeared in the ADA solution. Roy has a similar profile as Granger's. Vasquez also peaks around the age of 26 like Granger, but his height is smaller (with around 10 GmSc). Stephenson's profile does not appear in the SiVM solution. For SMRS (with the regularization parameter equals to 0.5) the representative players are: Greivis Vasquez, Andris Biedrins, Stephen Jackson and LeBron James. Now Biedrins has a profile similar to Stephenson's. As the sample size of this data set is larger than in the previous examples, the computational times are greater. They can be seen together with the RSS in Table 4.

Table 4 RSS for the player career trajectory analysis with ADA+FDA, with $k = 4$. The computational times are for: AA (from 1 to 10) 137 sec.; ADA with $k = 4$, 245 sec. (beginning from $cand_{ns}$), 128 sec. (beginning from $cand_\alpha$), 125 sec. (beginning from $cand_\beta$); SiVM $\ll$ 0.1 sec.; SMRS, 6 minutes (for this regularization parameter, but several regularization parameters were tested).

Method	ADA and $cand_\alpha$	SiVM	SMRS (0.5 regularization parameter)
RSS	0.08874	0.24421	0.08486

If we compute ADA with $k = 2$, to establish a unique performance ranking, the two resulting archetypoids are Darius Songaila and Kobe Bryant. In this example, a 'worst-best' direction is found, since Bryant has very high GmSc values, in contrast to Songaila. The first five players would be: LeBron James, Kobe Bryant, Kevin Durant, Anthony Davis and Chris Paul. This is a good result because these players are some of the best players of the league and actually, we could build a starting line-up with them: Paul as the point guard, Bryant as the shooting guard, James as the small forward, Durant as the power forward and David as the center. It would be a very strong team.

Table 5 describes the players in terms of the awards and titles obtained by them. All players have got a lot of individual awards. They have been

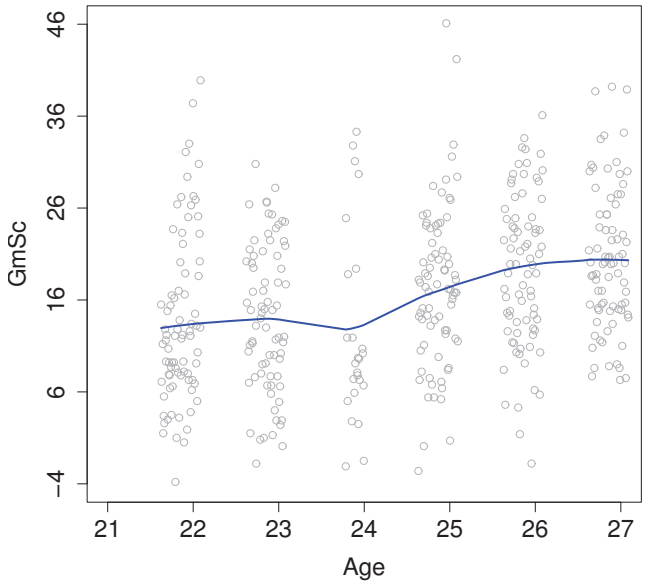

Fig. 6 GmSc observed (circles) and a *loess* regression smoother (solid line) for Stephen Curry. This figure reflects the Curry's archetypoid composition: a great percentage of a strong performer (Lebron), combined with a late bloomer profile.

very successful from their rookie season. From the season 2005-2006, LeBron James is the player who has won more NBA titles among them and also who has been chosen more times in the All-NBA First Team, as the league MVP and to participate in the All-Star Game. Kobe Bryant has a similar number of awards. It is worth noting that he was neither the rookie of the year nor a member of the All-Rookie First Team. However, he has been of the best players in history. Like LeBron and Kobe, Chris Paul is another an all-around player because he has also been selected both in the All-NBA First Team and in the All-NBA Defensive Team a lot of times. Kevin Durant is mainly a scorer, one of the greatest nowadays. Anthony Davis is currently one of the best players of the league and he is still very young.

5.3 Team performance analysis with ADA+h-plot

Let us show the procedure with the results from the 2014-2015 Spanish football league (20 teams in total). The data consist of a table with the pairwise results [4]. Teams do not usually perform identically at home and away, so each team will have a home and visiting profile. The team profile at home (and similarly

[4] Obtained from www.linguasport.com/futbol/nacional/liga/Liga_15.htm

Table 5 Awards and titles obtained by the players from the 2005-2006 season to the 2014-2015 season. Values in parentheses refer to the year they entered the league.

Players	Rookie season	All-NBA Team		All-NBA Defensive Team		Titles	NBA MVP	All-Star Game
		First team	Second team	First team	Second team			
LeBron James (03-04)	Rookie of the Year All-Rookie First Team	9	1	5	1	3	4	11
Kobe Bryant (96-97)	All-Rookie Second Team	8	0	6	1	2	1	10
Kevin Durant (07-08)	Rookie of the Year All-Rookie First Team	5	0	0	0	0	1	6
Anthony Davis (12-13)	All-Rookie First Team	1	0	0	1	0	0	2
Chris Paul (05-06)	Rookie of the Year All-Rookie First Team	4	2	5	2	0	0	8

as a visitor) is an ordinal vector that compiles the game results with all other teams, where -1 means that the team lost the match, 0 if it drew, and 1 if it won (0 is imputed in a hypothetical match with itself). In other words, this vector is a 20-dimensional vector of an ordered categorical variable. We have computed the dissimilarities between the profiles of each team, both at home and away. For each of the 20 teams, the Gower's coefficient (Gower, 1971) is computed as implemented in the R package `cluster` (Maechler et al., 2015), both for the home and visitor profile, returning a 40×40 dissimilarity matrix $\mathbf{D}$.

Figure 7 displays the h-plot representation for this data and Table 6 shows the team codes. The goodness-of-fit is 93%, which is good. If two team profiles are similar, they will be represented near each other in the 2D h-plot. For a specific team, the greater the distance between its home and visitor profiles, the more different its behavior is at home and as a visitor (it is more asymmetric). The first dimension (88% of the fit) is related to the number of wins. The teams that achieved most points are located on the left of the panel, whereas the teams in the lowest positions appear on the right. The second dimension refers to a pattern of wins that is different from other teams. The most remarkably opposite profiles in this dimension are the home profiles of RSC and RAY. RSC at home was a strong rival for the best teams and a weaker one when playing against those at the bottom of the classification. On the other hand, RAY behaved as "expected" at home, in the sense that it did not defeat any of the top five teams, but it defeated three of the last six in the ranking. Regarding teams with a different behavior at home and away, the most asymmetric teams were, in this order: RSC, MAL, GRA and VAL. On the contrary, the most symmetric teams, with a similar profile both at home and away, were, in this order: ELC, COR, ALM and BAR.

Let us now obtain the archetypoid teams. Each team has two profiles, but both are represented in the same configuration. Therefore, we apply ADA to a 20×4 matrix $\mathbf{X}$ made up of the combination of the representation of the two (home and away) h-plot profiles. Incidentally, in this small example with only 4 variables, 18 teams are vertices of the convex hull generated from the 20 teams. An elbow appears at 5 archetypoids, corresponding to RAY, MAL,

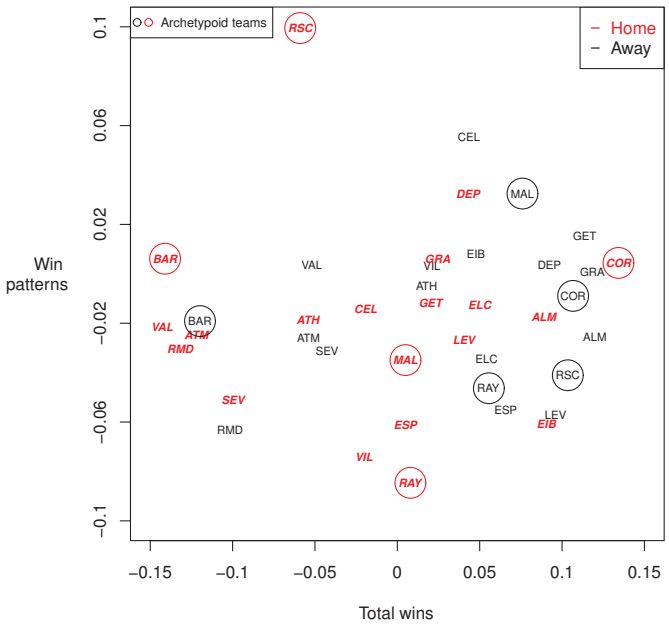

Fig. 7 H-plot representation of the 20 teams in the 2014-2015 Spanish football league. The home (visiting) profiles are in red (black). The names of the axes are based on intuition about the meaning of the dimensions 1 and 2.

Table 6 Teams in the 2014-2015 Spanish football league, with their abbreviations in brackets.

Almería (ALM)	Ath. Bilbao (ATH)	At. Madrid (ATM)	Barcelona (BAR)
Celta (CEL)	Córdoba (COR)	Deportivo (DEP)	Eibar (EIB)
Elche (ELC)	Espanyol (ESP)	Getafe (GET)	Granada (GRA)
Levante (LEV)	Málaga (MAL)	Rayo (RAY)	R. Madrid (RMD)
R. Sociedad (RSC)	Sevilla (SEV)	Valencia (VAL)	Villarreal (VIL)

BAR, COR and RSC. BAR was champion of the league and COR came last in the ranking. RSC, MAL and RAY were in the middle of the classification table, but their behavior was different. RSC and MAL were the most asymmetric teams, and the RAY at home profile was the opposite in dimension 2 to RSC's. As a visitor RAY was not able to defeat any of the top ranked teams; in fact, it only defeated teams which were classified below it (from 12th place to last). The alpha values tell us the contribution of each archetypoid to each team. In Figure 8, the alpha values for each archetypoid are displayed with a star plot. For each case, the 5 alpha values in this example are represented starting on the right and going counter-clockwise around the circle. The size of each alpha is shown by the radius of the segment representing it. The teams which are

similar to the archetypoids can be clearly seen (for example, ESP is similar to RAY, CEL to MAL, RMD to BAR, ALM to COR), as can the teams which are a mixture of several archetypoids (for example, ATH).

Results with $cand_\alpha$ (the solution given by Eugster (2012)), SiVM and SMRS (with the regularization parameter equal to 40) are similar to the archetypoids obtained: RAY, CEL, BAR, COR and RSC. CEL appears instead of MAL. However, the RSS with archetypoids is less than the RSS for the other methods. They can be seen together with the runtimes in Table 7.

Table 7 RSS for the team performance analysis with ADA+h-plot, with $k = 5$. The computational times are: AA (from 1 to 10), 2 sec.; ADA with $k = 5$, SiVM and SMRS return results instantaneously.

Method	ADA	$cand_\alpha$, SiVM and SMRS (40 regularization parameter)
RSS	0.0041	0.0044

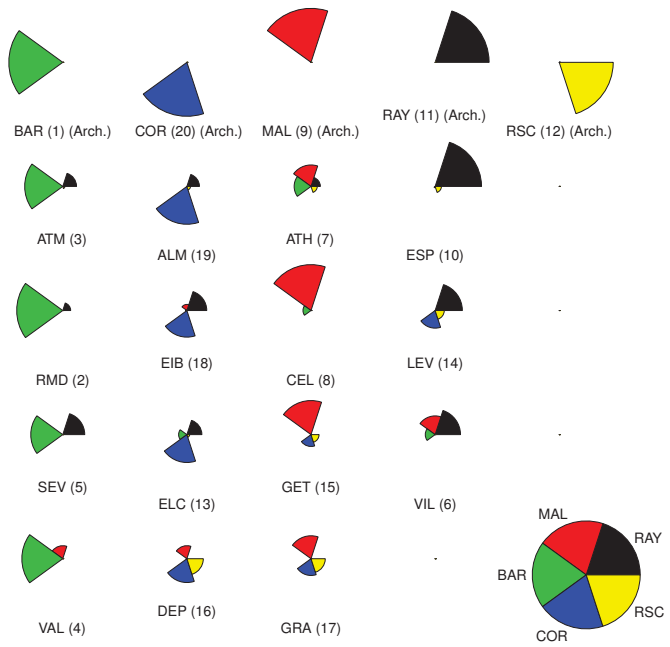

Fig. 8 Star plot of alpha values for the 5 archetypoid teams (RAY in black, MAL in red, BAR in green, COR in blue and RSC in yellow) in the 2014-2015 Spanish football league. The final league classification appears in brackets.

The team ranking obtained using ADA with $k = 2$ is quite similar to the final classification (the only remarkable difference is that RSC and MAL change positions from 9 to 12). In this ranking, several teams have the same alpha as the best team (in brackets): BAR (1), RMD (1) , VAL (1) , ATM (1), SEV (0.92) , ATH (0.64), VIL (0.54), CEL (0.48), RSC (0.45), RAY (0.39), ESP (0.37), MAL (0.34), ELC (0.26), GET (0.21), LEV (0.19), DEP (0.19), GRA (0.18), EIB (0.17), ALM (0), COR (0). For a more discriminative ranking without draws between teams in the top positions, alphas from two archetypes could be used, which gives the same ranking except that the positions of GET and LEV are interchanged.

6 Conclusions

One of the most hotly debated issues in any sport is that of who can be considered the most valuable player (or team in a league). ADA can be used to explore and discuss this question that is of interest to coaches, scouts and fans. Unlike AA, which was used by Eugster (2012), the archetypoid algorithm always identifies a number of real extreme subjects, thereby facilitating their analysis. This new statistical approach is a simple and useful way of looking at sports data. Furthermore, it is a data-driven approach. The rationale for using ADA is to overcome the limitations of using subjective observation alone and to achieve a greater understanding of performance. Another important contribution is the possibility of working with ADA when dissimilarities are available rather than features (even when they are asymmetric). In addition, we have shown how to compute archetypoids with sparse functional data. In particular, to the best of our knowledge, this is the second attempt to use FDA with sports data. Results in all cases are quite intuitive and consistent with the general opinion held by "classical" sports analysts. This study shows how ADA can be a useful mathematical tool to analyze sporting performance and to assess the value of players and teams in a league. This approach is not a definitive measure of sports value, but it provides some interesting indicators, which can be valuable for making educated decisions about trades or strategy.

Future work Although in the multivariate case the statistics of only one season are used, following the examples in Eugster (2012), statistics of more seasons could easily be used at the same time by simply combining the statistics from different seasons by columns. In case of missing values, the objective function could be modified analogously as done by Mørup and Hansen (2012) for AA. Moreover, ADA could also be adapted to deal with weighted observations or outliers, as Eugster and Leisch (2011) did with AA. With the recent developments in data collecting, such as the spatial-tracking data gathering, the traditional box score is being expanded with new features. We aim to use our methodology with them to discover new player patterns.

In the example about sparse longitudinal data, only one function per player was considered. However, the extension for dealing with more than one func-

tion at the same time is immediate if bases are orthonormal (see Epifanio (2016)). Multivariate ADA could be applied to the matrix composed by joining the functional principal component scores for each function. For future work, ADA could be used with players' trajectories in the field, in a similar way as Feld et al. (2015) did for routes in buildings. Furthermore, ADA could be extended to mixed data, with functional and vector parts. We have also used functional principal component scores for obtaining the functional archetypoids. Another interesting problem would be to predict the entire functions based on the estimated scores, as explained in Section 5.2, which would predict careers.

As regards interpretation of the results, awards and titles have been mainly used, but other features could be considered such as player salaries as in Schulte et al. (2015).

Acknowledgements The authors would like to thank the Editors and three reviewers for their very constructive suggestions, which have led to improvements in the manuscript.

References

Bauckhage C, Thurau C (2009) Making archetypal analysis practical. In: Denzler J., Notni G., Süsse H. (eds) Pattern Recognition. 31st annual pattern recognition symposium of the German Association for Pattern Recognition, 2009. Lecture Notes in Computer Science, vol 5748. Springer, Berlin, Heidelberg, Germany, 272–281

Bhandari I, Colet E, Parker J, Pines Z, Pratap R, Ramanujam K (1997) Advanced scout: Data mining and knowledge discovery in NBA data. Data Mining and Knowledge Discovery 1(1):121–125

Canhasi E, Kononenko I (2013) Multi-document summarization via archetypal analysis of the content-graph joint model. Knowledge and Information Systems, 1–22,

Canhasi E, Kononenko I (2014) Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. Expert Systems with Applications 41(2):535–543

Chan B, Mitchell D, Cram L (2003) Archetypal analysis of galaxy spectra. Monthly Notices of the Royal Astronomical Society 338:1–6

Cleveland W, Grosse E, Shyu W (1992) Statistical models in S, Wadsworth & Brooks/Cole, chap Local regression

Cutler A, Breiman L (1994) Archetypal analysis. Technometrics 36(4):338–347

Davis T, Love B (2010) Memory for category information is idealized through contrast with competing options. Psychological Science 21(2):234–242

D'Esposito M R, Palumbo F, Ragozini G (2012) Interval archetypes: A new tool for interval data analysis. Statistical Analysis and Data Mining 5(4):322–335

D'Esposito M R, Ragozini G (2008) A new R-ordering procedure to rank multivariate performances. Quaderni di Statistica 10:5–21

Donoghue O, Harrison A, Coffey N, Hayes K (2008) Functional data analysis of running kinematics in chronic Achilles tendon injury. Medicine and Science in Sports and Exercise 40(7):1323–1335

Elhamifar E, Sapiro G, Vidal R (2012) See all by looking at a few: Sparse modeling for finding representative objects. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1–8

Epifanio I (2013) H-plots for displaying nonmetric dissimilarity matrices. Statistical Analysis and Data Mining 6(2):136–143

Epifanio I (2014) Mapping the asymmetrical citation relationships between journals by h-plots. Journal of the Association for Information Science and Technology 65(6):1293–1298

Epifanio I (2016) Functional archetype and archetypoid analysis. Computational Statistics & Data Analysis 104:24–34

Epifanio I, Ávila C, Page Á, Atienza C (2008) Analysis of multiple waveforms by means of functional principal component analysis: normal versus pathological patterns in sit-to-stand movement. Medical & Biological Engineering & Computing 46(6):551–561

Epifanio I, Vinué G, Alemany S (2013) Archetypal analysis: Contributions for estimating boundary cases in multivariate accommodation problem. Computers & Industrial Engineering 64:757–765

Eugster M (2012) Performance profiles based on archetypal athletes. International Journal of Performance Analysis in Sport 12(1):166–187

Eugster M, Leisch F (2009) From Spider-Man to hero - Archetypal analysis in R. Journal of Statistical Software 30(8):1–23

Eugster, M, Leisch, F (2011). Weighted and robust archetypal analysis. Computational Statistics & Data Analysis 55(3):1215–1225.

Feld S, Werner M, Schönfeld M, Hasler S (2015) Archetypes of alternative routes in buildings. In: Proceedings of the 6th International Conference on Indoor Positioning and Indoor Navigation (IPIN), 1–10

Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315:972–976

Glossary of basketball (2016) `http://www.basketball-reference.com/about/glossary.html`

Gower J (1971) A general coefficient of similarity and some of its properties. Biometrics 27(4):857–871

Gruhl J, Erosheva EA (2014) A Tale of Two (Types of) Memberships. In: Handbook on Mixed-Membership Models, Chapman & Hall/CRC, 15–38

Harrison A (2014) Applications of functional data analysis in sports biomechanics. In: 32 International Conference of Biomechanics in Sports, 1–9

Harrison A, Ryan W, Hayes K (2007) Functional data analysis of joint coordination in the development of vertical jump performance. Sports Biomechanics 6(2):199–214

Hoopdata - NBA Statistics and Analysis (2009-2013). Retrieved from http://www.hoopdata.com/regstats.aspx

James G (2010) The Oxford handbook of functional data analysis, Oxford University Press, chap Sparse Functional Data Analysis

James G, Hastie T, Sugar C (2000) Principal component models for sparse functional data. Biometrika 87(3):587–602

Kaufman, L, Rousseeuw, P J, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley, New York

Kersting K, Bauckhage C, Thurau C, Wahabzada M, (2012) Matrix Factorization as Search. In: Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases, Bristol, UK, 850–853

Krein, M, Milman, D (1940) On extreme points of regular convex sets, Studia Mathematica 9:133-138

Kubatko J, Oliver D, Pelton K, Rosenbaum D (2007) A starting point for analyzing basketball statistics. Journal of Quantitative Analysis in Sports 3(3):1–10

Levitin D, Nuzzo R, Vines B, Ramsay J (2007) Introduction to functional data analysis. Canadian Psychology 48(3):135–155

Li S, Wang P, Louviere J, Carson R (2003) Archetypal analysis: A new way to segment markets based on extreme individuals. In: ANZMAC 2003, Conference Proceedings, Australia and New Zealand Marketing Academy Conference (ANZMAC), Adelaide, Australia, 1674–1679

Lutz D (2012) A cluster analysis of NBA players. In: MIT Sloan Sports Analytics Conference, MIT, Boston, USA, 1–10

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2015) `cluster`: Cluster analysis basics and extensions. R package version 2.0.1 — For new features, see the 'Changelog' file (in the package source)

Midgley D, Venaik S (2013) Marketing strategy in MNC subsidiaries: Pure versus hybrid archetypes. In: Proceedings of the 55th Annual Meeting of the Academy of International Business, AIB, Istanbul, Turkey, 215–216

Mohamed S, Heller K, Ghahramani Z (2014) A simple and general exponential family framework for partial membership and factor analysis. In: Handbook on Mixed-Membership Models, Chapman & Hall/CRC, 67–88

Mørup M, Hansen L (2012) Archetypal analysis for machine learning and data mining. Neurocomputing 80:54–63

O'Donoghue P (2010) Research methods for sports performance analysis. Routledge, Taylor & Francis Group, New York, NY

Peng J, Paul D (2009) A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. Journal of Computational and Graphical Statistics 18(4):995–1015

Peng J, Paul D (2011) `fpca`: Restricted MLE for functional principal components analysis. `https://CRAN.R-project.org/package=fpca`, R package version 0.2-1

Porzio G, Ragozini G, Vistocco D (2008) On the use of archetypes as benchmarks. Applied Stochastic Models in Business and Industry 24:419–437

R Development Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org/`, ISBN 3-900051-07-0

Ragozini, G, Palumbo, F, D'Esposito, MR (2017) Archetypal analysis for data-driven prototype identification. Statistical Analysis and Data Mining: The

ASA Data Science Journal 10(1):6–20
Ramsay J, Silverman B (2002) Applied functional data analysis. Springer
Ramsay J, Silverman B (2005) Functional data analysis, 2nd edn. Springer
Schulte, O, Zhao, Z Routley, K (2015) What is the Value of an Action in Ice Hockey? Learning a Q-function for the NHL. In: MLSA 2015: Machine Learning and Data Mining for Sports Analytics (MLSA 15), 1–10
Seiler C, Wohlrabe K (2013) Archetypal scientists. Journal of Informetrics 7:345–356
Shea S (2014) Basketball analytics: Spatial tracking. Createspace, Lake St. Louis, MO
Shea S, Baker C (2013) Basketball analytics: Objective and efficient strategies for understanding how teams win. Advanced Metrics, LLC, Lake St. Louis, MO
Theodosiou T, Kazanidis I, Valsamidis S, Kontogiannis S (2013) Courseware usage archetyping. In: Proceedings of the 17th Panhellenic Conference on Informatics, ACM, New York, NY, USA, PCI '13, 243–249
Thurau C, Kersting K, Wahabzada M, Bauckhage C (2012) Descriptive matrix factorization for sustainability adopting the principle of opposites. Data Mining and Knowledge Discovery 24(2):325–354
Ullah S, Finch C (2013) Applications of functional data analysis: A systematic review. BMC Medical Research Methodology 13(43):1–12
Vinué G (2014) Development of statistical methodologies applied to anthropometric data oriented towards the ergonomic design of products. PhD thesis, Faculty of Mathematics. University of Valencia, Spain, `http://hdl.handle.net/10550/35907`
Vinué G, Epifanio I, Alemany S (2015) Archetypoids: A new approach to define representative archetypal data. Computational Statistics and Data Analysis 87:102–115
Vinué G (2017) `Anthropometry`: An R package for analysis of anthropometric data. Journal of Statistical Software 77(6):1–39
Vinué G, Epifanio I, Simó A, Ibáñez M, Domingo J, Ayala G (2017) `Anthropometry`: An R package for analysis of anthropometric data. `https://CRAN.R-project.org/package=Anthropometry`, R package version 1.8
Wakim A, Jin J (2014) Functional data analysis of aging curves in sports, `http://arxiv.org/abs/1403.7548`
Williams C, Wragg C (2004) Data analysis and research for sport and exercise science. Routledge, Taylor & Francis Group, New York, NY
Winston W (2009) Mathletics : How gamblers, managers, and sports enthusiasts use mathematics in baseball, basketball, and football. Princeton University Press, Princeton, New Jersey
Yao F, Müller H-G, Wang, JL (2005) Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association 100(470):577–590