Forecasting basketball players’ performance using sparse functional data

G. Vinué(1), I. Epifanio(2)

(1) Department of Statistics and O.R., University of Valencia, 46100 Burjassot, Spain.
(2) Dept. Matemátiqes and Institut de Matemátiqes i Aplicacions de Castelló. Campus del Riu Sec. Universitat Jaume I, 12071 Castelló, Spain.

Abstract

There is an increasing importance of statistics and analytic methods in basketball. In particular, predicting players’ performance using past observations is a big challenge. The purpose of this study is to forecast the future behavior of basketball players. The available data are sparse functional data, which are very common in sports. So far, however, no forecasting method designed for sparse functional data has been used in sports. A methodology based on two methods to handle sparse and irregular data, together with the analogous method and the functional archetypoid analysis is proposed. Results in comparison with traditional methods show that our approach is very competitive and additionally provides prediction intervals. The methodology can also find use in other sports when sparse longitudinal data are available.

Keywords: Forecasting, Functional data analysis, Archetypal analysis, Functional sparse data, Basketball

1. Introduction

Sports forecasting is a fast-growing field. Proof of this is the increasing number of forecasting methods developed covering several sports and sportive topics (Stekler, H. O., Vaughan Williams, L. (Editors), 2010). Certainly, in professional sports, getting any notion in advance about the future performance of players can have important consequences on the roster composition.
in terms of renewing or buying players, or in terms of establishing levels of remuneration in function of that performance, for instance. Furthermore, countries invest economical and structural resources to train athletes, who are able to win medals in the Olympic Games. This is the reason why there have been a number of studies to find out which socio-economic variables are reliable predictors of how well a country will do in the Olympics. It is shown in Bernard and Busse (2004) that the country’s gross domestic product (GDP), the type of country’s political system or the condition of hosting country are strong predictors of the number of medals. In Forrest et al. (2010) other variables are founded reliable, especially the public spending on recreation.

Over the last decade, statistics in basketball has developed in a very positive way. Basketball analytics started to attract attention with the publications of Oliver (2004) and Hollinger (2005). In further years, other papers and books have been released (Kubatko et al., 2007; Shea and Baker, 2013b; Shea, 2014). Technological advances have made it possible to collect more data than ever about what is happening on the field, requiring new methods of analysis. There is currently a need for innovative methods that exploit the full potential of the data and that allow to generate additional value for athletes and technical staffs. One of the main challenges in basketball analytics is to use past performance to predict future performance (Shea, 2014). To address this open question, some forecasting methods have been developed. Following (Hyndman and Athanasopoulos, 2013, Chapter 1.4), two main approaches can be distinguished based on the type of data used: On time-series forecasting and cross-sectional forecasting. On the one hand, forecasting using data collected over time describes the likely outcome of the time series in the immediate future, having knowledge of the recent outcomes. On the other hand, cross-sectional forecasting methods use data collected at a single point in time. The goal here is to predict a target variable using some explanatory variables which are related to it.

Two well elaborated methods can be found using historical time data: College Prospect Rating (CPR) is a score assigned to college basketball players that approximates their NBA potential (Shea, 2014; Shea and Baker, 2013a). This methodology starts by averaging the player’s top ten game performances in a number of chosen stats (with the exception of three points percentage and free throws percentage, which are averaged using all the games available to take into account the shooting efficiency). These stats are previously standardized. Then, the values are summed to get what is called the
raw total (RT). To make RT always positive and to put it in a scale where only integer differences matter, a new raw total is calculated as \( NRT = 5 \times (1.2)^{RT} \). Finally CPR is computed as \( CPR = (1.7)(1-Y)^{NRT} \), where \( Y \) is a factor taking values \( Y = \{1, 2, 3, 3.5\} \) depending on the player’s age and college year.

A methodology similar in its conception to ours is the Career-Arc Regression Model Estimator with Local Optimization (CARMELO) algorithm. CARMELO identifies similar players throughout NBA history and uses their careers to develop a probabilistic forecast for the current player’s future (Silver, 2013). Similarity scores are obtained using a version of a nearest neighbor algorithm. CARMELO applies this process for 19 stats, some of which are weighted more heavily than others (weights are used arbitrarily).

Finally, a player’s CARMELO projection is formed just by averaging the career tracks of all historical players with a positive similarity score.

Regarding cross-sectional models, the most used method is regression. For example, in some entries of the blog Shea and Baker (2013a) a simple regression is used to claim that a team’s effective use of the corner 3 (the three point shot taken from the corner of a basketball court) can be very predictive of its overall shooting efficiency. A Weibull-Gamma with covariates timing model is proposed in Hwang (2012) to predict the points scored by players over time. In this case, the time variable is years played in NBA. Another interesting approach is presented in Salador (2011), where correlations and regression models are computed to figure out which foreign players will be successful in the NBA, by using their previous statistics in international competitions.

In addition to the effort of predicting individual performance, there have also been other approaches focused on teams and other features of the game. For instance, models using simulation are developed to forecast the outcome of a basketball match (Strumbelj and Vračar, 2012; Vračar et al., 2016). A comparison between predictions on NCAAB and NBA match data is discussed in Zimmermann (2016). A dynamic paired comparison model is described in Cattelan et al. (2013) for the results of matches in two basketball and football tournaments. In Benov (2015), an interactive win probability calculator based on locally weighted logistic regression, provides the probability of winning for each team. Furthermore, in Cervone et al. (2016) a process model is used with player tracking data for predicting possession outcomes, while in Puranmalka (2013) several statistical tools are presented using play-by-play data aimed at making predictions and at quantifying unexpected events.
We wish to consider a new perspective by using Functional Data Analysis (FDA) in sports. FDA is a relatively new branch of statistics that analyses data that are drawn from continuous underlying processes, often time, i.e. a whole function is a datum. An excellent overview of FDA is found in Ramsay and Silverman (2005). A continuous function lies behind these data even though functions are sampled discretely at certain points. The FDA framework is highly flexible since the sampling time points do not have to be equally spaced and both the argument values and their cardinality can vary across cases. When functions are observed over a relatively sparse set of points, we have sparse functional data. An excellent survey on sparsely sampled functions is provided by James (2010).

As regards the forecasting of functional time series, there is a body of research, such as Shang and Hyndman (2017); Aue et al (2015); Hyndman and Shahid Ullah (2007), where functions are measured at a fine grid of points. However, only a few works deal with the problem of forecasting sparse functional data (Dokumentov and Hyndman, 2016). Notice that when functions are observed over a dense grid of time points, it is possible to fit a separate function for each case using any reasonable basis. Nevertheless, in the sparse case, this approach fails and the information from all functions must be used to fit each function.

Sports data are sparse and irregular. They are sparse because most players do not have a very long career in the same league. And they are irregular because each player’s career is developed for a different time period. Despite the fact that time series data or movement trajectories are very common in sports, FDA has been mostly used in sport biomechanics or medicine (Epifanio et al., 2008; Harrison, 2014). To the best of our knowledge, there are only two references about sports analytics using FDA. In Wakim and Jin (2014), FDA was introduced to the study of players’ aging curves and hypothesis testing and exploratory analysis were performed. In Vinné and Epifanio (2017), the archetypoid analysis (ADA) was extended for sparse functional data (see also Vinné et al. (2015); Epifanio (2016)), showing the potential of FDA in sports analytics. In particular, it was demonstrated that the advanced analysis with FDA reveals patterns in the players’ trajectories along the years that could not be discovered if data were simply aggregated (averaged, for example).

In this paper, we aim to propose a methodology to predict player’s performances, represented by their Game Score (GmSc), using sparse functional data. To that end, we will focus on two existing methods designed
to handle sparse and irregular data: (i) Regularized Optimization for Prediction and Estimation with Sparse data (ROPES), originally developed by Alexander Dokumentov and Rob Hyndman (Dokumentov and Hyndman, 2016; Dokumentov, 2016); (ii) Principal components Analysis through Conditional Expectation (PACE), originally developed by Fang Yao, Hans-Georg Müller and Jane-Lin Wang (Yao et al., 2005). Our methodology will also involve using the method of analogues based on the functional archetypoid analysis (FADA), which will allow us to refine predictions for the players of interest and to get a more reliable forecasting, in line with the expectations of basketball analysts. We will apply them to a very comprehensive database of NBA players. Results will be obtained with the R software (R Core Team, 2016). The main advantage of our methodology is that we are building on a more sophisticated statistical theory, which also enables to obtain prediction intervals. Prediction intervals are very helpful and important because they express how much uncertainty is associated with the forecast. Results obtained show that our approach is preferable than traditional methods, especially for players who show a performance better than the average. CARMELO is giving forecasts only season by season. However, our approach can also forecast for more detailed time periods, so we can use not only per-season metrics, but also per-age metrics.

GmSc is a formula that quantifies how good or bad a player’s individual performance is in a given game. Thus, GmSc can be considered the type of production that is indicative of future success or decline. GmSc belongs to the class of encompassing metrics (the idea of reducing everything to one number), such as Win-Shares or Player Efficiency Rating (PER) (see Fromal (2012) for an explanation of these and other metrics), which are the culmination of a myriad of numbers.

Forecasting future performance is also very relevant in any other sport (see for instance Arndt and Brefeld (2016). We would like to emphasize that our methodology can also be used in other sports when sparse longitudinal data are available. The databases and R code (including a web application created with the R package shiny (Chang et al., 2013)) to reproduce the results can be freely accessed on https://www.uv.es/vivigui/software.html. The rest of the paper is organized as follows: Section 2 reviews ROPES, PACE, ADA and FADA. Section 3 will be concerned with the data and input variables used. Section 4 presents three analysis: (i) A validation study is done to choose an optimal blend of tuning parameters which ROPES depends on; (ii) ROPES and PACE are compared with each other and with standard
benchmarks; (iii) The reliability of ROPES predictions for current players using the method of analogues with FADA is discussed. A comparison with CARMEL0 results is also provided. The paper ends with some conclusions in Section 5.

2. Methodology

2.1. ROPES

The method ROPES (Regularized Optimization for Prediction and Estimation with Sparse data), proposed by Alexander Dokumentov and Rob Hyndman ([Dokumentov and Hyndman, 2016; Dokumentov, 2016]), solves problems involving sparse functional data that can be encompassed within the framework of this optimization problem:

\[
\{(U, V)\} = \arg\min_{U,V} \left( ||W \odot (Y - UV^T)||^2 + \lambda||KU||^2 + \theta||LV||^2 \right)
\]

where:

- \( Y \) is an \( n \times m \) matrix of two-dimensional data.
- \( U \) is an \( n \times k \) matrix of “scores” (“coefficients”), \( k = \min(n,m) \).
- \( V \) is a \( k \times m \) matrix of “features” (“shapes”).
- \( \lambda > 0 \) and \( \theta > 0 \) are smoothing parameters.
- \( K \) and \( L \) are “complexity” matrices which transform multivariate “scores” \( U \) and “features” \( V \) into the corresponding “complexity” matrices.
- \( ||.|| \) is the Frobenius norm.
- \( \odot \) is the element-wise matrix multiplication.
- \( W \) is an \( n \times m \) matrix of weights.

In order to deal with sparse functional data, the matrices \( Y \) and \( W \) are defined as follows: The rows of \( Y \) are measurements related to each individual taken at a given temporal point and columns refer to the temporal point themselves. Many cells in \( Y \) are missing. Likewise, \( W \) has the value 1 at the
places where \( Y \) has observations and 0 otherwise. Both \( Y \) and \( W \) have the same dimensions. The goal is to find a set of matrices \( Z \) which are of the same dimensions as \( Y \), are close to \( Y \) at the points where measurements are available, and are not very “complex”.

Note that each individual is measured at an irregular and sparse set of time points which differ widely across subjects, so the data set contains different classes of curves with considerably different shapes. Thus, \( Z \) is represented as the product of two matrices \( Z = UV^T \), where \( V \) is the set of “shapes” and \( U \) are the coefficients which these “shapes” are mixed with. \( Z \) is defined as not being “complex” if the “shapes” in \( V \) are “smooth” and the “coefficients” in \( U \) are small.

In practical terms, where the aim is to interpolate and extrapolate the sparse longitudinal data presented over the time dimension, the following optimization problem is solved:

\[
\{(U, V)\} = \text{argmin}_{U,V} \left( ||W \odot (Y - UV^T)||^2 + ||U||^2 + ||\text{DIFF}_2(\lambda_2)V||^2 + ||\text{DIFF}_1(\lambda_1)V||^2 + ||\text{DIFF}_0(\lambda_0)V||^2 \right)
\]

(2)

where \( \text{DIFF}_i(\alpha) \) is a linear operator which represents differentiation \( i \) times and multiplication of the result to the conforming vector \( \alpha \): \( \text{DIFF}_i(\alpha) = \alpha \odot (D^{(1)} \ldots D^{(i)}) \), where \( D^{(i)} \) are conforming differentiation matrices:

\[
D^{(i)} = \begin{bmatrix}
1 & -1 & 0 & \ldots & 0 & 0 \\
0 & 1 & -1 & \ldots & 0 & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & \ldots & 1 & -1
\end{bmatrix}
\]

This problem can be reduced to Eq. (1) by setting in Eq. (1) the matrix \( K \) to the identity matrix and by stacking matrices \( \text{DIFF}_2(\lambda_2) \), \( \text{DIFF}_1(\lambda_1) \) and \( \text{DIFF}_0(\lambda_0) \) into matrix \( L \) as:

\[
L = \begin{bmatrix}
\text{DIFF}_2(\lambda_2) \\
\text{DIFF}_1(\lambda_1) \\
\text{DIFF}_0(\lambda_0)
\end{bmatrix}
\]

ROPES is equivalent to maximum likelihood estimation with partially observed data, which allows the calculation of confidence and prediction in-
tervals. They are estimated using a Monte-Carlo style method, see specifical details in (Dokumentov and Hyndman, 2016).

2.2. PACE

Functional Principal Components Analysis (FPCA) is a common tool to reduce the dimension of data when the observations are random curves. The usual computational methods for FPCA based on discretizing the functions or basis expanding the functions are inefficient when data with only few repeated and sufficiently irregularly spaced measurements per subject are available. 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.

A version of FPCA, in which the FPC scores are framed as conditional expectations, was developed by Fang Yao, Hans-Georg Müller and Jane-Lin Wang to overcome this issue (Yao et al., 2005). This method was referred to as Principal components Analysis through Conditional Expectation (PACE) for sparse and irregular longitudinal data. In practice, the prediction for the trajectory $X_i(t)$ for the $i$th subject, using the first $K \phi_k$ eigenfunctions, is:

$$\hat{X}_i^K(t) = \hat{\mu} + \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\phi}_k(t)$$

where $\hat{\mu}$ is the estimate of the mean function $E(X(t)) = \mu(t)$ and $\xi_{ik}$ are the FPC scores. PACE and its implementation in the R library `fdapace` (Dai et al. (2016)) use local smoothing techniques to estimate the mean and covariance functions of the trajectories, specifically a local weighted bilinear smoother is used for estimating the covariance. Generalized Cross Validation is used for bandwidth choice, which is the default method for the function FPCA in the R library `fdapace` (parameters by default are considered, for example 10 folds are used and a Gaussian kernel). The number of components $K$ is determined by using the Fraction-of-Variance-Explained threshold (0.9999 by default) computed during the SVD of the fitted covariance function.

The eigenfunctions $\hat{\phi}_k(t)$ and $K$ are estimated with the training set, and they are used in the estimation of the scores for the test set. This is the procedure we will follow in Section 4.2. With the scores and the estimated eigenfunctions, we obtain an approximation of the trajectories and can be
used to predict unobserved portions of the functions. In Yao et al. (2005), it is also explained the construction of asymptotic pointwise confidence intervals for individual trajectories and asymptotic simultaneous confidence bands.

2.3. ADA

Archetypoid analysis (ADA) was presented in Vinué et al. (2015) and is an extension of archetypal analysis defined by Cutler and Breiman (1994) (see D’Esposito et al. (2012); Ragozini et al. (2017), for other derived methodologies). In ADA, archetypes correspond to real observations (the so-called archetypoids). Let $X$ 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 ADA is to find a $k \times m$ matrix $Z$ that characterizes the archetypal patterns in the data. In ADA, the optimization problem is formulated as follows:

$$RSS = \sum_{i=1}^{n} \| x_i - \sum_{j=1}^{k} \alpha_{ij} z_j \|^2 = \sum_{i=1}^{n} \| x_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l \|^2, \quad (4)$$

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.

Archetypoids are computed with the R package Anthropometry (Vinué, 2017).

2.3.1. ADA for sparse data with FDA

ADA was defined for functions in Epifanio (2016), where it was shown that functional archetypoids can be obtained as in the multivariate case if the functions are expressed in an orthonormal basis, just by applying ADA to the basis coefficients. When functions are measured over some sparse points, we have sparse functional data.

The basic idea of the functional archetypoid analysis (FADA) is as follows. Let us assume that $n$ smooth functions, $x_1(t), \ldots, x_n(t)$, are observed, with the
i-th function measured at \( t_{i1}, \ldots, t_{im} \) points. Based on the Karhunen-Loève expansion, the functions are approximated by:

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

where \( \xi_{ij} \) is the \( j \)th principal component score for case \( i \), \( \phi_j(t) \) represents the \( j \)th principal component function (eigenfunction), and \( m \) is the number of principal components used in the estimation. Because the eigenfunctions are orthonormal, to obtain FADA we can apply ADA to the \( n \times m \) matrix \( X \), with the scores (the coefficients in the Karhunen-Loève basis).

3. Data

Our starting database is the same as in [Vinue and Epifanio, 2017, Section 5.2. It contains the NBA players and their statistics per game, including their age (year and day) when they played each game and their GmSc for that game, from the season 2005-2006 to the season 2014-2015. There are in total 247577 rows. All the data were downloaded from http://www.basketball-reference.com/play-index/pgl_finder.cgi?lid=front_pi. We chose to focus on predicting GmSc because this is a formula that attempts to combine several player’s contributions into one number, but we could have easily chosen any other statistic available in the data. GmSc measures how well a player performed during a single game by looking only at box score stats. Its exact formula is as follows:

\[
\begin{align*}
PTS &+ 0.4 \times FGM - 0.7 \times FGA - 0.4 \times (FTA - FTM) + 0.7 \times ORB \\
&+ 0.3 \times DRB + STL + 0.7 \times AST + 0.7 \times BLK - 0.4 \times PF - TOV
\end{align*}
\]

where PTS indicates points, FGM field goals made, FGA field goals attempted, FTA free throws attempted, FTM free throws made, ORB offensive rebounds, DRB defensive rebounds, STL steals, AST assists, BLK blocks, PF personal fouls and TOV turnovers.

All stats are weighted differently and according to the frequency with which they happen. Positive actions receive positive coefficients and negative actions receive negative ones. The scale is similar to that of points
scored, (40 would be the performance of a superstar, around 10 is the average performance, etc). GmSc is currently one of the most common used advanced metrics. There are technically no lower and upper bounds. A player can have negative GmSc values. In practice, scores bigger than 40 are uncommon. The general rule of thumb is that integer differences matter in GmSc (especially differences larger than 5), while decimal differences are not that significant.

It is worth emphasizing that there may be many observations for the same player in a short period of time, but none for many other player ages. In fact, these many measurements in close ages can provide redundant information, because it is expected that the player’s performance is similar and stable in short time periods, such as two consecutive months. However, there may be a few players, especially young players or players who have just entered the league, with some fluctuations in their productivity (inside a reasonable range, such as a difference of 10 points in consecutive GmSc). This is what is called players’ inconsistency (Shea, 2014).

In FDA it is assumed that $n$ smooth functions, $x_1(t), \ldots, x_n(t)$, are observed, with the $i$-th function ($i = 1, \ldots, n$) measured at $t_{i1}, \ldots, t_{in}$, points, i.e. $x_i(t_{ij})$ ($j = 1, \ldots, n_i$). In our study, $x_i(t)$ represents the GmSc of player $i$ for a certain age $t$. As mentioned in Silver (2015), the most important attribute of all, in terms of determining a player’s future career trajectory, is his age. Fig. 1 illustrates the type of data we are working with. It shows the observations of some players, whose values are represented as connected points.
Figure 1: GmSc values for five players at their corresponding ages (colors in the online version).

Regarding data preprocessing, we have removed firstly the entries where the players played for less than 5 minutes. The ROPES method is intended to work with more individuals than temporal points, so it is more convenient to reduce the number of the available temporal points. This also allows the method not to be computationally very intensive, taking also into account that adding more temporal points would not be providing new information. To that end, we have averaged the players’ stats for each month to produce a single raw total. In this way, a data frame is created with the players’ GmSc values in the six months of the ten seasons considered. The player’s age is in the 15th day of each month. We have removed the entries where players were below 19 years old or over 40 years old (we have not removed the players, it is simply that the age range considered is from 19 to 40 years old). There are initially 1100 players. In order to define the number of temporal points, we have created a sequence from the lowest age to the highest age with an increment determined by a granularity value. Data granularity refers to the level of detail represented by the data, so granularity allows us to reduce the number of temporal points. By choosing a granularity equal to 0.25 we have obtained 84 age points. Due to the computational complexity of the ROPES
method, we considered that 84 was a suitable number to obtain results in a reasonable amount of time, while having enough temporal information.

4. Results

In Section 4.1 a validation study is proposed to choose an appropriate ROPES tuning parameters combination, which is crucial for good predictive activity. Section 4.2 contains a comparison analysis between ROPES, PACE and two benchmark methods. In Section 4.3 the type of projections obtained are detailed.

4.1. Selection of parameters

ROPES depends on three tuning parameters \((\lambda_2, \lambda_1, \lambda_0)\), which have to be chosen to guarantee that the model itself returns predictions with enough accuracy. We evaluate the precision of the model’s prediction in terms of the mean squared error (MSE). MSE measures the average of the squares of the differences between the predicted values \(\hat{y}\) and the true values \(y\) across all individual estimates \(i\), as shown in Eq. (7).

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2
\]

We adopt MSE since ROPES uses it to measure the error term. In order to select the parameters, we proceed as follows: Our goal will be to predict the GmSc in the season 2014-2015 for those players who played at least in one season before the season 2014-2015 and also played in the season 2014-2015 itself. The justification for doing this is related to sportive reasons. In sports, when coaches and managers are building their rosters, it is highly important for them to have a basic idea about how players will perform just the following season. Of course, they would also like to know the players’ performance in the long term, but most rosters are built according to the most immediate season. This would allow them to decide whether the current roster should remain the same for the next season or whether some players should be replaced. We are using here the season 2014-2015 since it is the last one of our database. This procedure makes sense because we will consider the previous performance of all the players selected, but we are only interested in predicting their GmSc for the next season, by taking into account each player’s data and the information of the others. This procedure is more computationally efficient than the leave-one-out approach.
Initially, both $Y$ and $W$ contain 946 players and 84 temporal points in total. These 946 players include both the players who didn’t play the season 2014-2015 and who did play that season. In order to improve the speed of calculations, we have chosen a subset of players: the first 200 players having most observation points. This will be the training set we use to select the optimal combination of $\lambda$ parameters for ROPES. The player with least entries has 17 GmSc observations recorded, whereas 27 are recorded for the player with most entries. Regarding players who played the season 2014-2015 and at least one season before, the data used by the algorithm contain all the available GmSc values for these players except their GmSc for the season 2014-2015. Therefore, in the $Y$ matrix the actual GmSc values related to the player ages in the season 2014-2015 are turned into NA and the same cells of $W$ are turned into 0. From the 200 players selected, 134 played the season 2014-2015 and at least one season before.

Firstly, we optimize only the parameter $\lambda_2$, setting $\lambda_1 = 0$ and $\lambda_0 = 10$. The parameter $\lambda_2$ takes values in a sequence from 0 to 1000 incrementing by 100. In this way, the first blend is $(\lambda_2 = 0, \lambda_1 = 0, \lambda_0 = 10)$, the second is $(\lambda_2 = 100, \lambda_1 = 0, \lambda_0 = 10)$ and so on. We are looking for smooth curves, so we give more emphasis to $\lambda_2$ because it is related to the second derivative and this derivative is strongly related to the smoothness of the curve. This is justified because if the second derivative is a smooth curve, both the first derivative and the original function will be also smooth. From the definition of derivative, it follows directly that if a function has first derivative existing at any point, then it doesn’t have a sharp bend (v-shape) at that point (the same can be said with the second derivative regarding the first derivative). See for example [Ramsay and Silverman, 2005. Section 5.2.2] for more insights. The opposite is not always true. Fig. 2 shows that the smallest MSE for the 134 players of the training set is for the combination with $\lambda_2 = 600$.

Once the optimal $\lambda_2$ has been found, we then adjust $\lambda_1$ and $\lambda_0$ as well. Both $\lambda_0$ and $\lambda_1$ take values in a sequence from 0 to 10 incrementing by 5. Fig. 3 shows that the most optimal combination in terms of MSE is $(\lambda_2 = 600, \lambda_1 = 5, \lambda_0 = 10)$. We also computed the Mean Absolute Error (MAE = $\frac{1}{n} \sum_{i=1}^{n} |e_i|$) and it is interesting to note that $(\lambda_2 = 600, \lambda_1 = 0, \lambda_0 = 10)$ got the smallest MAE. The choice of grid search was dictated by the low speed of calculations in ROPES.
Figure 2: MSE values associated with every combination of lambdas when $\lambda_2$ is moving and $\lambda_0, \lambda_1$ are fixed. It is highlighted in blue the combination of lambdas that provides the smallest MSE (colors in the online version).

Figure 3: MSE values associated with every combination of lambdas when $\lambda_2$ is fixed and $\lambda_0, \lambda_1$ are moving. It is highlighted in blue the combination of lambdas that provides the smallest MSE (colors in the online version).
4.2. Comparison with other methods

In order to evaluate the usefulness of ROPES and PACE, we carry out a comparison with each other and with two benchmark methods, such as the average method and the naïve method. In the average method, the forecast of the next value is the mean of the previous values. In the naïve option, the forecast is the value of the last observation. They are two common simple alternatives to more advanced techniques (Hyndman and Athanasopoulos, 2013, Section 2.3).

In order to check the performance of all methods, we have applied them to the test set containing all the players different from the 200 players selected in the previous Section. Table 1 reports an extract of the database that contains the following information: (i) only the players who played at least one season before the season 2014-2015 and also played in the season 2014-2015 itself, with their ages; (ii) their actual GmSc values in 2014-2015; (iii) the predictions with ROPES (with the optimal $\lambda$ parameters), PACE and the simple methods; (iv) the squared difference between actual values and predictions (denoted as Dif.$^2$); (v) the resulting total MSE (highlighted in bold). Players are ordered alphabetically starting from the season 2005-2006. Players who had a missing GmSc in their corresponding age were removed because the difference between the prediction and the actual value was also a missing value and MSE could not be computed. There are 640 rows and 269 players. PACE gets the smallest MSE, followed by ROPES. The GmSc mean got by PACE is practically the same as the actual one. The standard deviations of the predictions with ROPES, PACE and the average method are smaller than the one for the actual values, so their predictions are less dispersed.

Fig. 4 displays the MSE values obtained with the four methods in different GmSc intervals. This plot allows us to assess the players for which the predictions are best and worst. ROPES returns the lowest MSE in the intervals $[10, 17)$ and $[17, 24)$, PACE in the interval $[3, 10)$, which is the one with most players, and the average method in the interval $[-4, 3)$. The average method is better than ROPES in this interval and also in $[3, 10)$. A plausible explanation is that these intervals only contain the players with a production close to the typical GmSc, which is around 10. These players rarely do great, so their performance is very similar in all their careers.

ROPES is being more influenced by the outstanding players, so it is able to predict their performance better than the other methods, at the expense of the prediction for players with lower GmSc. Note that there are only 16
Table 1: Actual and predicted GmSc values with ROPES, PACE, the average method and the naïve method, for the season 2014-2015 of each player who played at least one season before the season 2014-2015 and also played in the season 2014-2015 itself, in the ages associated with the temporal points defined by the level of granularity. Differences between actual and predicted values and MSE are also provided. MSE is highlighted in bold.

<table>
<thead>
<tr>
<th>Player</th>
<th>Age</th>
<th>GmSc</th>
<th>GmSc_{pr}</th>
<th>Dif</th>
<th>GmSc_{pr}</th>
<th>Dif</th>
<th>GmSc_{pr}</th>
<th>Dif</th>
<th>GmSc_{pr}</th>
<th>Dif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alan Anderson</td>
<td>32.05</td>
<td>4.14</td>
<td>5.05</td>
<td>0.81</td>
<td>5.22</td>
<td>1.15</td>
<td>4.34</td>
<td>0.04</td>
<td>4.34</td>
<td>0.06</td>
</tr>
<tr>
<td>Alan Anderson</td>
<td>32.27</td>
<td>5.64</td>
<td>5.03</td>
<td>0.67</td>
<td>5.35</td>
<td>0.37</td>
<td>4.34</td>
<td>1.00</td>
<td>3.34</td>
<td>0.34</td>
</tr>
<tr>
<td>Jonas Jerebko</td>
<td>27.52</td>
<td>4.15</td>
<td>3.95</td>
<td>0.20</td>
<td>4.32</td>
<td>0.42</td>
<td>5.59</td>
<td>2.00</td>
<td>5.54</td>
<td>1.93</td>
</tr>
<tr>
<td>Jonas Jerebko</td>
<td>27.77</td>
<td>5.73</td>
<td>3.82</td>
<td>1.92</td>
<td>4.27</td>
<td>2.13</td>
<td>5.59</td>
<td>0.00</td>
<td>5.54</td>
<td>0.04</td>
</tr>
<tr>
<td>Jonas Jerebko</td>
<td>28.02</td>
<td>6.02</td>
<td>3.69</td>
<td>2.62</td>
<td>4.23</td>
<td>3.23</td>
<td>5.59</td>
<td>0.18</td>
<td>5.54</td>
<td>0.28</td>
</tr>
<tr>
<td>Quincy Acy</td>
<td>24.02</td>
<td>4.28</td>
<td>3.1</td>
<td>1.18</td>
<td>4.99</td>
<td>0.79</td>
<td>5.28</td>
<td>1.00</td>
<td>4.92</td>
<td>0.41</td>
</tr>
<tr>
<td>Quincy Acy</td>
<td>24.27</td>
<td>4.56</td>
<td>3.01</td>
<td>1.55</td>
<td>5.01</td>
<td>0.20</td>
<td>5.28</td>
<td>0.52</td>
<td>4.92</td>
<td>0.13</td>
</tr>
<tr>
<td>Quincy Acy</td>
<td>24.52</td>
<td>7.22</td>
<td>2.74</td>
<td>4.48</td>
<td>5.05</td>
<td>4.71</td>
<td>5.28</td>
<td>3.76</td>
<td>4.92</td>
<td>5.29</td>
</tr>
<tr>
<td>Victor Oladipo</td>
<td>22.52</td>
<td>11.04</td>
<td>9.16</td>
<td>1.88</td>
<td>9.37</td>
<td>2.79</td>
<td>9.2</td>
<td>3.4</td>
<td>9.13</td>
<td>3.65</td>
</tr>
</tbody>
</table>


Mean ± Sd (2.54,11.32) (3.03,10.99) (3.41,10.43) (2.74,9.32) (2.65,11.77)

players who have a value greater or equal than 17. Outstanding players are more interesting in sportive terms than regular players, since outstanding players are going to improve, to get worse or to remain stable. On the contrary, regular players are not going to develop, so predictions for special players are more valuable, and ROPES seems to perform better with them. For these reasons, we will use ROPES to forecast future players’ activity in next Section.
4.3. Projections of future performance with ROPES and the method of analogues

4.3.1. Case of study: Giannis Antetokounmpo

Giannis Antetokounmpo is a Greek player for the Milwaukee Bucks. He made his debut in the NBA the 2013-2014 season. Currently, he is on the way to be one of the most important superstars of the league in the coming seasons. In fact, he has become the youngest international player to be selected to play an All-Star game (NBA All Star Weekend 2017). Hence, it is interesting to see the forecasting for him.

In a first attempt to compute predictions using all the players of the data set, we realised that the ROPES method had some pull towards the mean of the entire sample (like the other methods discussed in Section 4.2 but not so strong as them). This gave unrealistic performance predictions for both the best and most promising players. For the particular case of Antetokounmpo, his future values were barely higher than 10, which was not
Table 2: Similarity of Antetokounmpo to the four archetypoids according to the $\alpha$ coefficients.

<table>
<thead>
<tr>
<th>Player</th>
<th>Archetypoids</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. Stephenson</td>
<td>0.42</td>
</tr>
<tr>
<td>LeBron</td>
<td>0.42</td>
</tr>
<tr>
<td>D. Granger</td>
<td>0.01</td>
</tr>
<tr>
<td>S. Jackson</td>
<td>0.15</td>
</tr>
<tr>
<td>G. Antetokounmpo</td>
<td>0.42</td>
</tr>
</tbody>
</table>

reliable at all. Therefore, in order to refine predictions, it is much suitable to use the so-called “method of analogues”. The idea is to find related players to the one of interest and then use their documented activity to get the predictions. We know how other players already performed, so we can use their information to have an approximate idea about future performances of others. The method of analogues has been used for years in fields such as climatology (Zorita and Von Storch, 1999) and epidemiology (Viboud et al., 2003). Recently, an R package has been released that contains analogue methods for palaeoecology (Simpson and Oksanen, 2016). The CARMELO method is also based on this scheme.

In order to find related players, we use the archetypoid analysis (see Vinué et al. (2015) for theoretical details). In this technique, the GmSc function of a player is approximated by a mixture of archetypoids, which are themselves functions of extreme players. Archetypoids are specific players and the $\alpha$ coefficients represent how much each archetypoid contributes to the approximation of each individual. We use the same procedure explained in Vinué and Epifanio, 2017, Section 5.2. The most comparable archetypoid should be the corresponding to the largest value of the $\alpha$ coefficients for the player of interest.

We are interested in looking at the players’ profiles regarding the league stars. To do this, the archetypoid analysis is more useful than clustering methods because the representative points obtained correspond to extreme players. As indicated in Vinué and Epifanio, 2017, Section 5.2 the archetypoids are Lance Stephenson, LeBron James, Danny Granger and Stephen Jackson. Table 2 shows the $\alpha$ values for Antetokoumpo. The star player here is LeBron James, so we focus on him as the analogue archetypoid due to the high expectations about the future Antetokoumpo’s performance.

Then, the total group of analogous players is made up of LeBron James, together with the other players whose largest $\alpha$ coefficient is also for LeBron and have an $\alpha$ value greater than the Antetokoumpo’s $\alpha$. Current stars such as Kevin Durant or Russell Westbrook and stars of previous seasons
such as Paul Pierce or Allen Iverson belong to this set.

The important point to note here is that we are working with data until the season 2014-2015 and it is well known that Antetokounmpo began to shine as the Bucks’ primary playmaker during the 2015-16 season. Until that season he was very young and his irruption was still to come. Hence his high similarity with Stephenson.

Fig. 5 shows the players with the six highest $\alpha$ values in the analogous group of the Greek player. As can be seen, they are six of the best players in the last years. In addition, results are also very interesting because Antetokounmpo is actually an athletic power forward such as Durant, with a similar height to Davis and Nowitzki (he stands at 2.11 m), and also can play effectively as a point guard due to his very good ball-handling skills, like other great point guards such as Chris Paul.

The ROPES algorithm (with the lambdas combination obtained in the validation study) is used to obtain $p$-forecasting intervals, where $p = 0.05$ is the selected significance level. Fig. 6 shows the forecasting obtained for Antetokounmpo using this set. He is going to improve his performance and his values seem to decline since it is difficult to maintain a high level of
Figure 6: Predictions for Giannis Antetokounmpo using only the set of analogue players (colors in the online version).

performance for a long period of time. However, these values are still fine until his early thirties. It is noticeable that this procedure that combines the archetypoid analysis and ROPES seems to be anticipating his success in coming seasons.

As can be seen, the curves are pretty smooth and the intervals are wide. These prediction intervals are very useful to assess the uncertainty in forecasts. In addition, the predictions are now much more reliable with the expectations of any basketball coach and fan.

4.3.2. Projections for more players

As mentioned, our initial database contains 1100 players. The important point to note here is that there are several players who played in the NBA some seasons ago and they did not play again, for example, the Spaniard Juan Carlos Navarro, who played in the NBA the season 2007-2008 and since then he has been playing back in Spain. Therefore, in sportive terms is much more reasonable to try to predict the performance of the players who are still in the league. As a consequence, we have filtered the players who belong to current 2016-2017 NBA rosters and who played at least one season in the
NBA (otherwise, we cannot know the most similar (analogous) archetypoid to them). We have got the players of the 2016–2017 NBA rosters by scraping from the website [http://basketball.realgm.com/nba/teams](http://basketball.realgm.com/nba/teams) using the R package XML (Lang and the CRAN Team, 2015).

In the interests of illustration, we have computed the predictions for the players who were selected among the first 15 picks in the 2011, 2012 and 2013 NBA drafts and who have a set of analogous players with less than 100 players. Again, the ROPES algorithm is used to obtain $p$-forecasting intervals where $p = 0.05$ is the selected significance level. An interactive web application available on [http://bayes2.ucd.ie:3838/gvinue/AppPredPer](http://bayes2.ucd.ie:3838/gvinue/AppPredPer) allows the user to represent the forecasting plots for each player. This app provides some basic information about the way it works and will be in constant developing. It can also be generated from R with these two commands: `library(shiny); runUrl('http://www.uv.es/vivigui/softw/AppPredPerf.zip')`.

Fig. 7 allows us to compare easily several players. It is interesting to see that Nerlens Noel has a bright future. Anthony Davis and Kyrie Irving are already great players and their performance is going to be constantly good for many years, although Irving may decline from his thirties. Regarding Enes Kanter, there may be a gradual serious reduction in his production.

Figure 7: Comparative plot of the past and future performance for four players drafted in the last few years: Enes Kanter (number 3, draft 2011), Anthony Davis (1, 2012), Kyrie Irving (1, 2011) and Nerlens Noel (6, 2013) (colors in the online version).
We have checked the CARMELO forecasting in order to compare our predictions with that system (accessing the second link in Silver (2015)). CARMELO predicts a variable called WAR (wins above replacement), which reflects a combination of a player’s projected playing time and his projected productivity while on the court. It is computed using the Box Plus/Minus variable (BPM), which is a metric for evaluating basketball players’ quality and contribution to the team. WAR is not in our database and cannot be computed with the stats available since BPM involves team stats. However, WAR and GmSc share the fact that they are two measures intended to give a total perspective on a player’s statistical performance. A great player will show both high GmSc and WAR. On the contrary, conventional players will show low values in both metrics.

We see that CARMELO is also identifying the breakout of Antetokounmpo. Enes Kanter shows a poor performance with low WAR values for all future seasons, which is partially in line with our decreasing predictions from age 26. Anthony Davis is considered by CARMELO as the most important player in the coming future. Our predictions for him are also very good and stable over time. For CARMELO, Irving will maintain his good level during two seasons, starting then a progressive decline. This is also what our predictions seem to suggest. Regarding Nerlens Noel, the CARMELO prediction says that his performance will slowly get better. Our method suggests here that he could become a great player, with an upper prediction even larger than 30. Predictions from CARMELO seem to agree with the perspectives provided by ROPES, maybe with a more conservative forecast for Noel. CARMELO gives predictions for single seasons, but ROPES can be adjusted for a smaller or bigger period of time, related to the players’ ages, providing then some more detailed information.

It is well known that players usually need time to develop their skills (see e.g. Hwang (2012)). Firstly, they have to get used to the league style (game pace, the way the referees perform). Then, players progressively reach peak fitness. As time passes, their ability usually decreases slowly, likely also decreasing their statistics. It is also worth pointing out that statistical models are not very reliable for long-term forecasting, because the assumption that the future looks similar to the past slowly breaks down the further you go into the future.

23
5. Conclusions

Basketball, like any other sport, contains a lot of uncertainty. A central issue is to predict future players’ performance using past observations. In spite of the fact that basketball data continues to expand and there is a constant demand of new techniques that provide objective information for understanding the game, there are not many publicly available projection systems. In this paper we have presented a methodology to deal with sparse functional data in order to forecast the performance of basketball players. This has been done by analyzing ROPES and PACE and by including the method of analogues together with the functional archetypoid analysis.

ROPES depends on several parameters, so we have done a validation study to choose an optimal combination that provides smooth curves and avoids overfitting. The combination obtained is working well to avoid narrow intervals and overconfident inferences. A comparison study has also been done to compare ROPES with PACE, and with simple alternatives, such as the average and naïve methods. In cases where the performances are very constant and poor, the average method could be used since the mean value approximates well a set of close numbers. PACE performed best with players whose performance is not bad but is not prominent either. ROPES seems to predict well even for players with non constant and very good performances. PACE was performing better than ROPES in runtime. However, unlike ROPES, it is not possible to get prediction intervals with its current computational implementation. In addition, from a practical point of view, it is more interesting to identify potential great players. Therefore, we have used ROPES to that end. ROPES is a method aimed to decompose, smooth and forecast sparse functional data. In particular, we have used one-dimensional data. Starting from a database that contains all the games played in the NBA regular seasons from 2005-2006 to 2014-2015, we have created the input matrices in the form required by the ROPES algorithm.

In the sparse case, information from all functions is used to fit each function, so all individuals contribute to a greater or lesser degree to form the estimations. In order to overcome this problem and to refine predictions, we have used the so-called “method of analogues”. The idea is to relate a player’s curve into one of possible players’ types and then to predict his performance using only the information of these comparable athletes. In our case, the players’ types are given by the archetypoids of the data set. It is remarkable that the results obtained with this procedure that combines the
functional archetypoid analysis and ROPES seems to be able to anticipate the activity of outstanding players accurately, taking also into account the forecast uncertainty by including prediction intervals. Prediction intervals are very important because they give an idea of the forecasting precision.

Once the computations are finished, an interactive web application can be used to represent the plots with the past and future behavior of every player. The variable involved in prediction has been GmSc because from the statistics available in the original database, is the one that gives a reasonable approximation about the player’s production. GmSc is of interest to basketball managers and fans. Any other variable (both per-season or per-age) can be used as well. The possibility of using per-age metrics enables to get a more detailed forecasting.

Player forecasting systems are important as a means of summarizing the overall match performance of individual players. Any forecasting method is limited because some aspects such as injury risk or work ethic, which influence future performance, are very difficult to quantify. However, coaches and experts can use these systems to review performances of their own players as well as tracking the performance levels of potential acquisitions. We expect that the approach presented here will provide valuable information about players’ overall ability to support decision making. Sparse functional data are very common in sports. Therefore, it is very reasonable to bring methods developed to deal with this kind of data to the sportive field. This methodology can serve as a baseline to further efforts in the same direction. Further work will involve using more previous seasons data to add historical players as archetypoids and analogue players. GmSc values are available from the NBA season 1983-1984. Data and all R code involved are freely available for reproducibility and further exploration of the results.

6. Acknowledgements

The authors wish to express their gratitude to Alexander Dokumentov and Rob Hyndman for kindly providing the R code to run the ROPES algorithm. Funding: This work was partially supported by Grant Ministerio de Economía y Competitividad DPI2013-47279-C2-1-R.

7. Data Accessibility

The authors are making the data associated with this paper available at https://www.uv.es/vivigui/software.html
References


