International Journal of Performance Analysis in Sport

ISSN: 2474-8668 (Print) 1474-8185 (Online) Journal homepage: https://www.tandfonline.com/loi/rpan20

Performance Profiles based on Archetypal Athletes

Manuel J. A. Eugster

To cite this article: Manuel J. A. Eugster (2012) Performance Profiles based on Archetypal Athletes, International Journal of Performance Analysis in Sport, 12:1, 166-187, DOI: 10.1080/24748668.2012.11868592

To link to this article: https://doi.org/10.1080/24748668.2012.11868592

Published online: 03 Apr 2017.

Submit your article to this journal

Article views: 86

View related articles

Citing articles: 3 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=rpan20

International Journal of Performance Analysis in Sport
2012, 12, 166-187.

Performance Profiles based on Archetypal Athletes

Manuel J. A. Eugster

Institut für Statistik, Ludwig-Maximilans-Universtät München, Munich, Germany.

Abstract

Performance indicators and, on their basis, performance profiles are one of the foundations of performance analysis in sports. Obviously, the crux is to develop performance profiles which allow to evaluate the subjects of interest accurately. The present paper contributes a further approach to the existing toolbox of profiling methods. Performance profiles based on archetypal athletes are not based on typical, i.e., mean, performances but on extreme performances—usually the most interesting aspect in sports. Archetypal athletes (outstanding—positive and/or negative—performers) are computed and performers are related to these archetypal athletes. As the archetypal athletes are interpretable, an easy interpretation of the performers' profiles follows. The method is demonstrated on basketball statistics and soccer skill ratings.

Keywords: archetypal analysis, convex hull, extreme value.

1. Introduction

"Dirk Nowitzki is the best basketball player. No, it's Kevin Durant! ". "Christiano Ronaldo is the number one, Lionel Messi is the number two soccer player in the world". "Ronaldinho is the better dribbler, but Zinǐdine Zidane is faster". These and similar statements can be found in almost all discussions on sports and athletes. They are interesting to debate, but they are also having a great impact on many (managerial) decisions—from a coach's tactical specification via engagements of new players through to a company's selection of a brand ambassador. The foundation of such statements are performance indicators and, on their basis, performance profiles. Obviously, the crux is to develop *good* performance profiles allowing to evaluate players and teams accurately. Literature reports great effort in developing meaningful performance profiles in sport. Hughes and Bartlett (2002) state the basic principle of relating performers' values for performance indicators to normative data or to opposition values. In this sense, Hughes et al. (2001), James et al. (2005) and O'Donoghue (2005) develop methods to compute stable normative profiles and to relate performers to them. In order to take the opposition into account (O'Donoghue, 2009, for example, shows this requirement), O'Donoghue and Cullinane (2011) present a regression based approach.

O'Donoghue (2005) describes three stages of performance profiling: (1) determining normative data; (2) describing the performances of the performers; and (3) relating the performances to the normative data. This paper presents a different approach for computing the reference data. Normative data allows to describe the *typical* (i.e., mean) performances of performers. However, in sports usually not the mean performances but the extreme performances (outstanding—positively and/or negatively—performers) are of interest. Therefore, the idea of the presented approach is to compute extreme performers (the archetypal athletes) and to profile all performers in relation to these archetypal athletes. A primary advantage of this method is the interpretability of the archetypal athletes (as different types of "good" and "bad" performers with extreme performance indicators) and the resultant easy interpretation of the performers' profiles.

Archetypal athletes are the result of a statistical methodology called archetypal analysis. In general, archetypal analysis has the aim to find a few, not necessarily observed, extremal observations (the archetypal athletes) in a multivariate data set such that all the data can be well represented as convex combinations of the archetypes (the performance profile). The archetypes themselves are restricted to being convex combinations of the individuals in the data set and lie on the data set boundary, i.e., the convex hull (see, e.g., Preparata and Shamos, 1990, Chapter 3). This statistical method was first introduced by Cutler and Breiman (1994) and has found applications in different areas, e.g., in economics (Li et al., 2003; Porzio et al., 2008), astrophysics (Chan et al., 2003) and pattern recognition (Bauckhage and Thurau, 2009).

The paper is organized as follows. In Section 2 we outline archetypal analysis by introducing the formal optimization problem. In Section 3 we illustrate the idea of performance profiles based on archetypal athletes using a two-dimensional subset of NBA player statistics from the season 2009/2010. In Section 4 we compare and justify performance profiling through archetypal analysis versus related *k*-prototypes methods like *k*-means. In Section 5 we then identify and discuss archetypal athletes and performance profiles for two popular sports. Section 5.1 extends the illustrative NBA example and computes performance profiles based on archetypal basketball players using common statistics from the season 2009/2010. Section 5.2 computes performance profiles based on archetypal soccer players of the German Bundesliga, the English Premier League, the Italian Lega Serie A, and the Spanish La Liga using skill ratings (at the time of September 2011). Finally, in Section 6 a discussion and a summary are given. Please note that the color specifications in visualizations refer to the colorized online version of the article; we use different line types, point symboles, and concrete position descriptions to make the grayscale version readable as well. All data sets and source codes for replicating our analyses are freely available (Section 8 on computational details).

2. Archetypal analysis

Consider an $n \times m$ matrix X representing a multivariate data set with n observations and m attributes (i.e., a database with n performers and m performance indicators). For a given k,

the archetypal problem is to find the matrix Z of k archetypes (i.e., archetypal athletes). More precisely, to find the two $n \times k$ coefficient matrices α (which we interpret as the performance profile) and β so that (1) the data are well represented as convex combinations of the archetypes, i.e., $X \approx \alpha Z^{\intercal}$, and (2) the archetypes are convex combinations of the data points, i.e., $Z = X^{\intercal}\beta$. Therefore, for a given number k, a suitable selection of archetypes minimizes the residual sum of squares (Formula 1).

$$RSS = \left\| X - \alpha Z^{T} \right\|_2 \; with \; Z = X^{T}\beta$$

subject to the constraints

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

$$\sum_{i=1}^{n} \beta_{ji} = 1 \; with \; \beta_{ji} \geq 0 \; and \; j = 1, \ldots, k.$$

$\left\| \cdot \right\|_2$ denotes the Euclidean matrix norm (also known as the Frobenius Norm). The norm of an $n \times m$ matrix A is defined as the square root of the sum of the absolute squares of the matrix elements a_{ij}, i.e., $\left\| A \right\|_2 = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} \left| a_{ij} \right|^2}$ (see, e.g., Golub and Loan, 1996).

Cutler and Breiman (1994) present an alternating constrained least squares algorithm to solve the problem: it alternates between finding the best α for given archetypes Z and finding the best archetypes Z for given α; at each step several convex least squares problems are solved, the overall RSS is reduced successively. Through the definition of the problem, archetypes lie on the boundary of the convex hull of the data. Let N be the number of data points which define the boundary of the convex hull, then Cutler and Breiman (1994) showed: if $1<k<N$, there are k archetypes on the boundary which minimize RSS; if $k=N$, exactly the data points which define the convex hull are the archetypes with RSS=0; and if $k=1$, the sample mean minimizes the RSS. In practice, however, these theoretical results can not always be achieved (Eugster and Leisch, 2009). Furthermore, there is no rule for the correct number of archetypes k for a given problem instance. A simple method to determine the value of k is to run the algorithm for increasing numbers of k and use the "elbow criterion" on the RSS where a "flattening" of the curve indicates the correct value of k. For detailed explanations we refer to Cutler and Breiman (1994, on the original algorithm), Eugster and Leisch (2009, on numerical issues, stability, and computational complexity), and Eugster and Leisch (2011, on robustness).

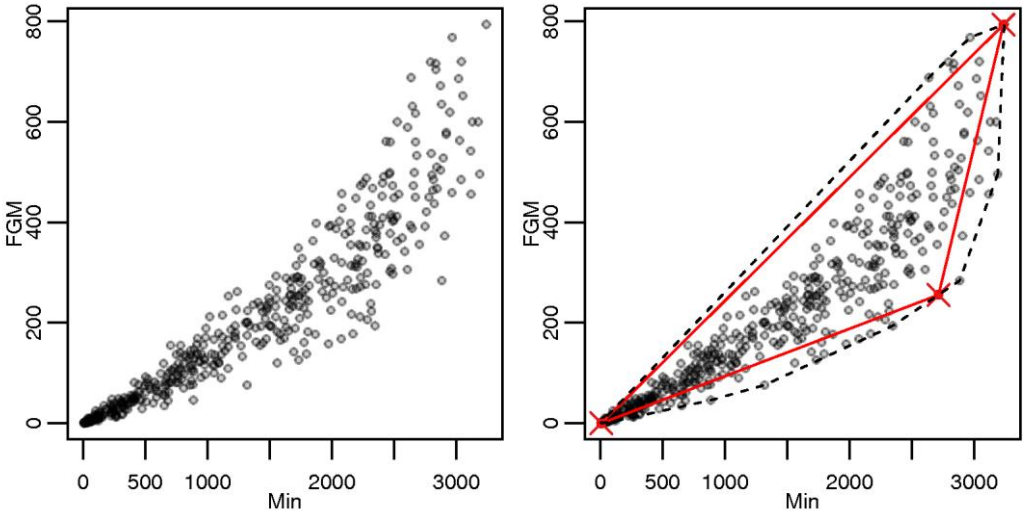

Figure 1: (a) Data set of two NBA player statistics from the season 2009/2010. (b) Convex hull (black, dashed) and the corresponding three archetypes solution (red, solid).

3. Performance profiling through archetypal analysis

In order to illustrate performance profiling based on archetypal athletes, we use a two-dimensional subset of the NBA player statistics from the season 2009/2010 which we analyze in detail in Section 5.1: the two performance indicators are *total minutes played* (*Min*) and *field goals made* (*FGM*) of 441 players, i.e., we investigate "the score efficiency". Figure 1 shows the data set. 42.86% of the players are in the range [0,1000] of *Min* and [0,200] of *FGM*. With increasing *Min*, the variance in *FGM* increases and the shape of the data set suggests the estimation of three archetypes. Figure 1 visualizes the k=3 archetypes solution (red, solid), together with the data's convex hull boundary (black, dashed). We see that this archetypes solution is a reasonable approximation of the convex hull (note that the archetypes do not have to be observed data points). Using this solution, the data points inside the archetypes solution are exactly approximated, the data points outside the archetypes solution are approximated with an error, as they are projected on the convex hull boundary of the archetypes solution.

The three archetypal athletes can be interpreted as follows. Concerning these two performance indicators *total minutes played* (*MIN*) and *field goals made* (*FGM*), three types of extreme scorers are detected based on the given data points:

Archetype 1 is the natural "maximum" with high values in both the performance indicators (*Min*=3234, *FGM*=793); this archetype represents a type of "*good*" scorer.

Archetype 2 is the natural "minimum" with low values in both the performance indicators (*Min*=7, *FGM*=0); this archetype represents a type of "*bad*" scorer.

Archetype 3 is another extreme value with a high number of *Min* but a (relatively) low number of *FGM* (*Min*=2713, *FGM*=256); this archetype represents another type of "*bad*" scorer (i.e., an ineffective scorer).

Note that the obtained archetypal solution and interpretation is dependent on the selection of the value of *k*. In fact, solutions are not nested for different values of *k*, i.e., the *k* archetypal athletes are not necessarily a subset of the *k*+1 archetypal athletes. A further important aspect of the interpretation is, that the performance indicators under investigation may only serve as a proxy for other performance indicators or player characteristics that are not available. For example, the number *field goals made* obviously is related to the position and tactical orientation of a player, but these information are not available in this illustrative data set and therefore cannot contribute to the interpretation of "*good*" and "*bad*" players.

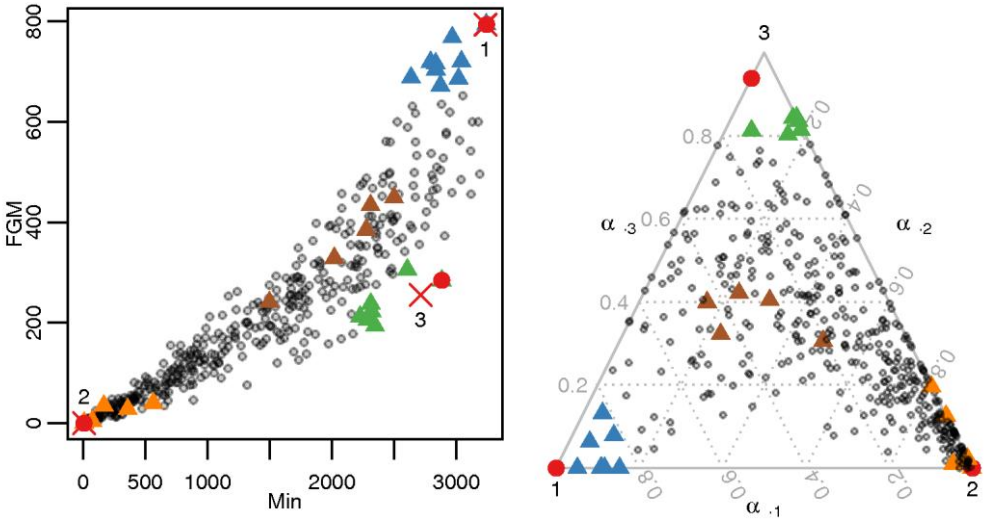

Figure 2: Visualizations of (a) the data set in case of the *k*=3 archetypes solution and (b) of the α coefficients using a ternary plot. The red points are the archetypes' nearest players; triangles colored with blue (near Archetype 1), orange (near Archetype 2), and green (near Archetype 3) are players where Archetype 1, 2, and 3 contribute more than 0.8.

Now, having identified the possible archetypal players within the given data set, the next step is to set the players in relation to them. The α coefficients of the archetypal problem (Formula 1) define how much each archetype contributes to the approximation of each individual player (as convex combination). In reversal this is the composition of each individual player based on the archetypal players—and therefore a reasonable performance profile.

The simplest interpretation of the α coefficients is the assignment of the players to their nearest archetypes and, consequently, the identification of the most archetypal observation(s). To support such an interpretation, Figure 2 shows visualizations of (a) the data set with the *k*=3 archetypes solution and (b) the corresponding α coefficients using a ternary plot. A ternary plot visualizes the membership in three groups (here the archetypes) by plotting the membership proportions (here the α coefficients) which sum to 1. Each side of the triangle represents one of the coefficients $\alpha_{\cdot 1}$, $\alpha_{\cdot 1}$, $\alpha_{\cdot 3}$; and the coordinates of a

point are defined by the gravity center of mass points interpreting the α coefficients as weights (Friendly, 2000, note that this plot is also known as trilinear plot).

In the data scatter plot (Figure 2a) the computed archetypes are shown as red crosses labeled with the archetypes' numbers; which are also the corners of the α coefficients' ternary plot (Figure 2b). In both figures the three players nearest to the respective archetypes are highlighted with red points. The three players are shown in Table 1.

Table 1. Data and performance profiles of the three players nearest to the respective archetypes.

	Name	Team	Position	Min	FGM	$\alpha_{\cdot 1}$	$\alpha_{\cdot 2}$	$\alpha_{\cdot 3}$
Archetype 1	Kevin Durant	OKL	SF	3241	794	1.00	0.00	0.00
Archetype 2	Dwayne Jones	PHO	C	7	0	0.00	1.00	0.00
Archetype 3	Jason Kidd	DAL	PG	2883	284	0.06	0.00	0.94

Note the corresponding performance profiles $\alpha_{\cdot 1}$, $\alpha_{\cdot 2}$ and $\alpha_{\cdot 3}$; as the selected players are the nearest players to one of the archetypal athletes, one coefficient is always near to one and all others near to zero. Archetype 1 and 2 have well-defined nearest players; Archetype 3, on the contrary, has a set of nearest players and the presented player identification should be considered as a "random" selection from the set of similar players.

We have identified Archetype 1 as the "*good*" archetypal athlete in this data setting—on this account, Kevin Durant can be considered as the best scorer. We can now use this as a tool to find other good scorers by defining a minimum threshold for $\alpha_{\cdot 1}$. Based on the specific task, this threshold (1) can be an already known value which holds for good scorers (domain knowledge, e.g., detected by analyses of previous seasons), (2) can represent the subjective opinion of the analyst (expert knowledge, e.g., when looking for a substitute), (3) or can simply be an arbitrary value (if used as a purely exploratory tool). The threshold obviously ranges between 1.0 (the athletes equal to the archetypal athlete) and 0.0 (all athletes), whereas values greater than 0.5 lead to either a unique or no assignment. Here, we use archetypal analysis as a purely exploratory tool—the players and their performance profiles where Archetype 1 contributes, for example, more than 0.8 (blue triangles near Archetype 1) are shown in Table 2.

Table 2: Data and performance profiles of players where Archetype 1 contributes more than 0.8.

Name	Team	Position	Min	FGM	$\alpha_{\cdot 1}$	$\alpha_{\cdot 2}$	$\alpha_{\cdot 3}$
Kevin Durant	OKL	SF	3241	794	1.00	0.00	0.00
Lebron James	CLE	SF	2967	768	0.95	0.05	0.00
Kobe Bryant	LAL	SG	2834	716	0.89	0.11	0.00
Dwyane Wade	MIA	SG	2793	719	0.89	0.11	0.00
Dirk Nowitzki	DAL	PF	3041	720	0.89	0.05	0.06
Amare Stoudemire	PHO	PF	2835	704	0.88	0.12	0.00
Carmelo Anthony	DEN	SF	2636	688	0.85	0.15	0.00
David Lee	NYK	C	3018	686	0.82	0.05	0.13
Derrick Rose	CHI	PG	2872	672	0.82	0.10	0.08

Note for example the performance profile of Dirk Nowitzki: Archetype 1 (the "*good*" scorer) contributes 0.89, Archetype 2 (the "*bad*" scorer) contributes 0.05, and Archetype 3 (the "*ineffective* scorer) contributes 0.06. We equivalently proceed for the other two archetypes: Players where the "*ineffective*" scorer Archetype 3 contributes more than 0.8 are Jason Kidd, Thabo Sefolosha, Earl Watson, Anthony Parker, Derek Fisher, Ron Artest, Marcus Camby (green triangles near Archetype 3). The "*bad*" scorer Archetype 2 contributes more than 0.8 to 118 players; five randomly selected players are Ryan Bowen, Sean Marks, Ian Mahinmi, Jamaal Magloire, Quinton Ross (orange triangles near Archetype 2).

Observations toward the center of the data set are not approximated by one archetype alone, but each archetype contributes a significant fraction. Table 3 shows five players that are randomly selected from the data sets' center (brown triangles in the center of the scatter and ternary plots).

Table 3: Data and performance profiles of five randomly selected players from the data sets' center.

Name	Team	Position	Min	FGM	$\alpha_{\cdot 1}$	$\alpha_{\cdot 2}$	$\alpha_{\cdot 3}$
Vince Carter	ORL	SG	2310	434	0.44	0.23	0.32
Anthony Morrow	GSW	SG	2019	329	0.28	0.31	0.40
C.j. Miles	UTA	SF	1497	241	0.21	0.49	0.31
Paul Millsap	UTA	PF	2275	385	0.35	0.23	0.42
Rodney Stuckey	DET	PG	2499	449	0.44	0.16	0.40

Based on the defined thresholds 0.8 for all three archetypes, no assignment to one of the archetypes is possible.

Besides this kind of analysis—the interpretation of highest α above a defined threshold—the interpretation of all α's of an observation is of interest as well. Suppose that, for example, the data set describes skill ratings of players, then the α's can be interpreted as the players' compositions of skills; see Section 5.2 for such an interpretation of performance profiles based on archetypal athletes.

4. Comparison with related methods

Archetypal analysis is part of the class of *k*-prototypes-like algorithms (*k*-means, *k*-median, fuzzy *k*-means, etc.; see, e.g., Steinley, 2006; Leisch, 2006). These are partitioning methods for segmenting a data set into *k* non-empty clusters. Each cluster is represented by a cluster center. In addition, coefficients indicate the degree of membership for each observation in each cluster. In general, the membership coefficients sum up to 1, whereas the possible values are dependent on the method. *k*-means, for example, is a hard clustering method and requires that each observation belongs to exactly one cluster. Therefore, the possible membership values are {0,1}. Fuzzy *k*-means, on the other hand, is a soft clustering method and allows that an observation can belong to more than one cluster. The possible membership values range in [0,1]. For detailed descriptions of the methods we refer to Steinley (2006).

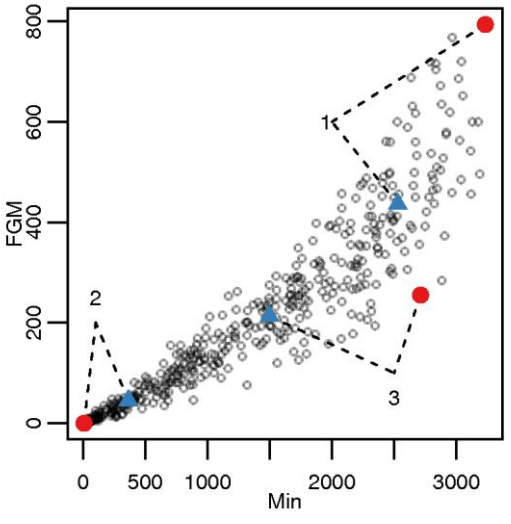

Figure 3: Visualization of the *k*=3 solutions for archetypal analysis (red points) and fuzzy *k*-means (blue triangles).

An obvious idea is to use cluster methods for performance profiling—the cluster centers are interpreted as prototypal observations, and the membership coefficients as performance profiles. Based on the used method we get performance profiles where the players are assigned to either one (e.g., *k*-means) or more (e.g., fuzzy *k*-means) prototypal athletes. In order to make a suitable comparison we focus on the latter one. Figure 3 shows the fuzzy *k*-

means solution with k=3 for the two-dimensional subset of the NBA player statistics from the season 2009/2010 (blue triangles). The archetypal solution discussed in the previous section is shown as well (red points). In our point of view, three major problems make common clustering methods less suitable for performance profiling than archetypal analysis.

The first major problem is the nonexistence of real clusters in the data. As we can see in Figure 3, there is no segmentation structure available. And—based on our experience—this circumstance is true for the majority of cases in sports data (however, we are not aware of any structured analysis of this statement). This, in fact, holds the danger of getting a random solution, rather than a reliable solution (Dolnicar and Leisch, 2010, discuss this problem for market segmentation). A similar problem can occur for archetypal analysis if the data are a perfectly round—this, however, is very unlikely for real data.

The second major problem is the nature of the prototypes and, consequently, the interpretation of the prototypes. In case of archetypal analysis the computed archetypes are extreme values on the boundary of the convex hull. This allows a distinct interpretation. The prototypes, on the other hand, are the mean, median, etc. values (based on the used clustering method) of the corresponding clusters and typically lie within the data set (as we can see in Figure 3). A clean interpretation is not possible anymore. For example, the values of the archetypes and prototypes seen in Figure 3 are shown in Table 4.

Table 4: Values of (a) the archetypes and (b) the prototypes.

	Min	FGM
Archetype 1	3234	793
Archetype 2	7	0
Archetype 3	2713	256

	Min	FGM
Prototype 1	2529	436
Prototype 2	363	45
Prototype 3	1497	213

We can only draw a very vague interpretation of the three prototypal scorers: Prototype 1 is the "*better*" scorer, Prototype 2 is the "*not so good*" scorer, and Prototype 3 is the "*acceptable*" scorer. These prototypal scorer are not really useful for performance profiling.

In consequence of the second problem, the third major problem is the interpretation of the membership coefficients. For example, we considered Kevin Durant as the best scorer in Section 3, his performance profile based on the fuzzy k-means prototypes is shown in Table 5.

Table 5: Data and prototype-based performance profile of the best scorer (see Section 3).

Name	Team	Position	Min	FGM	$u_{\cdot 1}$	$u_{\cdot 2}$	$u_{\cdot 3}$
Kevin Durant	OKL	SF	3241	794	0.79	0.06	0.15

From this profile we only see that he is for the most part a "*better*" scorer. In summary, the vague interpretations of the prototypal athletes and the performance profiles make

performance profiling through cluster-based prototypes very imprecise. This problem gets more and more relevant the higher the number of performance indicators is—and where no easy interpretation in the context of the data set via a visualization is possible anymore.

5. Profiling basketball and soccer players

Using archetypal analysis for performance analysis in sports provides two interesting results: (1) k extremal (positive and/or negative) performers , i.e., the archetypal athletes, and (2) a performance profile α for each performer defining its composition based on the archetypal athletes. This enables performance profiling based on extreme performances—typically of great interest in analyzing sport performances.

In this section we determine archetypal athletes and performance profiles for two popular sports and their representative leagues. Section 5.1 extends the illustrative two-dimensional example and computes archetypal basketball players and corresponding performance profiles with common statistics from the NBA season 2009/2010. Section 5.2 determines archetypal soccer players and performance profiles of the German Bundesliga, the English Premier League, the Italian Lega Serie A, and the Spanish La Liga using skill ratings (at the time of September 2011).

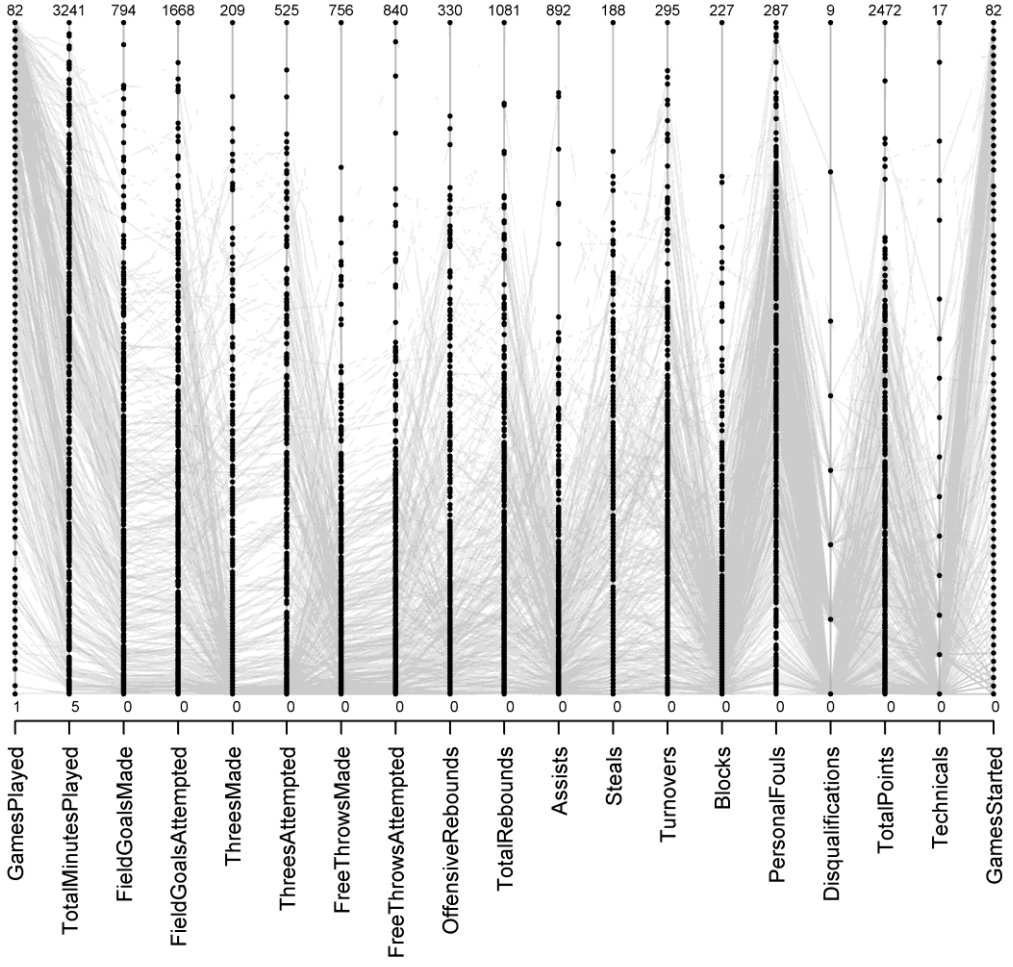

Figure 4: Parallel coordinates plot of the performance indicators of 441 players from the NBA season 2009/2011.

Table 6: Five-number summary plus mean and standard deviation of the performance indicators of 441 players from the NBA season 2009/2011.

Statistic	0%	25%	50%	75%	100%	Mean	SD
GamesPlayed	1	38	65	76	82	56.25	24.02
TotalMinutesPlayed	5	502	1318	2163	3241	1348.15	919.84
FieldGoalsMade	0	56	171	314	794	210.27	177.13
FieldGoalsAttempted	0	131	388	688	1668	455.74	373.36
ThreesMade	0	0	13	63	209	35.88	45.58
ThreesAttempted	0	3	43	178	525	101.19	121.00
FreeThrowsMade	0	22	68	149	756	103.90	111.68
FreeThrowsAttempted	0	31	93	200	840	136.91	140.32
OffensiveRebounds	0	16	39	82	330	61.11	62.46
TotalRebounds	0	79	178	343	1081	232.73	204.04
Assists	0	21	72	161	892	118.51	137.01
Steals	0	12	33	60	188	40.26	34.02
Turnovers	0	22	63	112	295	75.65	62.83
Blocks	0	5	14	36	227	27.09	33.19
PersonalFouls	0	54	115	174	287	116.35	73.01
Disqualifications	0	0	0	1	9	0.73	1.24
TotalPoints	0	146	453	841	2472	560.32	479.47
Technicals	0	0	1	2	17	1.66	2.65
GamesStarted	0	1	12	58	82	27.89	30.37

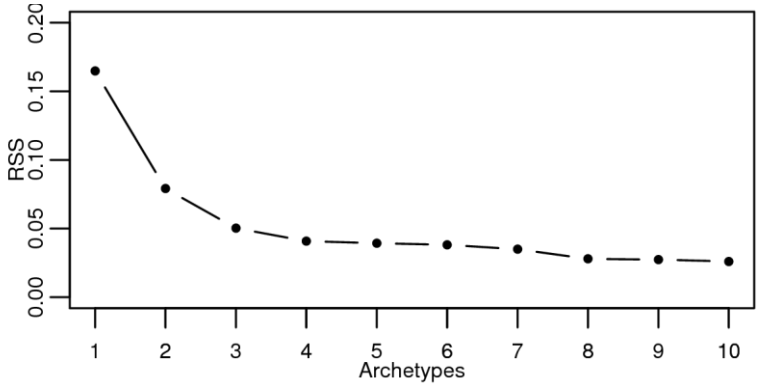

Figure 5: Scree plot of the residual sum of squares for 1 to 10 archetypal basketball players.

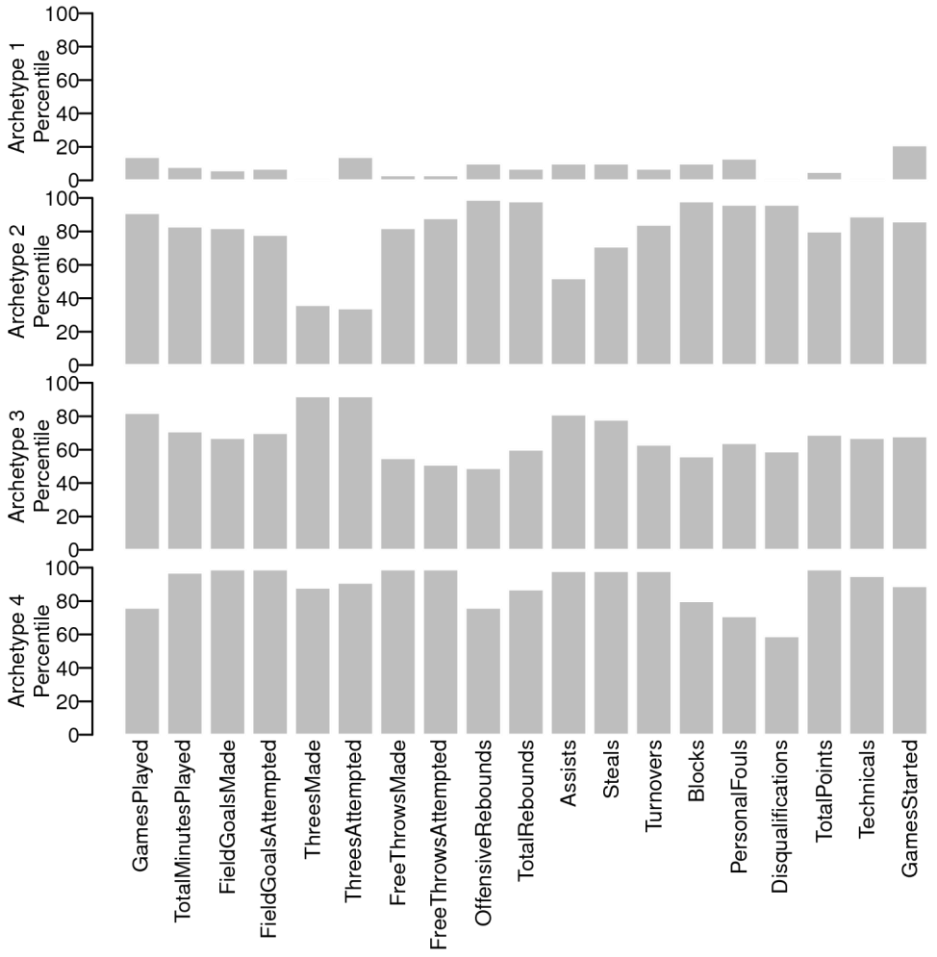

Figure 6: Percentile plot of the four archetypal basketball players solution.

5.1. Profiling basketball players

We determine the archetypal basketball players and the players' performance profiles of the NBA season 2009/2010. Kubatko et al. (2007) define basic statistics used in what is now the mainstream of basketball statistics. Following their suggestion we use a data set provided by Steele (2011) with 19 statistics of 441 players as database of performance indicators. In order to have some evidence of reliability of the data we randomly cross checked with the player statistics provided by ESPN NBA (ESPN.com, 2012). Table 6 shows the five-number summary plus the mean and the standard deviation for each of the statistics.

Figure 4 visualizes the data using a parallel coordinates plot. Parallel coordinates (introduced by Inselberg, 1985) is a common way to visualize high dimensional data sets (see, e.g., Wegman, 1990). For an *m* dimensional data set *A*, the coordinate system consists of *m* equally spaced parallel lines. An observation $a_{i.}$ is represented as a polyline whereas the position of $a_{i.}$'s polyline on the parallel axis *j* is the value a_{ij}. For readability the

j=1,…,m variables are often scaled. In Figure 4 we also indicate the observations' values using additional points on the axes. In comparison to the two-dimensional illustrative example no structure is easily observable; and there is, for example, no player which is the maximum over all statistics. We fit k=1,…,10 archetypes; Figure 5 shows the corresponding scree plot: the first "elbow" is at k=4 (*RSS*=0.04), the second one at k=8 (*RSS*=0.03). The additional error reduction between k=4 and k=8 is marginal and we decide on k=4 archetypal basketball players.

Figure 6 displays the percentile plots of the four archetypal basketball players available in the NBA season 2009/2010. For each archetype it shows the percentile value of each statistic as compared to the data. In case of the statistic *GamesPlayed*, for example, the height of the bar in Archetype 1 is 14 and in Archetype 2 the height is 91. This indicates that the *GamesPlayed* value in Archetype 1 is in the 14nd percentile and in Archetype 2 in the 91nd percentile of the data. We use this plot to establish the particular characteristics of the four archetypal basketball players:

Archetype 1 is the archetypal "benchwarmer" with few games played and therefore low values in all statistics.

Archetype 2 is the archetypal rebounder and defensive player with high values in the rebounds, blocks and foul-related statistics, and low values in the three-pointers.

Archetype 3 is the archetypal three-point shooter with high values in the three-pointer statistics and low values in the free throws and rebounds.

Archetype 4 is the archetypal offensive player with high values in all throw-related statistics and low values in foul-related statistics.

Archetype 1 represents a type of "*bad*" basketball player while all others represent different types of "*good*" players. The four basketball player nearest to one of the four archetypes are shown in Table 7.

Table 7: Data and performance profiles of the four players nearest to the respective archetypes.

	Name	Team	Position	$\alpha_{\cdot 1}$	$\alpha_{\cdot 2}$	$\alpha_{\cdot 3}$	$\alpha_{\cdot 4}$
Archetype 1	Dwayne Jones	PHO	C	1.00	0.00	0.00	0.00
Archetype 2	Taj Gibson	CHI	SF	0.00	1.00	0.00	0.00
Archetype 3	Anthony Morrow	GSW	SG	0.00	0.00	0.96	0.04
Archetype 4	Kevin Durant	OKL	SF	0.00	0.00	0.00	1.00

On this account, Taj Gibson, Anthony Morrow, Kevin Durant can be considered as the best basketball players of the season 2009/2010 with respect to the characteristics of their corresponding nearest archetypes. However, note that in case of Archetype 3 the player is not exactly the archetype. In order to find further good players, we look at the players' performance profiles $\alpha_{\cdot 1}$, $\alpha_{\cdot 2}$, $\alpha_{\cdot 3}$, and $\alpha_{\cdot 4}$ which are near to one of the three "*good*"

archetypes. As the highest value for $\alpha_{\cdot 3}$ is 0.96 we define 0.95 as a reasonable minimum threshold; the selected players are shown in Table 8.

Table 8: Data and performance profiles of players where the three "*good*" archetypes contribute more than 0.95.

Archetype	Name	Team	Position	$\alpha_{\cdot 1}$	$\alpha_{\cdot 2}$	$\alpha_{\cdot 3}$	$\alpha_{\cdot 4}$
Archetype 2	Taj Gibson	CHI	SF	0.00	1.00	0.00	0.00
	Andrew Bogut	MIL	C	0.00	1.00	0.00	0.00
	Samuel Dalembert	PHI	C	0.02	0.98	0.00	0.00
	Jason Thompson	SAC	PF	0.03	0.96	0.00	0.00
Archetype 3	Anthony Morrow	GSW	SG	0.00	0.00	0.96	0.04
	Steve Blake	LAC	PG	0.02	0.00	0.96	0.02
Archetype 4	Kevin Durant	OKL	SF	0.00	0.00	0.00	1.00
	Lebron James	CLE	SF	0.00	0.00	0.00	1.00
	Dwyane Wade	MIA	SG	0.00	0.00	0.00	1.00
	Kobe Bryant	LAL	SG	0.03	0.00	0.00	0.97

The equal coefficients for the first two players in case of Archetype 2 occur due to rounding to two decimal places. Note that we are interested in an exploratory analysis and the threshold 0.95 is not based on previous studies or expert knowledge. It is a starting point and in further analysis steps we can, for example, lower the minimum threshold and investigate the distribution of players around the "*good*" archetypes. However, this minimum threshold is, in fact, the only subjective decision one has to make when discussing the quality of performers using performance profiles based on archetypal athletes.

5.2 Profiling soccer players

We determine the archetypal soccer players and performance profiles of soccer players playing in four European top leagues. The skill ratings are from the PES Stats Database (2011, PSD), a community-based approach to create a database with accurate statistics and skill ratings for soccer players (originally for the video game "Pro Evolution Soccer" by Konami). The extracted data set consists of 25 skills (performance indicators) of 1658 players (all positions—Defender, Midfielder, Forward—except Goalkeepers) from the German Bundesliga, the English Premier League, the Italian Serie A, and the Spanish La Liga. The skills are rated from 0 to 100 and describe different abilities of the players: physical abilities like balance, stamina, and top speed; ball skills like dribble, pass, and shot accuracy and speed; and general skills like attack and defence performance, technique, aggression, and teamwork. Note that we assume that the differences are interpretable, i.e., the ratings are on a ratio scale.

Table 9 shows a summary of the data set. Most skills range between 50 and 100; this is due to the fact that PSD describes soccer players of all hierarchy levels of a league system. We fit k=1,…,15 archetypes and decide to use k=4 archetypal soccer players. The complete

decision process (with visualization of the data, scree plot, etc.) is available in the online supplement (Section 8 on computational details explains how to reproduce the results).

Table 9: Five-number summary plus mean and standard deviation of the performance indicators of 1658 players from the German Bundesliga, the English Premier League, the Italian Serie A, and the Spanish La Liga.

Skill	0%	25%	50%	75%	100%	Mean	SD
Attack	51	68	73	78	95	72.79	6.91
Defence	30	47	65	74	94	61.34	15.18
Balance	69	78	81	84	96	81.15	4.01
Stamina	76	82	83	85	96	83.72	2.51
TopSpeed	72	79	82	84	97	81.37	3.59
Acceleration	70	78	82	84	98	81.51	4.08
Response	73	78	80	82	96	80.46	2.69
Agility	68	77	80	83	97	80.34	4.15
DribbleAccuracy	65	76	78	82	97	78.50	4.48
DribbleSpeed	66	77	80	83	97	79.97	4.41
ShortPassAccuracy	67	73	75	78	96	75.44	3.93
ShortPassSpeed	67	73	75	78	96	75.92	3.58
LongPassAccuracy	62	73	76	79	96	75.88	4.27
LongPassSpeed	65	74	77	80	96	76.97	3.97
ShotAccuracy	58	67	71	76	94	71.62	5.59
ShotPower	72	81	82	83	95	82.16	2.52
ShotTechnique	54	68	73	78	95	73.27	6.23
FreeKickAccuracy	50	65	68	74	89	69.86	5.77
Curling	52	71	75	79	95	75.07	5.27
Header	63	72	76	81	95	76.06	5.57
Jump	65	76	78	82	96	78.74	3.87
Technique	67	76	79	82	97	79.15	4.33
Aggression	55	72	80	83	99	77.51	7.49
Mentality	68	77	80	83	95	79.92	3.93
Teamwork	67	77	79	81	98	79.10	3.18

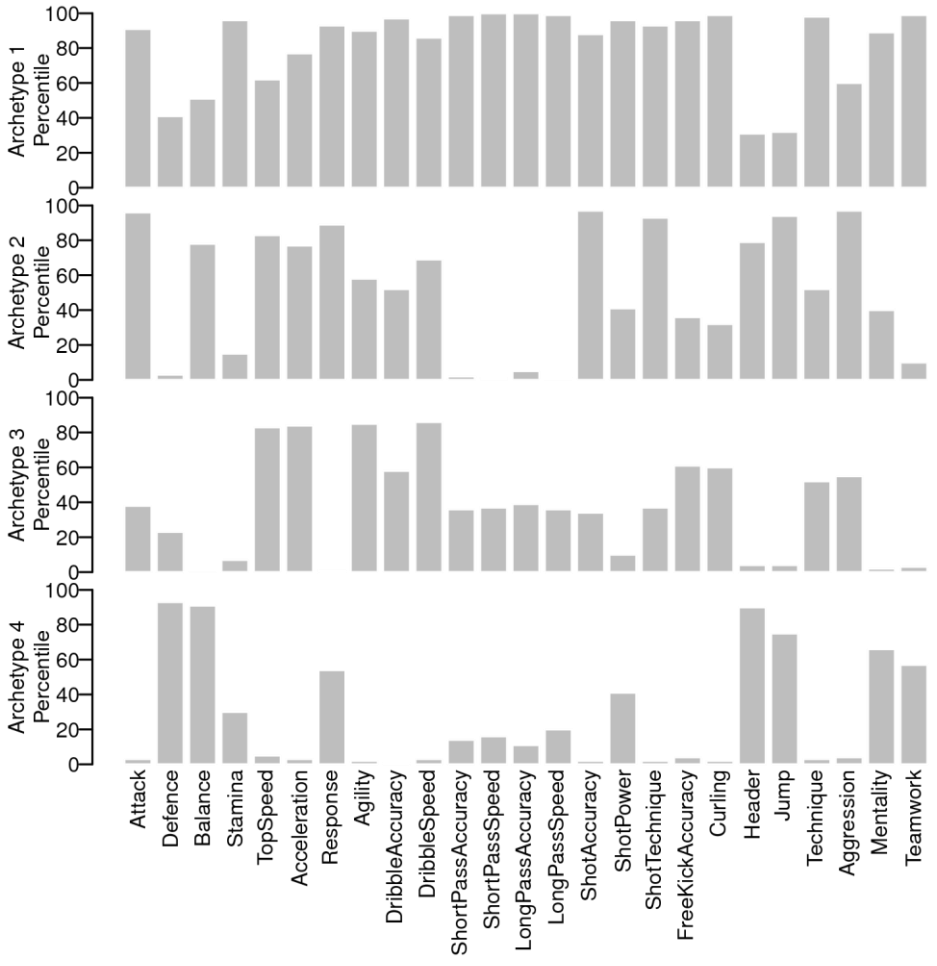

Figure 7: Percentile plot of the four archetypal soccer players solution.

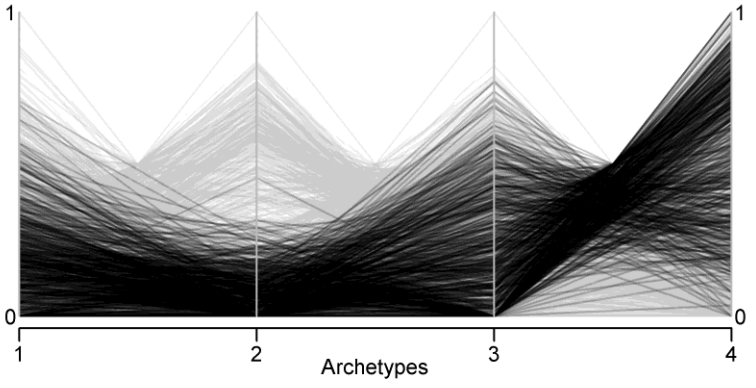

Figure 8: Parallel coordinates plot of α coefficients of the four archetypal soccer players solution with highlighted defenders (black).

Figure 7 displays the percentile plots of the four archetypal soccer players. The particular characteristical skills are:

Archetype 1 is the archetypal offensive player with all skills high except the defense, balance, header, and jump.

Archetype 2 is the archetypal center forward with high skills in attack, shot, acceleration, header and jump, and low passing skills.

Archetype 3 is the archetypal weak soccer player with high skills in running, but low skills in most ball related skills.

Archetype 4 is the archetypal defender with high skills in defense, balance, header, and jump.

To verify this interpretation we look at the α coefficients (the performance profiles) in combination with the players' position; Figure 8 exemplarily shows the parallel coordinates plot of the performance profiles with the "Defender" position highlighted (black). As we can see, most of the defenders have a high α coefficient for Archetype 4.

The performance profiles now can be used to find and to compare players. If we are, for example, looking for a defender with offensive qualities, we are in fact looking for a player composed of Archetype 4 (the defender) and Archetype 1 (the offensive). In numbers we can express such a query for example with a performance profile with $\alpha_{\cdot 4}>0.65$ and $\alpha_{\cdot 1}>0.33$; the selected players are shown in Table 10.

Table 10: Data and performance profiles of players which we consider as defenders with offensive qualities.

Name	Team	$\alpha_{\cdot 1}$	$\alpha_{\cdot 2}$	$\alpha_{\cdot 3}$	$\alpha_{\cdot 4}$
Lorik Cana	SS Lazio	0.34	0.00	0.00	0.66
Mats Hummels	Borussia Dortmund	0.33	0.00	0.00	0.67
John Terry	Chelsea FC	0.33	0.00	0.00	0.67

The selection is then a set of players with characteristics described by the corresponding fractions of the archetypal soccer players. Note that this can be especially useful for talent profiling.

Performance profiles based on archetypal athletes also allow to investigate the question of the best soccer player. In order to do this, we have to make a (subjective) definition of "*the best*" in terms of the four archetypes. For us, the best player is a combination of Archetype 1 and Archetype 2 with Archetype 1 contributing more than Archetype 2 (according to the common sense that offensive players are match-winning). Table 11 shows soccer players who apply to the definition (ordered according to $\alpha_{\cdot 1}$).

Table 11: Data and performance profiles of players which we consider as "*the best*" players.

Name	Team	$\alpha_{\cdot 1}$	$\alpha_{\cdot 2}$	$\alpha_{\cdot 3}$	$\alpha_{\cdot 4}$
Wayne Rooney	Manchester United FC	0.82	0.18	0.00	0.00
Leo Messi	FC Barcelona	0.79	0.21	0.00	0.00
Cristiano Ronaldo	Real Madrid CF	0.68	0.32	0.00	0.00
Antonio Di Natale	Udinese Calcio	0.67	0.33	0.00	0.00
Carlos Tivez	Manchester City FC	0.66	0.34	0.00	0.00
Diego Forlan	FC Internazionale Milano	0.64	0.36	0.00	0.00
Dimitar Berbatov	Manchester United FC	0.60	0.40	0.00	0.00
Adrian Mutu	AC Cesena	0.60	0.40	0.00	0.00
Zlatan Ibrahimovic	AC Milan	0.54	0.46	0.00	0.00
Luis Suarez	Liverpool FC	0.53	0.47	0.00	0.00
Mladen Petric	Hamburger SV	0.53	0.47	0.00	0.00
Xavi Hernandez	FC Barcelona	0.52	0.48	0.00	0.00
Didier Drogba	Chelsea FC	0.52	0.48	0.00	0.00
Giuseppe Rossi	Villarreal CF	0.51	0.49	0.00	0.00

Based on such definition the best player is Wayne Rooney, followed by Leo Messi and Christiano Ronaldo. The performance profiles $\alpha_{\cdot 1}$, $\alpha_{\cdot 2}$, $\alpha_{\cdot 3}$, and $\alpha_{\cdot 4}$ show the players' composition; note that after the first half of the table the values between $\alpha_{\cdot 1}$ and $\alpha_{\cdot 2}$ are nearly balanced. When we interpret the performance profiles, we see for example, that Wayne Rooney is mostly described by Archetype 1—all skills high except defence, balance, header, and jump. He gets some header and jump skills from Archetype 2, but stays low in the defence skills.

6. Discussion and summary

The present paper introduces performance profiles based on archetypal athletes. Archetypal athletes are the result of a statistical methodology called archetypal analysis. The archetypal athletes are data-driven extreme values, i.e., (artificial) performers lying on the boundary of the data set defined by the performance indicators. Additionally, for each performer a performance profile defining its composition based on the archetypal athletes is computed. The proposed profiling way is (1) to estimate the archetypal athletes, then (2) to identify and characterize the athletes as different types of "*good*" and "*bad*" athletes, and finally (3) to set all performers in relation to the archetypes using the α coefficients as performance profiles. This strategy—introduced by Porzio et al. (2008) for finding benchmark performers in business performance analysis—is in fact a general recipe for using archetypal analysis for performance analysis in any application field.

The purpose of the proposed method is not to replace well-known profiling methods (e.g., Hughes et al., 2001; James et al., 2005; O'Donoghue, 2005), but to provide a further tool to

profile from a different point of view. The key aspects of performance profiles based on archetypal athletes are:

1. The method is based on extreme performances, usually the most interesting fact in sports.
2. The archetypal athletes define an interpretable basis for the performance profiles.
3. The complexity of the performance profiles is reduced because usually less archetypal athletes are computed than the number of performance indicators.
4. The results are dependent on two decisions made by the analyst:
   (a) The number of archetypes *k* used (supported, for example, by a scree plot).
   (b) The threshold values used for analyzing the performance profiles (supported by domain or expert knowledge, or arbitrary if used as an exploratory tool).

These aspects make the method interesting for profiling (finding and comparing) performers in a very interpretable way.

Future work contains the integration and consideration of performers' variability. The presented examples are based on summarized performance indicators, i.e, one value per performance indicator for each performer. One idea to express the possible variability of performers between individual matches is to compute performance profiles based on archetypal analysis for each match and compare the match-related archetypal athletes as well as the match-related performance profiles of individual performers.

Furthermore, there are connecting factors to (at least) two interesting topics which we want to investigate. First, the α coefficients form a dual space with barycentric coordinates (see, e.g., Coxeter, 1989). This could be of (theoretical) interest when comparing archetypal analysis with other *k*-prototypes-like methods (e.g., Fuzzy *k*-means). Second, the α coefficients can be interpreted as compositional data. This introduces appropriate norm and distance measures and, in further consequence, enables enhanced statistical analyses, e.g., statistical testing and principal component analysis (see, e.g., Aitchison, 1982; Pawlowsky-Glahn and Buccianti, 2011).

7. Acknowledgements

The author thanks Stephan Hable for data checking in case of the NBA season 2009/2010 data set. The author also thanks the reviewers for very helpful comments.

8. Computational details

All computations and graphics have been done using the statistical software R 2.13.1 (R Development Core Team, 2011), the archetypes package (Eugster, 2010), and the SportsAnalytics package (Eugster, 2011). R itself and all packages used are freely available under the terms of the General Public License from the Comprehensive R Archive Network at http://CRAN.R-project.org/.

Data sets and source codes for replicating our analyses are available in the SportsAnalytics package. An individual analysis is executed via (replace `***` with `nba-2d`, `nba` and `soccer`):

```
R> demo("archeplayers-***", package = "SportsAnalytics")
```

The source code file for a demo is accessible via:

```
R> edit(file = system.file("demo", "archeplayers-***.R",
+                          package = "SportsAnalytics"))
```

9. References

Aitchison, J. (1982), The statistical analysis of compositional data. **Journal of the Royal Statistical Society, Series B** (Statistical Methodology), 44 (2): 139–177.

Bauckhage, C. and Thurau, C. (2009), Making archetypal analysis practical. In Proceedings of the 31st DAGM Symposium on Pattern Recognition, 272–281.

Chan, B.H.P., Mitchell, D.A. and Cram, L.E. (2003), Archetypal analysis of galaxy spectra. Monthly Notice of the Royal Astronomical Society, 338: 790–795.

Coxeter, H.S.M. (1989), **Introduction to Geometry**. NY: John Wiley & Sons.

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

Dolnicar, S. and Leisch, F. (2010), Evaluation of structure and reproducibility of cluster solutions using the bootstrap. **Marketing Letters**, 21: 83–101.

Eugster, M.J.A. (2010) archetypes: Archetypal Analysis, URL http://cran.r-project.org/package=archetypes. R package version 2.0-2.

Eugster, M.J.A. (2011), SportsAnalytics: Sports Analytics, URL http://cran.r-project.org/package=SportsAnalytics. R package version 0.1.

Eugster, M.J.A. and Leisch, F. (2009), From Spider-man to Hero – Archetypal analysis in R. **Journal of Statistical Software**, 30 (8): 1–23.

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

Friendly, M. (2000), Visualizing Categorical Data. SAS Institute.

Golub, G.H. and Van Loan, C.F. (1996), Matrix Computations. Johns Hopkins University Press.

Hughes, M.D. and Bartlett, R.M. (2002), The use of performance indicators in performance analysis. **Journal of Sports Sciences**, 20 (10): 739–754.

Hughes, M.D., Evans, S. and Wells, J. (2001), Establishing normative profiles in performance analysis. **International Journal of Performance Analysis in Sport**, 1 (1): 1–26.

Inselberg, A. (1985), The plane with parallel coordinates. **Visual Computer**, 1: 69–91.

James, N., Mellalieu, S. and Jones, N. (2005), The development of position-specific performance indicators in professional rugby union. **Journal of Sports Sciences**, 23 (1): 63–72.

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

Leisch, F. (2005), A toolbox for k-centroids cluster analysis. **Computational Statistics and Data Analysis**, 51 (2): 526–544.

Wang, S.L.P., Louviere, J. and Carson, R. (2003), Archetypal analysis: A new way to segment markets based on extreme individuals. In **A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution**. Proceedings of the ANZMAC 2003 Conference, December 1-3, 2003, pp. 1674–1679.

O'Donoghue, P.G. (2005). Normative profiles of sports performance. **International Journal of Performance Analysis in Sport**, 5 (1): 104–119.

O'Donoghue, P.G. (2009). Interacting performances theory. **International Journal of Performance Analysis in Sport**, 9 (1): 26–46.

O'Donoghue, P.G. and Cullinane, A. (2011), A regression-based approach to interpreting sports performance. **International Journal of Performance Analysis in Sport**, 11 (2): 295–307.

Pawlowsky-Glahn, V. and Buccianti, A. (2011), **Compositional Data Analysis: Theory and Applications**. John Wiley & Sons.

PES Stats Database (2011). PSD – PES Stats Database. http://pesstatsdatabase.com/; visited on 10/12/2011.

Porzio, G.C., Ragozini, G. and Vistocco, D. (2008), On the use of archetypes as benchmarks. **Applied Stochastic Models in Business and Industry**, 24 (5): 419–437.

Preparata, F.P. and Shamos, M.I. (1990), **Computational Geometry: An Introduction**. Springer-Verlag, 1990.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Steele, D. (2011), Doug's NBA & MLB statistics home page. http://dougstats.com/; visited on 10/12/2011.

Steinley, D. (2006), k-means clustering: A half-century synthesis. **British Journal of Mathematical and Statistical Psychology**, 59(1), 1–34.

Wegman, E.J. (1990), Hyperdimensional data analysis using parallel coordinates. **Journal of the American Statistical Association**, 85 (411): 664–675.