Erasmus University Rotterdam

Clustering soccer players to find the drivers of soccer team performance

Master Thesis
Business Analytics & Quantitative Marketing

Econometrics & Operations Research
Erasmus School of Economics

Author:
Eline van de Ven
373298

Supervisor:
Dr. Andreas Alfons

April 12, 2018

ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

Contents

Abstract

In the professional soccer business, technical directors are under constant pressure to perform well with their teams. The choices made in assembling a team lineup are of crucial importance to achieve satisfying results. This paper aims to answer three interconnected research questions to investigate how team lineups in soccer matches should be formed to achieve positive match outcomes. By first assigning soccer players to different player types based on the outcomes of several cluster algorithms, the attributes distinguishing one soccer player from another are identified. Subsequently, the presence of combinations of the resulting player types are found with the use of rare correlated pattern mining. Finally, by estimating match outcomes with an ordered probit model, the effect of the presence of a certain combination of player types in the lineup of a team is investigated. The results of the match outcome estimation show that a good goalkeeper is particularly important for a soccer team in away matches. Another interesting finding is that although player types on their own can negatively influence the match outcome, a combination of them may positively influence the match outcome.

Keywords: Professional Soccer; Clustering; Pattern Mining; Ordered Probit Model

1 Introduction

What are the drivers of a good performing soccer team? This must be the vital question of which every soccer coach wants to know the answer. Several determinants of soccer performance have been proposed in the current literature. Ingersoll et al. (2013) for example, show that cultural diversity within a team leads to better performances in the world's top soccer league: the UEFA Champions League. The results of Franck and Nüesch (2010) indicate that talent disparity within the entire squad is beneficial for the team's league standing at the end of the season. Frick and Simmons (2008) found that hiring a better quality coach reduces technical inefficiency and can lead to higher league points. In this paper a new approach is proposed to discover the drivers of match results. The influence of the presence of particular types of players in team lineups on match results will be investigated.

In soccer there are many different types of players in terms of their technical strengths and weaknesses in the field. Different types of players may enhance each other's strengths resulting in better team performances. Alternatively, some types of players may be individually strong, but do not combine very well in the field, leading to poor team performances. First, this paper aims to identify what technical characteristics can be found to categorize players into player types in order to answer the following question:

Research Question 1 *How can players be categorized in terms of their technical strengths and weaknesses?*

Second, this paper aims to find combinations of player types that are present in the lineups of soccer teams, that significantly influence the match result. This will answer the second research question:

Research Question 2 *What combinations of player types that are present in professional soccer team lineups, significantly influence the match result?*

After having identified the combinations that influence the match results, the final goal is to investigate what combinations of player types lead to good team performances, i.e. good match results, to answer the main research question:

Main Research Question *What combinations of player types lead to good results in soccer matches?*

The answers to these questions are relevant for the billion dollar professional soccer business as it can help technical directors make choices in what types of players to invest in, given their current selection. Instead of investing in new players, the results can also be useful in determining what technical qualities of current players should be improved by extra attention during training sessions.

The current literature on soccer focuses mainly on predictions of match outcomes to test the efficiency of fixed odds betting markets (Kuypers (2000), Goddard and Asimakopoulos (2004), Graham and Stott (2008)) and on efficiency of the players and coaches (Frick and Simmons (2008), Kern and Süssmuth (2005)). This paper will contribute to the current literature by exploring the drivers of the match results rather than aiming to predict the results. In addition, this paper pioneers the use of detailed data about technical characteristics of individual players to estimate soccer match outcomes.

2 Literature

2.1 Player Clustering

The first part of this research consists of identifying different types of players by assigning players to different groups based on their technical strengths and weaknesses. Although there exist a vast amount of literature on soccer data, up to the best of my knowledge there has been no research conducted on the clustering of soccer players. Clustering techniques in general however, have received great attention with applications in numerous scientific fields, from the segmentation of images in computer vision (Shi and Malik (2000)) to the clustering of gene expression data in biology (Yeung and Ruzzo (2001)).

According to Rai and Singh (2010), by far the most popular clustering tool used in scientific and industrial applications is the K-means algorithm proposed by Hartigan and Wong (1979). Rai and Singh (2010) state that the most popular and simple algorithm K-means is still widely used in spite of the fact that it was proposed over 50 years ago and thousands of clustering algorithms have been published since then.

A stream of literature where the clustering of players has caught the interest of researchers can be found in the computer gaming research (Drachen et al. (2012)). For example, Drachen et al. (2014) compared different techniques to develop clusters of players in terms of their gaming behavior. They applied K-means, C-means, Non-negative Matrix Factorization (NMF), Principal Component Analysis (PCA) and Archetypal Analysis (AA) on game metrics data of players of the highly popular computer game World of Warcraft. Their results indicate the drawback of K-means and C-means that the clusters tend towards being similar, making it difficult to identify the differences between different clusters of players. Their results of NMF and PCA are intuitively hard to interpret in terms of the player behavior. They conclude that AA presents intuitively interpretable behavioral profiles and therefore is a potentially useful algorithm for the classification of player behavior.

To mitigate the curse of dimensionality, dimension reduction techniques are often applied when clustering large scale multivariate data. Although PCA is a popular technique for dimension reduction, one key drawback is the lack of sparseness in the solution, making the representation of the data hard to interpret. This drawback can be redressed with the use of Sparse Principal Component Analysis (SPCA). SPCA aims to find principal components that are constrained to have sparse loadings. This algorithm has been applied for clustering and feature selection on gene expression data and has proven to efficiently isolate relevant genes while maintaining most of the original clustering power of PCA (Luss and d'Aspremont (2010)).

This paper will contribute to the current literature by applying existing techniques such as K-means, Archetypal Analysis, and Sparse Principal Component Analysis to a new type of data.

2.2 Combinations of Player Types

The second part of this paper aims at finding combinations of types of players that significantly influence the match result. First, combinations of player types that are present in the lineups of soccer matches will be identified. We are looking for patterns like: if players from type X and Z are present in the lineup, than it is likely that also a player from type Y will be present in the lineup. A subfield of data mining concerned with finding correlations in binary relations is known as *correlated pattern mining.* This is often formulated as a *market basket* problem which was originally introduced by Agrawal et al. (1993) to find association rules between frequent items in a large database of customer transactions, a task also known as *frequent pattern mining.* For example, if customers are buying milk, then they are likely to also buy bread in the same transaction. Since its introduction, frequent pattern mining has found broad applications in for example, frequent term-based text clustering (Beil et al. (2002)) and finding frequently repeated paths in spatio-temporal data (Cao et al. (2005)).

As mining frequent patterns in large databases often results in a high number of generated association rules, including rules uninteresting to the user, several measures have been introduced in the literature to evaluate the interestingness of association rules (Tan et al. (2004)). The use of a measure of correlation to filter the extracted patterns and thus reduce the number of discovered rules results in *correlated pattern mining.*

Nevertheless, mining frequent correlated patterns in large databases may not be very attractive for some applications. As stated by Szathmary et al. (2010), the frequent patterns and rules targeted by conventional pattern mining reflect the globally valid trends and regularities in the data which are in many cases known beforehand or predictable. They argue that regularities of local scope may be of greater value as they convey less well-known phenomena or contradictions to the general beliefs. Therefore several researchers focus on mining *rare association rules*, which refers to association rules between either frequent and rare items or only rare items in the data. Similar to conventional association rule mining, various objective measures can be used to assess the interestingness of rare association rules (Kiran and Reddy (2010)).

As the aim of this paper is to find combinations of player types that lead to good match results, we are looking for combinations of players that convert an average lineup into an excellent lineup. We are thus looking for rare combinations of players in the lineups of the matches which are interesting. The integration of a correlation measure in the task of finding rare association rules leads to the problem of *rare correlated pattern mining* (Bouasker and Ben Yahia (2015)). Up to the best of my knowledge, this paper will be the first in which the method of finding rare correlated patterns is applied to data containing sports team lineups.

2.3 Match Results Estimation

For the final goal of investigating the influence of the found player combinations on team performance, the match results will be estimated. Soccer performances have gained the interest of economic researchers in recent years. A popular stream of literature is the estimation of goal scoring in soccer matches. Using an independent Poisson model, Maher (1982) estimates the home and away team scores of matches from four English soccer league divisions during three seasons from 1971 to 1974. Although goodness-of-fit tests show that an independent Poisson model gives a reasonably accurate description of the data, the assumption of independence between the home and away team scores seems to be invalid as it causes an underestimation of the number of draws in the matches. To improve the fit, extensions of this model by using a bivariate Poisson model have been proposed by among others, Dixon and Coles (1997), and Karlis and Ntzoufras (2003).

A second stream of literature focuses on the direct estimation of match results (home team win, draw, away team win) rather than goals scored by the individual teams. The direct estimation of match results is advocated by several researchers because of its simplicity; it requires fewer parameters and the estimation procedures are simpler compared to the bivariate Poisson model (Goddard and Asimakopoulos (2004)). Koning (2000) proposes an ordered probit model to estimate the outcomes of soccer matches in the Premier League of the Netherlands during the seasons from 1956 to 1996. Goddard and Asimakopoulos (2004) use an extension of the ordered probit model by adding a variety of explanatory variables to forecast the results of soccer matches in England and Wales.

Because the ordered probit model allows for an easy inclusion of a variety of explanatory variables, it will be used in this paper to estimate the match results. New explanatory variables related to team lineups will be used to identify the drivers of match results.

3 Data

The data that will be used is an SQLite database obtained from kaggle.com. It contains information about more than 25,000 soccer matches with over 10,000 players in 11 European Countries' lead championships from the seasons in 2008 to 2016. It contains player and team attributes, team lineups, betting odds, and detailed match events. The data was sourced from different sites: the scores, lineup, team formation and events are obtained from http://football-data.mx-api.enetscores.com/, the betting odds from http://www.football-data.co.uk/ and the players and team attributes from EA Sports FIFA games http://sofifa.com.

3.1 Matches

Information on the played matches are included in the *Match* table. It contains detailed match events, the players in the lineups of both teams, the betting odds of several bookmakers, and the number of goals scored by the home and away team. Although many information is available, for this research only the match outcomes, and players present in both team lineups are of interest. As an example we take 'El Clásico' between Real Madrid and FC Barcelona on 21-11-2015. Table 3.1 shows the relevant information of the data instance from the *Match* table of the corresponding match. Note that of both lineups only 11 players are given, there is no data available on substitutions during the match.

Table 3.1: Example of a data instance of the *Match* table.

Column	**Value**	**Label**
league	21518	Spain LIGA BBVA
season	2015/2016	
date	21-11-2015	
match_id	2030194	
home_team_id	8633	Real Madrid CF
away_team_id	8634	FC Barcelona
home_team_goal	0	
away_team_goal	4	
home_player_1	51949	Keylor Navas
home_player_2	208077	Danilo
home_player_3	230982	Raphael Varane
home_player_4	30962	Sergio Ramos
home_player_5	28467	Marcelo
home_player_6	31097	Luka Modric
home_player_7	95078	Toni Kroos
home_player_8	164684	James Rodriguez
home_player_9	31921	Gareth Bale
home_player_10	26166	Karim Benzema
home_player_11	30893	Cristiano Ronaldo
away_player_1	37421	Claudio Bravo
away_player_2	33988	Daniel Alves
away_player_3	37482	Gerard Pique
away_player_4	38818	Javier Mascherano
away_player_5	150739	Jordi Alba
away_player_6	25773	Ivan Rakitic
away_player_7	154257	Sergio Busquets
away_player_8	30955	Andres Iniesta
away_player_9	212472	Sergi Roberto
away_player_10	40636	Luis Suarez
away_player_11	19533	Neymar

In addition to the information shown in Table 3.1, X and Y coordinates are given of all the 22 players in the lineups of both teams, representing the position of the player on the field in the corresponding match. The X and Y coordinates range from 1 to 9 and 1 to 11 respectively where all goalkeepers have coordinates (1,1). A distinction between defenders, midfielders, and forwards is made based on the Y coordinate, where players with an Y coordinate between 2 and 5 are defined as defenders, between 6 and 8 as midfielders, and between 9 and 11 as forwards.

Of the more than 25,000 matches in the data, 4,618 matches have missing data of variables essential for the analysis and are excluded from the sample. This leaves us with some teams being present in only a few matches. If a team was present in less than 15 matches, all its matches will be excluded from the sample. The Poland Ekstraklasa will be entirely excluded from the sample as only limited data from the last three seasons from this league is left. Table 3.2 shows the number of matches per league of the entire dataset and of the final sample.

Table 3.2: Number of matches per European soccer league contained in the entire data and final sample.

League	**Number of matches**	
	Entire data	**Final sample**
Belgium Jupiler League	1728	1203
England Premier League	3040	2962
France Ligue 1	3040	2864
Germany 1. Bundesliga	2448	2374
Italy Serie A	3017	2735
Netherlands Eredivisie	2448	2034
Poland Ekstraklasa	1920	0
Portugal Liga ZON Sagres	2052	1251
Scotland Premier League	1824	1541
Spain LIGA BBVA	3040	2694
Switzerland Super League	1422	1178
Total	25979	20836

3.2 Player Attributes

Two tables in the database contain detailed information of the individual players. The *Player* table consists of 11,060 rows, each containing the name, day of birth, height, and weight of a distinct soccer player. The *Player Attributes* table contains ratings in the range of 0 to 100 on different soccer skills of each player at a certain date. The player attributes are expert ratings used in the FIFA video games published by EA Sports.[1] Through a player id the *Player* and *Player Attributes* tables can be linked to each other and to the *Match* table. 871 players in the *Player* table are excluded from the analysis as they were not present in any of the remaining matches in the final sample discussed in the previous section. Another 364 players are excluded because of missing values for one or more attributes from the *Player Attributes* table or data inaccuracies, leaving us with 9,825 distinct players. For every player, the most recent rating is obtained that will be used to cluster the players. As an example of a data instance from the *Player Attributes* table, the most recent rating of the FC Barcelona player Andres Iniesta is presented in Table 3.3.

[1] https://www.easports.com/fifa

Table 3.3: Example of a data instance of the *Player Attributes* table.

Column	**Value**
player id	30955
date	16-10-2015
overall rating	88
potential	88
preferred foot	right
attacking work rate	high
defensive work rate	medium
crossing	79
finishing	73
heading_accuracy	54
short_passing	92
volleys	74
dribbling	90
curve	80
free_kick_accuracy	70
long_passing	86
ball_control	92
acceleration	76
sprint_speed	75
agility	83
reactions	88
balance	87
shot_power	65
jumping	54
stamina	64
strength	59
long_shots	74
aggression	58
interceptions	68
positioning	85
vision	92
penalties	71
marking	57
standing_tackle	57
sliding_tackle	56
gk_diving	6
gk_handling	13
gk_kicking	6
gk_positioning	13
gk_reflexes	7

Based on the Y coordinate described in the previous section, the players are assigned to four different positions: goalkeepers, defenders, midfielders, and attackers. For example in the match on 21-11-2015, Andres Iniesta had the Y coordinate 7, meaning that he was positioned as a midfielder. Note that it is possible for a player to have several positions as he can be placed at different positions in different matches. In figures A.1 to A.4 in Appendix A, boxplots of the numerical player attributes are shown for the players at each of the four positions. In Figure A.1 it is striking that some players assigned to goalkeepers have low scores for the goalkeeper-related attributes *gk_diving*, *gk_handling*, *gk_kicking*, *gk_positioning*, and *gk_reflexes*. After identifying the names of the corresponding players and consulting a web search engine, I conclude that 12 players are incorrectly assigned to goalkeepers. Furthermore, in figures A.2 to A.4, some players have unexpected high scores for the goalkeeper-related attributes, while they are assigned to defenders, midfielders, or attackers. Further inspection of these corresponding players reveals that 13 of them are actually goalkeepers. All 25 incorrectly assigned players are removed from the inaccurate positions. Table 3.4 shows the final number of players per position.

In addition to the numerical player attributes shown in figures A.1 to A.4, the ordinal variables *attacking work rate* and *defensive work rate*, and the categorical variable *preferred foot* are present in the *Player Attributes* table. Both the *attacking work rate* and *defensive work rate* variables contain many inaccurate records (around 6%). Inspection of the *preferred foot* attribute shows that 75.4% of the players prefer right but the reliability of this attribute can be questioned because in the data each player is assigned one preferred foot while some players are known to be two-footed. Because the *preferred foot* attribute is dubious and the *attacking work rate* and *defensive work rate* attributes contain a substantive proportion of data inaccuracies, they will not be considered in the analysis.

Table 3.4: Number of players used for analysis per position.

Position	**Number of players**
goalkeeper	833
defender	4736
midfielder	6131
forward	4133

4 Methodology

To answer the main research question there are several steps to be taken that require different types of methods. In this section the methods are described that will be used to perform the different steps. The first step is to categorize the players based on the player attributes. The second step is to find rare correlated combinations of player types in the team lineups and the last step is to investigate the influence of the found combinations of player types on the match results.

4.1 Clustering

The goal of cluster analysis is to divide data into groups that are meaningful and interpretable. For the first step, three different clustering algorithms, explained in sections 4.1.1, 4.1.2, and 4.1.3, will be used. The results of the different algorithms will be presented and compared in section 5.1.

Evaluating the performance of a clustering algorithm is one of the challenges of cluster analysis. It is essential for two reasons. First, to select the clustering algorithm among different (configurations of) clustering algorithms that best fits the data. Second, to determine the optimal number of clusters to be computed. Two criteria are proposed for evaluating a clustering scheme: compactness and separation. Members of the same cluster should be close to each other (compact) and different clusters should be distant from each other (separate).

A large number of methods have been proposed in the literature to measure the cluster validity. The performance of different methods vary considerably and may be dependent on the nature of the data. Researchers are therefore encouraged to use one or more of the better performing methods (Milligan and Cooper (1985)). Arbelaitz et al. (2013) conducted an extensive comparative study of 30 cluster validity indices in different environments with different characteristics. Although their results do not show sufficiently strong evidence to distinguish a small set of indices as being significantly better than the rest, they identify a group of about 10 indices that seem to be recommendable. As Silhouette, Davies-Bouldin, and Calinski-Harabasz are in the top of this group, these indices will be used to compare the performances of the clustering algorithms in this paper.

Following Arbelaitz et al. (2013), we define a dataset X as a set of n objects represented as vectors in an M-dimensional space: $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^m$. A clustering in X is a partitioning of the instances in X into K disjoint groups: $C = \{C_1, C_2, ..., C_K\}$ where $\bigcup_{C_k \in C} C_k = X$, $C_k \cap C_l = \emptyset \ \forall \ k \neq l$. The centroid of a cluster C_k is the vector $\mathbf{c_k}$ and the centroid of the whole dataset is the vector $\bar{\mathbf{x}}$.

- The Silhouette (Sil) index (Rousseeuw (1987)) is the overall average silhouette width of the clustering, where the silhouette of a cluster is based on its compactness and separation. The compactness is measured based on the distances between all the points of the same cluster and the separation is measured based on the distances between all points of a cluster to all points of the nearest neighbor cluster. It is defined as:

$$\mathrm{Sil}(C) = 1/n \sum_{C_k \in C} \sum_{\mathbf{x_i} \in C_k} \frac{\mathrm{b}(\mathbf{x_i}, C_k) - \mathrm{a}(\mathbf{x_i}, C_k)}{\max\{\mathrm{a}(\mathbf{x_i}, C_k), \mathrm{b}(\mathbf{x_i}, C_k)\}}, \tag{1}$$

where

$$\mathrm{a}(\mathbf{x_i}, C_k) = 1/|C_k| \sum_{\mathbf{x_j} \in C_k} \mathrm{d_e}(\mathbf{x_i}, \mathbf{x_j}), \tag{2}$$

$$\mathrm{b}(\mathbf{x_i}, C_k) = \min_{C_l \in C \setminus C_k} \{1/|C_l| \sum_{\mathbf{x_j} \in C_l} \mathrm{d_e}(\mathbf{x_i}, \mathbf{x_j})\}, \tag{3}$$

with $\mathrm{d_e}(\mathbf{x_i}, \mathbf{x_j})$ denoting the Euclidean distance between object $\mathbf{x_i}$ and $\mathbf{x_j}$. A larger value of the Silhouette index means a better partitioning of the data.

- The Davies-Bouldin (DB) index (Davies and Bouldin (1979)) measures the compactness of a clustering based on the distance from the members of a cluster to its centroid and the separation based on the distance between centroids. It is defined as:

$$\mathrm{DB}(C) = 1/K \sum_{C_k \in C} \max_{C_l \in C \setminus C_k} \left\{ \frac{\mathrm{S}(C_k) + \mathrm{S}(C_l)}{\mathrm{d_e}(\mathbf{c_k}, \mathbf{c_l})} \right\}, \tag{4}$$

where

$$\mathrm{S}(C_k) = 1/|C_k| \sum_{\mathbf{x_i} \in C_k} \mathrm{d_e}(\mathbf{x_i}, \mathbf{c_k}). \tag{5}$$

A smaller value of the Davies-Bouldin index means a better partitioning of the data.

- The Calinski-Harabasz (CH) index (Caliński and Harabasz (1974)) measures the compactness of a clustering based on the distance from the members of a cluster to its centroid and the separation based on the distance from the centroids of the clusters to the centroid of the whole dataset. It is defined as:

  $$\mathrm{CH}(C) = \frac{(n-K)\sum_{C_k \in C} |C_k|\, \mathrm{d_e}(\mathbf{c_k}, \bar{\mathbf{x}})}{(K-1)\sum_{C_k \in C} \sum_{\mathbf{x_i} \in C_k} \mathrm{d_e}(\mathbf{x_i}, \mathbf{c_k})}. \tag{6}$$

  A larger value of the Calinski-Harabasz index means a better partitioning of the data.

4.1.1 K-means

The first clustering algorithm that will be used is K-means. K-means is a prototype-based, partitional clustering technique that attempts to find a user-specified number of clusters (K), which are represented by their centroids (Tan et al. (2006)). Considering a dataset X, a set of n objects represented as vectors in an M-dimensional space: $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^m$, the goal of K-means is to minimize the sum of the squared error (SSE) according to some proximity function:

$$SSE = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \mathrm{dist}(\mathbf{c}_k, \mathbf{x}_i)^2. \tag{7}$$

Although several distance metrics can be specified as the proximity function, K-means is typically used with the Euclidean distance (Jain (2010)). Wu et al. (2007) defined a family of functions that can be used for K-means clustering as ‘The K-means distance’. They revealed that irrespective of proximity function used, K-means tends to produce clusters with relatively uniform sizes. Applying K-means to the player attributes data, it is assumed that the players are approximately evenly distributed among the different player types.

To obtain the first clustering outcome, the original player attributes data will be standardized after which K-means will be applied for K ranging from 2 to 15. Different K-means algorithms have been presented in the literature and although they will always converge, they are subject to finding a local minimum solution instead of a global one (Morissette and Chartier (2013)). The efficient algorithm of MacQueen et al. (1967) allows the use of 1000 different initial cluster centroids for each K, to find the optimal cluster centroids. To choose an appropriate value of K for partitioning the players, the obtained cluster validity indices as described in section 4.1 will be computed.

4.1.2 Sparse Principal Component Analysis with K-means

As a second approach to cluster the players, K-means will be applied on the data in a reduced dimension. As seen in section 3.2, each player is described by 35 numerical player attributes that measures the player’s technical skills. Instead of applying K-means on all these attributes as described in section 4.1.1, K-means will now be applied on a projection of the original data onto a space of lower dimension. The technique used to represent the original data in a reduced dimension is Sparse Principal Component Analysis (SPCA). The goal of PCA is to represent the data in a lower dimension ($d < m$), by a set of attributes that captures maximum variance of the data. The new attributes (principal components) are orthogonal to each other and ordered with respect to how much of the variance they explain.

Consider a dataset $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^m$. We define the matrix $\mathbf{X}$, that contains the objects of X as rows and is standardized. The goal of PCA is to find d components that capture maximum variance of the data. Each principal component is a linear combination of the original attributes:

$$\mathbf{u}_j = w_{j,1}X_1 + w_{j,2}X_2 + w_{j,3}X_3 + ... + w_{j,m}X_m. \tag{8}$$

The variance of a component can be expressed as:

$$Var(\mathbf{u}_j) = Var(\mathbf{Xw}) = (n-1)^{-1}\mathbf{w}'\mathbf{X}'\mathbf{Xw} = \mathbf{w}'\mathbf{Rw} \tag{9}$$

with

$$\mathbf{R} = (n-1)^{-1}\mathbf{X}'\mathbf{X}, \tag{10}$$

representing the sample covariance matrix of $\mathbf{X}$, as $\mathbf{X}$ is standardized. The variance of $\mathbf{u}_j$ is maximized to capture maximum variance of the data. To avoid the trivial solution of choosing the elements of the loadings vector $\mathbf{w}$ extremely large, $\mathbf{w}$ is constraint to be a unit length vector:

$$\begin{aligned} \underset{w}{\text{argmax}} \quad & \mathbf{w}'\mathbf{Rw}, \\ \text{subject to} \quad & \mathbf{w}'\mathbf{w} = 1. \end{aligned} \tag{11}$$

Using the technique of Lagrange multipliers the maximization problem can be written as:

$$\underset{w}{\text{argmax}} \quad \mathbf{w}'\mathbf{Rw} - \lambda(\mathbf{w}'\mathbf{w} - 1). \tag{12}$$

Differentiating (12) with respect to $\mathbf{w}$ results in:

$$\mathbf{Rw} = \lambda\mathbf{w}. \tag{13}$$

This can be recognized as the eigenvector equation with $\mathbf{w}$ an eigenvector of $\mathbf{R}$ and λ the associated eigenvalue. Using (9), the variance explained by the component corresponds to:

$$Var(\mathbf{u}_j) = \mathbf{w}'\mathbf{Rw} = \mathbf{w}'\lambda\mathbf{w} = \lambda\mathbf{w}'\mathbf{w} = \lambda. \tag{14}$$

Maximizing the variance of the first principal component results in choosing as loadings vector the eigenvector of $\mathbf{R}$, with the largest associated eigenvalue. Successive components should have maximimum variance under the constraint of being orthogonal to all the previous components. In a similar fashion as above, the loadings vector of the j^{th} principal component corresponds to the j^{th} eigenvector of the sample covariance matrix $\mathbf{R}$. The variance explained by the j^{th} principal component equals the associated eigenvalue.

To determine the number of components (d) to be used, several approaches have been proposed in the literature. Jackson (1993) compared some of these approaches with PCA on real and simulated datasets of different nature. Consistent results were obtained with the broken-stick method, it provides a good combination of simplicity and accurate evaluation of dimensionality. The broken-stick method proposed by Frontier (1976) assumes that the variance of random data is divided randomly among all the components. The expected distribution of the variance explained by each component will follow a broken-stick distribution, which can be easily calculated as:

$$b_j = \sum_{i=j}^{m} \frac{1}{i}, \tag{15}$$

where m is the number of variables and b_j is the additional variance explained by the j^{th} principal component under the broken-stick model. The additional variance explained by each component can be plotted against j in a scree-plot together with the broken-stick model. A component should be used if the additional variance explained by the component exceeds the generated variance explained by the broken-stick model.

Having applied PCA, the original data can be represented by the principal components that explain most of the variance. Subsequent analysis, such as clustering, can be conducted on this representation of the data in a reduced dimension. Although in many applications PCA can explain a great part of the variance with only a few components, a drawback of this approach is that it results in the principal components being a linear combination of all original attributes, making it hard to interpret the different components. A preferred solution is to have principal components being a linear combination of only a subset of the original attributes, which is obtained by applying Sparse Principal Component Analysis.

Several approaches for SPCA have been proposed in the literature. A simple thresholding approach to obtain sparse components is replacing the 'near-zero' loadings by exact zeroes, but this procedure is shown to be misleading (Cadima and Jolliffe (1995)). Some approaches control the degree of sparsity by adding a penalty on the number of non-zero loadings (l_0 norm) of $\mathbf{w}$ in the maximization problem (11). Adding the l_0 penalty makes the problem NP-hard and heuristics have been proposed by Moghaddam et al. (2006) and d'Aspremont et al. (2007) to approximate the optimal solution. Other approaches add a penalty on the sum of the absolute values of the loadings (l_1 norm) of $\mathbf{w}$, which leads to some exactly zero loadings in $\mathbf{w}$. An example is SCoTLASS, proposed by Jolliffe et al. (2003), which uses the idea of the LASSO constraint from Tibshirani (1996). Although SCoTLASS derives sparse components, the computational cost is high. The degree of sparsity is controlled by a penalty parameter and a good rule to determine its value is missing.

An efficient algorithm similar to SCoTLASS is proposed by Croux et al. (2013). Their algorithm maximizes an estimate of the variance, using projection-pursuit to find the loadings vectors of the components. A penalty on the l_1 norm of $\mathbf{w}$ is incorporated in the objection function to control the sparsity. In addition, they propose the use of an information criterion to determine the optimal level of the penalty parameter. The loadings vector of the first sparse principal component is given by:

$$\tilde{\mathbf{w_1}} = \underset{||\mathbf{w}||=1}{\operatorname{argmax}} V(\mathbf{w'x_1}, \mathbf{w'x_2}, ..., \mathbf{w'x_n}) - \rho_1 ||\mathbf{w}||_1, \tag{16}$$

where V is a variance measure. The l_1 norm of $\mathbf{w}$ is incorporated in the optimization with $||\mathbf{w}||_1 = \sum_{j=1}^{m} |\mathbf{w_j}|$. The sparsity of the loadings vector of the first component is controlled by the parameter ρ_1. The loadings vectors of subsequent components are defined by:

$$\tilde{\mathbf{w_j}} = \underset{||\mathbf{w}||=1, \mathbf{w} \perp \mathbf{w_1}, ..., \mathbf{w} \perp \mathbf{w_{j-1}}}{\operatorname{argmax}} V(\mathbf{w'x_1}, \mathbf{w'x_2}, ..., \mathbf{w'x_n}) - \rho_j ||\mathbf{w}||_1, \tag{17}$$

for $1 < j \leq m$, with ρ_j possibly different from ρ_1.

To find the optimal loadings, the projection-pursuit approach searches for directions that maximize a variance measure of the data projected on it. Croux et al. (2013) extend the Grid algorithm of Croux et al. (2007) and the implementation of the algorithm is available in the R package pcaPP (Filzmoser et al. (2009)).

The penalty parameter ρ_j controls the degree of sparsity of the loadings vector of the j^{th} principal component. Larger values of ρ_j increase the importance of sparseness relative to the variance measure V in (17). To have a similar degree of sparsity over the different principal components Croux et al. (2013) suggest to take:

$$\rho_j := \rho \mathcal{V}(\mathbf{X}^{(j)}), \tag{18}$$

where $\mathbf{X}^{(j)}$ contains the data projected on the orthogonal complement of the space spanned by the first $j-1$ optimal loadings vectors. $\mathcal{V}$ denotes the total estimated variance of a data matrix, defined as:

$$\mathcal{V}(\mathbf{Y}) = \sum_{i=1}^{m} V(\mathbf{y_i}), \tag{19}$$

for any n by m matrix $\mathbf{Y}$ with $\mathbf{y_i}$ the i^{th} column of $\mathbf{Y}$ and V the variance estimator. Croux et al. (2013) propose the use of an information criterion to determine the optimal level of the penalty parameter. The parameter ρ is determined by minimizing a BIC type criterion:

$$BIC(\rho) = \frac{\widetilde{RV}}{RV} + df(\rho)\frac{log(n)}{n}, \tag{20}$$

with $\widetilde{RV}$ and RV referring to the total estimated variance of the residuals matrix obtained by a sparse PCA and an unconstrained PCA respectively. $df(\rho)$ is the number of non-zero loadings when using ρ as the penalty parameter. The first term in the BIC criterion measures the fit of the model and the second term penalizes for model complexity.

In addition to the BIC criterion, Croux et al. (2013) propose the *tradeoff curve* to select a value for each ρ_j. The *tradeoff curve* is a graphical tool that shows the variance explained by the sparse component plotted against ρ_j. An increase of sparseness (higher ρ_j) will in general decrease the explained variance. The value of ρ_j should be chosen such that a sharp decline in explained variance happens just afterwards.

Before applying SPCA, the number of components (d) to be computed must be determined. Similar to Croux et al. (2013), d will be selected based on the scree-plot of the unconstrained PCA. In this paper, the broken-stick method will be used to determine the value of d. Having chosen the number of components to be computed, the Grid algorithm explained in Croux et al. (2013) will be used to compute the sparse components. 100 different values of ρ will be considered, with 100 iterations each. The empirical variance will be used as a variance estimator. The parameter ρ will be determined by minimizing the BIC type criterion

Having computed the sparse principal components, the original data will be projected on the components to reduce the dimension. A clustering of the players will be obtained by applying K-means on the data in the reduced dimension in a similar fashion as described in section 4.1.1.

4.1.3 Archetypal Analysis

As a third approach to cluster the players, archetypal analysis (AA) will be applied. As introduced by Cutler and Breiman (1994), "Archetypal analysis represents each "individual" in a dataset as a mixture of "individuals of pure type", or "archetypes"." Each data point is approximated by a convex combination of a set of archetypes. The archetypes themselves are restricted to being mixtures of individual data points, making them easy to interpret. AA thus aims to find 'pure types' (archetypes) in a set of multivariate observations, such that all the data can be well represented as convex combinations of the archetypes (Eugster and Leisch (2009)). To keep the archetypes 'pure' and preserve all information in the data, AA will only be applied to the original data. Applying a dimension reduction technique, such as SPCA, to the data before applying AA will lose information about the extreme data points and will make the archetypes less 'pure'.

Following Bauckhage and Thurau (2009), consider a dataset $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^m$. The matrix $\mathbf{X}$ contains the objects of X as rows. The goal of AA is to find a set of archetypes $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_p\}$ with $p \ll n$ that are linear combinations of the data points

$$\mathbf{z}_j = \sum_{i=1}^{n} \mathbf{x}_i b_{ij}, \tag{21}$$

where the coefficients $b_{ij} \geq 0$ and $\sum_i b_{ij} = 1$. These two constraints force the archetypes to resemble the data and to be convex mixtures of the data instances. For a given set of archetypes, AA minimizes

$$\| \mathbf{x}_i - \sum_{j=1}^{p} \mathbf{z}_j a_{ji} \|^2 \tag{22}$$

to determine coefficients a_{ji} that allow the data $\mathbf{x}_i$ to be well described by the archetypes. Again the constraints $a_{ji} \geq 0$ and $\sum_j a_{ji} = 1$ are imposed to force the data points to be mixtures of archetypes. The set of archetypes is chosen to minimize the residual sum of squares

$$RSS(p) = \sum_{i=1}^{n} \| \mathbf{x}_i - \sum_{j=1}^{p} \mathbf{z}_j a_{ji} \|^2 = \sum_{i=1}^{n} \| \mathbf{x}_i - \sum_{j=1}^{p} \sum_{k=1}^{n} \mathbf{x}_k b_{kj} a_{ji} \|^2 \; . \tag{23}$$

Cutler and Breiman (1994) show that if $p = 1$, choosing the archetype to be the sample mean minimizes RSS and for $p > 1$, the archetypes that minimize the RSS are located on the convex hull of X. They propose an optimization algorithm that iterates over:

1. finding the coefficients a_{ji} that minimize the RSS for a given set of $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_p\}$,
2. computing the intermediate archetypes $\tilde{\mathbf{z}}_j$ given the updated a_{ji},
3. determining the coefficients b_{ij} that minimize the p problems $\| \mathbf{X}\mathbf{b}_j - \tilde{\mathbf{z}}_j \|^2$,
4. computing the RSS.

Each step requires the solution of several convex least squares problems, for details of the implementation of the several steps I refer to Eugster and Leisch (2009).

As stated by Eugster and Leisch (2009), there is no rule for the correct number of archetypes p. They use the "elbow criterion" in the plot of the RSS against p to determine the number of archetypes to use. This simple rule of a flattening of the curve indicating the correct value of p will be used in this paper. AA will be applied on the standardized data for p ranging from 1 to 15. For each value of p, the algorithm is started 50 times with different random initial archetypes to help avoid getting stuck in local minima. Having chosen the optimal value of p, each player can be approximated by a convex combination of the computed archetypes. The players will be clustered by assigning each player to the archetype with the highest loading (a_{ji}).

4.2 Rare Correlated Pattern Mining

Assuming that in the first step all the players are assigned to one of K types of players, the goal of the second step is to find rare correlated combinations of player types in the lineups of the matches. The final goal is to investigate what combinations of player types significantly influence the match results. By performing rare correlated pattern mining, the number of possible combinations of player types will be reduced, leaving us with the unpredictable combinations that may be of great value as they convey less well-known phenomena (Szathmary et al. (2010)).

The team lineups of all matches can now be represented as binary vectors with the K player types as columns and a value of 1 indicating the presence of at least one player of the corresponding player type in the team lineup. To clarify this, the example of 'El Clásico' is shown in Table 4.1. Both lineups of the home and away team are turned into binary vectors indicating the presence of a certain player type in the lineup of the corresponding match. To find interesting combinations of different player types, *rare correlated pattern mining* will be performed on all team lineups.

Table 4.1: Example of the binary vector indicating the presence of player types in the lineups of the teams in a match.

league	**season**	**date**	**match_id**	**team_id**	**type 1**	**type 2**	**type 3**	**. . .**	**type K**
21518	2015/2016	21-11-2015	2030194	8633	0	1	1	. . .	0
21518	2015/2016	21-11-2015	2030194	8634	1	1	0	. . .	1

The problem of association rule mining was introduced by Agrawal et al. (1993) to find interesting associations among items in customer transactions, where each transaction consists of the items purchased by a customer in a visit. In our case, the problem is to find interesting associations among player types in team lineups where each team lineup consists of the player types present in the lineup. Consider a database of lineups D and a set of all player types $I = \{i_1, i_2, ..., i_K\}$. Each lineup $t \in D$ is represented as a binary vector with $t[p] = 1$ if player type i_p is present in lineup t, and $t[p] = 0$ otherwise. Consider a set α containing m different player types from I. The conjunctive support of the m-player combination α, is defined as the number of lineups in D, containing all player types of α:

$$\textbf{Conjunctive support: } Supp(\wedge\alpha) = \mid \{t \in D \mid \forall i_p \in \alpha,\ t[i_p] = 1\} \mid. \tag{24}$$

The disjunctive support of α is defined as the number of lineups in D, containing at least one player type of α:

$$\textbf{Disjunctive support: } Supp(\vee\alpha) = \mid \{t \in D \mid \exists i_p \in \alpha,\ t[i_p] = 1\} \mid. \tag{25}$$

α is frequent in D if $Supp(\wedge\alpha) \geq minsupp$, with *minsupp* the minimum support threshold specified by the user. If $Supp(\wedge\alpha) < minsupp$, α is infrequent or rare. Depending on the value of *minsupp* an (in)frequent pattern mining algorithm may generate a huge number of patterns satisfying the minimum support threshold. To overcome this problem, different correlation measures have been introduced to filter the extracted patterns.

In this paper, we will follow Bouasker and Ben Yahia (2015) by using the *bond* correlation measure proposed by Segond and Borgelt (2011), defined as:

$$bond(\alpha) = \frac{Supp(\wedge\alpha)}{Supp(\vee\alpha)}. \tag{26}$$

The bond measure is a straightforward generalization of the Jaccard index which ranges between 0 and 1 with a higher value indicating a higher correlation between the player types in α. The set of correlated patterns associated with the *bond* measure are all player combinations satisfying $bond(\alpha) \geq minbond$ with *minbond* the minimum correlation threshold specified by the user.

We are interested in finding the set of *rare correlated patterns (RCP)* which is defined as:

$$RCP = \{\alpha \subseteq I \mid Supp(\wedge\alpha) < minsupp \ \& \ bond(\alpha) \geq minbond\}. \tag{27}$$

In this paper, the CORI algorithm introduced by Bouasker and Ben Yahia (2015) will be used to mine the set of rare correlated patterns according to the *bond* measure.

The resulting set of *rare correlated patterns*, which in our case contains the player combinations, is heavily dependent on the values for the *minsupp* and *minbond* thresholds set by the user. Different combinations of the *minsupp* threshold ranging from 50% to 80% of all lineups and the *minbond* threshold ranging from 0.25 to 0.50 will be examined. The goal of this paper is to find what player combinations significantly influence the match results. The optimal threshold values will therefore be chosen based on the ability of the resulting set of player combinations to fit the model of match outcomes that will be explained in the following section.

4.3 Match Results Estimation

Having identified combinations of player types that are present in the team lineups, the last step is to discover what their effect is on the match results. This will be done by estimating the results of matches with an ordered probit model. The direct estimation of match results is advocated by several researchers because of its simplicity; it requires fewer parameters and the estimation procedures are simpler compared to most models estimating exact goal scoring of the home and away team. Because the ordered probit model allows for an easy inclusion of a variety of explanatory variables, it will be used in this paper to estimate the match results. By incorporating the presence of certain player combinations in the model, I will examine if they have a significant influence on the result.

Suppose for a certain combination of the *minsupp* and *minbond* thresholds, g rare correlated player combinations are found and $F = \{PC_1, PC_2, ..., PC_g\}$ is the set containing the g player combinations (PC). Every match can now be represented as a vector with twice the g different player combinations as columns indicating the presence of the player combination in the home and away team. Continuing with the example of 'El Clásico': if we assume that the combination of player type 2 and type 3 is found and called PC1, and the combination of player type 1 and type K is found and called PCg, then 'El Clásico' will be represented as shown in Table 4.2.

Table 4.2: Example of the binary vector indicating the presence of a rare correlated player combinations in the lineups of both teams in a match.

match_id	**home_team_id**	**away_team_id**	**PC1_h**	**PC2_h**	**...**	**PCg_h**	**PC1_a**	**PC2_a**	**...**	**PCg_a**
2030194	8633	8634	1	0	...	0	0	0	...	1

Similar to Koning (2000), we assume that the match result in season s between home team i

and away team j is determined by the unobserved latent variable y^*_{ijs}:

$$y^*_{ijs} = \alpha_i + \sum_{PC_f \in F} \beta_{fh} * PC_{fis} - \alpha_j + \sum_{PC_f \in F} \beta_{fa} * PC_{fjs} + \eta_{ijs}. \tag{28}$$

The strength of team i is measured by the parameter α_i and is assumed to be constant over different seasons. The variables PC_{fis} and PC_{fjs} are dummy variables indicating the presence of player combination PC_f in season s in the team lineups of home team i and away team j respectively. The parameters β_{fh} and β_{fa} indicate the effect of the presence of player combination PC_f in the home and away team respectively on y^*_{ijs}. Similar to Koning (2000), the identifying restriction $\sum_i \alpha_i = 0$ is imposed such that the parameters α_i can be interpreted as deviations from a hypothetical team with strength 0. The strength of a team (α_i) is kept constant over different seasons because it is assumed that during the season the lineup of a certain team will not vary greatly. If in the next season the lineup has changed and the team performs better, this can be due to the presence of a new combination of player types which will be captured by the corresponding β. The random component η_{ijs} captures other determinants of the result of the match.

The latent variable y^*_{ijs} is linked to the observed match result (y_{ijs}) by:

$$y_{ijs} = \begin{cases} 1 & \text{for } y^*_{ijs} > \tau_2, \\ 0 & \text{for } \tau_1 < y^*_{ijs} \leq \tau_2, \\ -1 & \text{for } y^*_{ijs} \leq \tau_1, \end{cases} \tag{29}$$

with $y_{ijs} = 1$ if team i wins, $y_{ijs} = 0$ if team i plays a draw and $y_{ijs} = -1$ if team i loses. Because the random component η_{ijs} is assumed to follow a normal distribution with variance σ^2, the probabilities of the possible match results are:

$$\begin{aligned} Pr(y_{ijs} = 1) &= 1 - \Phi\{(\tau_2 - \gamma_{ij})/\sigma\}, \\ Pr(y_{ijs} = 0) &= \Phi\{(\tau_2 - \gamma_{ij})/\sigma\} - \Phi\{(\tau_1 - \gamma_{ij})/\sigma\}, \\ Pr(y_{ijs} = -1) &= \Phi\{(\tau_1 - \gamma_{ij})/\sigma\}, \end{aligned} \tag{30}$$

with

$$\gamma_{ij} = \alpha_i + \sum_{PC_f \in F} \beta_{fh} * PC_{fis} - \alpha_j + \sum_{PC_f \in F} \beta_{fa} * PC_{fjs}, \tag{31}$$

and Φ the standard normal distribution function. The parameters can be estimated using maximum likelihood estimation. To evaluate and compare the models, Aikake's Information Criterion (AIC) will be computed. In using AIC to select the best model, the emphasis is on the goodness of fit while an allowance is made for parsimony (Bozdogan (1987)).

5 Results

5.1 Clustering

In this section, the results of the three different clustering algorithms, explained in section 4.1, will be presented. Clustering all the players together leads to groups separating the players from different positions in the field. The clustering algorithms are therefore applied to players at each of the four positions (goalkeepers, defenders, midfielders, and forwards) separately to obtain different player types for each position.

5.1.1 K-means

After standardizing the data, K-means is applied for K ranging from 2 to 15. For each K, 1000 different initial cluster centroids are used to find the optimal cluster centroids. To choose an appropriate value of K for partitioning players at each of the four positions, the obtained cluster validity indices as described in section 4.1 are shown in Table B.1 of Appendix B.1.

The table shows that different values of K are suggested by the different indices. It is more important that the members of a cluster are close to its centroid than that the members of the cluster are all close to each other. As both the Davies-Bouldin index and the Calinski-Harabasz index measure the compactness based on the distances between the points in a cluster to the corresponding centroid, I prefer these above the Silhouette index. For measuring the separation, the Calinski-Harabasz index uses the distances between the cluster centroids and the centroid of the whole dataset, while I consider it acceptable for a cluster to represent an average player, and thus be close to the centroid of the whole dataset. Therefore, the analysis will be continued with the values of K suggested by the Davies-Bouldin index to see if they lead to interpretable clusters. This means that 2 types of goalkeepers, 5 types of defenders and midfielders, and 7 types of forwards will be computed.

Figures B.1 to B.4 in Appendix B.1 show visualizations of the centroids of the clusters obtained for each of the four positions. The centroids show what an average player of the corresponding cluster looks like. Table 5.1 presents the number of players assigned to each cluster together with labels of how each player type is interpreted based on the cluster centroid. Note that the descriptions of the player types as presented in Table 5.1 are interpretations of the cluster centroids and aim to summarize the information shown in Figure B.1 to B.4 in Appendix B.1.

Table 5.1: Interpretation of player types (PT) of the clusters obtained with K-means followed by the size of the cluster in number of players assigned to the cluster.

PT	Label	Size
	Goalkeepers	
1	overall average player with good goalkeeper attributes but bad finishing, positioning and marking	678
2	bad goalkeeper attributes but several excellent player attributes	155
	Defenders	
3	good strength, marking, tackles and interceptions, and excellent heading accuracy	882
4	overall poor player	1284
5	excellent finishing, positioning, and volleys but bad interceptions, marking and tackles	379
6	overall good player with excellent passing, ball control, and long shots	1251
7	overall bad player with good strength	940
	Midfielders	
8	overall good player with excellent finishing, dribbling, and ball control but bad marking and tackles	1373
9	overall excellent player with excellent passing, reactions, interceptions and tackles	1028
10	overall bad player with bad finishing, dribbling, long shots and positioning but good marking and tackles	871
11	overall poor player with good marking and tackles	1577
12	overall bad player with bad interceptions, marking and tackles	1282
	Forwards	
13	overall bad player with bad finishing, volleys and positioning but good interceptions, marking and tackles	433
14	overall average player	851
15	overall good player with excellent interceptions, marking and tackles	664
16	overall poor player with bad acceleration, speed, agility and balance but excellent heading accuracy and strength	315
17	overall excellent player with excellent dribbling and ball control	617
18	overall good player with excellent heading accuracy	679
19	overall bad player	574

5.1.2 Sparse Principal Component Analysis with K-means

Before applying SPCA on each of the four positions, I select appropriate values for the number of components (d) to be computed. Similar to Croux et al. (2013), d will be selected based on the scree-plot of the unconstrained PCA. The broken-stick method described in section 4.1.2 will be used to determine the value of d for each of the four positions. Figure B.5 of Appendix B.2 shows the four obtained scree-plots together with the broken-stick model. Based on the plot, 3 components will be computed for the goalkeepers, defenders and midfielders, and 4 components for the forwards.

Having chosen the number of components to be computed for each position, the Grid algorithm explained in Croux et al. (2013) will be used to compute the sparse components. For each position, 100 different values of ρ are considered, with 100 iterations each. The empirical variance is used as a variance estimator. The parameter ρ is determined by minimizing the BIC type criterion explained in section 4.1.2. The found ρ's are reviewed by inspecting the tradeoff curves, shown in Figure

B.6 of Appendix B.2. In all four cases, the optimal value of ρ according to the BIC type criterion, is found just after a sharp decline in the total variance explained by the components. This is in contrast with the desire of choosing ρ just before such a decline. As a consequence, the resulting cumulated explained variance (CEV) is disappointing. I therefore consider an alternative approach for determining the value of ρ, proposed by Filzmoser et al. (2009): Tradeoff Product Optimization (TPO). TPO chooses a (possibly different) value ρ_j for each component by maximizing the explained variance multiplied by the number of zero loadings of the corresponding component. Figures B.8 to B.10 of Appendix B.2 show the tradeoff curves again, now with the optimal values of ρ based on TPO. The explained variance of each component is plotted against the number of zero loadings on the x-axis. The resulting values of ρ lead to a substantial increase in cumulated explained variance with little sacrifice of sparseness (high numbers of zero loadings). I therefore choose to set the values for ρ based on TPO.

The obtained sparse components are visualized in terms of their loadings in figures B.11 to B.14 of Appendix B.2. Interestingly, the components obtained for the different positions look similar, indicating that irrespective of the position, the players distinguish from each other on the same attributes.

An interpretation of what the principal components represent is given in Table 5.2. The first principal component of the goalkeepers is highly correlated with both attributes that belong to a good attacker (such as finishing, dribbling, and positioning), as well as with some defensive player attributes (such as marking and tackling). Goalkeepers with a high score on this component will have high values for these player attributes. The second component is highly correlated with the overall rating and all the goalkeeper attributes, indicating that a high score on this component will be found for excellent goalkeepers. The third component is negatively correlated with the overall rating and some goalkeeper attributes, while it has a strong positive correlation with acceleration, sprint speed and agility.

The first principal components of the defenders, midfielders and forwards are positively correlated with the overall rating, and many other attributes. Players with a high score on this component will thus be overall good players. The second components of the defenders, midfielders and forwards are all positively correlated with marking and the tackle attributes. This shows that players who are good at marking, also tend to be good at standing and sliding tackles. For defenders and forwards this component also has a high correlation with interceptions, and for defenders and midfielders a high correlation with aggression. Good defenders will have a high score on this component. The third component of defenders and midfielders is similar to the fourth component of the forwards, it has strong positive correlations with the goalkeeper attributes. The third component of the forwards has a high positive correlation with acceleration, spring speed, agility and balance while it has a negative correlation with strength and heading accuracy. Players with a high score on this component will thus be nimble players who are not very strong while players with a negative score on this component are likely to be very strong but slow.

Table 5.2: Interpretation of the sparse components obtained with SPCA.

	Goalkeepers	**Defenders**	**Midfielders**	**Forwards**
PC 1	good offense and defense	overall good	overall good	overall good
PC 2	excellent goalkeeping	good defense	good defense	good defense
PC 3	bad goalkeeping but high speed	good goalkeeping	good goalkeeping	nimble but not strong
PC 4				good goalkeeping

Now we have identified sparse principal components that explain a large part of the variation in the data, we can represent the players in terms of the components. K-means can then be applied to the data in the reduced dimension. Again, K-means is applied for K ranging from 2 to 15, with 1000 different initial cluster centroids for each K. The cluster validity indices obtained with K-means on the reduced data is shown in Table B.2 of Appendix B.2. Based on similar reasoning as in section 5.1.1, we continue with the optimal number of clusters K according to the Davies-Bouldin index. The final cluster centroids found are shown in figures B.15 to B.18 of Appendix B.2. The interpretation of the obtained clusters are given in Table 5.3 together with the number of players assigned to each cluster. Note that the interpretation of the cluster centroids is based on the interpretation of the sparse components. This means, that if a cluster centroid is found to have a low value for the agility and speed component, the players in this cluster are likely to have high values for strength and heading accuracy, as these attributes have a negative correlation with the agility and speed component.

Table 5.3: Interpretation of player types of the clusters obtained with K-means on projected data on the sparse principal components followed by the size of the cluster.

PT	Label	Size
	Goalkeepers	
1	average goalkeeper but slow	85
2	good goalkeeper	88
3	bad goalkeeper with excellent offensive and defensive attributes	60
4	good goalkeeper with high speed	105
5	average goalkeeper with high speed	40
6	bad goalkeeper	80
7	poor goalkeeper with high speed	53
8	poor goalkeeper with excellent offensive and defensive attributes	20
9	overall bad player	38
10	poor goalkeeper with excellent offensive and defensive attributes but slow	31
11	excellent goalkeeper with very high speed	28
12	overall average player	112
13	average goalkeeper with excellent offensive and defensive attributes	44
14	excellent goalkeeper	49
	Defenders	
15	excellent player with bad defensive attributes	484
16	good player with good defensive attributes but bad goalkeeper	999
17	good player with excellent defensive attributes and goalkeeping	1531
18	overall poor player	1722
	Midfielders	
19	good player with bad goalkeeping	98
20	poor player with poor defensive attributes	965
21	excellent goalkeeper with poor defensive attributes	789
22	poor player with good defensive attributes	1100
23	good player with excellent defensive and goalkeeping attributes	1127
24	excellent player with poor defensive attributes	896
25	good player with excellent defensive attributes	1156
	Forwards	
26	overall good player with excellent defensive attributes and good goalkeeping	339
27	bad player with good defensive attributes	301
28	excellent goalkeeper attributes	5
29	overall poor player with good goalkeeper attributes	418
30	average player with poor defensive attributes	396
31	overall good player with poor nimbleness but good strength	341
32	overall poor player with bad nimbleness but excellent strength	269
33	overall poor player with excellent defensive attributes and good strength	202
34	overall poor player with poor deensive attributes but very nimble	420
35	overall good player with bad goalkeeper attributes	63
36	overall bad player	295
37	overall good player with excellent defensive attributes	342
38	average player with poor defensive attributes but very nimble and good goalkeeper	402
39	overall excellent player with poor defensive attributes but very nimble	340

5.1.3 Archetypal Analysis

As explained in section 4.1.3, AA will only be applied on the original data. After standardizing the data, AA is applied for different number of archetypes (p) ranging from 1 to 15. For each p, the algorithm is started 50 times with different random initial archetypes to help avoid getting stuck in local minima. Because AA does not directly cluster the data instances into different groups, the cluster validity indices discussed in section 4.1 cannot be computed directly. Instead, the number of archetypes to be used will be determined using the 'elbow criterion' based on a plot of the RSS obtained for each value of p. In Figure B.19 in Appendix B.3, the best obtained RSS are plotted against p, for each of the four positions. The points on each line indicate where an elbow is identified and the corresponding values for p are used in consecutive analysis. Based on the plot, 6 archetypes for the goalkeepers, 4 archetypes for the defenders, and 7 archetypes for the midfielders and forwards will be computed. Note that the choice of p is somewhat subjective and may be arguable.

The obtained archetypes at each position are visualized in figures B.20 to B.23 of Appendix B.3. As explained in section 4.1.3, archetypes are mixtures of individual data points and can thus be seen as an example of a player. Each player is approximated by a convex combination of the seven archetypes, so we can cluster the players by assigning each player to the archetype with the highest loading (a_{ji}). Table 5.4 shows the resulting cluster sizes together with labels of how each player type is interpreted based on the archetypes.

Table 5.4: Interpretation of player types (PT) of the clusters obtained with AA followed by the size of the cluster in number of players assigned to the cluster.

PT	Label	Size
	Goalkeepers	
1	excellent goalkeeper with excellent reactions	273
2	poor goalkeeper with good shot power, passing and stamina	92
3	poor goalkeeper with bad sprint speed	193
4	good goalkeeper with good finishing, positioning marking and tackles	62
5	bad goalkeeper with good finishing, positioning, marking and tackles	106
6	average goalkeeper with bad short passing and high aggression	107
	Defenders	
7	overall poor player with bad shot power but good acceleration and sprint speed	1694
8	good finishing and free kick but bad interceptions, marking and tackles	759
9	excellent player with excellent reactions and ball control	1327
10	good strength and heading accuracy but bad acceleration, sprint speed, agility and balance	956
	Midfielders	
11	good marking and tackles but bad finishing and positioning	730
12	overall average player with bad goalkeeper skills	312
13	excellent player with excellent reactions	1294
14	poor player with bad acceleration, sprint speed and agility	573
15	excellent player with excellent finishing, reactions and positioning	750
16	bad player with bad short passing, ball control and reactions	1458
17	good player with good ball control and balance	1014
	Forwards	
18	excellent player with excellent short passing, dribbling, ball control, reactions and vision	642
19	overall average to good player	747
20	poor player with bad finishing, volleys, shot power, long shots, positioning and penalties	228
21	overall good player with good interceptions, marking and tackles	946
22	overall average player with excellent goalkeeper skills	23
23	poor player with bad acceleration, sprint speed, agility and balance	384
24	overall bad player	1163

5.1.4 Comparing Clustering Outcomes

Although the clustering outcomes of the different methods seem to vary substantially, there are some similarities. The number of different types of goalkeepers of the AA and SPCA with K-means clustering outcomes is relatively big, considering that only 5% of all the players in the dataset are goalkeepers. Not surprisingly, goalkeepers mainly distinguish themselves on the goalkeeper attributes. In addition, the attributes finishing, positioning and marking have shown to be important factors distinguishing goalkeepers from one another in all cluster outcomes. In the outcomes of AA and SPCA with K-means, speed is also found to be a distinctive feature for a goalkeeper.

Considering the different types of defenders, tackling, marking, interceptions, and heading accuracy are important attributes in all clustering outcomes. In addition, defenders are found to have a good strength, while this is often associated with a lower rating for acceleration, speed, and agility.

Also for the midfielders, tackling and marking are shown to be distinctive features. In addition the attributes ball control, dribbling, and finishing are notable in some types of midfielders. In general, player types with good ball control are found to be overall good players.

For forwards, the attributes finishing, volleys, and dribbling, among others are important in all clustering outcomes. The forwards can also be distinguished based on their acceleration, speed, and agility. Similar to defenders, high values of the latter attributes are often associated with a lower strength. Therewithal, a forward with a good strength is found to have a good heading accuracy.

Having found what attributes are important in defining a certain player type, it is also interesting to know what attributes are present in the data, but not often found to be characteristic for a type of soccer player. Some examples of these attributes are crossing, curve, and jumping. Surprisingly, the seemingly important attributes free kick accuracy and penalties are both found to be important in only two player types.

5.2 Rare Correlated Pattern Mining

Now we have clustered the players into different player types, we aim to find combinations of player types in the team lineups that significantly influence the match result. To find a candidate set of player combinations, the set of rare correlated patterns in all team lineups will be mined using the CORI algorithm introduced by Bouasker and Ben Yahia (2015). A pattern (combination of player types) is defined rare if its conjunctive support is less than the *minsupp* threshold, and defined correlated if its *bond* correlation measure equals or exceeds the *minbond* threshold.

The set of rare correlated patterns will be mined for each of the three clustering outcomes presented in section 5.1. Different combinations of the *minsupp* threshold ranging from 50% to 80% of all lineups and the *minbond* threshold ranging from 0.25 to 0.50 are examined, each resulting in a different set of rare correlated patterns. The obtained patterns for each combination of *minsupp* and *minbond* are then used as input for the ordered probit model to estimate the match outcomes, as explained in section 4.3. The match outcomes are estimated for the seasons from 2008/2009 till 2014/2015. Based on the best fitted model according to the AIC criterion, the optimal thresholds are chosen for each of the three clustering outcomes. Table 5.5 shows the found optimal combinations of the *minsupp* and *minbond* thresholds for each model, together with the obtained AIC.

Table 5.5: Optimal threshold values *minsupp* and *minbond* of probit models based on AIC.

	minsupp	***minbond***	**AIC**
K-means player types	75%	0.5	35842
SPCA with K-means player types	55%	0.4	35966
AA player types	80%	0.4	35931

The obtained rare correlated patterns (player combinations) for each of three clusterings are shown in Table C.1 of Appendix C. In the first column, an ID of each player combination is given to which will be referred in the following section. The second column shows the player types contained in the pattern. The numbers refer to the player type numbers given in tables 5.1, 5.3, and 5.4. The conjunctive support and *bond* correlation measure of the player combinations are given in the third and fourth column respectively.

5.3 Match Results Estimation

In this section, the results of the estimation of the ordered probit models explained in section 4.3 will be presented. In the previous section, the best fitted models are chosen based on the AIC criterion. To evaluate how well the models are able to predict unseen match outcomes, the obtained parameter estimates are now used to predict the outcomes in the test set, that contains the matches of season 2015/2016. The predictions are compared to predictions of the test set obtained with a baseline model where only the α_i's are contained as explanatory variables in equation 28. As the α_i's measure the overall strength of team i, this model predicts the match outcome solely based on the overall strength of the home and away team. In addition, the prediction accuracy of a random generation of one of three match outcomes is computed for the test set. Table 5.6 shows the proportion of correctly predicted match outcomes in the test set for the different models. The best prediction results are obtained with the model containing the combinations of player types obtained with K-means on the original data. All three models that include the presence of player combinations as explanatory variables outperform the baseline model with only the α_i's, and the random match outcomes.

Table 5.6: The proportion of correctly predicted match outcomes in the test set obtained with the different models.

Model	**Proportion correct predictions**
Player combinations with K-means player types	0.501
Player combinations with AA player types	0.492
Player combinations with SPCA with K-means player types	0.491
Baseline including only α_i's	0.468
Random match outcome	0.333

The parameter estimates of main interest are the β_{fh} and β_{fa} of all the found player combinations, indicating the effect of the presence of a certain player combination in the home and away team respectively, on the match outcome. All significant estimates of β_{fh} and β_{fa} on the 5% level obtained with maximum likelihood estimation are shown in tables 5.7 to 5.9 together with the description of the player types contained in the corresponding player combination. The presence of a good player combination in the lineup of the home team will lead to a positive estimate of the corresponding β_{fh}, as it will increase the probability of a home win. The result will be amplified if the same player combination has a negative β_{fa}, as the same combination in the away team will decrease the probability of a home win. It is thus expected to find opposite signs for the two betas of the same player combination. This is often the case as can be seen in the tables.

Table 5.7 shows the parameter estimates from the model with the K-means player types. While the presence of a single poor or bad player has a negative effect on the match outcome, the combination of an overall poor player as defender (player type 4) and an overall poor player with good marking and tackles as midfielder (player type 11) is found to have a positive effect on the match outcome. Not surprisingly, having an overall excellent player as forward in the home team with excellent dribbling and ball control will lead to a positive effect on the probability of a home win, and a negative effect if it is present in the away team. The last three estimates show that a good goalkeeper (player type 1) combines well with a good defender (player type 3) and with a good midfielder (player type 8), while the combination of the three of them does not lead to a better match outcome.

The estimates of the model with player types obtained with SPCA with K-means are shown in Table 5.8. Almost all types of goalkeepers have a negative influence on the probability of a home win if they are present in the away team. This means that the goalkeeper is particularly important when playing an away match. Remarkable is the finding that defenders (player types 15, 16, 17 and 18) only influence match outcomes in combinations. The positive β_{fh} estimate of player combination 38, with defender type 17 and midfielder type 24 shows that an excellent midfielder with poor defensive attributes can be complemented with a good defender. If the same type of midfielder is complemented with a an overall poor player as a defender, this can negatively influence the match outcome, as shown by the positive β_{fa} of player combination 39.

Finally, Table 5.9 shows the significant β_{fh} and β_{fa} estimates of the model with AA player types. While all estimates belonging to a single player type are not surprising given the player descriptions, some interesting results are found with combinations of player types. An overall poor player with bad shot power but good acceleration and sprint speed (player type 7) as a defender is shown to negatively influence the match outcome. A defender with good strength but bad speed (player type 10) is also shown to negatively influence the match outcome, while the combination of these types of defenders has a positive effect on the match outcome. Having a fast but not so strong defender can thus be compensated with a strong, but not so agile defender. Surprising is the positive β_{fa} coefficient for the combination of defender type 9 and midfielder type 17. While both types have good attributes, the results indicate that they only have a positive influence on the match outcome if also an excellent player with excellent reactions as midfielder (player type 13) is present in the lineup.

Table 5.7: Player combinations of player types obtained with K-means with a significant β_{fh} or β_{fa} coefficient on the 5% level.

PC	β_{fh}	β_{fa}	**PT**	**Description**
1		-0.17	2	bad goalkeeper attributes but several excellent player attributes
2	-0.12	0.27	4	overall poor player
3	-0.11		5	excellent finishing, positioning, and volleys but bad interceptions, marking and tackles
5	-0.09	0.19	7	overall bad player with good strength
8	-0.14	0.17	10	overall bad player with bad finishing, dribbling, long shots and positioning but good marking and tackles
9	-0.15	0.29	11	overall poor player with good marking and tackles
10	-0.06	0.10	12	overall bad player with bad interceptions, marking and tackles
11	-0.13		13	overall bad player with bad finishing, volleys and positioning but good interceptions, marking and tackles
14		0.10	16	overall poor player with bad acceleration, speed, agility and balance but excellent heading accuracy and strength
15	0.06	-0.11	17	overall excellent player with excellent dribbling and ball control
17		0.17	19	overall bad player
18	0.09	-0.16	4-11	overall poor player overall poor player with good marking and tackles
27		-0.23	1-8	overall average player with good goalkeeper attributes but bad finishing, positioning and marking overall good player with excellent finishing, dribbling, and ball control but bad marking and tackles
33		-0.28	1-3	overall average player with good goalkeeper attributes but bad finishing, positioning and marking good strength, marking, tackles and interceptions, and excellent heading accuracy
35		0.27	1-3-8	overall average player with good goalkeeper attributes but bad finishing, positioning and marking good strength, marking, tackles and interceptions, and excellent heading accuracy overall good player with excellent finishing, dribbling, and ball control but bad marking and tackles

Table 5.8: Player combinations of player types obtained with SPCA with K-means with a significant β_{fh} or β_{fa} coefficient on the 5% level.

PC	β_{fh}	β_{fa}	**PT**	**Description**
1		-0.29	1	average goalkeeper but slow
2		-0.35	2	good goalkeeper
4		-0.36	4	good goalkeeper with high speed
5		-0.24	5	average goalkeeper with high speed
7		-0.25	7	poor goalkeeper with high speed
8		-0.36	8	poor goalkeeper with excellent offensive and defensive attributes
10		-0.24	10	poor goalkeeper with excellent offensive and defensive attributes but slow
11		-0.58	11	excellent goalkeeper with very high speed
13		-0.35	13	average goalkeeper with excellent offensive and defensive attributes
14	0.29	-0.50	14	excellent goalkeeper
16		-0.16	19	good player with bad goalkeeping
17	-0.06	0.13	20	poor player with poor defensive attributes
18		0.07	21	excellent goalkeeper with poor defensive attributes
19	-0.15	0.19	22	poor player with good defensive attributes
20		-0.33	24	excellent player with poor defensive attributes
22		0.13	27	bad player with good defensive attributes
24		0.13	29	overall poor player with good goalkeeper attributes
27		0.08	32	overall poor player with bad nimbleness but excellent strength
30	0.11	-0.18	35	overall good player with bad goalkeeper attributes
31		0.22	36	overall bad player
32		-0.08	37	overall good player with excellent defensive attributes
34		-0.15	39	overall excellent player with poor defensive attributes but very nimble
38	0.18		17-24	good player with excellent defensive attributes and goalkeeping excellent player with poor defensive attributes
39		0.20	18-24	overall poor player excellent player with poor defensive attributes
42		-0.11	16-17	good player with good defensive attributes but bad goalkeeper good player with excellent defensive attributes and goalkeeping
46		0.11	17-18-25	good player with excellent defensive attributes and goalkeeping overall poor player good player with excellent defensive attributes

Table 5.9: Player combinations of player types obtained with AA with a significant β_{fh} or β_{fa} coefficient on the 5% level.

PC	β_{fh}	β_{fa}	**PT**	**Description**
4		-0.26	4	good goalkeeper with good finishing, positioning marking and tackles
7	-0.18	0.64	7	overall poor player with bad shot power but good acceleration and sprint speed
8		0.10	8	good finishing and free kick but bad interceptions, marking and tackles
10	-0.22	0.19	10	good strength and heading accuracy but bad acceleration, sprint speed, agility and balance
11		0.10	11	good marking and tackles but bad finishing and positioning
13		0.11	14	poor player with bad acceleration, sprint speed and agility
15	-0.09	0.16	16	bad player with bad short passing, ball control and reactions
16		-0.19	17	good player with good ball control and balance
17		-0.12	18	excellent player with excellent short passing, dribbling, ball control, reactions and vision
18		-0.31	19	overall average to good player
22		0.10	23	poor player with bad acceleration, sprint speed, agility and balance
23		0.13	24	overall bad player
26		0.30	9-17	excellent player with excellent reactions and ball control good player with good ball control and balance
29	0.14	-0.33	7-10	overall poor player with bad shot power but good acceleration and sprint speed good strength and heading accuracy but bad acceleration, sprint speed, agility and balance
33		0.27	10-19	good strength and heading accuracy but bad acceleration, sprint speed, agility and balance overall average to good player
34		0.27	13-19	excellent player with excellent reactions overall average to good player
37		-0.34	1-13	excellent goalkeeper with excellent reactions excellent player with excellent reactions
40	0.19		10-13	good strength and heading accuracy but bad acceleration, sprint speed, agility and balance excellent player with excellent reactions
41		-0.35	9-13-17	excellent player with excellent reactions and ball control excellent player with excellent reactions good player with good ball control and balance
43		-0.30	10-13-19	good strength and heading accuracy but bad acceleration, sprint speed, agility and balance excellent player with excellent reactions overall average to good player

6 Conclusion

This paper aimed to answer three interconnected research questions to investigate how team lineups in soccer matches should be formed in order to achieve positive match outcomes. By first assigning soccer players to different player types based on the outcomes of several cluster algorithms, the attributes distinguishing one soccer player from another are identified. Subsequently, the presence of combinations of the resulting player types are found with the use of rare correlated pattern mining. Finally, by estimating the match outcomes with an ordered probit model, the influence of the presence of a certain combination of player types in the lineup of a team is investigated.

The outcomes of the different clustering methods show that some technical attributes are particularly important in categorizing soccer players. In all outcomes, the sprint speed, acceleration, and agility have shown to be important factors that distinguish one player from another. A high score on these attributes is found to be associated with a low score on strength. On the other hand, players with low sprint speed, acceleration, and agility often are very strong and have a good heading accuracy. In addition, the tackle attributes, interceptions, and marking are found to be important and associated with each other, not only for the defenders, but on all positions in the field. Ball control has also shown to be an important factor in defining the player types at all positions. A player type with a good ball control is found to be an overall good player. Attributes that have not shown to be of big importance in defining a certain player type are crossing, curve, jumping, free kick accuracy, and penalties.

The mined rare correlated player combinations show what interesting player combinations are present in the team lineups, and together with the parameter estimates of the ordered probit model, the player combinations that lead to good results in soccer matches are identified. In all three outcomes, a goalkeeper with some good player attributes is found to positively influence the match outcome for the away team, meaning that goalkeepers are particularly important for the team in away matches. Another interesting finding is that although player types on their own can negatively influence the match outcome, a combination of them may positively influence the match outcome. For example the combination of a fast but poor defender with a strong but slow defender is found to increase the probability of a win.

When interpreting the results, some limitations of the research must be taken into account. First, although match outcomes are estimated of the eight seasons from 2008/2009 till 2015/2016, only the most recent rating of the attributes of a player is used. This means that the players are assigned to clusters based on their most recent player attributes, irrespective of the date the match was played. If the ratings of attributes of a player vary greatly within the time frame, an improved approach could use the most recent rating up untill the match, to assign the player to a player type in the corresponding match. A decision has then to be made about which ratings to include in computing the player clusterings. If all available ratings will be included, the dataset may be unbalanced due to some players having more ratings than others.

Further research could also take into account substitutions during the match. In the current approach, only the eleven players in the team lineup are considered while during the match this might change because of injuries and substitutes. Changes in lineups during the match may have a significant impact on the final outcome which is currently not accommodated for.

Another possible extension for further research is the use of quantitative frequent pattern mining in finding the rare correlated player combinations. In this paper, only the presence of at least one player of a certain player type is considered while it might also be of importance how many players of that certain player type are present in the lineup.

References

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., and Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256.

Bauckhage, C. and Thurau, C. (2009). Making archetypal analysis practical. In *Joint Pattern Recognition Symposium*, pages 272–281. Springer.

Beil, F., Ester, M., and Xu, X. (2002). Frequent term-based text clustering. In *Proceedings of the* *eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 436–442. ACM.

Bouasker, S. and Ben Yahia, S. (2015). Key correlation mining by simultaneous monotone and anti-monotone constraints checking. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pages 851–856. ACM.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370.

Cadima, J. and Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, 22(2):203–214.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Cao, H., Mamoulis, N., and Cheung, D. W. (2005). Mining frequent spatio-temporal sequential patterns. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE.

Croux, C., Filzmoser, P., and Fritz, H. (2013). Robust sparse principal component analysis. *Technometrics*, 55(2):202–214.

Croux, C., Filzmoser, P., and Oliveira, M. R. (2007). Algorithms for projection–pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225.

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

d'Aspremont, A., Bach, F. R., and Ghaoui, L. E. (2007). Full regularization path for sparse principal component analysis. In *Proceedings of the 24th international conference on Machine learning*, pages 177–184. ACM.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

Drachen, A., Sifa, R., Bauckhage, C., and Thurau, C. (2012). Guns, swords and data: Clustering of player behavior in computer games in the wild. In *Computational Intelligence and Games (CIG), 2012 IEEE Conference on*, pages 163–170. IEEE.

Drachen, A., Thurau, C., Sifa, R., and Bauckhage, C. (2014). A comparison of methods for player clustering via behavioral telemetry. *arXiv preprint arXiv:1407.3950*.

Eugster, M. and Leisch, F. (2009). From spider-man to hero-archetypal analysis in r.

Filzmoser, P., Fritz, H., and Kalcher, K. (2009). pcapp: Robust pca by projection pursuit. *R package version*, 1.

Franck, E. and Nüesch, S. (2010). The effect of talent disparity on team productivity in soccer. *Journal of Economic Psychology*, 31(2):218–229.

Frick, B. and Simmons, R. (2008). The impact of managerial quality on organizational performance: evidence from german soccer. *Managerial and Decision Economics*, 29(7):593–600.

Frontier, S. (1976). Étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le moddle du bâton brisé. *Journal of Experimental Marine Biology and Ecology*, 25(1):67–75.

Goddard, J. and Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1):51–66.

Graham, I. and Stott, H. (2008). Predicting bookmaker odds and efficiency for uk football. *Applied Economics*, 40(1):99–109.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Ingersoll, K., Malesky, E. J., and Saiegh, S. M. (2013). Heterogeneity and group performance: Evaluating the effect of cultural diversity in the world's top soccer league.

Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, 74(8):2204–2214.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393.

Kern, M. and Süssmuth, B. (2005). Managerial efficiency in german top league soccer: an econometric analysis of club performances on and off the pitch. *German Economic Review*, 6(4):485–506.

Kiran, A. and Reddy, K. (2010). Selecting a right interestingness measure for rare association rules. *Management of Data*, page 115.

Koning, R. H. (2000). Balance in competition in dutch soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):419–431.

Kuypers, T. (2000). Information and efficiency: an empirical study of a fixed odds betting market. *Applied Economics*, 32(11):1353–1363.

Luss, R. and d'Aspremont, A. (2010). Clustering and feature selection using sparse principal component analysis. *Optimization and Engineering*, 11(1):145–157.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

Moghaddam, B., Weiss, Y., and Avidan, S. (2006). Spectral bounds for sparse pca: Exact and greedy algorithms. In *Advances in neural information processing systems*, pages 915–922.

Morissette, L. and Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in mathematica. *Tutorials in Quantitative Methods for Psychology*, 9(1):15–24.

Rai, P. and Singh, S. (2010). A survey of clustering techniques. *International Journal of Computer Applications*, 7(12):1–5.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Segond, M. and Borgelt, C. (2011). Item set mining based on cover similarity. *Advances in Knowledge Discovery and Data Mining*, pages 493–505.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905.

Szathmary, L., Valtchev, P., and Napoli, A. (2010). Generating rare association rules using the minimal rare itemsets family. *International Journal of Software and Informatics (IJSI)*, 4(3):219–238.

Tan, P., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Pearson International Edition. Pearson Addison Wesley.

Tan, P.-N., Kumar, V., and Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Wu, J., Xiong, H., Chen, J., and Zhou, W. (2007). A generalization of proximity functions for k-means. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 361–370. IEEE.

Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774.

Appendices

A Data

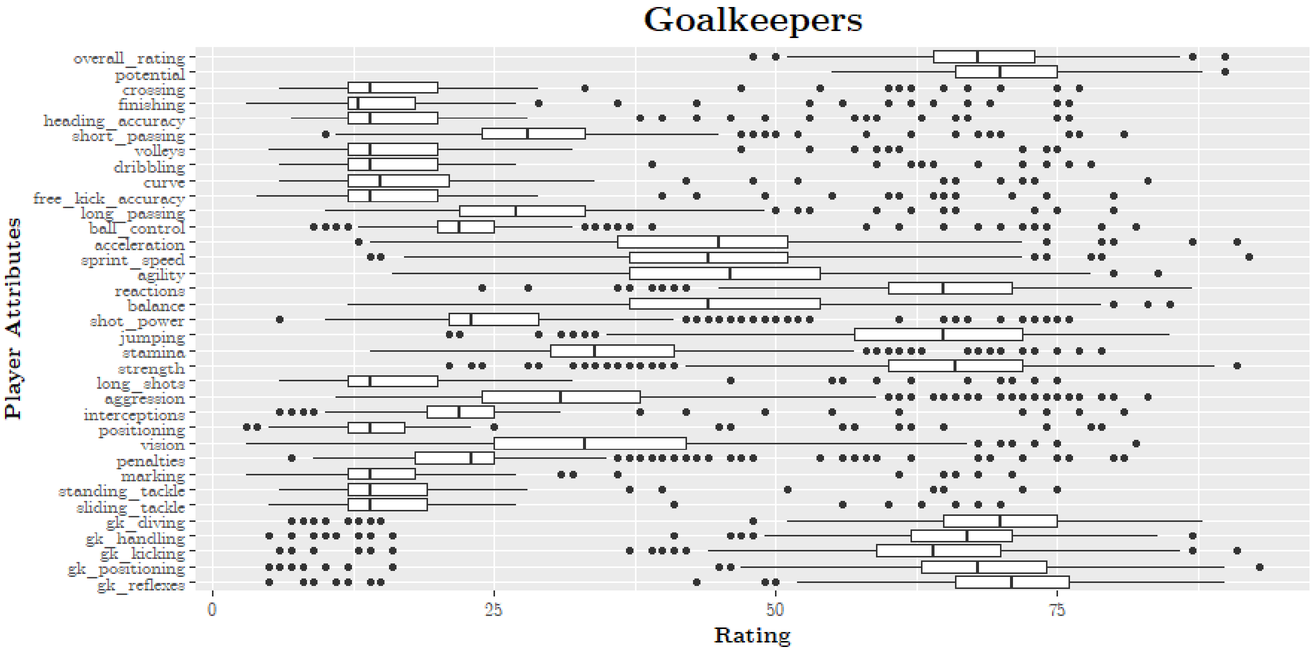


Figure A.1: Boxplots of the player attributes of all goalkeepers.

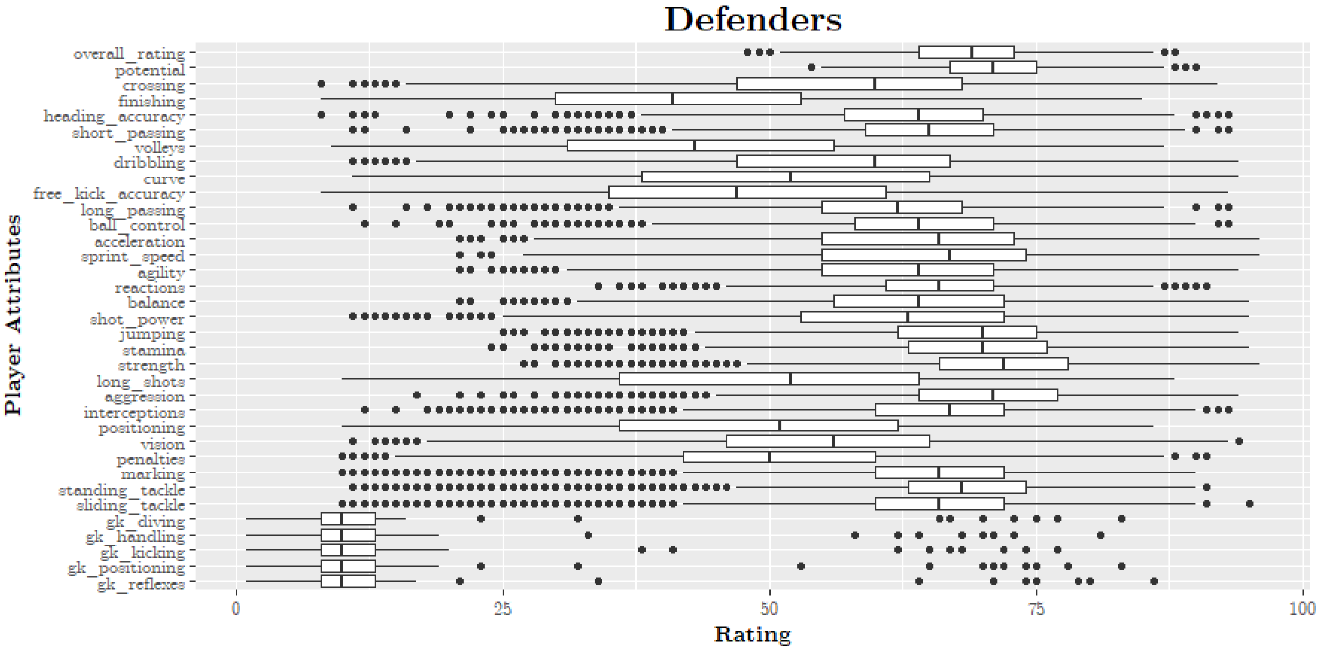


Figure A.2: Boxplots of the player attributes of all defenders.

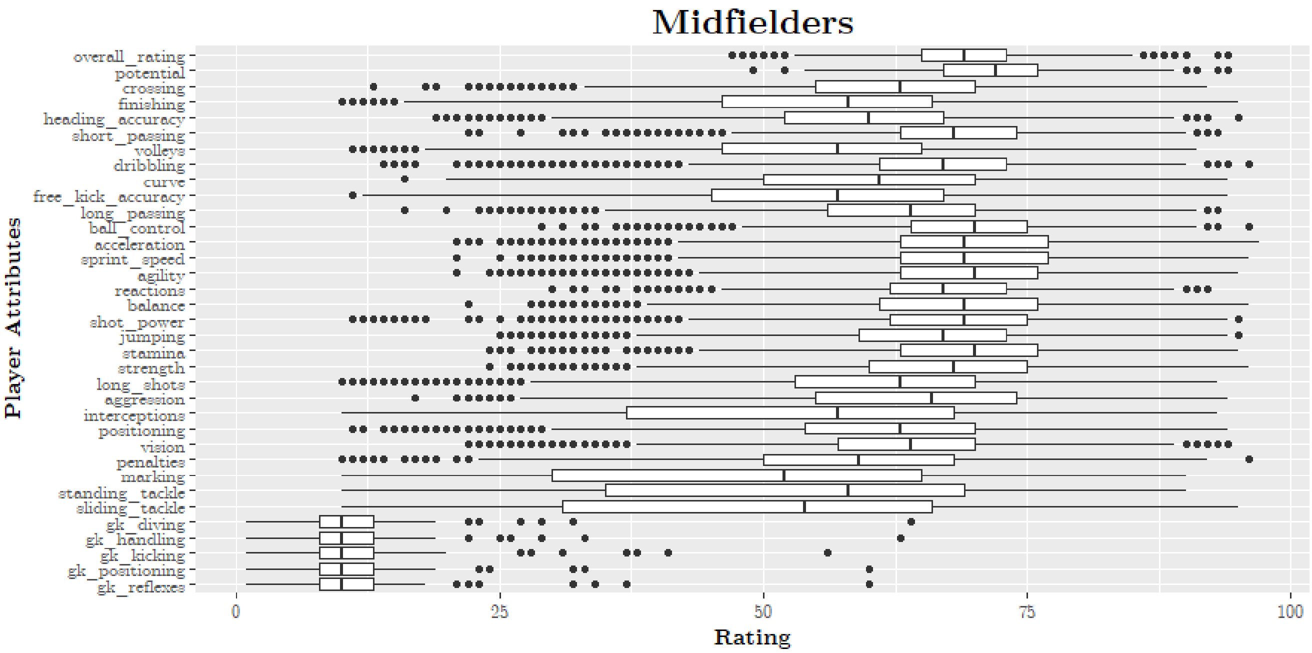


Figure A.3: Boxplots of the player attributes of all midfielders.

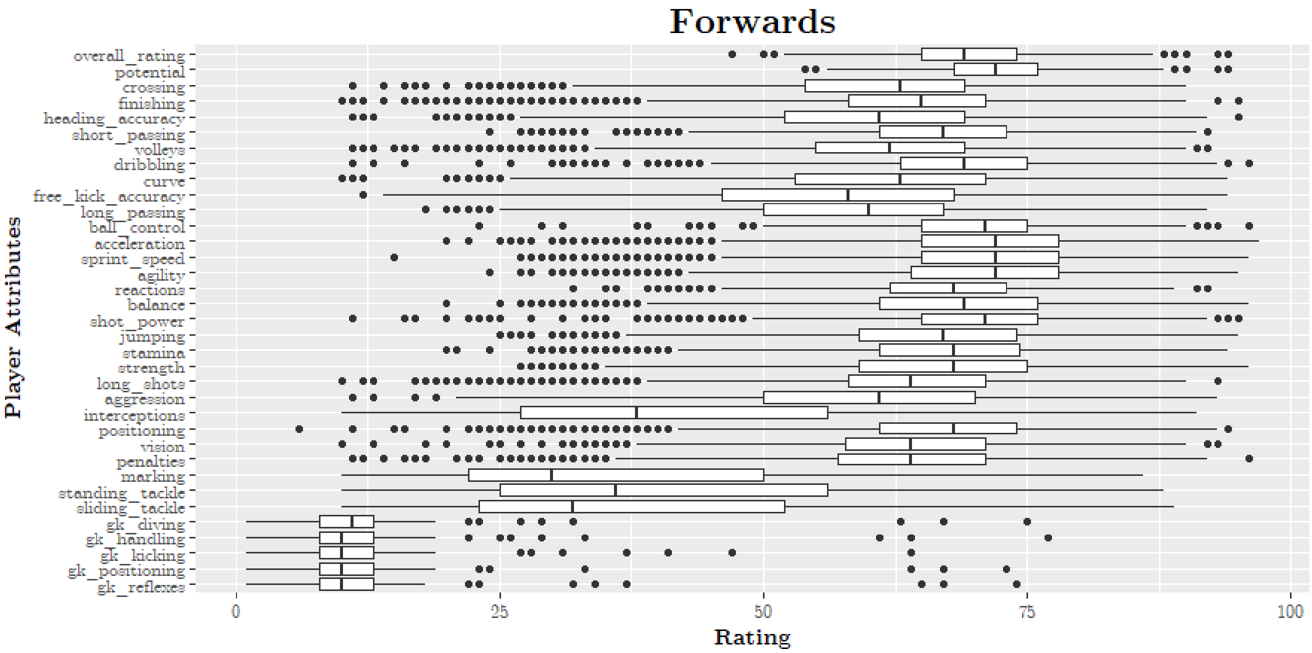


Figure A.4: Boxplots of the player attributes of all forwards.

B Clustering

B.1 K-means

Table B.1: Cluster validity indices Silhouette (Sil), Davies-Bouldin (DB), and Calinski-Harabasz (CH) obtained with K-means for different values of K. Numbers in bold show for each position the best result according to the corresponding index.

	Goalkeepers			**Defenders**			**Midfielders**			**Forwards**		
K	**Sil**	**DB**	**CH**	**Sil**	**DB**	**CH**	**Sil**	**DB**	**CH**	**Sil**	**DB**	**CH**
2	**0.363**	**1.171**	**301**	**0.194**	1.795	**1379**	**0.165**	1.987	**1457**	**0.163**	2.014	**965**
3	0.235	1.856	232	0.165	1.700	1028	0.148	2.070	1218	0.142	1.963	745
4	0.172	2.356	176	0.141	1.794	881	0.132	1.989	1059	0.113	2.206	631
5	0.149	2.027	145	0.130	**1.307**	781	0.109	**1.780**	904	0.118	1.765	577
6	0.133	2.748	124	0.114	2.320	684	0.105	1.901	795	0.116	1.938	529
7	0.115	2.623	109	0.104	1.474	614	0.100	2.281	718	0.105	**1.744**	477
8	0.081	2.552	99	0.094	2.140	559	0.093	2.307	659	0.098	2.096	436
9	0.073	2.335	90	0.092	1.463	513	0.091	1.927	608	0.091	2.058	402
10	0.072	2.212	83	0.083	2.326	474	0.084	2.009	565	0.089	2.342	374
11	0.073	2.513	78	0.081	1.860	443	0.083	2.216	528	0.084	1.910	352
12	0.070	2.570	73	0.078	2.454	415	0.079	2.002	496	0.083	2.178	330
13	0.066	1.482	69	0.073	2.433	389	0.079	2.107	470	0.079	2.104	311
14	0.089	2.303	65	0.073	1.882	369	0.076	1.950	444	0.075	1.891	294
15	0.085	1.641	62	0.073	2.071	350	0.073	1.954	423	0.071	2.249	279

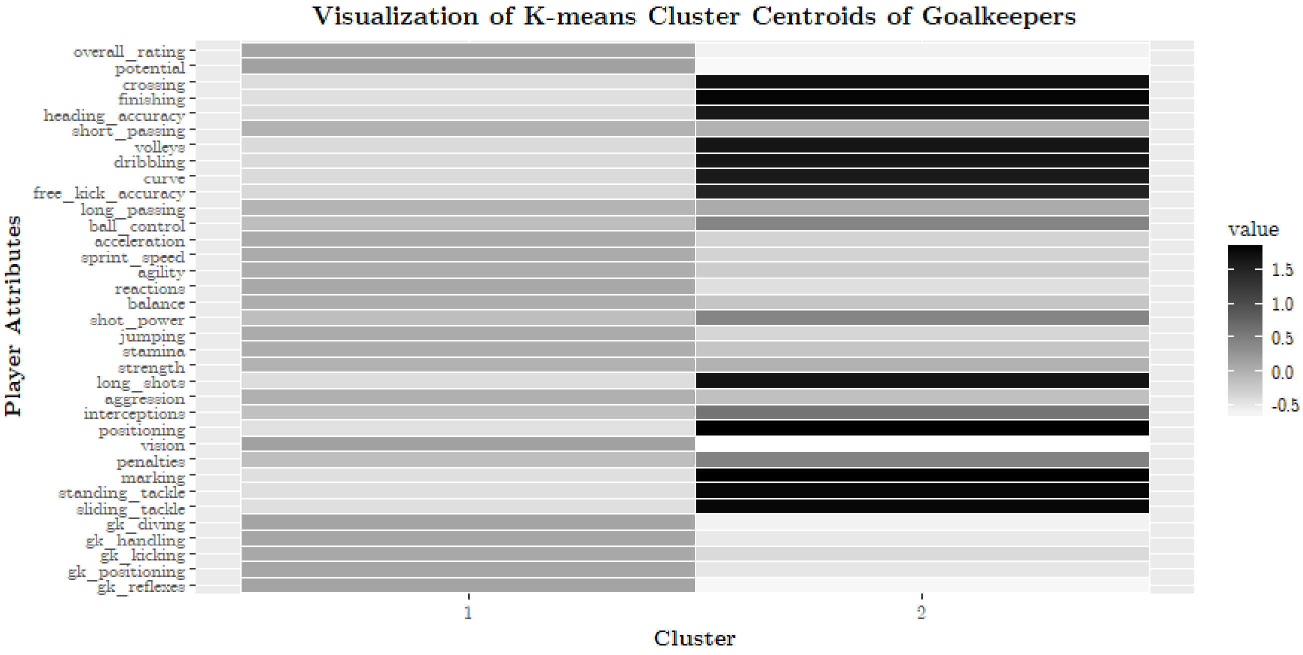


Figure B.1: Visualization of the centroids of goalkeeper clusters obtained with K-means.

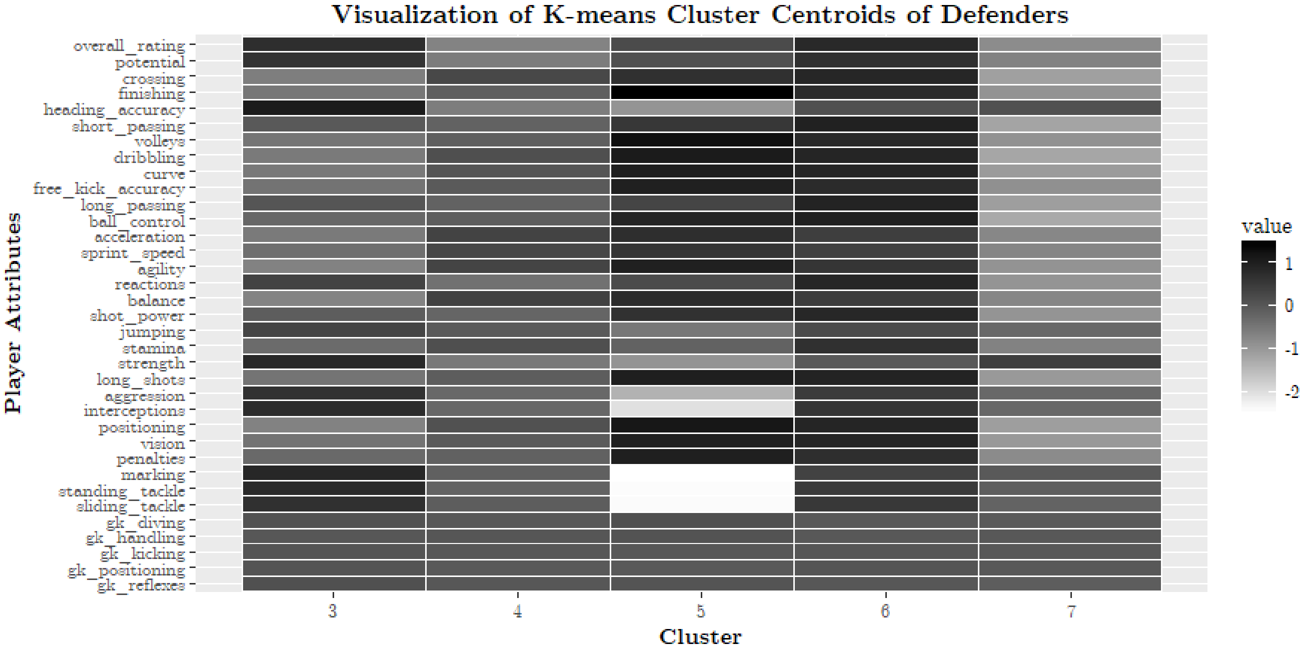


Figure B.2: Visualization of the centroids of defender clusters obtained with K-means.

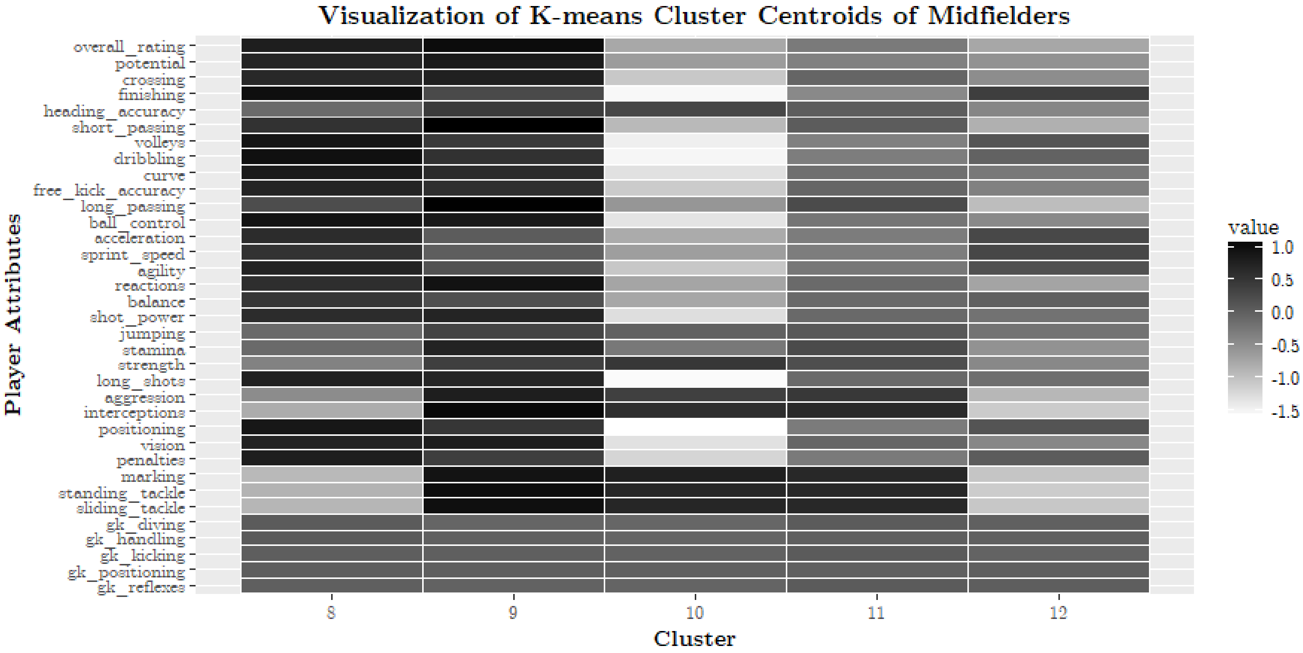


Figure B.3: Visualization of the centroids of midfielder clusters obtained with K-means.

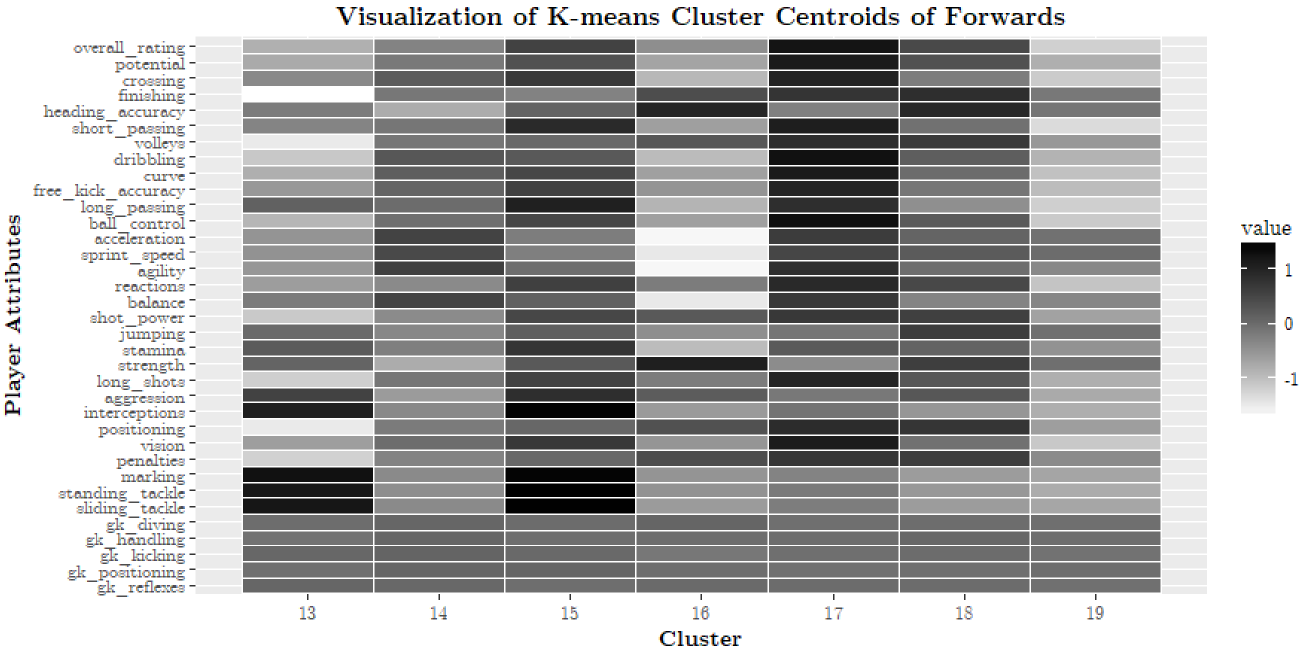


Figure B.4: Visualization of the centroids of forward clusters obtained with K-means.

B.2 Sparse Principal Component Analysis with K-means

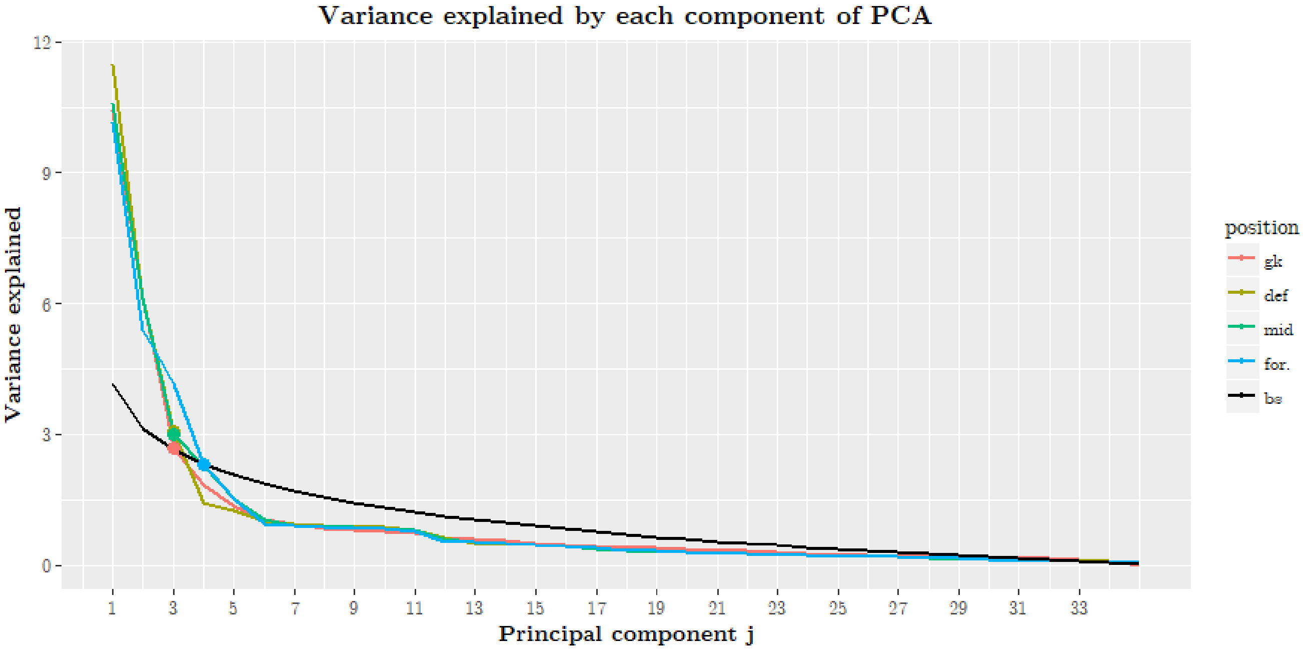


Figure B.5: Variance explained by each component obtained with unconstrained PCA for each position. The points indicate the chosen number of components based on comparison with the broken-stick (bs) model shown in black.

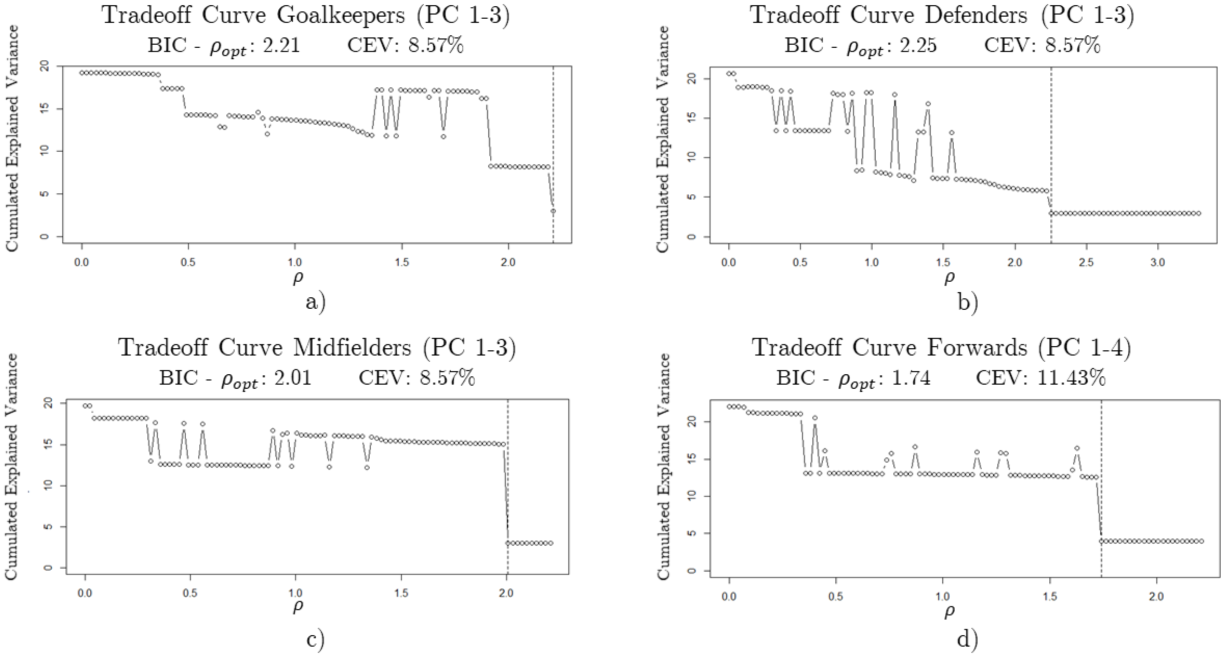


Figure B.6: Tradeoff curves obtained with SPCA on goalkeepers (a), defenders (b), midfielders (c) and forwards (d) for each value of ρ. The dashed line indicates the optimal ρ based on the BIC type criterion.

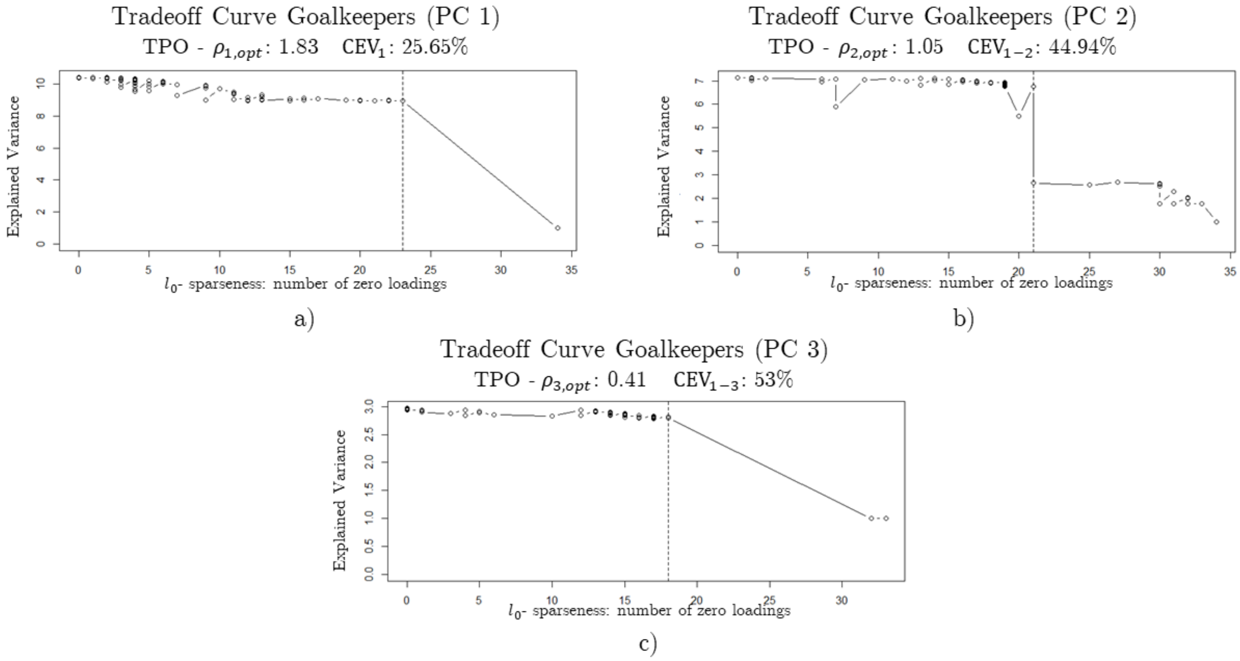
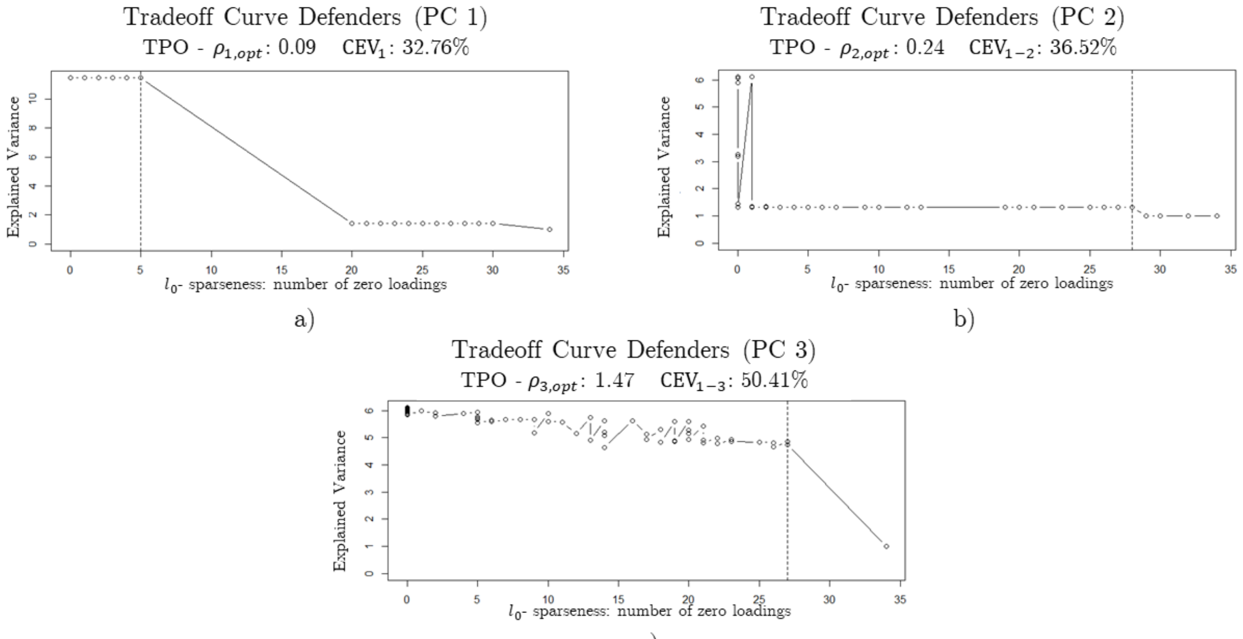
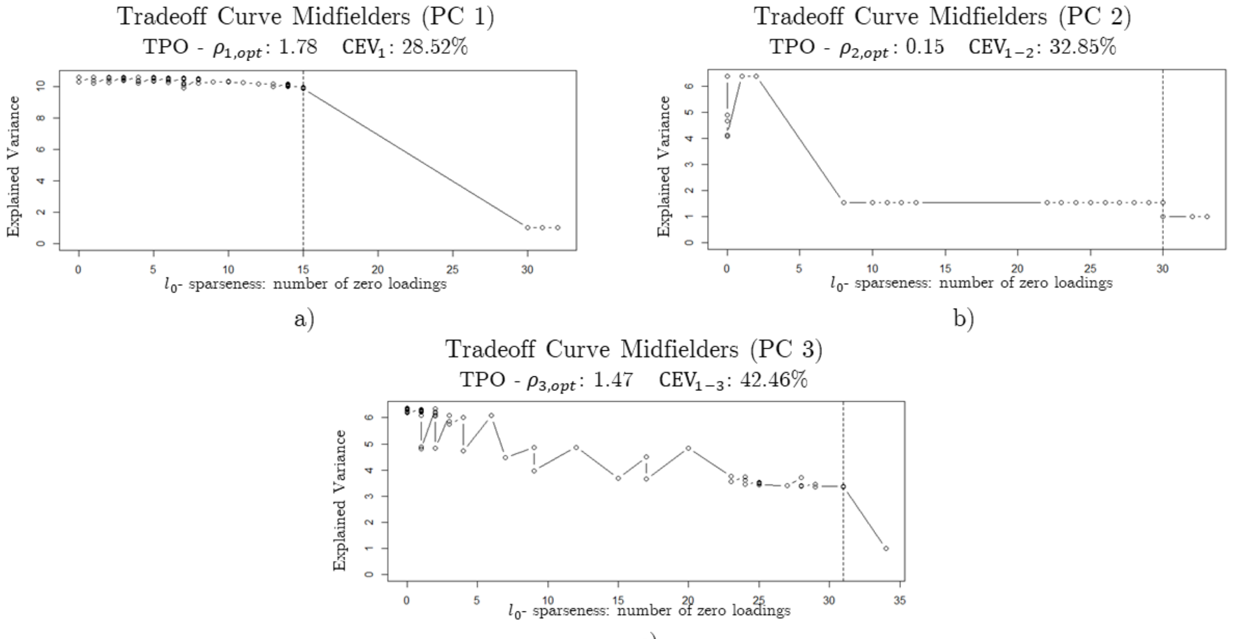


Figure B.7: Tradeoff curves of each component obtained with SPCA on goalkeepers with TPO. The dashed line indicates the optimal ρ for each component based on the TPO criterion.



a) b) c)

Figure B.8: Tradeoff curves of each component obtained with SPCA on defenders with TPO. The dashed line indicates the optimal ρ for each component based on the TPO criterion.



a) b) c)

Figure B.9: Tradeoff curves of each component obtained with SPCA on midfielders with TPO. The dashed line indicates the optimal ρ for each component based on the TPO criterion.

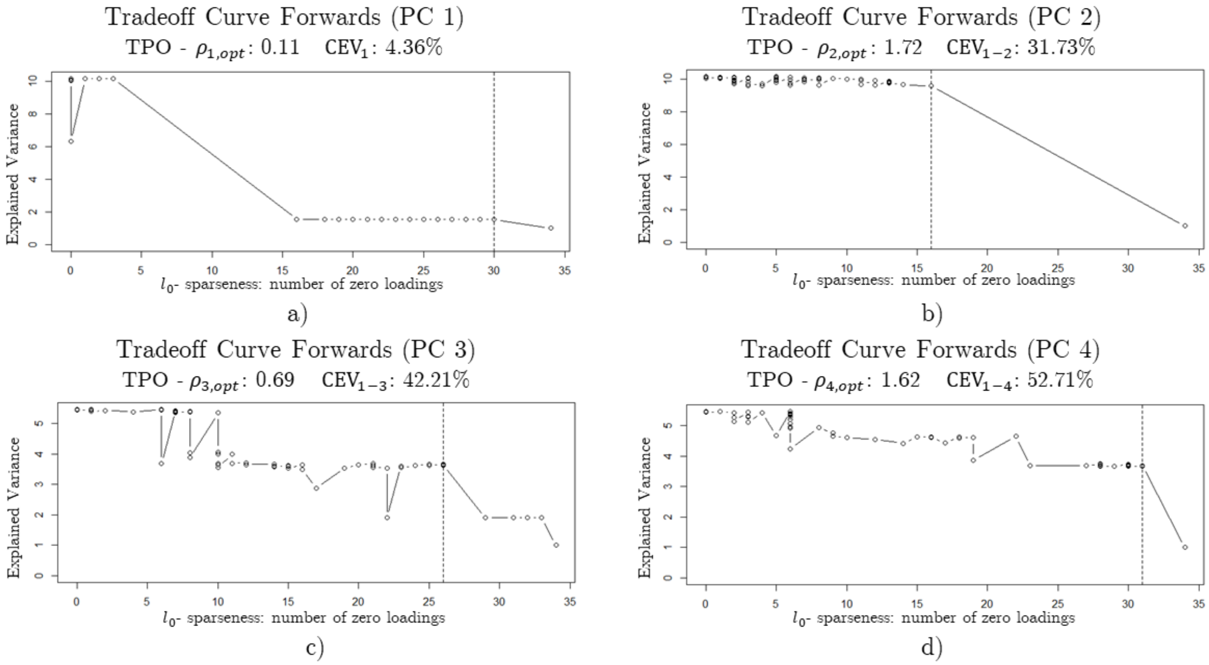


Figure B.10: Tradeoff curves of each component obtained with SPCA on forwards with TPO. The dashed line indicates the optimal ρ for each component based on the TPO criterion.

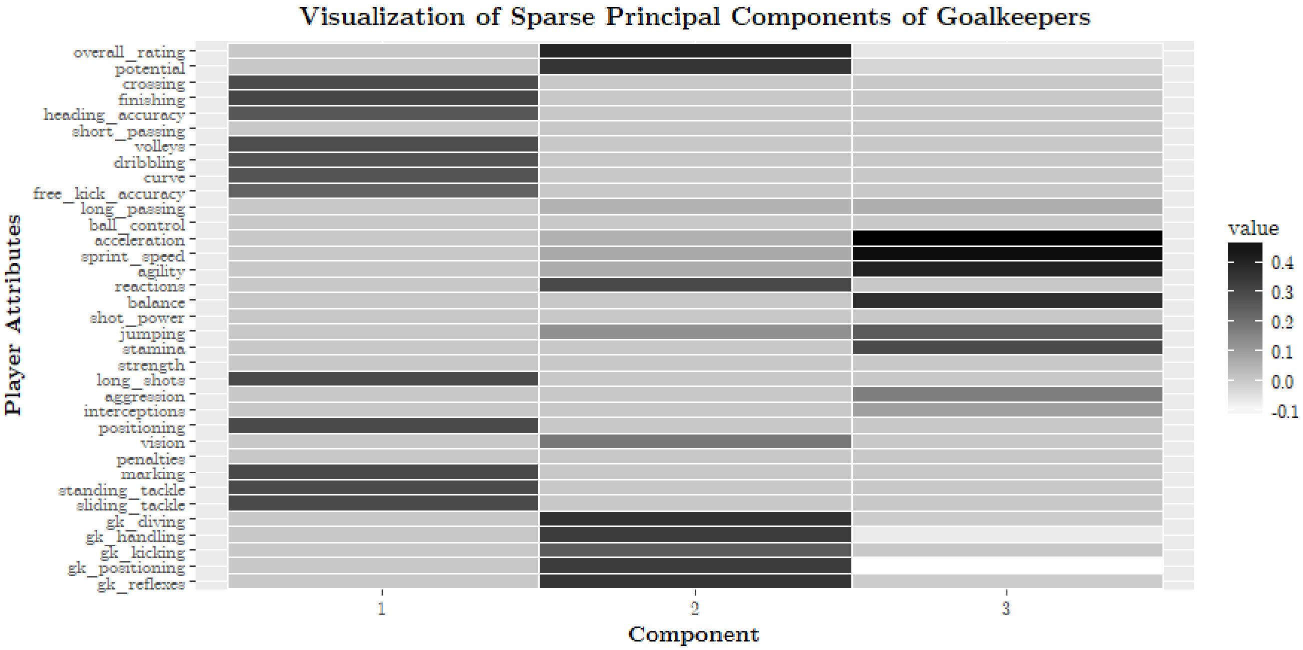


Figure B.11: Visualization of the sparse principal components of goalkeepers.

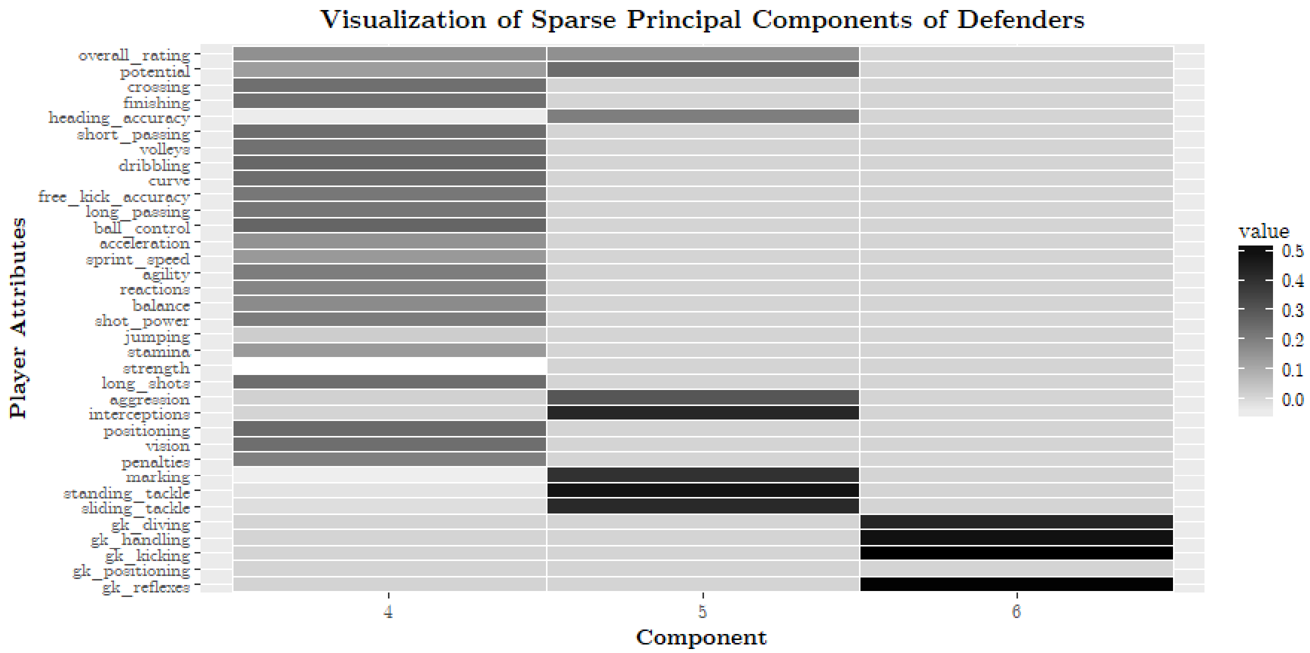


Figure B.12: Visualization of the sparse principal components of defenders.

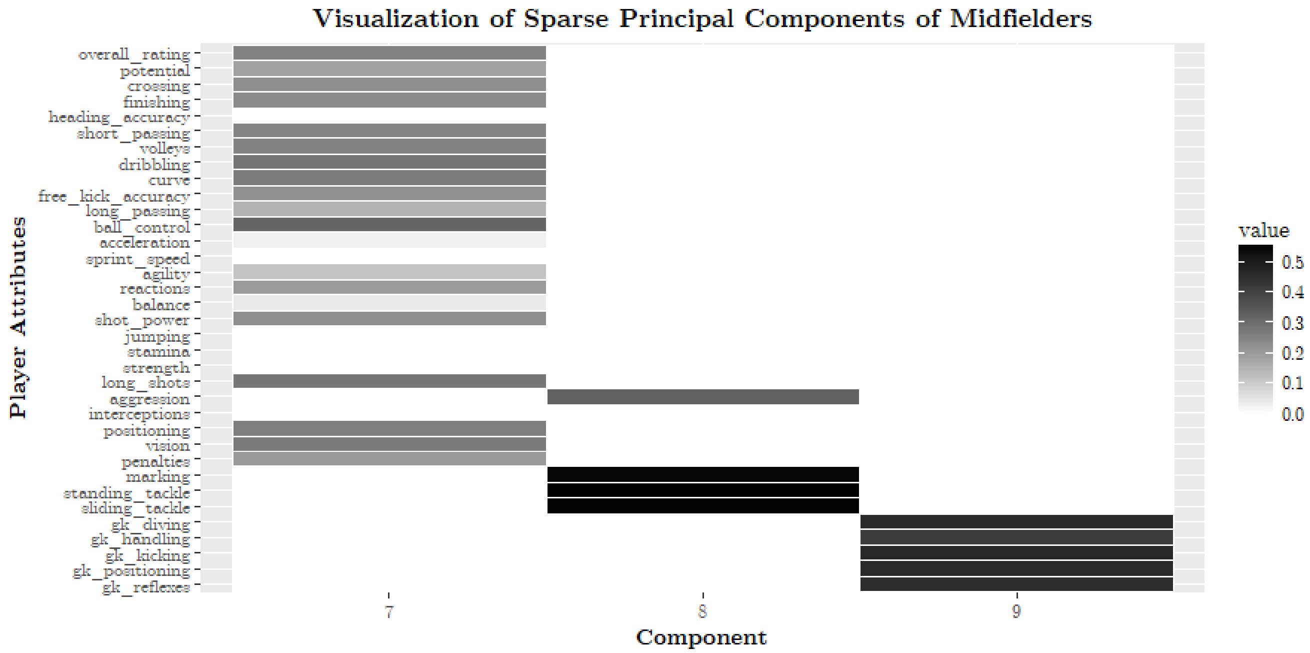


Figure B.13: Visualization of the sparse principal components of midfielders.

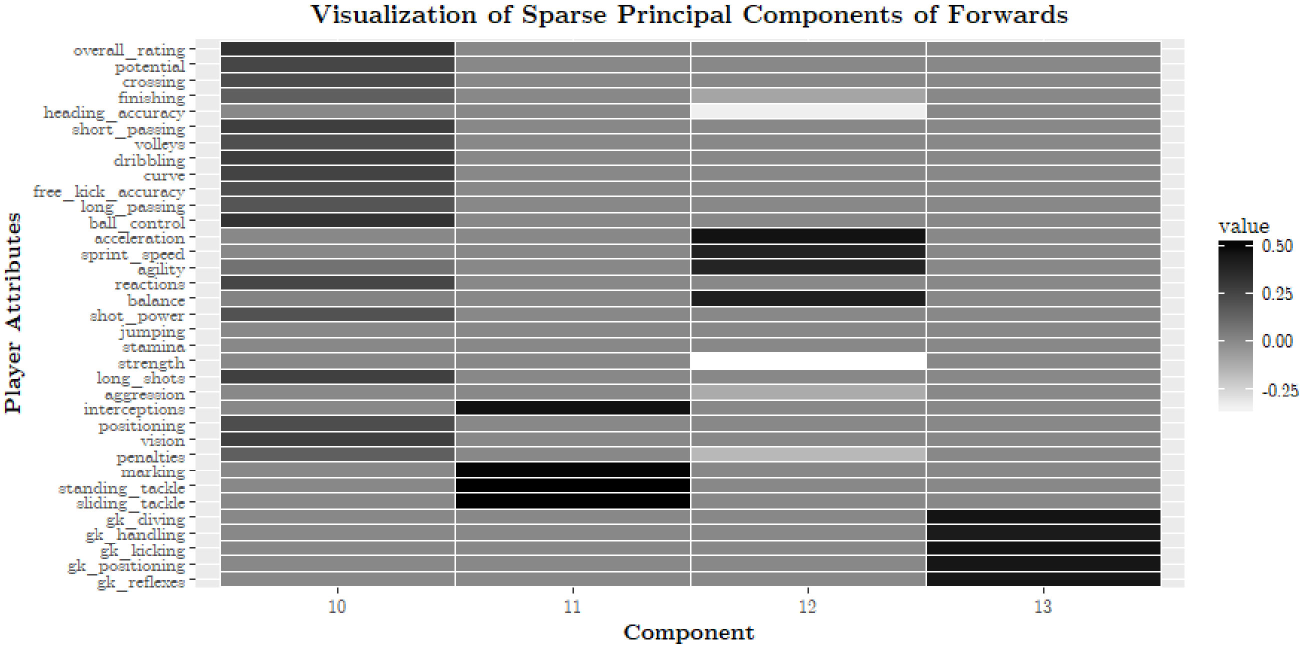


Figure B.14: Visualization of the sparse principal components of forwards.

Table B.2: Cluster validity indices Silhouette (Sil), Davies-Bouldin (DB), and Calinski-Harabasz (CH) obtained with K-means on projected data on the sparse principal components for different values of K. Numbers in bold show best result according to corresponding index.

	Goalkeepers			**Defenders**			**Midfielders**			**Forwards**		
K	**Sil**	**DB**	**CH**	**Sil**	**DB**	**CH**	**Sil**	**DB**	**CH**	**Sil**	**DB**	**CH**
2	**0.440**	0.958	421	0.243	1.576	1480	**0.319**	1.301	**2705**	0.237	1.682	1044
3	0.369	1.148	**509**	**0.284**	1.420	1616	0.287	1.271	2392	0.227	1.437	**1121**
4	0.341	1.139	477	0.273	**0.876**	**1754**	0.269	1.110	2276	0.203	1.388	1005
5	0.313	0.680	452	0.248	1.135	1665	0.249	1.356	2139	0.197	1.304	951
6	0.296	1.073	436	0.235	1.083	1557	0.275	1.019	2070	0.197	1.213	918
7	0.298	0.761	438	0.229	1.235	1505	0.268	**0.709**	2031	0.210	1.227	872
8	0.292	0.911	423	0.237	1.068	1485	0.265	1.046	2008	0.200	1.161	840
9	0.283	1.068	413	0.247	1.015	1432	0.305	1.017	2030	**0.248**	1.308	815
10	0.288	0.898	409	0.239	0.981	1387	0.288	1.019	1981	0.242	1.249	801
11	0.285	0.973	401	0.234	1.056	1347	0.278	1.128	1947	0.235	1.316	789
12	0.306	0.852	397	0.234	0.982	1313	0.270	1.043	1916	0.233	1.279	770
13	0.301	0.691	396	0.236	0.992	1283	0.269	0.838	1878	0.227	1.125	756
14	0.299	**0.670**	391	0.238	0.968	1257	0.266	0.947	1843	0.220	**1.064**	740
15	0.295	0.953	386	0.230	0.976	1231	0.263	0.864	1814	0.221	1.183	727

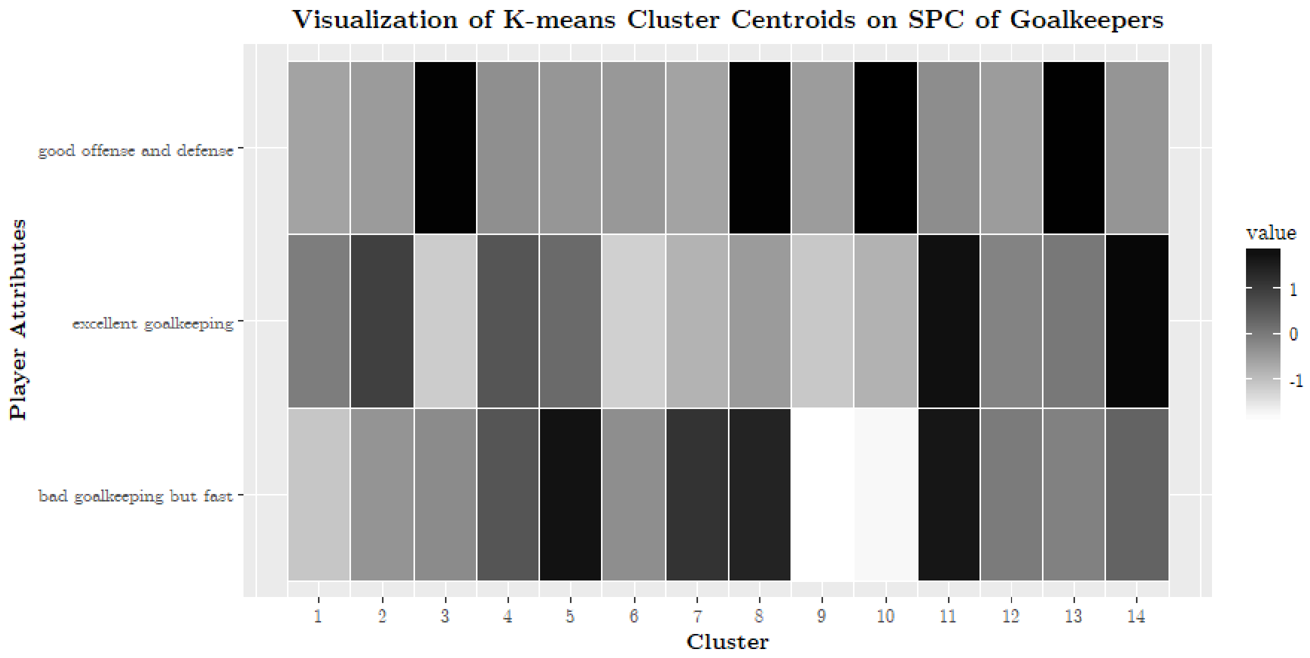


Figure B.15: Visualization of the centroids of goalkeeper clusters obtained with K-means on sparse principal components.

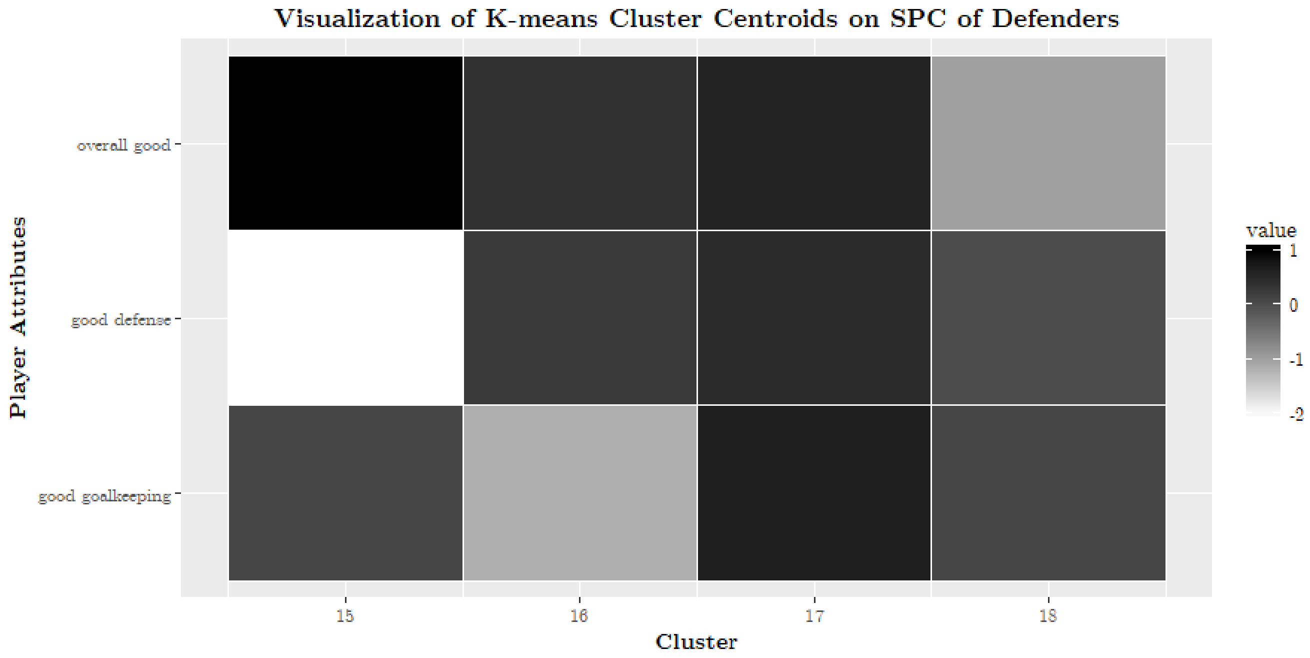


Figure B.16: Visualization of the centroids of defender clusters obtained with K-means on sparse principal components.

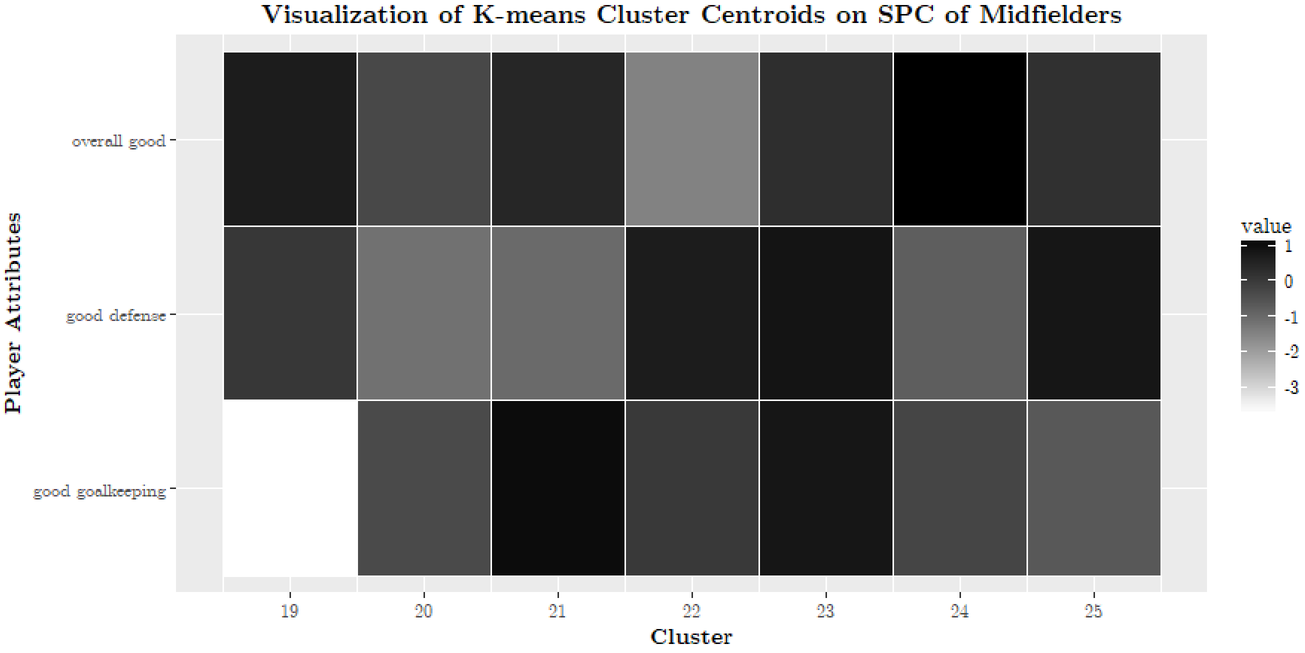


Figure B.17: Visualization of the centroids of midfielder clusters obtained with K-means on sparse principal components.

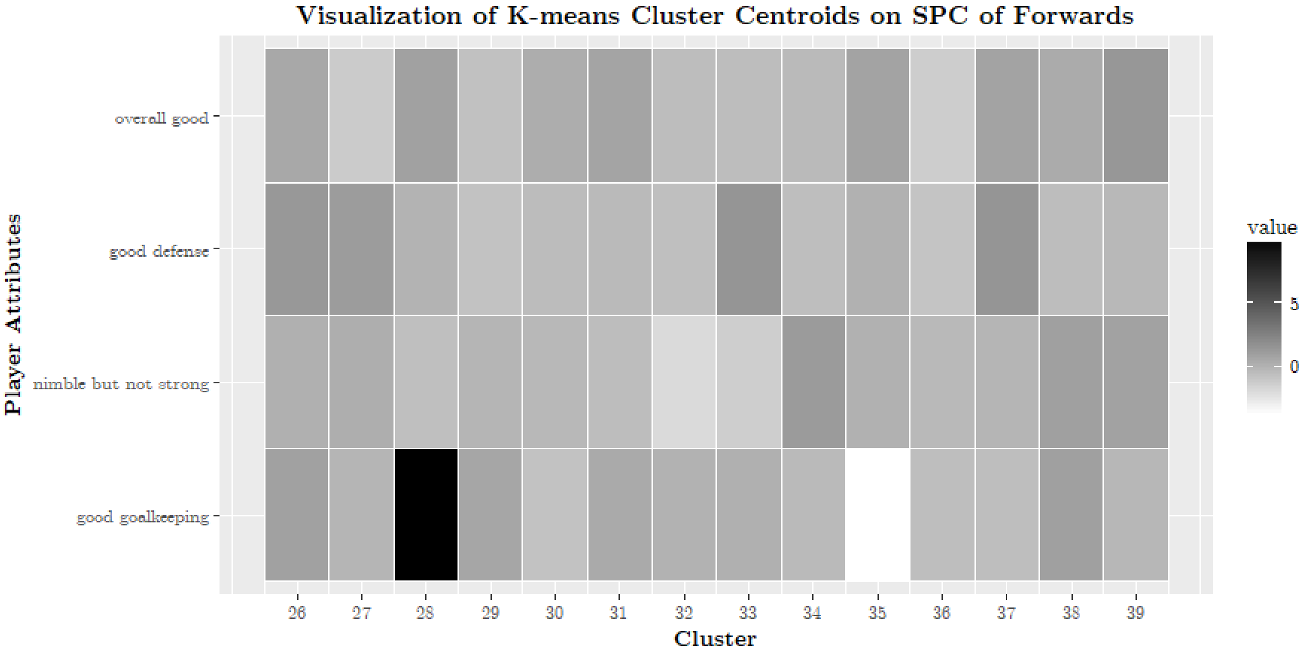


Figure B.18: Visualization of the centroids of forward clusters obtained with K-means on sparse principal components.

B.3 Archetypal Analysis

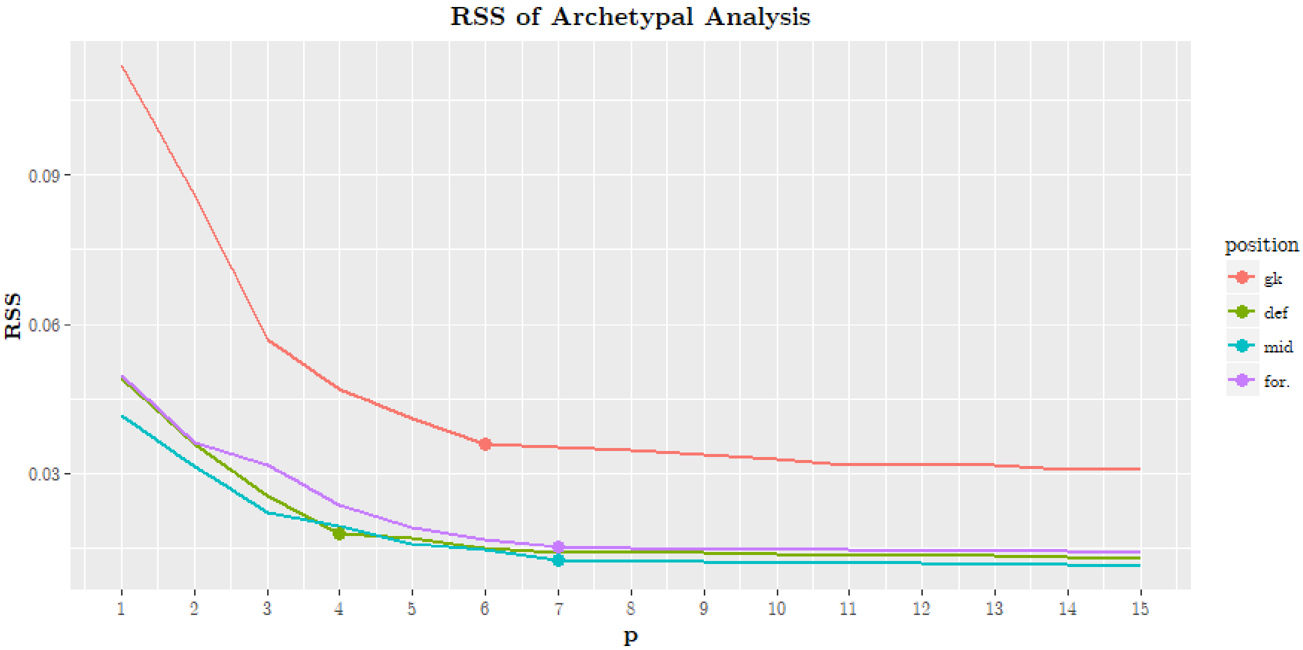


Figure B.19: Best RSS obtained with 50 trials of AA for each position, with points indicating identified elbows.

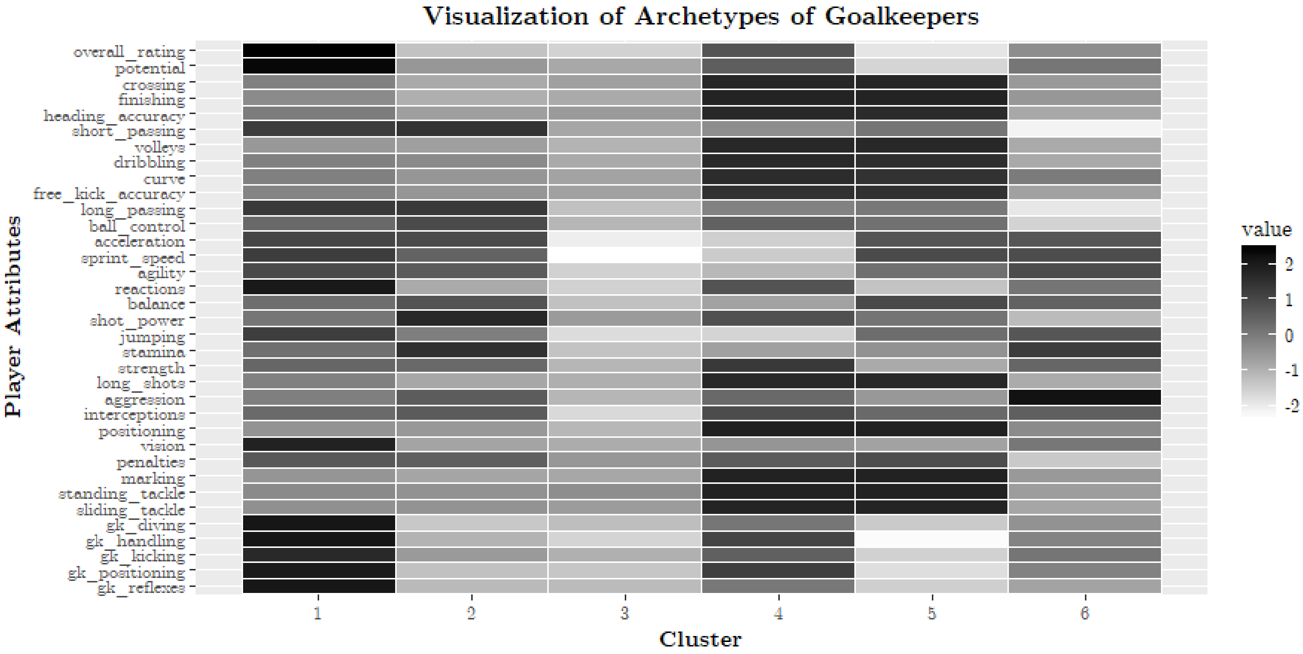


Figure B.20: Visualization of the archteypes of goalkeepers.

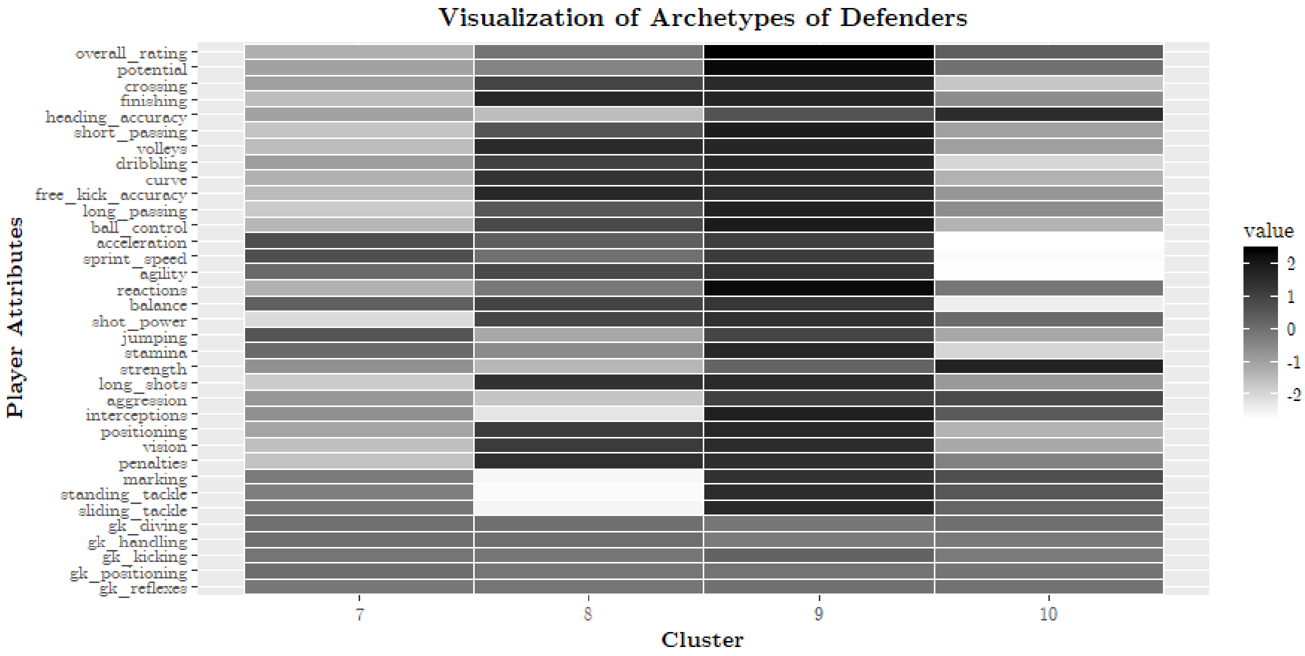


Figure B.21: Visualization of the archteypes of defenders.

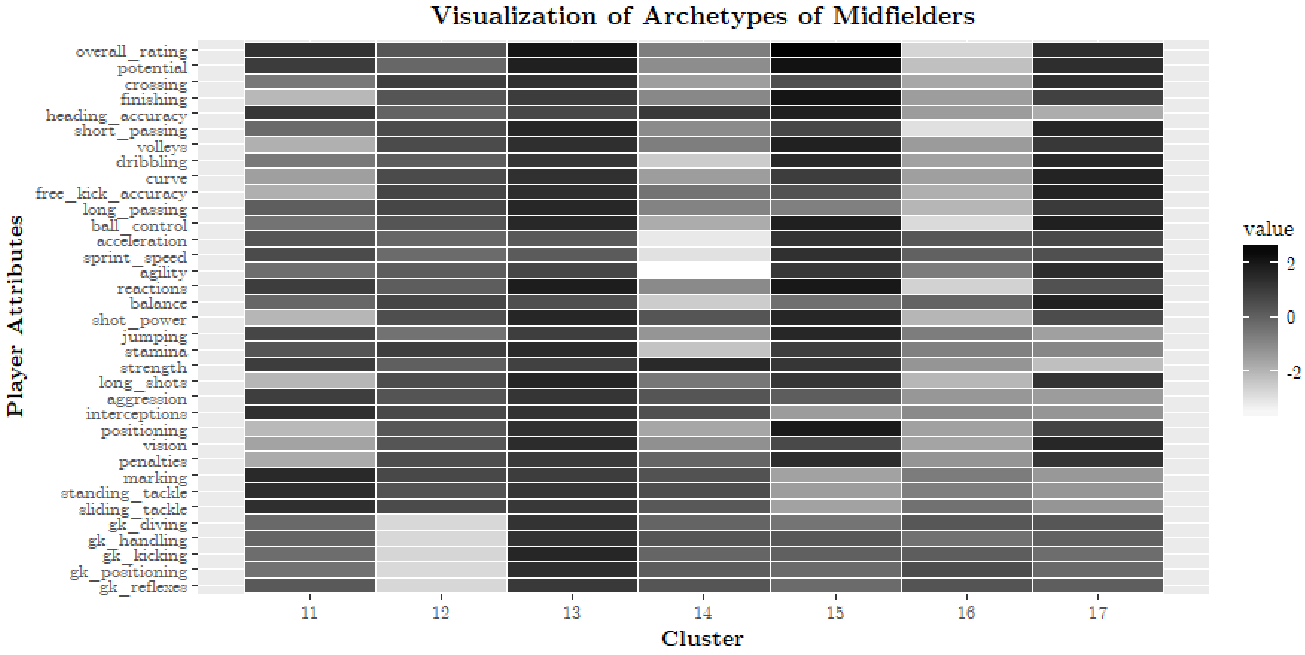


Figure B.22: Visualization of the archteypes of midfielders.

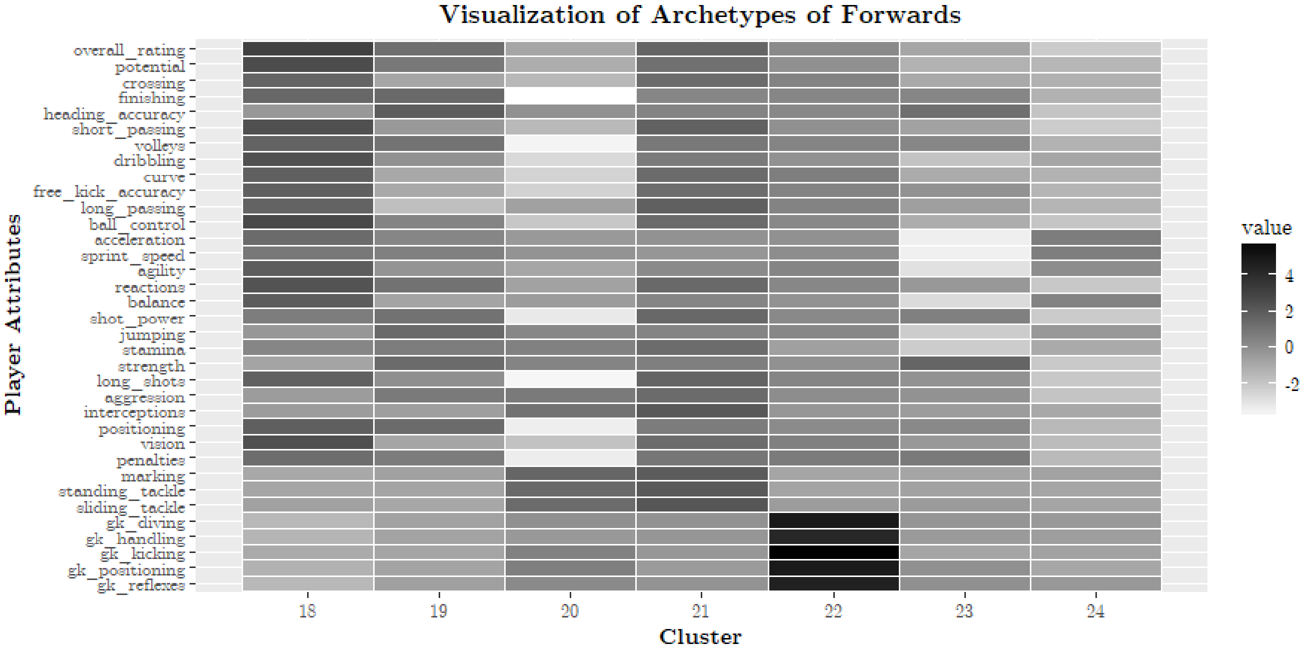


Figure B.23: Visualization of the archteypes of forwards.

C Rare Correlated Player Combinations

Table C.1: Support and *bond* measure of the mined rare correlated player combinations (PC) of the player types (PT) obtained with different 'optimal' *minsupp* and *minbond* thresholds for the three different clusterings.

PC	PT	Support	Bond
1	2	11%	1.00
2	4	46%	1.00
3	5	3%	1.00
4	6	70%	1.00
5	7	40%	1.00
6	8	64%	1.00
7	9	74%	1.00
8	10	16%	1.00
9	11	57%	1.00
10	12	23%	1.00
11	13	3%	1.00
12	14	21%	1.00
13	15	9%	1.00
14	16	19%	1.00
15	17	35%	1.00
16	18	58%	1.00
17	19	12%	1.00
18	4-11	35%	0.52
19	1-11	49%	0.51
20	6-18	44%	0.53
21	9-18	47%	0.55
22	3-18	51%	0.58
23	1-18	52%	0.55
24	6-8	50%	0.59
25	8-9	50%	0.57
26	3-8	56%	0.63
27	1-8	56%	0.59
28	6-9	59%	0.69
29	3-6	62%	0.71
30	1-6	63%	0.66
31	3-9	67%	0.77
32	1-9	67%	0.70
33	1-3	72%	0.75
34	3-8-9	46%	0.50
35	1-3-8	50%	0.51
36	3-6-9	55%	0.60
37	1-6-9	54%	0.55
38	1-3-6	57%	0.58
39	1-3-9	61%	0.62
40	1-3-6-9	50%	0.51

(a) K-means player types with *minsupp* 75%, *minbond* 0.50.

PC	PT	Support	Bond
1	1	10%	1.00
2	2	19%	1.00
3	3	3%	1.00
4	4	12%	1.00
5	5	3%	1.00
6	6	3%	1.00
7	7	3%	1.00
8	8	1%	1.00
9	9	3%	1.00
10	10	2%	1.00
11	11	8%	1.00
12	12	9%	1.00
13	13	5%	1.00
14	14	18%	1.00
15	15	5%	1.00
16	19	9%	1.00
17	20	18%	1.00
18	21	33%	1.00
19	22	22%	1.00
20	24	54%	1.00
21	26	6%	1.00
22	27	2%	1.00
23	28	1%	1.00
24	29	13%	1.00
25	30	29%	1.00
26	31	33%	1.00
27	32	19%	1.00
28	33	2%	1.00
29	34	11%	1.00
30	35	5%	1.00
31	36	6%	1.00
32	37	5%	1.00
33	38	17%	1.00
34	39	22%	1.00
35	16-24	36%	0.44
36	24-25	35%	0.41
37	23-24	38%	0.43
38	17-24	48%	0.55
39	18-24	42%	0.44
40	16-25	42%	0.47
41	16-23	45%	0.50
42	16-17	50%	0.53
43	16-18	49%	0.50
44	23-25	44%	0.47
45	17-25	54%	0.58
46	17-18-25	44%	0.44
47	17-18-23	46%	0.47

(b) SPCA with K-means player types with *minsupp* 55%, *minbond* 0.40.

PC	PT	Support	Bond
1	1	60%	1.00
2	2	5%	1.00
3	3	14%	1.00
4	4	6%	1.00
5	5	6%	1.00
6	6	8%	1.00
7	7	58%	1.00
8	8	11%	1.00
9	9	75%	1.00
10	10	79%	1.00
11	11	23%	1.00
12	12	24%	1.00
13	14	26%	1.00
14	15	26%	1.00
15	16	26%	1.00
16	17	57%	1.00
17	18	35%	1.00
18	19	60%	1.00
19	20	1%	1.00
20	21	17%	1.00
21	22	2%	1.00
22	23	20%	1.00
23	24	23%	1.00
24	17-19	35%	0.42
25	1-17	37%	0.46
26	9-17	47%	0.56
27	10-17	46%	0.51
28	13-17	47%	0.51
29	7-10	44%	0.47
30	7-13	45%	0.47
31	1-19	40%	0.49
32	9-19	48%	0.56
33	10-19	48%	0.53
34	13-19	51%	0.56
35	1-9	52%	0.62
36	1-10	47%	0.51
37	1-13	54%	0.61
38	9-10	59%	0.62
39	9-13	65%	0.71
40	10-13	66%	0.69
41	9-13-17	40%	0.42
42	9-13-19	43%	0.45
43	10-13-19	41%	0.43
44	1-9-10	40%	0.41
45	1-9-13	47%	0.50
46	1-10-13	42%	0.44
47	9-10-13	51%	0.52

(c) AA player types with *minsupp* 80%, *minbond* 0.40.