Pattern identification in Sport Data

Nikoletta Louca
s1303824@sms.ed.ac.uk

October 21, 2016

Contents

1 Introduction and Aims of project

This project looks to discover and analyze patterns that arise from the motion of players during a game of football (soccer). A team's quality is often judged by sparse events during a football game, such as shots on target, fouls and corners. More importantly, the winner in a football game is the team that has scored the most goals. Recent research [1, 2] has shown that this is a poor indicator of a team's performance, and a better indicator is the team's formation and can be identified by spatiotemporal analysis of players' positions. This project examines these properties, interpreting the data obtained from players' positions, drawing conclusions from that.

2 Methods

Throughout the duration of the project I have been reading relevant papers and applied methods that they used to a dataset that I had obtained. I also used analytics methods [5, 6] as a basis to try new things on my own.

One of the techniques used is SVD (Singular Value Decomposition). This technique enables one to see which were the most important movements, the ones that dominated the team's game. However, as mentioned before, the winner in a game will be judged by sparse events - goals. The run that will end up in a goal may or may not be included in the principal movements. Still, according to the research mentioned in section 1, this gives a lot of information about the strategies of the team, which is what should be studied.

The method of SVD is used as follows: Let

$$A = \begin{bmatrix} x_{t_1}^1 & y_{t_1}^1 & x_{t_1}^2 & y_{t_1}^2 & \cdots \\ x_{t_2}^2 & y_{t_2}^1 & \ddots & & \\ \vdots & & & \ddots & \end{bmatrix}$$

be the matrix containing all the players' information at all times, where $x_{t_i}^j$ is the x position of the j-th player at time t_i and $y_{t_i}^j$ the respective y position. Now SVD gives $A = U\Sigma V^T$, where U and V contain the left and right singular vectors of A respectively, and Σ has the singular values on its diagonal, arranged by highest to lowest. This enables one to extract the most important information with respect to time (given by the columns of V in decreasing significance) and position (given by the columns U in decreasing significance).

Another method used is k-means clustering. This is a method of clustering the data, and I have used 10 clusters (one for each player). I have used this together with confusion matrices, which was a technique described in [7, 8]. As mentioned in [8], assigning roles to each player is more insightful to a team's formation when one is only viewing the spatiotemporal data, as players tend to change positions between them during the game. The confusion matrix then examines to what extend the players were consistently following a formation. If they do, the confusion matrix should have its highest values along the main diagonal.

One more technique used is the Katz centrality for dynamically evolving networks [5]. Considering the players as nodes of a network depending on time, we can study how some properties between players vary through time. Two properties examined are the distance between players and the direction they are moving towards. These give quantitative results, so we can compare whether sets of players are more related with respect to distance or direction.

An adjacency matrix is formed for regular time intervals, creating a set of adjacency matrices A_{t_j}. Now the matrix we are interested in is given by

$$Q = \Pi_{j=1}^{n}(I - \alpha A_{t_j})^{-1} \tag{1}$$

where n is the number of adjacency matrices, and α a constant which has to be less than the spectral radius of the A_{t_j} for all j.

From this, two vectors are created, $\mathbf{b} = Q^T\mathbf{s}$ and $\mathbf{r} = Q\mathbf{s}$, where $\mathbf{s}$ is the column vector of ones. These two vectors represent the transmitters and receivers of direction/distance depending on what is being measured.

3 Results

I have been working on a set of three games from a Norwegian football team, made available through [3]. The data can be found in [4]. The dataset consists of a timestamp, a player tag, x and y positions as shown in figure 1 , direction, speed, distance covered and heading for each player. The sampling rate of the data is 20Hz.

For this report I have only included the results obtained from one of the games for consistency and in order to be able to draw sensible conclusions.

3.1 Full-team analysis

A rather insightful image is the occupation density or more commonly known as heat map of a team. This indicates the more popular areas of the pitch, where players spend most of the time. The pitch is divided into boxes, and each one has a number assigned to it according to the amount of times any player is in that box. Then each box is given a colour according to the scale shown on the right of the figure. Figure 2 shows a heat map produced for the first and second halves of the game.

Using the method of SVD, I have recreated the heat map for the first half, only now I am using the first singular value to reconstruct the data matrix, then adding the second and so on until the seventh, as shown below with A_k being

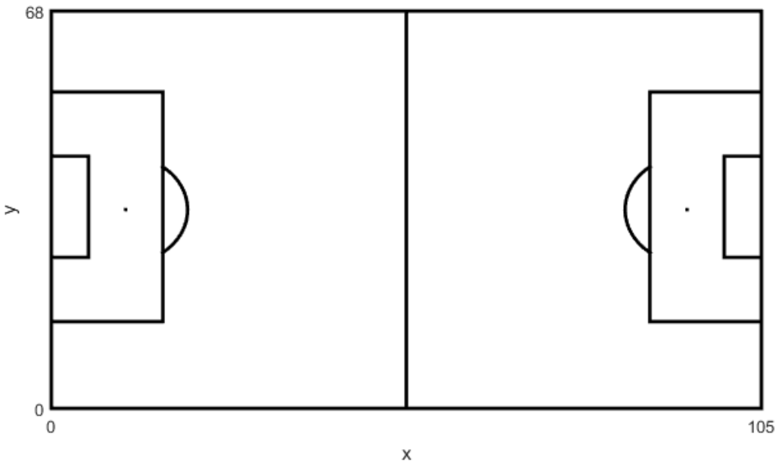

Figure 1: Pitch dimensions.

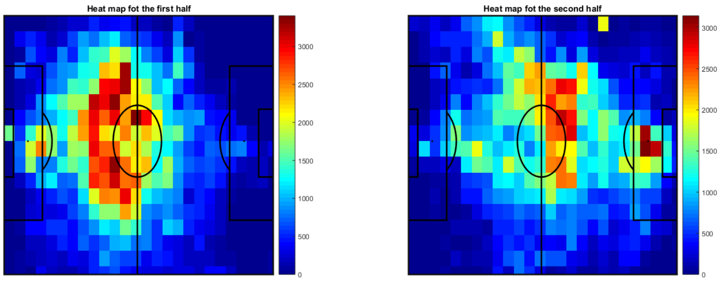

Figure 2: Heat maps for the two halves of the game with red indicating the areas where players spent most of the time and blue the least. On the left, during the first half, the team was attacking from left to right and on the right, during the second half, the team was attacking from right to left.

the matrix including the first k singular values λ.

$$A_{20} = U \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & & \ddots & \\ \vdots & & & \lambda_{20} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} V^T$$

$$A_1 = U \begin{bmatrix} \lambda_1 & 0 & \cdots & \\ 0 & 0 & \cdots & \\ \vdots & & \ddots & \\ \vdots & & 0 & \\ \vdots & \vdots & \vdots & \end{bmatrix} V^T$$

$$A_k = U \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & & \ddots & \\ \vdots & & \lambda_k & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} V^T$$

This is shown in an animation in [9]. After the seventh singular value there is a relatively big drop in the singular values, which means that we can represent the original data with a matrix with a much smaller rank (7 instead of 20) without losing a lot of information.

In figures 3 and 4 are also shown the four first main directions in which the team moved for the first and the second halves respectively. These directions are the four right singular vectors corresponding to the four highest singular values of the matrix containing all the players positions. Comparing these we can see that in the first half the team was more defensive whereas in the second half the team showed to have more possession in the opposing team's half. However, there seems to be a less structured style of play in the second half, and also a strong favouring of the right side.

Studying this long-term, and with more games, we could uncover as well as improve team strategies.

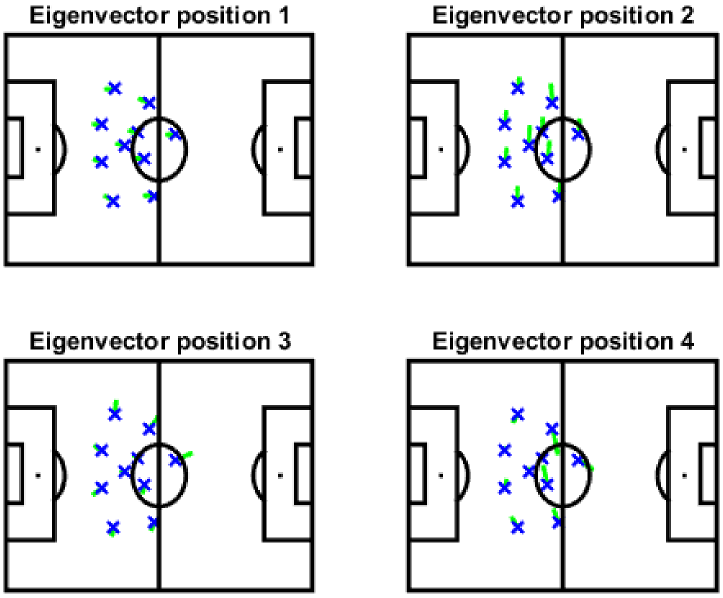

Figure 3: The main directions of the team during the first half, with blue crosses indicating the mean position of the player and green the direction of the corresponding player. The first eigenvector is the one corresponding to the highest eigenvalue, and the fourth corresponds to the fourth highest.

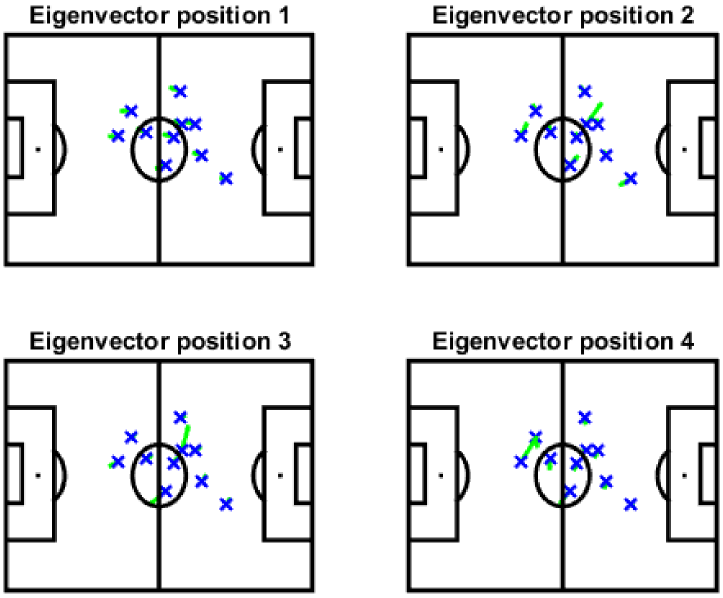

Figure 4: The main directions of the team during the second half, with blue crosses indicating the mean position of the player and green the direction of the corresponding player. The first eigenvector is the one corresponding to the highest eigenvalue, and the fourth corresponds to the fourth highest.

3.2 Relationships between pairs of players

In order to study relations between groups of players I have used the Katz centrality for dynamically evolving networks. The two things to be compared for the groups were their direction and the distance between them. The groups used in this case were the two center-backs, the left and right defenders, and the two center midfielders.The results are listed in the table below.

Position	Highest Relationship
Center-backs	Direction
Left and right defenders	Direction
Center Midfielders	Distance

3.3 Players and roles

As mentioned in section 2, assigning players to roles reveals the true formation of a team, not affected by players switching roles. I have created clusters using k-means clustering every 5 minutes. An animation showing how the centroids of the clusters moved is shown in [10].

4 Discussion

The very first difficulty encountered was extracting the data in the desired format, and then due to the fact that it was such a large piece of data, I had to find ways to manipulate it so that it was both informative but also not time-consuming. A true obstacle in this was the fact that the data was given with no further information on the details of the game, so in fact very little was known about what was happening in the game and had to be extracted through the data - which is precisely what I am claiming that these methods are doing. Moreover, there were some anomalies in the data which can be explained by measurement errors which had to be taken account of. Another objective was to see how this was applied to a larger scale, but unfortunately a larger dataset could not be obtained.

5 Personal Statement

Reading similar papers gave me an idea about what experts in the field have already done.

Experimenting and trying out new things was also very beneficial as I got a glimpse of how it is to perform research.

This project has also greatly enhanced my Matlab skills. Spending so much time working on it and trying to solve problems that arose, I came to find it rather exciting, and would be glad to tackle similar projects in the future.

6 Summary

In this project I tried to understand a team's style of play solely based on its players movement on the pitch. The ultimate aim is to help a team better structure itself to optimise its strength against its opponents, and at the same time they could be able to read the opponents strategies.

7 Acknowledgments

My two supervisors for this project were Professor Jacques Vanneste and Dr. Michal Branicki both of who guided me very well during this project.

References

[1] Bialkowski A. et al, "Identifying Team Style in Soccer using Formations Learned from Spatiotemporal Tracking Data", in the ICDM Workshop on Spatial and Spatiotemporal Data Mining (SSTDM), 2014.

[2] Lucey P. et al, " "Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data", in MIT Sloan Sports Analyrics Conference (MITSSAC), 2015.

[3] Pettersen S.A. et al, "Soccer Video and Player Position Dataset", in Proceedings of the International Conference on Multimedia Systems (MMSys), Singapore, March 2014.

[4] Soccer Video and Player Position Dataset. `http://home.ifi.uio.no/paalh/dataset/alfheim/`, 2013.

[5] Grindord P., "Mathematical Underpinnings of Analytics", Oxford, 2015.

[6] Shlens J, "A tutorial on Principal Component Analysis", Google Research, 2003.

[7] Lucey P. et al, "Representing and Discovering Adversarial Team Behaviors using Player Roles", in IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[8] Bialkowski A. et al, "Large-Scale Analysis of Soccer Matches using Spatiotemporal Tracking Data", in International Conference on Data Mining (ICDM), 2014.

[9] The animation with 7 singular values can be found in `https://s25.postimg.org/x6zfq5gxb/singularheat.gif`.
An animation with all 20 singular values can be found in `https://s12.postimg.io/ffuhmed59/singularheat.gif`.

[10] The animation can be found in `http://gifyu.com/images/citampougiou2.gif`