How to rate the performance of a soccer team? An application of Principal Components Analysis

For this analysis, I will use a database that contains information of matches form La Liga and the Premier League, from 2011 to 2016.

As the title might give you a clue I will use principal components for this task, unlike other posts, this time I will highlight some technical points around the analysis and code lines as well.

First let’s start loading the libraries and the data as well as some options to make our life easier with the data wrangling.

CODE

The data contains stats associated to 4142 matches from the Premier League and La Liga, each column has also the home team associated.

Analysis

Once the data is in a nice format we can start running the Principal component analysis (PCA). The Principal component analysis can use either covariance matrix or the correlation matrix to be performed, in this case I will use the correlation matrix because there are some variables that have a wider range such as Fouls or Possession than others like Goals or Red Cards, this gives right away a huge impact on the analysis, and that is the interpretability of the Components, as you will see below the Components have a fairly intuitive explanation fort this dataset using the correlation matrix for the analysis.

CODE

PCA is a dimensional reduction technique, this allow us to plot the first 2 components, making it easier to present them. The biplot is a great tool to show how variables are related and how much they impact on each component.

On the table below we can see the loadings of each component, from this table is simple to understand that the first component rate the team performance, while the second one explains how often the referee was involved during the match.

Loadings on the components can switch the signs indistinctly if you change the sign on all the component values, on the first component if a team performed great the value will be negative, but that’s not intuitive so, to make things more natural let’s change the direction of the first component.

Now let’s take 4 teams to show their first component time series.

PLOT (GRANDA< REAL, BARCA< MANCHESTER )

Form here is possible to see that teams as Barcelona or Manchester City have consistently positive rates provided by the first component on their matches performance, while smaller teams such as Granada or Stoke City have mixed values (Sorry Granada and Stoke City fan, but that’s the truth).

As you can see the first component rate the performance of a team, this would lead naturally to create a table that compare all the teams rate. The following table presents the average performance rate of each team for the last six years (up to January 2017).

TABLE

If you are a Barca fan you might agree with the table, although if you are a Real Madrid fan you will disagree, just remember that variables such as players’ handsomeness are not included in the analysis.

Now let’s take the average of the last 15 games of each team to obtain an ‘updated’ rate for both components. So far I haven’t done any analysis regarding the second component, although this is about to change.

Even though clustering is not the main goal of PCA it is possible to create groups according to the first components. As you might have seen on the previous plot the dotted lines show where the mean of each component lies, these lines create 4 groups.

* **The Artists**. - Teams with good performance and fluent matches, typically teams that have entertaining matches. (Real Madrid, Barcelona, Liverpool)
* **The Street Fighters**. - Teams with poor performance and matches that include tons of fouls and referee interruptions (Espanyol, Sunderland, Celta de Vigo)
* **The Rugby Teams**. - Teams with good performance and a lot of fouls and interruptions.  They might like better rugby rules. (Manchester United, Chelsea, Real Sociedad)
* **No Fouls, No Goal Teams**. - Teams that have bad performances but fluent matches. (Hull City, Villarreal, Everton)

To end I would like to point one curios behavior shown on the last image, most of the Premier League teams with good performance tend to have a lot of misconduct interruptions, while the opposite happens to La Liga teams. Is this something that has to do with the way referees apply the rules different in each league? Or is it just a way how team styles variate between this two leagues? Perhaps that could be a topic for a new post.