# Principal Component Analysis on Football Competitions via Linear Model in China Football Association Super League Tournament

Weinan Li[1, a], Guopeng Sang[2, b, *]

[1]Department of sports training, Jilin Sport University, Changchun, China;

[2]Department of sports training, Xi'an Physical Education University, Xi'an, China.

[a]liweinan@qq.com, [b]1016103072@qq.com

**Abstract.** Technical and tactical performance evaluation relies on statistical analysis. Principal component analysis was before established in bioinformatics in a decade. Big data collection has been increasing applied in football industry while principal component methods therefore can be introduced in the region to calculate the performance parameters that acquired during the single football competition. During the work, Tactical and technical data with the rank of 2017 China Football Association Super League Tournament were then assessed by linear regression model integrated by the method and confirmed by multiple correlation coefficient statistical analysis. It revealed that the principal component analysis is available to be used for football match performance evaluation and the result is correspondence to the final rank of the season. Furthermore, both attack and defense parameters for the teams have a positive correlation to the outcome of game. Therefore, the established model is a useful tool for coaches and group member to adjust their training program and evaluate their team competence.

**Keywords:** Principal component analysis; tactical; parameters.

## 1. Introduction

Big data has been widespread in industrial and academic areas. With the increasing development in internet network and hardware data collection system, it facilitates us to analyze and evaluate how single or multiple items influence on the final results and achievements of social activities, for instances, football competitions. Football club and train groups consumed decades to assess the accomplishment during the competition. Multiple parameters such as shots, shot on goal, scored goal, fouls, etc., have been applied into the matches to evaluate the competition results [1, 2]. However, the global judgment in competition series or comparison the performance of one football in different years is normally impeded by these independent standards. Individual parameters sometimes even interfere with each other [3, 4].

During one season of China Football Association Super League Tournament, the result of one hundred and twenty matches from sixteen football clubs were integrated by scores. Although the list via points acquired from win or failure can be a representative of team's performance. The details of tactical and technical elements were hidden behind. The coaches and executives require comprehensive analysis on each team players performance and individual standard parameters during the competition. Therefore, an integrated method which allows for examination of different individual parameters is essential to the analysis and record on tactical behaviors and technical points during the football training and matches.

Principal component analysis (PCA) is the common tool for the application in predictive models, which has been used in biological area for example, investigating the genetic distance between family groups or detecting the relatedness between populations [5-7]. This statistical procedure utilizes orthogonal transformation to adapt a wide range of potential correlated variables during the observation into a list of values of linearly uncorrelated variables. The whole process created principal components. When n observations can be defined with p variables, the number of distinct principal components is representative of min (n-1, p). Moreover, the first component is given into the largest possible variance and subsequent item has the highest variance possibility under the orthogonal matrix or arrays [7]. Therefore, the single individual performance parameter from big data collection system can be converted into such principal components to analyze their influences on the competition results. Meanwhile, multiple correlation coefficient statistical analysis was introduced to confirm the correlation of each competition-oriented performance parameters during PCA [8].

## 2. Material and Method

### 2.1 Samples and Variables

The samples of study applied for 120 games and 200 sets of technical and tactical statistical data from the 2017 season China Super League. The variables evaluated in the work consist of results of single match (win, flat, failure) and the technical and tactical parameters during the matches. 12 technical and tactical parameters were chosen. It includes goals scored, shots on goal, shooting, control ball rate, pass, offside, corner, steals, head off foul, yellow and red cards.

### 2.2 Principal Component Analysis

Within football performance data matrix, the rows N represents different matches in whole season and P columns reflect to particular parameters during the competition. The orthogonal transformation of each performance parameters can be describes as equation 1:

$$X_{j(i)}=N_{(i)}\times P_{(j)} \tag{1}$$

Where X means new principal components, for is=1, 2 and j=1, 2.

In order to acquire the maximum variance, P (j) can be described as

$$P_{(j)} =\arg\ \max\left\{\frac{NP^{T}PN^{T}}{P^{T}P}\right\} \tag{2}$$

And the correlation between P and X can be linearly plotted as equation 3 while the full components decomposed to X

$$X=N\times P \tag{3}$$

### 2.3 Multiple Correlation Coefficient Statistical Analysis

Multiple correlation coefficient measures the linear correlation of one variable to multiple variables. However, this statistical method merely allows for indirect correlation.

For instances, To determine the correlation coefficient between a variable y and other variables X1, X2, Xu, it is available to build a linear combination of X1, X2, , Xu toy. The simple correlation coefficient is used to describe the correlation between the variable y and X1, X2, Xu. The process is as follows in equation 4

$$\hat{y} = \widehat{a_0} + \widehat{a_1}X_1 + \cdots \widehat{a_k}X_k \tag{4}$$

Where is regard to be constant, therefore, the multiple correlation coefficient are deduced as equation 5

$$R=\frac{\sum(y-\bar{y})(\hat{y}-\bar{y})}{\sqrt{\sum(y-\bar{y})^2\sum(\hat{y}-\bar{y})^2}} \tag{5}$$

R is the complex correlation coefficient and the square of R represent to the linear regression equation coefficient. Equation 5 is described by equation 6.

$$R^2=\frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} \tag{6}$$

### 2.4 Statistical Analysis

The PCA and the calculation of multiple correlation coefficient statistical analysis are conducted by the data statistics software SPSS 20.0. The magnitude-based inference is calculated by Excel 2007. Cluster analysis was applied to the final result from PCA method [5], X scores of the 16 teams were served as average technical and tactical indicators. The clustering result was divided into 4 groups (excellent, good, medium, and poor). Z standard differentiation method was adapted to the clustering, the formula 7 is as follow:

$$Z= \frac{(t-t')}{s} \tag{7}$$

Where represent to the original score, is the average value of the scores and s is the standard deviation of the original score. Since Z value can be negative, we use an estimated equation to convert it back to positive t standard values. The whole clustering processing then described as equation 8.

$$t = 20 + 5 \times Z \tag{8}$$

Table 1. Result of PCA method on China football association super league tournament in 2017 season

| Football clubs | Goals[a] | Shot | Shot in goals | Pass | Corner | Ball control rate | Offside | Steal | Fouls | Interception | Red card | Yellow card | category |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Guangzhou Evergrande | 2,3 | 14,4 | 3,57 | 422,5 | 4,77 | 0,545 | 2,67 | 18,10 | 16,33 | 10,60 | 0 | 2,6 | 1 |
| Shanghai SIPG | 2,4 | 15,7 | 3,67 | 372,8 | 5,27 | 0,525 | 2,27 | 18 | 14,9 | 10,1 | 0,03 | 1,77 | 1 |
| Tianjin Quanjian | 1,53 | 13,1 | 2,93 | 343,5 | 4,67 | 0,482 | 2,83 | 17,20 | 15,93 | 10,60 | 0,03 | 2,7 | 1 |
| Hebei Fortune | 1,83 | 12 | 2,7 | 400,8 | 4,97 | 0,523 | 1,93 | 15,1 | 15,7 | 11,2 | 0,03 | 1,9 | 2 |
| Guangzhou R&F | 1,97 | 14 | 4 | 458,5 | 4,97 | 0,561 | 2,5 | 16,2 | 13,93 | 10,9 | 0,07 | 1,9 | 2 |
| Shandong Luneng | 1,63 | 14,3 | 3,17 | 403,5 | 6,13 | 0,513 | 1,87 | 19,20 | 17,57 | 13,30 | 0 | 2,3 | 2 |
| Changchun Yatai | 1,53 | 14,1 | 3,5 | 369,1 | 5,93 | 0,514 | 2,97 | 16,5 | 18,33 | 12,30 | 0,07 | 2,23 | 2 |
| Guizhou Hengfeng | 1,3 | 13,2 | 3,03 | 375,4 | 4,9 | 0,517 | 2,23 | 16,5 | 17,67 | 10,20 | 0,1 | 2,33 | 2 |
| Beijing Guoan | 1,4 | 14 | 3,37 | 419 | 5,27 | 0,551 | 2,87 | 17,6 | 16,3 | 12,70 | 0,07 | 2,23 | 3 |
| Chongqing Dangdai | 1,23 | 12,2 | 3 | 337,3 | 4,53 | 0,464 | 1,53 | 17,3 | 12,9 | 12,00 | 0 | 1,33 | 3 |
| Shanghai Shenhua | 1,73 | 12,4 | 2,57 | 245,9 | 5,03 | 0,491 | 2,5 | 15,2 | 13,77 | 10,60 | 0,1 | 2,17 | 3 |
| Jiangsu Suning | 1,33 | 11,8 | 2,93 | 358,8 | 4,87 | 0,478 | 1,63 | 16,70 | 16,63 | 11,50 | 0,03 | 2,53 | 3 |
| Tianjin Teda | 1 | 13,2 | 3,37 | 381,2 | 5,33 | 0,493 | 1,73 | 16,4 | 15,4 | 11,10 | 0 | 2,23 | 3 |
| Henan Jianye | 1,13 | 11,2 | 2,3 | 317,6 | 3,6 | 0,434 | 1,77 | 18,5 | 14,4 | 12,60 | 0,03 | 1,63 | 4 |
| Yanbian Funde | 1,1 | 11,3 | 2,9 | 426,6 | 4,07 | 0,468 | 1,9 | 17,7 | 13,17 | 11,10 | 0 | 1,43 | 3 |
| Liaoning Whowin | 1 | 10,1 | 2,2 | 328,7 | 3,6 | 0,442 | 1,63 | 16,8 | 16,3 | 11,50 | 0,03 | 1,97 | 4 |

A: all parameters data were divided by the game.

## 3. Results

The technical and tactical performance data among 16 teams in the 2017 China Super League were applied for PCA method. As shown in Table 1, Guangzhou ever Grande football club accomplish the leadership among the competitions with higher individual performance parameters such as goals, steals, ball possession rate, and passing. In contrast, such key parameters in Liaoning Kaolin team (Liaoning Who win) are different from other 15 groups. According to the analysis of both offensive and defensive parameters data, Guangzhou ever Grande took control the whole matches during each events. Therefore, it is undoubtedly correspondence to the result of championship in 2017 season that ever Grande club wins

### 3.1 Defensive Parameters Analysis

During the analysis, defensive technical indicators consist of steals, fouls, interceptions, red cards, and yellow cards. In comparison to these five parameters, we found that steals, fouls, interceptions and yellow card data have an impact on the outcome during the matches while red cards does not contribute any statistical significance. Teams ranked from one to five were better at steals, fouls, interceptions and yellow cards than other teams. It can be seen that good defense plays a vital role in the outcome. Data is shown in Table 2

### 3.2 Attack Parameters Analysis

Statistical analysis of offensive indicators revealed that football clubs belonging of top 5 represent higher offensive parameters value than the team of last three in the list of table 1. It indicated that such parameters such as shot, shot on goal, corner, pass, offside and ball possession rate play a significant role in the outcome of the football competition. Therefore, initiative and attack during the game is vital for the team to control the field and win.

Table 2. PCA method and cluster analysis of attack and defensive tactical parameters

| Rank | Attack tactical parameters | | | | | Defensive tactical parameters | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Shot[a] | Shot in goals | Pass | Corner | BPR[b] | Steal | Fouls | Interception | Yellow card |
| 1-5 | 13,84 | 3,37 | 399,62 | 4,9276 | 0,5272 | 16,92 | 15,3598 | 10,68 | 2,1734 |
| 6-11 | 13,37 | 3,11 | 358,37 | 5,30 | 0,51 | 17,05 | 16,089 | 11,85 | 2,099833 |
| 12-16 | 11,52 | 2,74 | 362,58 | 4,29 | 0,463 | 17,22 | 15,18 | 11,56 | 1,9598 |

A: all parameters data were divided by the game. B: ball possession rate

### 3.3 Multiple Correlation Coefficient Calculation

Multiple correlation coefficient () was calculated as 0.976 from the statistic model by SPSS 20.0. the linear regression model correlated with goals scored, shots on goal, shooting, control ball rate, pass, offside, corner, steals, head off foul, yellow and red cards. The probability value of the regression is 0.010 (P<0.05), validated the linear model. The linear regression then can be described as below. The final result from the multiple correlation coefficient calculation is consistent to the cluster analysis conducted by PCA method. Therefore, the PCA method can be applied to study the performance parameters collected during the competition in China Football Super League 2017 season.

$$R = 22.837 - 0.041 * h + 0.142 * l - 0.221 * s - 0.27 * a - 0.019 * i - 0.063 * o + 0.027 * c \\ - 0.02 * f + 0 - 01 * y + 0.744 * r$$

Where R equals to rank, h: ball possession rate; l: Shot; s: Shot in goals; a: Pass; I: Corner; c: Steal; f: Fouls; y: Yellow card; r: Interception.

## 4. Discussion

The result of clustering analysis with PCA method is consistent with the final ranking of the China Football Super League. It revealed that the teams that present strong attack potential competence with high success rate and defensive technique showed a higher wins and rank.

Guangzhou Ever Grande was in the first class during the season. The value of the indicators, such as 69 goals in total games, including averaged 2.3 goals per game, averaged 14.4 goals per game, averaged 3.57 shot in goals per game and 54.5% ball possession rate, are comparatively higher than other teams' in the Super League, in which indicated its aggressive attack capacity during the competition. In contract, the team as well presented robust offensive capability as it shown on the data of technical and tactical parameters which are likely to steal, foul, intercept and yellow card. This is mainly due to the initial construction of the club, which included the investment on substitute team, hiring of professional foreign player to assist a competitive atmosphere around the groups and clear tactical idea and advanced training program.

In second category, the average goals are among one to two among five groups. It is noteworthy, the attack indicators of Guangzhou R&F are in high quality and the team was gathered in the second class although it ranked in fifth. Guangzhou R&F team possessed robust attack tactical competence during the football competition, however, the weakness on defense parameters were observed by the PCA analysis. The imbalance between attack and defense faded the whole team's progress.

Two teams, Henan Jeanie and Liaoning Kaolin, were clustered in the last group. The average number of scored goals was 34 and 30, respectively. The poor and incompetence performance on both attack and defensive tactical and technical parameters leaded to the result. It can be seen that there are 4 teams in the 1st and 4th categories in the 2017 season that accounts for one-third of the total number of Super League teams. The top group has excellent performance both on the offensive and defensive side. It is obvious that vigorous tactical and technical plan with good cooperation within the team players has an impact on the outcome of the game.

Based on the analysis of the principal components and the established complex correlation model, we found a positive correlation between the football tactical parameters and the rank of China Super League Tournament in 2017 season. Offensive and defensive technical indicators involved in the success of the game.

Taken together, how to improve the capability on attack and strengthen the defensive technique are two major problems that the teams in the Super League intended to contemplate. Based on the two analysis methods, it is apparent that ball possession, ability for ball control during exchange in high speed mode, communication and cooperation during the game and passing accuracy are key issues for all the teams planning their new training program among the gap of each season.

## References

[1]. C. Lago-Peña's, J. Lago-Ballesteros, A. Della, et al., Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league, J. Sports Sci. Med. 9(2) (2010) 288.

[2]. H. Liu, M.-Á. Gomez, C. Lago-Peña's, et al., Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup, J. Sports Sci. 33(12) (2015) 1205-1213.

[3]. R. Mackenzie, C. Cushion, Performance analysis in football: A critical review and implications for future research, J. Sports Sci. 31(6) (2013) 639-676.

[4]. E. Repining, A. Coutts, C. Castagno, et al., Variation in top level soccer match performance, Int. J. Sports Med. 28(12) (2007) 1018-1024.

[5]. H. Abdi, L. J. Williams, Principal component analysis, Wiley interdisciplinary reviews: computational statistics 2(4) (2010) 433-459.

[6]. S. P. Broglie, B. Schnabel, J. J. Sarnoff, et al., The biomechanical properties of concussions in high school football, Med. Sci. Sports Exec. 42(11) (2010) 2064.

[7]. I. T. Joliffe, Principal component analysis and factor analysis, Principal component analysis, Springer1986, pp. 115-128.

[8]. R. J. Aleman, D. L. Brown, A correlation coefficient for modal vector analysis, Proceedings of the 1st international modal analysis conference, Orlando: Union College Press, 1982, pp. 110-116.