Cricketers Ranking using PCA

Amulya Rachel (ee14btech11036@iith.ac.in)[1],Wasim Akram(ee14btech11037@iith.ac.in)[2]

***Abstract*— In this project we used PCA to rank the cricketers who played in the 2012 Indian Premier League (IPL) competition as referred in the paper[1].We have attached the required codes in the reference for further understanding[3].**

I. INTRODUCTION

We are trying to reciprocate the ranking done by Mr.Ramakrishnan, the official ranker for ESPN sports.We are using the data of IPL 2012 for both the batsman and bowlers considering 6 parameters for batsman and 4 parameter for the bowlers.After we collect the data we find the correlation matrix and then find the eigen values and corresponding eigen vectors. We observe that one of eigen value has more variability that the others hence we can rank the cricketers based on this eigen vector alone which we call as principal component which gives us the required results making the analysis to be called principal component analysis.

II. PARAMETER DETAILS

In the world of cricket there are lot of ways we can rank the performance of the cricketers using different parameters.There are 2 main divisions of cricketers called BATSMEN and BOWLERS who can be ranked in the same list.Hence we follow different ranking schemes for them.For the analysis, there are a set of widely-recognized variables/parameters that can be used to measure the quality of each player.

A. Parameters-Batting

1) RUNS :The total number of runs scored by a player in the IPL 2012 season. Higher values indicate stronger performance.:

2) BATTING AVERAGE(Ave) : The total number of runs a batsman has scored divided by the total number of times he has been called out in the IPL 2012 season.Higher values indicate stronger performance. However, for a batsman with several not out cases, this number overrates the batsman, which is a weakness in this measure, and this is why it should not be used as the only variable for batting performance analysis.:

3) BATTING STRIKE RATE (SR) :It is defined as the number of runs scored per 100 balls faced by a batsman in the IPL 2012 season.Higher values indicate stronger performance.:

4) FOURS :The total number of boundaries (fours = four runs) made in the IPL 2012 season by a batsman.Higher values indicate stronger performance.:

5) SIXES :The total number of sixes (= six runs) made in the IPL 2012 season by a batsman.Higher values indicate stronger performance.:

6) HALF CENTURY (HF's) :It consists of the number of Centuries (100 or more runs in an innings) together with the number of Fifties.Higher values are indicative of exceptional performance.:

B. Parameters-Bowling

1) WICKETS :The number of wickets taken by a bowler.The goal of a bowler is to get the maximum number of wickets by using a minimum number of balls while simultaneously conceding a minimum number of runs.:

2) BOWLING AVERAGE(Ave) :The average number of runs conceded per wicket.Lower values are preferred since a bowlers goal is to concede the minimum number of runs while simultaneously earning the maximum number of wickets.:

3) STRIKE RATE(SR) :The average number of balls bowled per wicket taken. Lower values are preferred since a bowler should try to bowl the minimum number of balls per wicket.:

4) ECONOMY RATE (Econ) :The average number of runs conceded per over. Lower values are preferred since this is the run-rate against a specific bowler for a batting team.:

III. BATSMEN

A. Data histogram

We have considered the data of 90 Batsmen from IPL 2012 for the above mentioned parameters.After we have considered the per parameter data of the batsmen we have plotted the histograms of the individual parameters with appropriate bin size as shown in the Fig.1.The attached code is in [3]

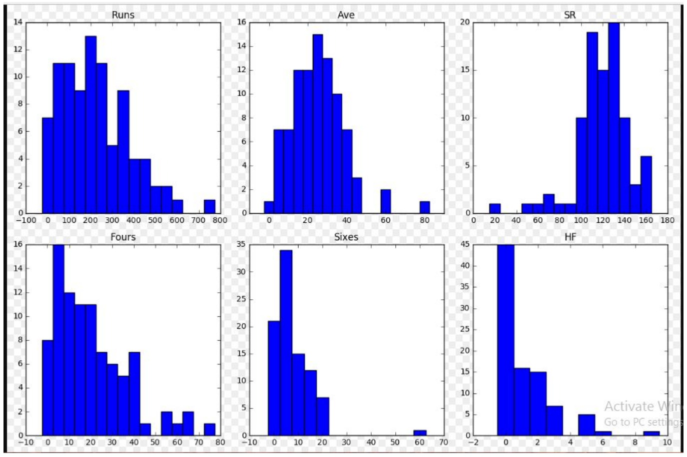


Fig. 1. Histogram of the parameter data

B. Correlation plots of the parametric data of batsmen

We plotted the parameter vs parameter data to find whether the data is correlated or not. The first row of the plot is Runs vs all the parameters and so on for better understanding.The attached code is in [3].

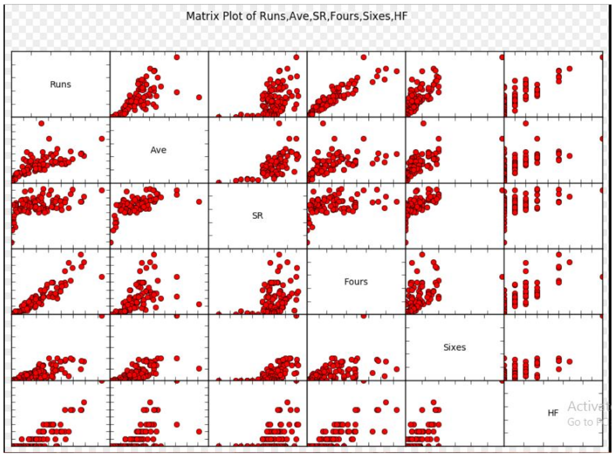
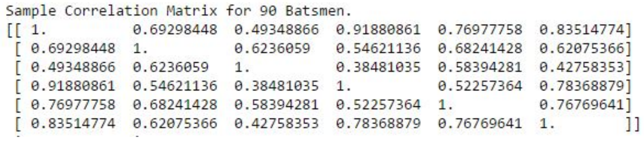


Fig. 2. The correlation of the parameters of batsmen

C. Correlation matrix

We normalized the data matrix of size 90x6 using Standardscaler and then we found the correlation matrix of size 6x6 for the same.The matrix is given below in the Fig.3. The code is in [3]

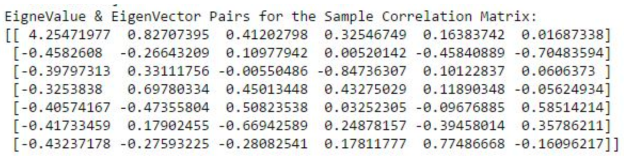


```
Sample Correlation Matrix for 90 Batsmen.
[[ 1.          0.69298448  0.49348866  0.91880861  0.76977758  0.83514774]
 [ 0.69298448  1.          0.6236059   0.54621136  0.68241428  0.62075366]
 [ 0.49348866  0.6236059   1.          0.38481035  0.58394281  0.42758353]
 [ 0.91880861  0.54621136  0.38481035  1.          0.52257364  0.78368879]
 [ 0.76977758  0.68241428  0.58394281  0.52257364  1.          0.76769641]
 [ 0.83514774  0.62075366  0.42758353  0.78368879  0.76769641  1.        ]]
```

Fig. 3. Correlation matrix of the parameters

D. Eigenvalues and Eigenvectors

As to reduce the dimensionality of the above matrix, we either do Eigen value decomposition or Singular value decomposition to the correlation matrix which results in only few significant eigen values and its corresponding eigen vectors.The eigen values and vector pairs are given below in the Fig.4.The code is in [3]

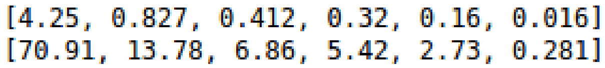


```
EigneValue & EigenVector Pairs for the Sample Correlation Matrix:
[[ 4.25471977  0.82707395  0.41202798  0.32546749  0.16383742  0.01687338]
 [-0.4582608  -0.26643209  0.10977942  0.00520142 -0.45840889 -0.70483594]
 [-0.39797313  0.33111756 -0.00550486 -0.84736307  0.10122837  0.0606373 ]
 [-0.3253838   0.69780334  0.45013448  0.43275029  0.11890348 -0.05624934]
 [-0.40574167 -0.47355804  0.50823538  0.03252305 -0.09676885  0.58514214]
 [-0.41733459  0.17902455 -0.66942589  0.24878157 -0.39458014  0.35786211]
 [-0.43237178 -0.27593225 -0.28082541  0.17811777  0.77486668 -0.16096217]]
```

Fig. 4. Eigen values and Eigen vector pairs

E. Variability of Eigenvalues

As mentioned earlier we don't need the whole data for finding the ranks as there is a lot of correlation in the data correlation matrix.Hence we can find the variability of the eigen values using the formula $Var_i = \frac{\lambda_i}{\sum \lambda_i}$ The different variabilities corresponding to the respective eigen values are computed as mentioned in the figure below i.e. Fig.5. The code is in [3]



```
[4.25, 0.827, 0.412, 0.32, 0.16, 0.016]
[70.91, 13.78, 6.86, 5.42, 2.73, 0.281]
```

Fig. 5. Percentage Variability of eigen values

F. Cumulative graph and Scree plot

We can observe that most of the data information can be obtained using the first eigen value alone. Hence,lets observe how significant they actually are using cumulative graph and the Scree plot as given in Fig.6 and 7.The code is in [3]

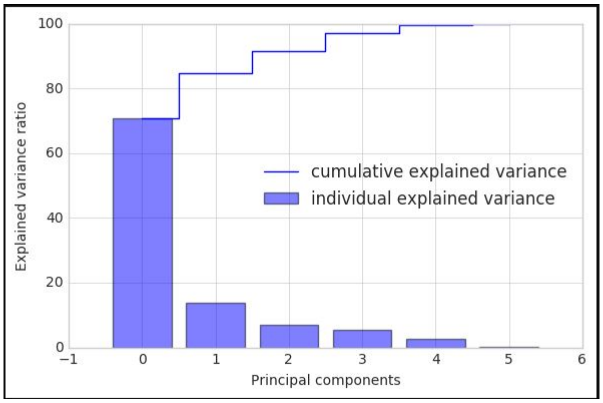


Fig. 6. Cumulative eigen value bar plot

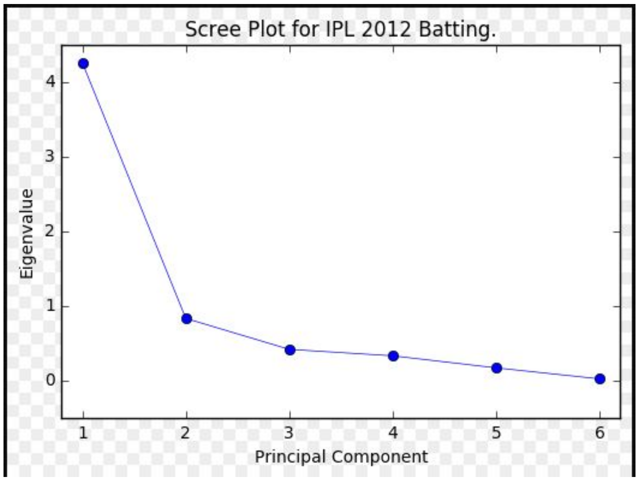


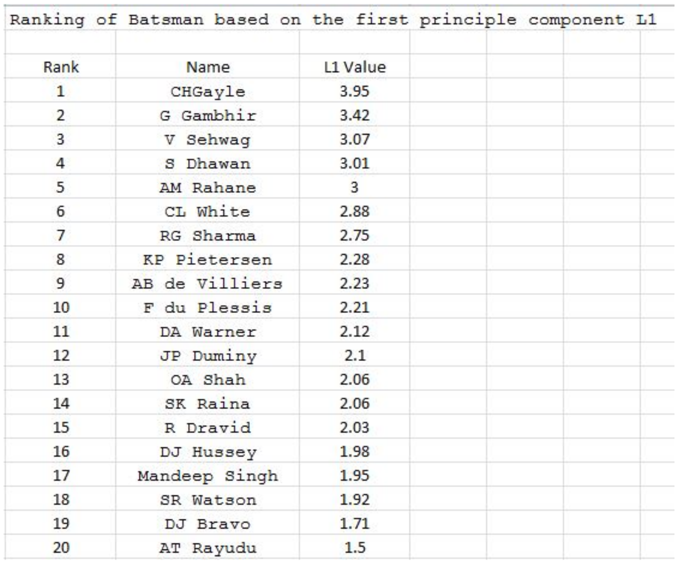
Fig. 7. Scree Plot

G. Ranking Criteria

Hence we considered only the first principal component's eigen value (4.25) and it's corresponding eigen vector as it constitutes more that 70 percent of the data necessary and found the L1 value as given below.
$L1 = 0.458*(Runs)+0.398*(Ave)+0.325*(SR)+0.406*(Fours)+0.417*(Sixes)+0.432*(HF)$
NOTE: We have to use the normalized values in calculating the L1 values and not the original data set.

H. Ranking Based on L1 values

We found the L1 values for all the 90 observations and have ranked them on the basis of decreasing values of L1 which are given in Fig.8 The code is in [3]



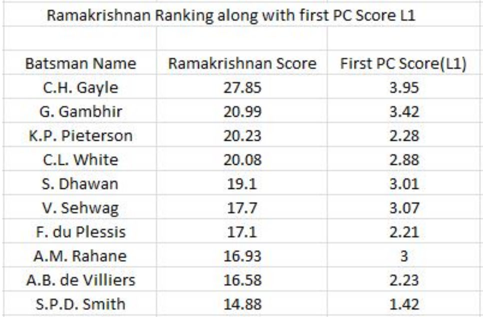
Ranking of Batsman based on the first principle component L1

Rank	Name	L1 Value
1	CHGayle	3.95
2	G Gambhir	3.42
3	V Sehwag	3.07
4	S Dhawan	3.01
5	AM Rahane	3
6	CL White	2.88
7	RG Sharma	2.75
8	KP Pietersen	2.28
9	AB de Villiers	2.23
10	F du Plessis	2.21
11	DA Warner	2.12
12	JP Duminy	2.1
13	OA Shah	2.06
14	SK Raina	2.06
15	R Dravid	2.03
16	DJ Hussey	1.98
17	Mandeep Singh	1.95
18	SR Watson	1.92
19	DJ Bravo	1.71
20	AT Rayudu	1.5

Fig. 8. Ranking of the top 20 Batsmen

I. Ranking comparison with Mr.Ramakrishnan

The comparison with the official ranking upto top 10 is given in the Fig.9.The details are given in the Observation section.



Ramakrishnan Ranking along with first PC Score L1

Batsman Name	Ramakrishnan Score	First PC Score(L1)
C.H. Gayle	27.85	3.95
G. Gambhir	20.99	3.42
K.P. Pieterson	20.23	2.28
C.L. White	20.08	2.88
S. Dhawan	19.1	3.01
V. Sehwag	17.7	3.07
F. du Plessis	17.1	2.21
A.M. Rahane	16.93	3
A.B. de Villiers	16.58	2.23
S.P.D. Smith	14.88	1.42

Fig. 9. Ranking of top 10 Batsmen comparison

IV. BOWLERS

A. Data histogram

We have considered the data of 83 Bowlers for the above mentioned parameters and plotted the histograms of the individual parameters with appropriate bin size in the Fig.10.The attached code is in [3]

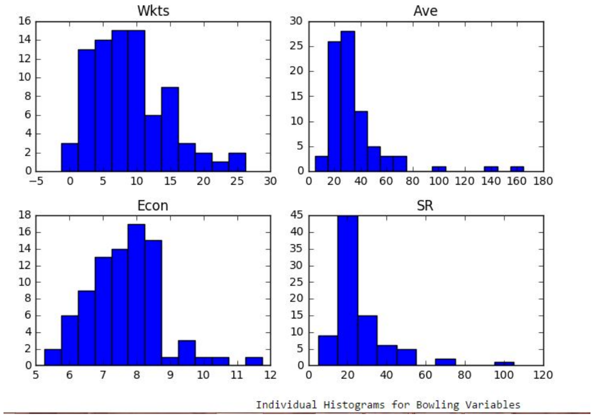


Fig. 10. Histogram of the parameter data for Bowlers

B. Correlation of the parametric data as plots

Lets plot the parameter vs parameter data to find whether the data is correlated or not. The first row of the plot is Wkts vs all the parameters and so on for better understanding as in Fig.11. The attached code is in [3]

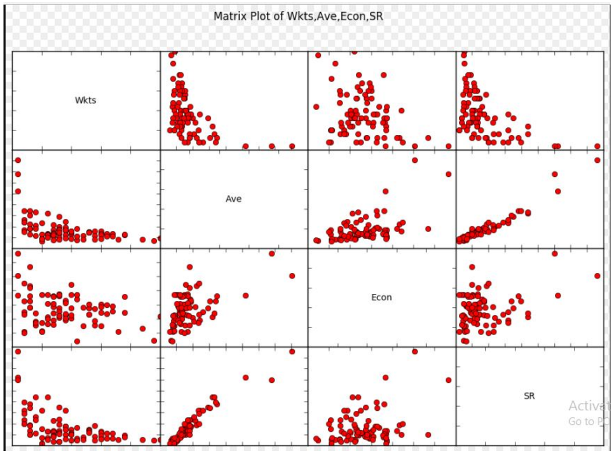
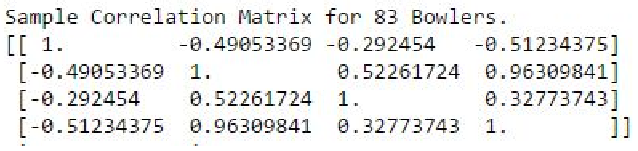


Fig. 11. The correlation of the parameters of bowlers

C. Correlation matrix

We normalized the data matrix of size 83x4 using Standardscaler and then we found the correlation matrix of size 4x4 for the same.The matrix is given below in the Fig.12.The code is in [3].

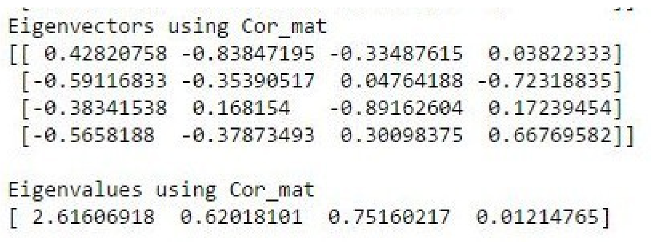


```
Sample Correlation Matrix for 83 Bowlers.
[[ 1.          -0.49053369 -0.292454   -0.51234375]
 [-0.49053369  1.           0.52261724  0.96309841]
 [-0.292454    0.52261724  1.           0.32773743]
 [-0.51234375  0.96309841  0.32773743  1.        ]]
```

Fig. 12. Correlation matrix of the parameters

D. Eigenvalues and Eigenvectors

So as to reduce the dimensionality of the above matrix we either do Eigen value decomposition or Singular value decomposition to the correlation matrix which results in only few significant eigen values and its corresponding eigen vectors. The eigen values and vector pairs are given below in the Fig.13. The code is in [3]



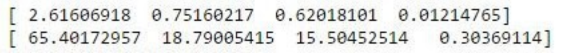
```
Eigenvectors using Cor_mat
[[ 0.42820758 -0.83847195 -0.33487615  0.03822333]
 [-0.59116833 -0.35390517  0.04764188 -0.72318835]
 [-0.38341538  0.168154   -0.89162604  0.17239454]
 [-0.5658188  -0.37873493  0.30098375  0.66769582]]

Eigenvalues using Cor_mat
[ 2.61606918  0.62018101  0.75160217  0.01214765]
```

Fig. 13. Eigen values and Eigen vector pairs

E. Variability of Eigenvalues

As mentioned earlier we don't need the whole data for finding the ranks as there is a lot of correlation in the data correlation matrix.Hence we can find the variability of the eigen values by the given formula $Var_i = \frac{\lambda_i}{\sum \lambda_i}$ The different variabilities are in the Fig.14. The code is in [3]



```
[ 2.61606918  0.75160217  0.62018101  0.01214765]
[ 65.40172957  18.79005415  15.50452514   0.30369114]
```

Fig. 14. Percentage Variability of eigen values

F. Cumulative graph and Scree plot

We can observe that most of the data can be obtained using the first eigen value alone.Lets observe how significant they actually are using cumulative graph and scree plot as given below in Fig.15,16.The code is in [3].

G. Ranking Criteria

Hence we considered only the first eigen value (2.61) and its eigen vector as it constitutes more that 65 percent of the data necessary and found the L1 value as given below.

$$L1 = -0.43*(Wkts) + 0.59*(Ave) + 0.38*(Econ) + 0.57*(SR)$$

NOTE:We have to use the normalized values in calculating the L1 values and not the original data set.

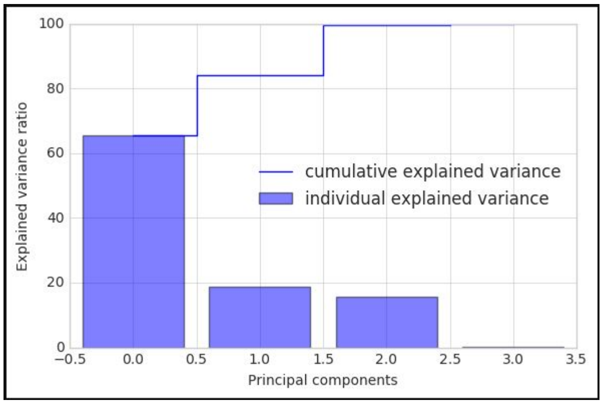


Fig. 15. Cumulative eigen value bar plot

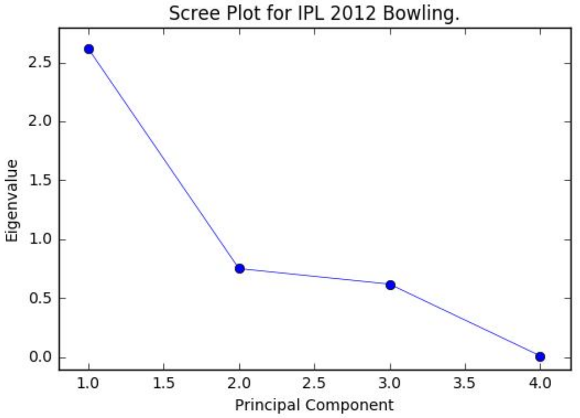
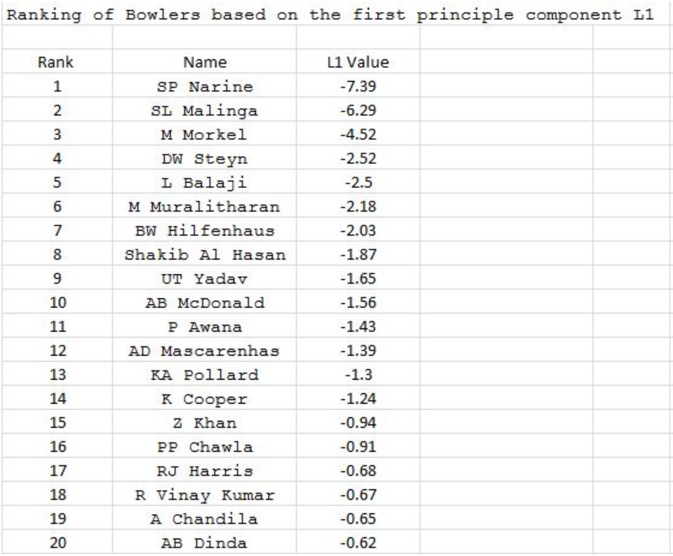


Fig. 16. Scree Plot for Bowler eigen values



Ranking of Bowlers based on the first principle component L1

Rank	Name	L1 Value
1	SP Narine	-7.39
2	SL Malinga	-6.29
3	M Morkel	-4.52
4	DW Steyn	-2.52
5	L Balaji	-2.5
6	M Muralitharan	-2.18
7	BW Hilfenhaus	-2.03
8	Shakib Al Hasan	-1.87
9	UT Yadav	-1.65
10	AB McDonald	-1.56
11	P Awana	-1.43
12	AD Mascarenhas	-1.39
13	KA Pollard	-1.3
14	K Cooper	-1.24
15	Z Khan	-0.94
16	PP Chawla	-0.91
17	RJ Harris	-0.68
18	R Vinay Kumar	-0.67
19	A Chandila	-0.65
20	AB Dinda	-0.62

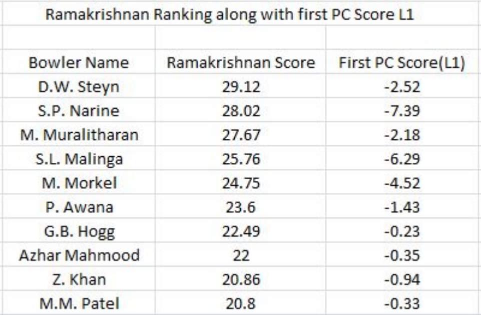
Fig. 17. Ranking of the top 20 Bowlers

H. Ranking Based on L1 values

We found the L1 values for all the 83 observations and have ranked them on basis of descending order which are given in the below Fig.17 The code is in [3]

I. Ranking comparison with Mr.Ramakrishnan

The comparison with the official ranking upto top 10 is given in the Fig.18.The details are given in the Observation section.



Ramakrishnan Ranking along with first PC Score L1

Bowler Name	Ramakrishnan Score	First PC Score(L1)
D.W. Steyn	29.12	-2.52
S.P. Narine	28.02	-7.39
M. Muralitharan	27.67	-2.18
S.L. Malinga	25.76	-6.29
M. Morkel	24.75	-4.52
P. Awana	23.6	-1.43
G.B. Hogg	22.49	-0.23
Azhar Mahmood	22	-0.35
Z. Khan	20.86	-0.94
M.M. Patel	20.8	-0.33

Fig. 18. Ranking of top 10 Batsmen comparison

V. Observations

We have recreated the ranking of the cricketers based on the L1 values we obtained from PCA which match with the ranking done by the authors in reference paper[1].Out of the top 10 ranking by Mr.Ramakrishnan out of which 7 of them match perfectly, with remaining having a slight variation.This may be due to the fact that the parameters we have considered are not the only major contributors but also parameters like number of overs in the case of bowlers and number of matches and innings in the case of batsmen also matter, which we have not considered in our PCA calculation. But still we can say that PCA can be pretty much efficient after all the dimension reduction we have done.

VI. Conclusions

Hence using 2012 Indian Premier League (IPL 2012) data,we have shown how to rank batsmen and bowlers based on their contributions to their teams and have demonstrated how PCA can be helpful and be applied to a coorelated-multivariate dataset.However,it might not be as efficient as that of Mr.Ramakrishna's official ranking but we could still match up to 70 percent of the top 10 players in either case i.e Batsman and Bowlers.

VII. References

[1]http://ww2.amstat.org/publications/jse/v21n3/scariano.pdf

[2]http://www.espncricinfo.com/indian-premier-league-2012/content/story/566523.html

[3]https://github.com/wasim-ee37/DSA-PH4130