Clustering Analysis on NBA draftees

Introduction

The question that we attempt to answer is how to predict NBA draftees' rankings in an NBA draft using their statitics when playing in college basketball teams. Along the way, we explore how these draftees "group" together based on their college performance.

Data

We scraped NBA draftees data from online websites (sources see reference). We choose five seasons between 2010 and 2015, and we are able to scrape 273 NBA draftees' college statistics. Having omitting the missing data, we are left with 244 observations. We use draftee's average college performance measurements as variables. We've only included players who played professionally in college basketball teams. We've also tried scraping earlier seasons such as years 2000 to 2008, however, many players' college data were not readily available online. We settle with 244 data observations of college basketball players who were drafted 2010 to 2015.

Below is a list of variables in the NBA draftee dataset:

- `G`: number of games played
- `MP`: minutes played per game
- `FG`: number of field goals per game
- `FGA`: number of field goal attempts per game
- `FG%`: field goal percentage FG/FGA
- `2P`: number of 2-point field goals per game
- `2PA`: number of 2-point field goal attempts per game
- `2P%`: 2-point field goals percentage 2P/2PA
- `3P`: number of 3-point field goals per game
- `3PA`: number of 3-point field goal attempts per game
- `3P%`: 3-point field goal percentage 3P/3PA
- `FT`: number of free throws per game
- `FTA`: number of free throws attempts per game
- `FT%`: free throw percentage FT/FTA
- `TRB`: total number of rebounds per game
- `AST`: number of assists per game
- `STL`: number of steals per game
- `BLK`: number of blocks per game
- `TOV`: number of turnover per game
- `PF`: number of personal fouls per game
- `PTS`: number of point scored per game
- `#`: player ranking in his draft class
- `H`: player height (cm)
- `P`: player position
- `draft_class`: the draft class to which a draftee belongs
- `PER`: a player's effective rating in the season right before he was drafted
- `label`: labels generated after discretizing PER by 4 quantiles

In Figure 1, we included two summary plots of our variables: a set of boxplots of continuous variables, and a histogram of player height distribution. We did not include `G` in the boxplot because it is large in scale, and since our data is per game, we exclude this variable from our dataset and analysis. Variables such as `MP`

and PTS have high values. Variables that indicates attempts, such as 3PA (3-point attempts) and FGA (field goal attempts) have higher average values than their respective goals made, 3P (3-point made) and FA (field goals made); this makes sense as no one, not even the best NBA players, can make 100%. Player heights are normally distributed with average around 200 cm, which is the normal height among basketball players.

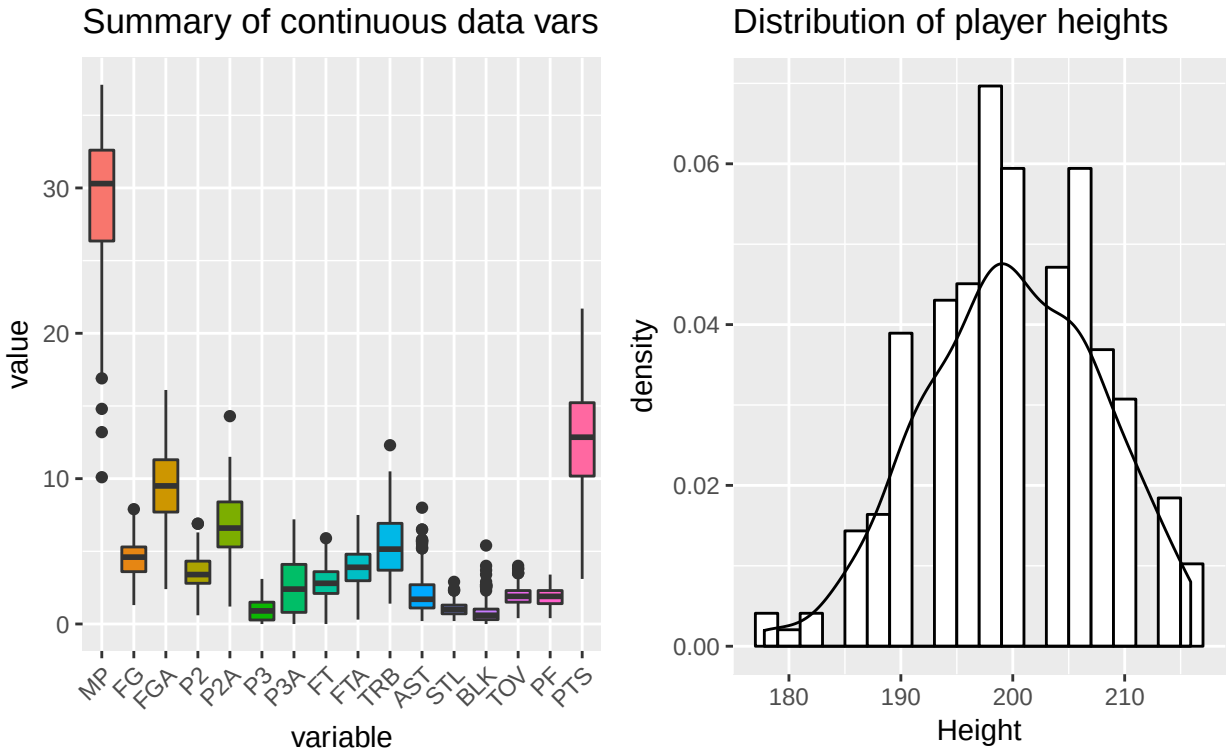

Figure 1: Summary plots of data variables

Methods

Since we have measurements of players per game performance, we first try to discover interesting things about these measurements. We perform PCA and clustering analysis to see if there are subgroups among the observations. We first use PCA to visualize the high-dimensional data, and then we apply clustering analysis based on the first two principal components we found. We found that PC1 roughly corresponds to a measure of players' average overall skills and PC2 is the component that discriminates players of different physical sizes.

Combining what we've learned in PCA, we look at clustering analysis, first using k-means and then Gaussian mixture models. Cluster 1 consists of omnipotent players. Cluster 2 consists of bad performance players who have low efficientcy in offense. Cluster 3 consists of functional players who specialize in blocking and rebounds. Cluster 4 consists of players who are most efficient in scoring, and this groups has a lot of super-stars playing in NBA.

Then, we attempt to classify players to a specific ranking group by discretizing players' PER by quantiles, where PER is a metric of a player's per minute performance and contributions in a particular season. For classification, we predict the quantile to which a player belongs. The best model is logistic regression with two-group label: top 50% and lower 50%. However, our overall classification performance is poor, it might be just that thre a myriad of factors that complicate the analysis which makes classification difficult, but this could be an indication for us to further modify our model and analysis.

Data Analysis

PCA

For PCA analysis, we drop variables `FG%`, `2P%`, `3P%` and `FT%` because they contain redundant information as we have already included in our dataset goal attempts and goals made. We also dropped variable `G`, games played, because our data is already per game, so it is not meaningful to include number of games played.

First we look at PCA scree plot (plot on the left in Figure 2), we can see that there is a sharp change at the value around 4; with four PCs, we can explain much of the variance in our data, approximately 84%. However, since four PCs are difficult to visualize, we plot the first two PCs and provide our interpretation of these two principal components. PC1 explains about 44.2% of the total variation and PC2 explains about 24.2% of the total variance, together they explain about 68.4% of the total variation, which is only a fair amount. We see from the biplot that the variable vectors have similar length which is a result of standardization.

We can see from the biplot (plot on the right in Figure 2) some groupings of variable vectors. `3PA` and `3P` variable loading vectors almost overlap, indicating a high correlation. This is may seem obvious because naturally the more you attempt throwing the ball at three-point line, the more you would score 3-points, regardless of your 3-point shoot skill. For example, an untrained person attempts throwing at 3-point line 10 times, and made only 1 goal, if he attempts 10 more times, he may score a few more like 3. However, when a professionally trained basketball player attempts throwing at 3-point line 10 times, he may be unlucky or haven't warmed up at first, so say he made only 2 goals out of 10 attempts. Since the player is trained, he may warm up in the next 10 attempts and score a lot more in the next round of attempts. This explains why these two variable vectors are so close to each other in our dataset of professional college basketball players. We see a similar grouping of `2PA` and `2P`; the more a player attempts at 2-point line, the more he scores 2-points.

We also see that `STL` (steal) and `AST` (assist) are also highly correlated with each other. Steal and assist measurements are the two main indicators of a player's agility and reflex. A player who is flexible is likely to have high measurements on steals and assists. Another grouping of variables consists of `FTA` (free throw attempts), `FT` (free throw made), `FGA` (field goal attempt), `FG` (field goal made), `TOV` (turnover), `PTS` (points scored), `MP` (minutes played). These variable vectors are close to one and other, indicating high correlations among them. For these college professional basketball players, the longer they stay on the court, the more points they score, the more occurrence of turnover, the more they are fouled by opposite team players, the more they get free throw attempts and free throw scores.

Variables vectors for `TRB`, `BLK` and `PF` are grouped together indicating high correlations among them. Tall and strong players are good at grabbing rebounds and because of their strong body built they are also good at blocks. These players move mostly in the area near the basket, and likely to have body contacts with other players, and consequently they are more likely to have personal fouls.

We then look at principal component loadings. We start by interpreting the second principal component. Looking at loadings of PC2, we discover that some variables have negative values and some variables have positive values, indicating a comparison. We examine this comparison by inspecting PC2's loadings, where `3P`, `3PA`, `AST`, and `STL` are negative while the rest are positive or near zero. Variables with negative PC2 values measure a players' agility and reflex. Short players who are usually points guards, shooting guards or small forwards are flexible and thus normally have high measurements in these variables. Variables with positive PC2 values can be interpreted as to represent those tall and big players who are usually centers, power forwards or sometimes small forwards. Since these players have strong physical built, they get rebounds more easily. They are also the players who like to drive to the pant and try to score with body contacts, and so they tend to be fouled a lot by opposite team players, which means that they get more chance to do free throws as a result of fouls.

Now we turn to interpreting the first principal component. The first PC puts similar weights to most of the variables such as `FTA`, `FT`, `FGA`, `FG`, `TOV`, `PTS` , `MP`. We interpret PC1 as a measure of the average overall skill of a player. PC1 puts less weights to variables such as `3PA`, `3P`, `TRB`, `BLK`, so specialized skills such as those

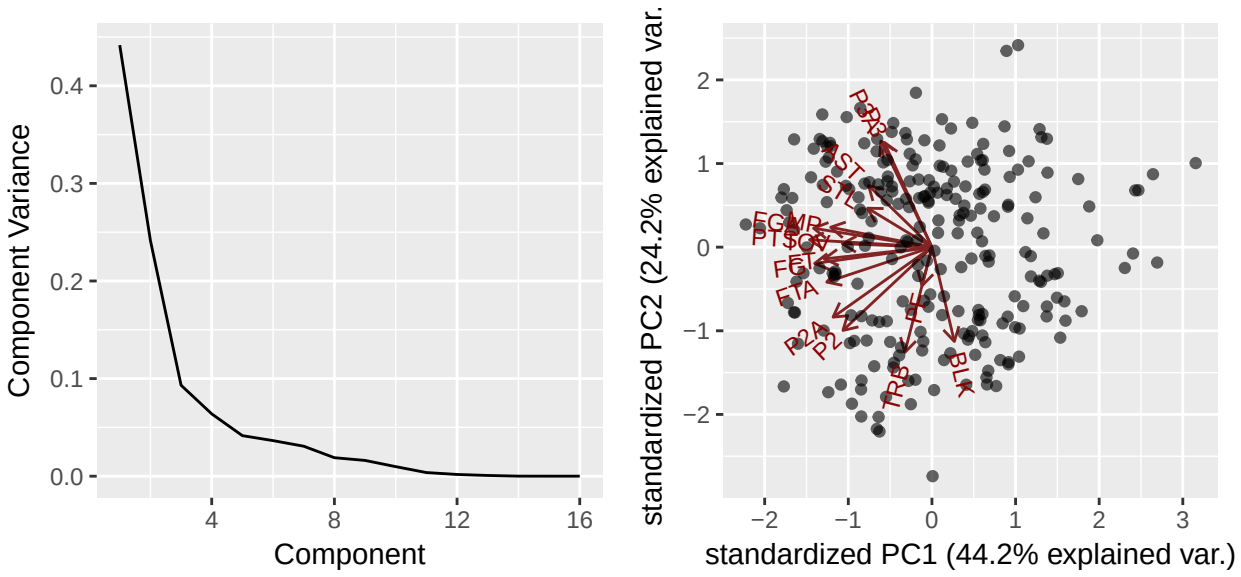

Figure 2: PCA Scree Plot and Biplot

possessed by short and flexible players or by tall and strong players are less weighted in PC1. In sum, we say that PC1 is roughly a measure of average overall skills and PC2 is the component that discriminates "small guys" and "big guys" on the court.

Clustering Analysis

K-means++

Instead of using the usual k-means clustering algorithm, which has the problem of being arbitrarily bad with respect to the objective function compared to the optimal clustering, we used the k-means++ algorithm for choosing the initial values. K-means partition the dataset into K non-overlapping clusters and to perform k-means clustering, we first need to specify the optimal number of clusters. In choosing the optimal number of clusters K, we relied on our knowledge of basketball as well as statistical methods such as silhouette and variance explained.

In basketball there are 5 positions: point guard, shooting guard, small forward, power forward, and center. Since there are 5 positions, we may expect to see 5 clusters where players of different positions falls into each distinct position groups neatly. However, note that players of different positions may have similar and overlapping responsibilities when playing on court. For example, a good small forward player can also be points guard and shooting guard. Notable players such as LeBron James could player positions ranging from point guard, shooting guard, small forward or even powerful forward. In emergency situations, he can even play centers.

First we look at the elbow plot (plot on the left in Figure 3), which plots percentage of variance explained as a function of the number of clusters. From the plot, we see the elbow occurs around 4 or 5 clusters, where the amount of change in percentage of variance explained changes the most. This implies that players' positions may be playing a significant role when partitioning the data into clusters.

Next we examine the silhouette scores of different clusters (plot on the right in Figure 3). The silhouette score computes the average distance between a player and all other players in the cluster that this player belongs, and this distance is scaled by the distance between this player and all players in the next nearest cluster. Silhouette scores range between -1 and 1; the higher the score, the better the clusters are separated. We see from the silhouette scores plot that $K = 2$ has the highest silhouette score, followed by $K = 3$ and

Table: PC loadings

	PC1	PC2
MP	-0.30	0.08
FG	-0.34	-0.06
FGA	-0.35	0.08
P2	-0.26	-0.33
P2A	-0.29	-0.28
P3	-0.14	0.42
P3A	-0.15	0.42
FT	-0.32	-0.05
FTA	-0.31	-0.14
TRB	-0.08	-0.42
AST	-0.19	0.24
STL	-0.19	0.15
BLK	0.07	-0.37
TOV	-0.26	0.01
PF	-0.03	-0.15
PTS	-0.36	0.03

$K = 4$. Silhouette scores for $K \geq 5$ drop significantly. Since elbow plot and silhouette scores do not give consistent results, we further explores the cluster behavior by partitioning PCA projections into clusters and see which K provides interpretability.

As we have first expected that the players should fall into subgroups based on 5 positions. As we look at $K = 5$ model (plot on the right in Figure 4), we focus on examining the two clusters on the lower left. These two clusters maybe a problem of overfitting, as if we look at the one cluster on the lower right from the $K = 4$ model (plot on the left in Figure 4), players in the cluster are mostly power guards and shooting guards. If we follow the $K = 5$ model and further partition this lower left region into two more groups, we are unable to interpret the extra clusters, which indicates redundancy.

We look specifically at the silhouette plot for 4-cluster k-means++ model (Figure 5). We see that each of the four clusters has above average silhouette scores. The size of clusters did not fluctuate extremely; the first three plots are similar in size, while the fourth cluster is larger and it also has the highest silhouette scores. In all, we say that based on silhouette analysis, $K = 4$ is an appropriate model.

Combining all the methods for finding the optimal K, we find $K = 4$ is the most intrepretable model with a decent silhouette score. We further our clustering analysis with $K = 4$.

K-means++ clustering results discussion

We visualize cluster analysis results using stack bar charts (Figure 6), which display players' statistics in each category. The longer the bar, the higher the player's ability in that category. From these stack bars, we interpret the $K = 4$ model.

We find that most positions of players in cluster 1 are PF (power forward) or SF (small forward) or both (PF/SF). Players in cluster 1 are the most muti-functional players. Since they are well-around, they have high average skills, and thus they do not outperform in a specific category. These players grab rebounds, assist their teammates, and score 2-points shots by themselves. This cluster is well expected, because small forwards or SF/PF are the omnipotent positions on the basketball court.

Cluster 2 is less homogeneous than cluster 1. It contains players of many positions such as point guard, shooting guard, small forwards, power forwards or some combinations of these positions. Players in the cluster 2 tend to use 3-point shoots as their main scoring weapon, as opposed to cluster 1 and cluster 3, but their `3P%` (percentage of 3-point goals made) is much lower than cluster 4, which is a cluster that contain players who rely heavily on 3-point shoots. These players like to finish scoring on their own, especially through 3-point

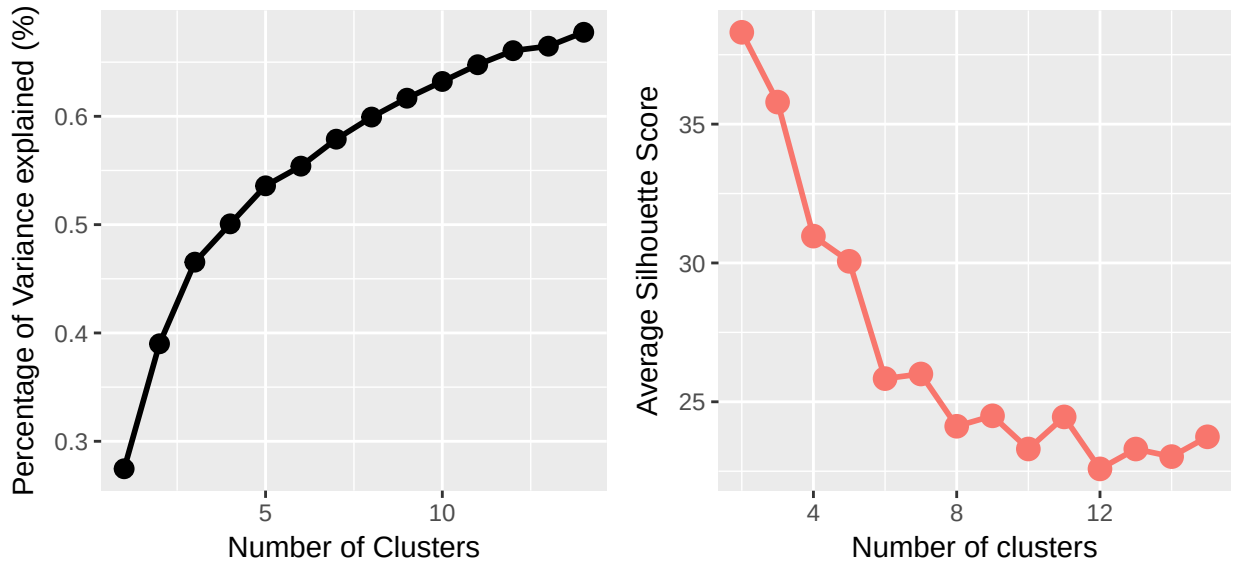

Figure 3: Elbow plot and silhouette scores plot

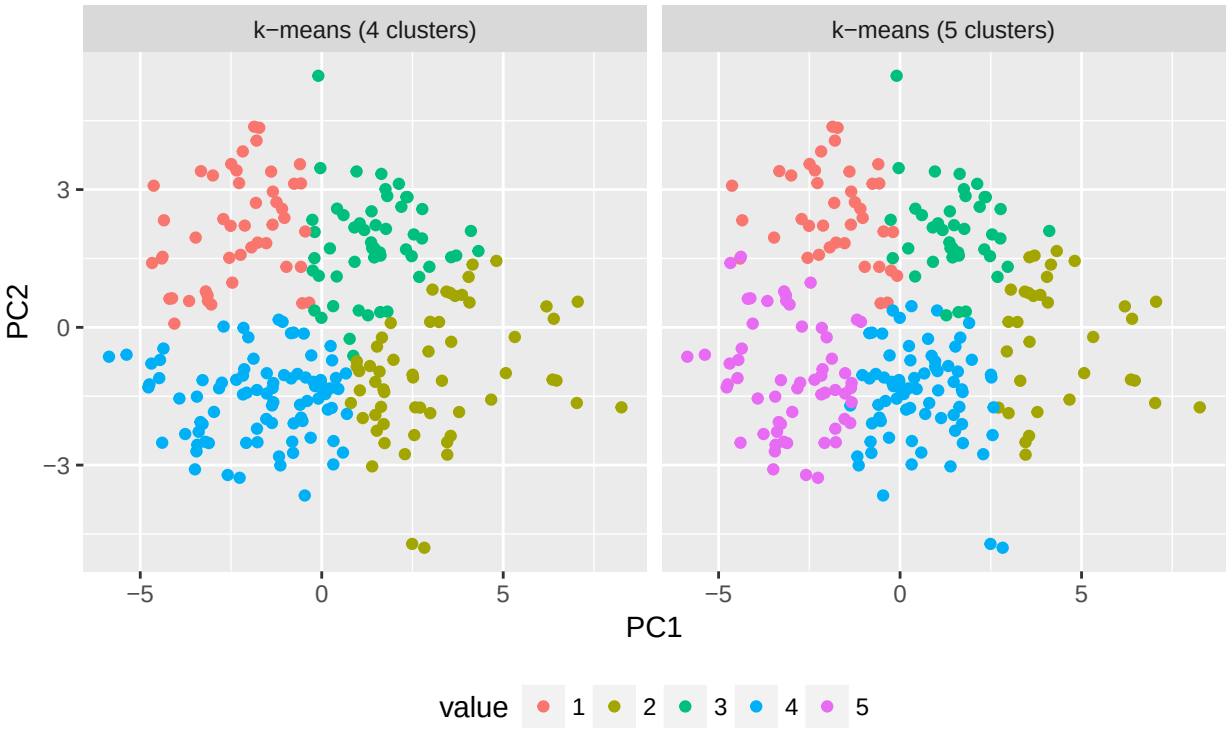

Figure 4: K-means++ clustering with K=4 and K=5

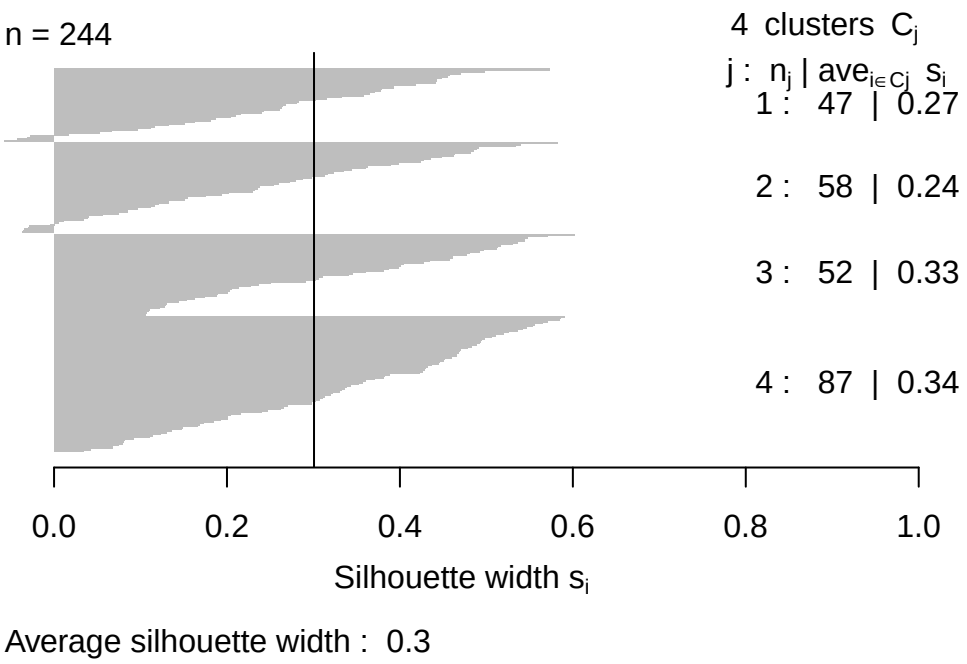

Figure 5: Silhouette plot of 4 clusters

shoots, but their 3-point shoot ability is not optimistic since they miss a lot of the shoots. This is indicated in cluster 2 stack bar chart, through their long `3PA` (3-point attempt) but short `3P` (3-point made).

Players in cluster 3 has the strongest rebound ability and their main scoring strategy is 2-points shoots. Most of these players are power forwards or centers and hey usually have strong physical built. These players have the highest average height as they are mostly centers. As we have mentioned in PCA, these players move mostly in the area near the basket, and likely to have body contacts with other players because they are good at blocks, which may lead to their tendency of having more personal fouls.

In cluster 4, players' positions are either power guards or small guards or both (PG/SG). We notice that some of the NBA super-stars are in this cluster, and we believe that this is a phenomenon corresponding college basketball game tactics. NCAA tend to use tactics that focus on 3-point shoots because for college players, the efficiency of scoring using 3-points is higher than that using 2-points. In a college basketball game, the 3-point line arc is relatively closer to that of the basket, so the college teams focus on scoring 3-points. NCAA playing strategy corresponds to the current NBA's game strategy trend. Recently in NBA games, the main trend of tactics focus on "Small Ball"; this is a tactic that would sacrifice height, physical strength and low post offense/defense in favor of a lineup of smaller players for speed, agility and increased scoring (often from the three-point line). Hence, the importance of "big guys" in a team is dramatically decreased, this means that most of stars in NBA would be "small guy".

We observe from the stacked bar chart that players in cluster 4 have the highest efficiency in scoring as shown by their average field goal percentage, which are on top among players compared to other 3 clusters. Players in cluster 4 also have the lowest average height, which is expected because players in cluster 4 consist mostly of PGs and SGs. These positions for the "small guys" on the basketball court.

Lastly, we comment that cluster 2 is too heterogeneous after looking at data statistics of specif players belonging to that cluster. For cluster 3, its PC1 values are positive (that is their overall performance is not good), but since these players are characterized by their large physical built (centers and power forwards), they specialize in grabbing rebounds and blocking others' shots. These players are supporting functional players in a team. Players in cluster 1 and 4 are the most efficient in scoring and some other abilities; and PC2 discriminate them into big size group and small size group, which coincides with their positions PG/SG

and SF.

Lastly, we examine a few notable players using radar charts (Figure 7). Radar chart is a chart that consists of a sequence of equi-angular spokes, called radii, where each spoke represents one of the variables. The data length of a spoke is proportional to the magnitude of the variable for the data point relative to the maximum magnitude of the variable across all data points. A line is drawn connecting the data values for each spoke. This gives the plot a star-like appearance and the origin of one of the popular names for this plot.

Take the player Kyrie Irving for example, we can see that his free throws made is the best among all the players. And his Points Per Game and 2 points Field Goals are also very outstanding among peers. His steals reaches about 50% in these players. His assists is also above 50% in those players. His rebounds and blocks are pretty bad from the radar chart. Base on interpretation provided above, we can say that Kyrie Irving would be a shorter player on the court, and would be a highly efficient player, and these characteristics correspond to those of players belonging to cluster 4.

Gaussian Mixture Model

In addition to K-means++, we further our clustering analysis employing Gaussian mixture model. In Gaussian mixture model, we are concerning both the covariance type of model and number of clusters. To obtain the best optimal of a Gaussian model as well as the optimal number of clusters, we took advantage of information-theoretic criteria (BIC) to conduct model selection. Figure 8 shows the BIC corresponding to various types of mixture models and the number of clusters. To compare with K-means++, we consider comparing $K = 4$ models. Among the 4-cluster models, the one with ellipsoidal, equal volume and orientation, and orientation (EVE) results in the largest BIC, so we choose EVE model with 4 clusters for analysis. The silhouette scores of EVE is 0.20, smaller than 0.3 in K-means++, which means the clustering performance of EVE is not as good as K-means++. The clustering results of EVE are also visualized in the first two PC projections in Figure 9. After examining players in each cluster, it can be found that the clusters produced by Gaussian mixture models are not intrepretable, as the clusters are too heterogeneous. Take Tim Hardaway Jr, Eric Bledsoe, Bradley Beal and Kyrie Irving as examples. They are all talented and famous players with similar positions but are split into different clusters. However, the Gaussian mixture model might be more flexible for fitting the data and worth for further exploration.

Classification

Here we attempt to classify players to a specific ranking group by discretizing players' PER by quantiles. `PER` is an indicator of a player's per minute performance in one particular season. The higher a player's PER, the better a player performances in a given season. In our dataset, we used the draftee's PER in the last season before he was drafted. During a player's career, his skills were immature and developing in earlier periods. A coach may not put a freshman player on the court for very long, and so even for players who have a lot of potentials, their earlier stats maybe not be outstanding. When players have accumulated skills and ready for NBA draft, their skills and mental strength would be the most mature, and so we believe that a player's college performance during his last season in college was the most pertinent to being drafted by NBA. Although PER is not a perfect metric of a players' performance, we include PER in our classification analysis because in the end, PER is the one number that sums up a player's statistical performance.

We attempt to classify players to a specific ranking group by discretizing players' PER by quantiles. We employ various classification techniques to forecast the quantile that each player belong in his draft class: top 25%, 25-50%, 50%-75%, 75%-100%. We use methods such as LDA, QDA, multinomial logistic regression, SVM, K-nn and decision tree and we use CV errors to evaluate these models. however, the performance among all the models are not satisfying. We also try to predict the player into upper 50% and lower 50%. The best predict CV error is 0.22 from logistic regression with only two labels. With only two labels, it is hard to discover interesting things about the data. It simply classify players to good and bad players, which is not informative. Since our error rate indicate a poor performance and that the dataset is not originally labeled, and that we have created these labels based on an imperfect metric `PER`, we say that classification

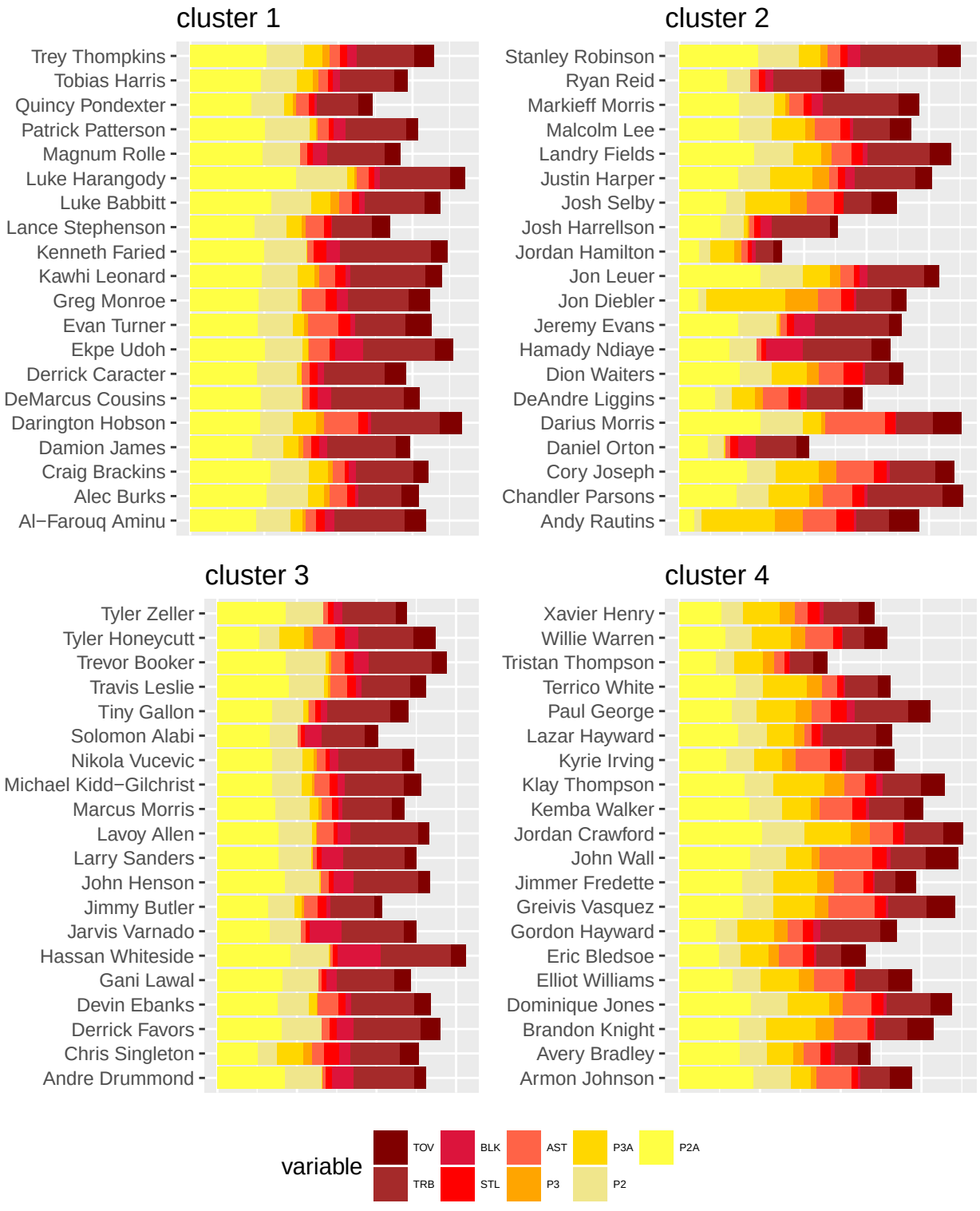

Figure 6: Stack Bar Charts by Clusters

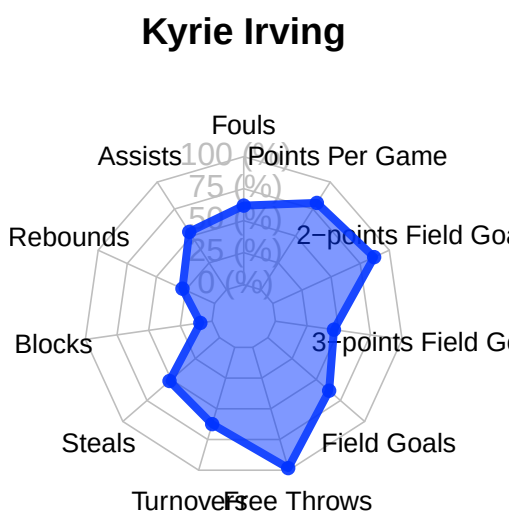

Figure 7: Radar Charts for notable player Kyrie Irving

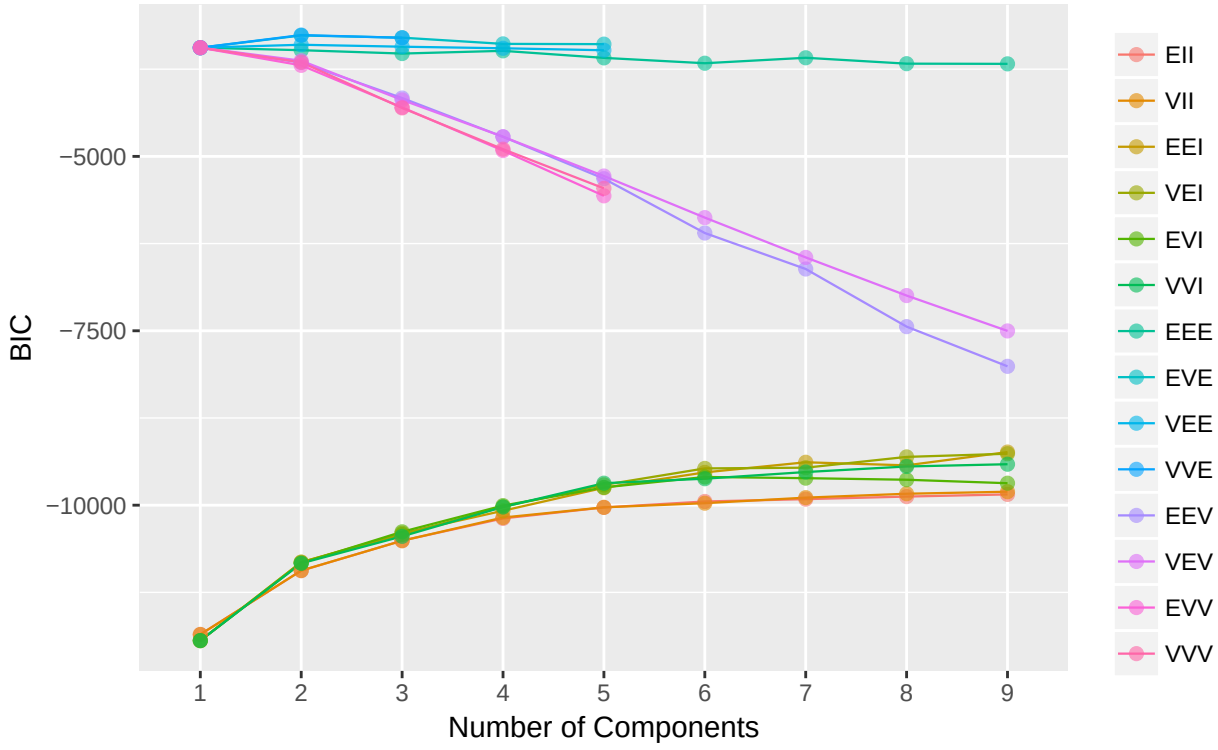

Figure 8: Gaussian mixture model selection by BIC

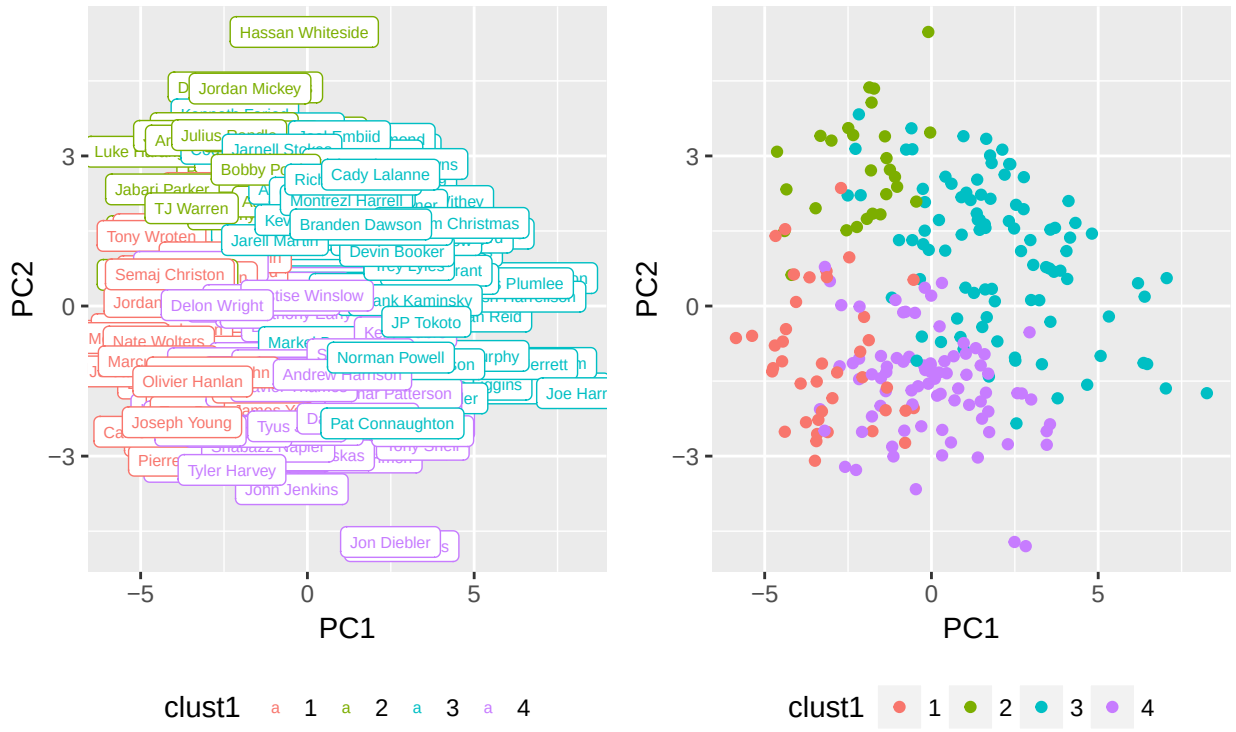

Figure 9: Guassian mixture model cluster plots

may not be a good method for predicting the players' rankings in a draft and that we may need to select a different metric when creating labels for players in the future.

Table: CV Errors

	Methods	4 classes	2 classes
1	logistic	0.48	0.22
2	lda	0.43	0.25
3	qda	0.55	0.31
4	knn	0.54	0.38
5	svm_linear	0.47	0.24
6	svm_radial	0.64	0.29
7	tree	0.58	0.33

Conclusion

We use unsupervised learning algorithms to partition data into subgroups, and we find that the 4-cluster model is most interpretable. We also try to classify players, but the results are poor, especially when we to create 4 labels based on PER metric. The reason is likely to be our data don't have the labels in the first place, but then we use our own method to add labels for each player. Since 4-label model is performs poorly, we try the 2-label model as an alternative method for classification. Because of the time limit, we do not put much emphasis on interpreting the Gaussian mixture model clustering. In the future, we will expland our analysis to mixture model algorithm and modify our model with classification methods that help us classify more accurately and efficiently.

Reference

Groothuis, P. Early Entry in the NBA Draft: The Influence of Unraveling, Human Capital, and Option Value . Retrieved May, 2005 .

Edwards, R. Using Pre-NBA Draft Data to Project Success in the NBA. Retrieved 2015.

Yang, Y. S. Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics. Retrieved May, 2015.

Torres, R. A. Prediction of NBA games based on Machine Learning Methods. Retrieved December, 2013.

LUTZ, D. A CLUSTER ANALYSIS OF NBA PLAYERS. Retrieved February, 2012.

Fischer, Using machine learning to predict the long-term value of NBA draftees http://tothemean.com/2014/06/17/machine-learning-predict-long-term-value-in-draft.html

McCool, Projecting NBA Draft Picks https://tartansportsanalytics.com/2017/03/09/projecting-nba-draft-picks/

Data Sources

College Player Individual Stats: http://www.sports-reference.com/cbb/players/damian-lillard-1.html College Team's statistics: http://www.sports-reference.com/cbb/seasons/2014-school-stats.html NBA Draft Results http://www.basketball-reference.com/draft/NBA_2015.html