**Assignment: multivariate analysis of a dataset in R**

I0P16a: Applied Multivariate Statistical Analysis

William Steijn

r0643267

2017 – 2018

**Problem Description**

Comparisons between players and classification of players is an important part of baseball analysis, and has implications in areas such as line-up construction and roster building. Player comparisons are usually based off outcomes, such as batting average or homeruns, which are more influenced by luck, and focus on careers rather than single seasons. In this report, I will investigate the ability to compare and cluster players using process-based rate statistics, which will hopefully yield more informative insights about individual players.

**Dataset Description**

The dataset used in this report is batted ball and plate discipline data for all players from the 2002-2014 seasons who had at least 250 plate appearances, downloaded in csv format from fangraphs.com. This gives 4020 observations which are uniquely defined by player name and season, and 19 variables. The 19 variables are as followed:

O.swing = the percentage of pitches a batter swings at outside the strike zone.

Z.swing = percentage of pitches a batter swings at inside the zone

Swing = percentage of total pitches a batter swings at

O.contact = pitches outside the zone on which contact was made /swings on pitches outside the zone

Z.contact = pitches outside the zone on which contact was made/swings on pitches inside the zone

contact = pitches on which contact was made/total swings

zone = pitches in the strike zone / total pitches

F.strike = first pitch strikes/ plate appearances

SwStr = swings and misses / total pitches

BB = walk percentage

K = strikeout percentage

ISO = isolated power = slugging percentage – batting average

BABIP = batting average on balls in play

LD = percentage of contact that results in a line drive

GB = percentage of contact that results in a ground ball

FB = percentage of contact that results in a fly ball

IFFB = percentage of contact that results in an infield fly ball

HR.FB = percentage of flyballs that are a homerun

IFH = infield hit percentage

The dataset does not contain any missing values. These variables are all percentages, so the scale is the same for all variables. However, the variance for each variable is different; for example, the Pull variable has a much greater variance than the infield hit percentage variable.

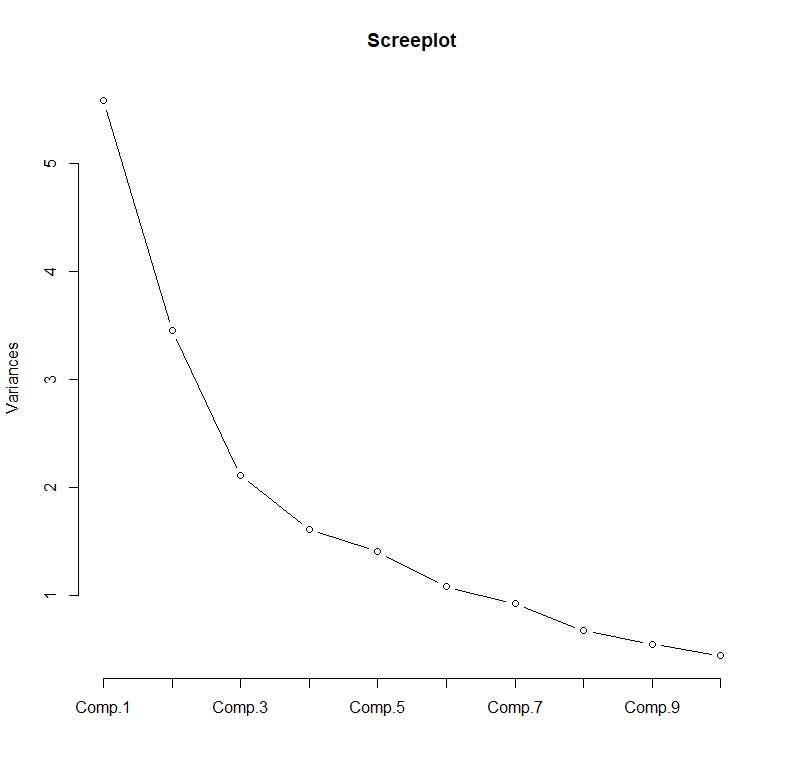
**Methodology**

The first technique used was principal component analysis to reduce the dimensionality of the dataset. The PCA was done on the correlation matrix to account for differing variable variance, to give each variable the same weight in the PC loadings, and a scree plot was generated to identify the optimal number of principal components. The principal components were then used to create different player archetypes, as each linear combination that is generated by PCA ideally explains a different type of skill that a player possesses. The PC scores for each player were converted into z-scores for each of the principal components. If a player has an especially high or low z-score for a specific principal component, that means that this player differs significantly from average along this principal component axis, and the player is either a strong positive or negative example of the player archetype that goes along with that principal component. Next, Mahalanobis distances were used to find the closest observations to a given player in the matrix of z-scores, and these observations represent the best player-comparisons for that given player.

Following PCA, hierarchal cluster analysis using complete linkage was performed to find outliers, which are players that would not have good comparisons. After removal of these outliers, K-means clustering was used on the PCA scores to see if the players can be clustered into groups similar to the player archetypes found previously. Finally, tree based modelling was used to find the significant variables that can be used to discriminate between players in different clusters.

**Results and Discussion**

I began the exploratory analysis by looking at the correlation between the 19 explanatory variables. The correlation matrix did not show evidence of high correlation between variables, which is supported by the low value of the proportion of variance explained by the first principal component (Table 1). However, the scree plot in Figure 1 shows seven principal components is sufficient for explaining the variance in the data. The first seven principal components explain over 85 percent of all player variation, and is a considerable dimension reduction from the original 19 variables (Table 1).



**Figure 1: Scree plot for the variance explained by the principal components**

**Table 1: Variance explained by first 7 principal components**

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7

Standard deviation 2.3615526 1.8584156 1.4526162 1.27008317 1.18659018 1.04102960 0.96357037

Proportion of Variance 0.2935227 0.1817741 0.1110576 0.08490059 0.07410507 0.05703909 0.04886673

Cumulative Proportion 0.2935227 0.4752968 0.5863544 0.67125495 0.74536002 0.80239911 0.85126584

Next, the loadings for the principal components were analysed for different types of hitters, and the variables that primarily explain each principle component are highlighted in bold (Table 2). To get a high z-score for component 1, a player would need to have low contact percentages and a high swinging strike percentage, K percentage, ISO, and HR/FB rate. These qualities propose that the players with high PC 1 z-scores are “power hitters”. Component 2 is defined by low swing rates and a high walk rate, which suggests a “patient hitter” type. The third player type is one that hits weak fly balls, as high scorers for component 3 have a low BABIP and high fly ball rates. The remaining player types can be seen in the last row of Table 2, and are explained by the component loadings in bold.

**Table 2: PC loadings for 19 variables with player types**



After computing the z-score matrix, the best 10 comparisons were found to an individual player during a specific season were found using Mahalanobis distances. For example, Table 3 shows the ten closest player comparisons to Ryan Howard’s 2008 season, along with the z-scores for each of the players’ seven components. Ryan Howard’s 2008 season has the highest absolute value z-score for component one at 2.7, meaning his projection onto the first component is 2.7 standard deviations above average, and therefore classifying him as a “power hitter”. As expected, all the ten closest comparisons have z-scores well above one for component one.

**Table 3: Top 10 comparisons for Ryan Howard 2008**

Name Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7

3518 Ryan Howard 2008 2.757526 0.6625547 -1.1922528 -0.001840556 1.8259057 0.12048987 0.1455996

822 Chris Davis 2014 2.800412 0.3839956 -0.9955996 -1.253185023 1.5563136 -0.32015151 -0.5758235

480 Brad Hawpe 2005 1.318335 0.9397336 -1.2534860 0.380369185 1.3418048 0.37150379 0.1933718

1893 Jeff Bagwell 2004 1.238428 1.1076899 -0.6871700 0.259629946 1.1526346 0.04906213 -0.1471460

1851 Jay Bruce 2013 1.773059 -0.3399181 -0.8812843 -0.460181960 1.0309655 -0.54382957 0.2878699

481 Brad Hawpe 2007 1.950240 0.5002984 -0.7499574 0.916638300 0.7098438 0.14583653 0.8143607

1989 Jim Edmonds 2004 2.360919 1.6346706 -0.3765275 0.423240756 1.0894044 1.05843085 -0.1063262

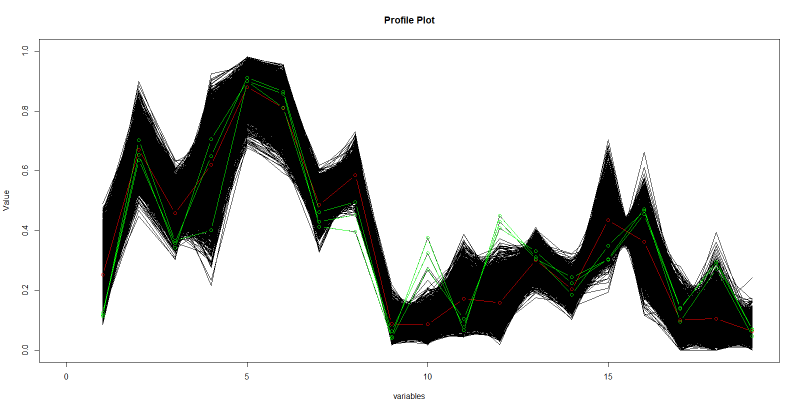
823 Chris Davis 2009 2.866220 -0.9928755 -1.5897948 -0.270606104 1.3941413 0.68258778 0.4295971

661 Carlos Delgado 2006 1.322750 0.8386666 -0.9660329 0.323105976 0.9054723 0.07118291 -0.5448139

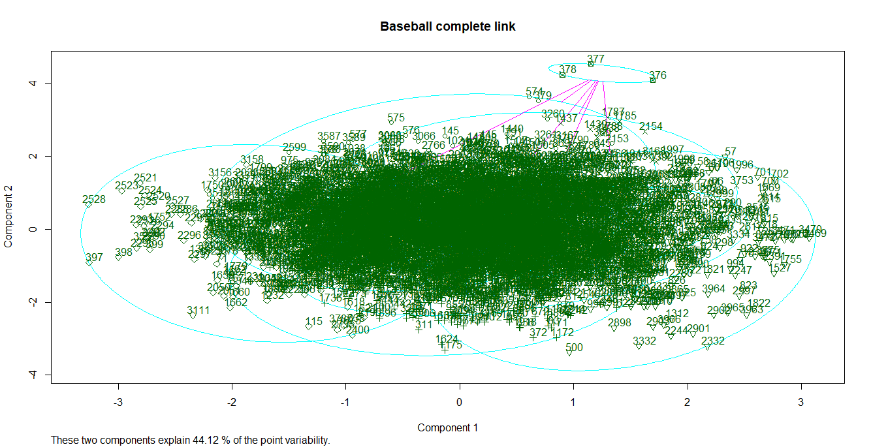
658 Carlos Delgado 2002 1.492273 1.7352278 -1.1702196 0.771599820 1.5623645 -0.29025335 0.1640053

2056 Joey Votto 2009 1.057024 0.8462754 -1.6482982 -0.412781476 1.5202067 0.35960402 -0.4023875

The profile plot for all the observations (Figure 2) does not show any obvious grouping structure, but since the profiles are all very similar, some potential outliers can be detected. There appears to be an observation with a significantly lower F.Strike percentage than the rest of the observations, as well as a couple values for O.contact, BB percentage, ISO, and GB percentage that seem to be significantly different than the rest of the observations. Complete linkage cluster analysis of seven clusters (Figure 3), chosen to match the seven principal components, one cluster with only three observations is discovered. These three observations, the 2002, 2004, and 2007 seasons for Barry Bonds, are characterized by extremely high (above 4) z-scores for component 2, and extremely low (below -3.5) z-scores for component 7. This suggests a player that is very patient, does not strike out, and has a lot of power, which correlates with the high ISO, BB rate, and HR/FB rate and low F.strike rate revealed in the profile plot.



**Figure 2: Profile plot for all the observations with the variables on the x-axis and their corresponding values on the y-axis. The red line is the average for each variable. The green lines are the profiles of the outliers found in complete linkage clusters analysis.**



**Figure 3: Complete linkage cluster analysis on the baseball data.**

Subsequently, K-means clustering (K = 7) was performed on the PCA scores, and the value for K was chosen by looking at the within group sum of squares for different numbers of clusters. The clusters were plotted in different colors on their first two principal components (Figure 4), and the average z-score for each of the seven principal components was calculated for each of the clusters (Table 4) to try to identify which player archetype is represented by each cluster. Out of the seven principal components, only the first three contained z-scores above 1 or below -1, which means that these clusters represent combinations of the first three player archetypes (e.g. cluster 5 has z-scores above 1 for comp. 1 and 2, which means these players are patient power hitters). Ideally, the cluster analysis would have created 7 clusters, each with a significant z-scores for a different principal component, but the proportion of variance for components 4-7 was likely too low for the formation of clusters based off these components. Cluster 2 only has a z-score with an absolute value above 1 for component 3, which might explain why there appears to be one cluster on top of another cluster in Figure 4, which only considers the first two components. To check the results of the clustering, I looked at the clustering of the 10 closest comparisons for Ryan Howard 2008 from Table 3. Eight of the ten players on this list were classified in cluster 5 alongside Ryan Howard 2008.

Table 4: Z-scores for each of the 7 components for the mean of cluster

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7

cluster1\_mean 0.07969585 0.8288817 -0.27341114 -0.376739217 0.15402075 0.009500024 -0.10518628

cluster2\_mean 0.16501717 -0.1705718 1.07120594 0.538202190 0.05337756 -0.142195647 0.04654950

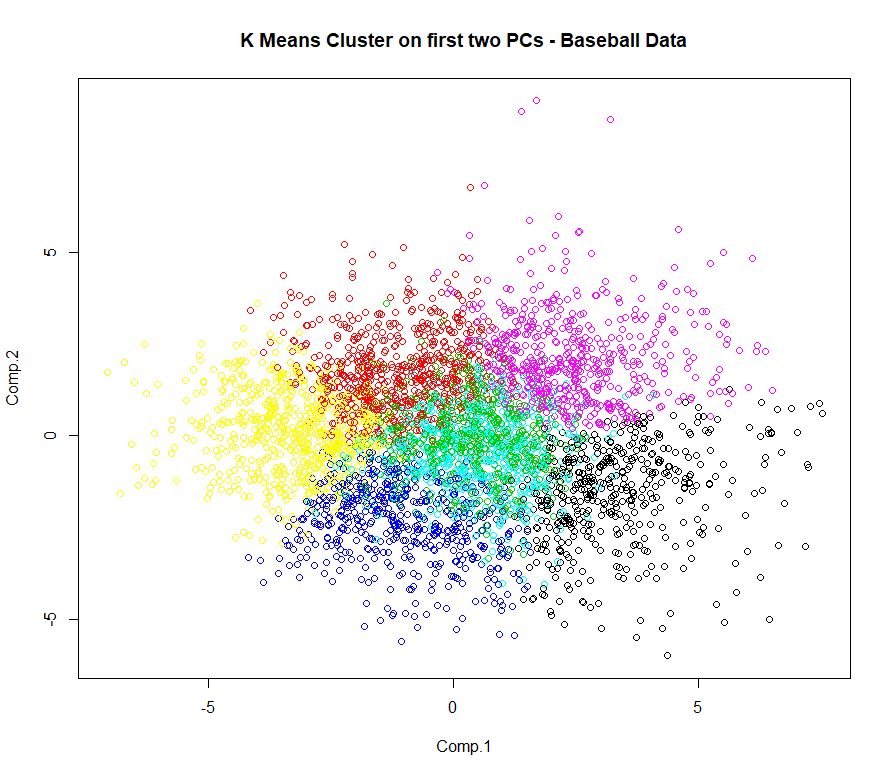
cluster3\_mean -0.31172172 -1.2161662 0.10058969 -0.309110195 0.50892813 0.075631818 -0.22969331

cluster4\_mean -1.13882999 0.5108808 0.34811924 -0.008541644 0.10152373 -0.085186994 0.25432707

cluster5\_mean 1.31531224 1.0041313 -0.09921669 0.195462423 -0.11979340 0.144280560 -0.06181338

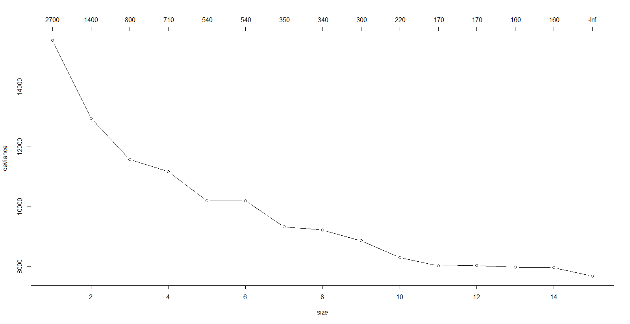
cluster6\_mean 1.14544971 -0.8914769 -0.37291899 -0.110330753 -0.20117294 -0.063016619 0.09686465

cluster7\_mean -1.13435053 -0.4030971 -0.97195121 0.192970025 -0.63718244 0.134828201 -0.05256907

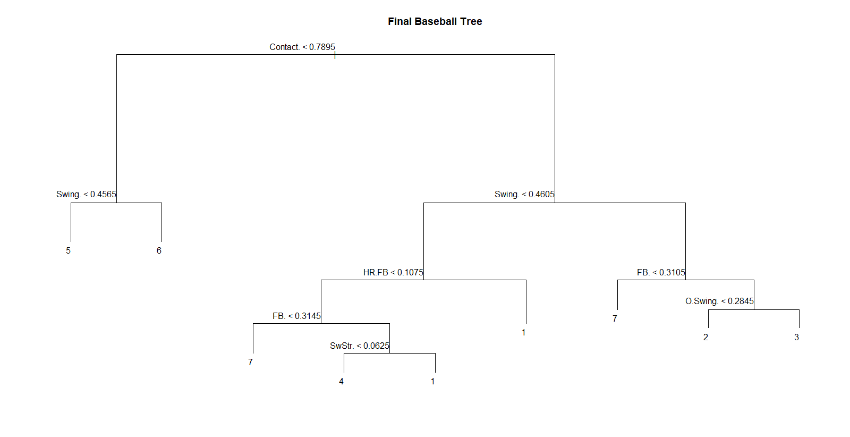


**Figure 4: K-means clusters of the PCA scores plotted on the first two principal components, with each cluster plotted using a different color.**

For each cluster to be discriminated from the others, at least seven end nodes were needed. After analysing the deviance plot (Figure 5), 10 end nodes were initially chosen. However, nine end nodes were discovered to give the same number of misclassification errors, so nine end nodes were chosen for the final tree (Figure 6). The misclassification error rate was quite high at nearly 37 percent, but this is expected due to the imperfect establishment of the clusters as well as a lack of any apparent grouping structure in the data. The first, fifth, and seventh end nodes showed particularly low discriminating power, suggesting clusters 1, 5, and 7 are hardest to classify. Interestingly, these three clusters are also the most ambiguous when the z-scores of the cluster means are considered, as these are not definitively characterized by a z-score above one or below negative one in any single cluster.



**Figure 5: Total deviance for the tree with the corresponding number of end nodes**



**Figure 6**: **Final tree model for the baseball data**

baseball=read.table("~/applied multivariate/assignment/baseball.final.csv",header=T, sep = ',')

names = read.csv("~/applied multivariate/assignment/PlayerNames.csv", header = T)

summary(baseball)

baseball2 = baseball[,1:19]

baseball2

pairs(baseball2)

cor(baseball2)

baseball.pca = princomp(baseball2, cor = T)

summary(baseball.pca)

plot(baseball.pca,type="l",main='Screeplot')

names(baseball.pca)

attributes(baseball.pca)

baseball.scores = (baseball.pca$scores[,1:7])

baseball.matrix = as.matrix(baseball2)

a = scale(baseball.scores, center = TRUE, scale = TRUE)

baseball.zscores = cbind(names, a)

baseball.zscoresmatrix = as.matrix(a)

#mahalanobis distances - uses pca scores

mahalanobis.comps = function(player = " "){

a = which(names == player)

d = baseball.zscoresmatrix[a,]

y = mahalanobis(baseball.zscoresmatrix, center = d, cov = var(baseball.zscoresmatrix))

e = order(y)

f = baseball.zscores[c(e[1:11]),]

return(f)

}

mahalanobis.comps(player = "Ryan Howard 2008")

#cluster analysis

dim(baseball.matrix)

# Set up blank plot for profiles

plot(c(0,19),c(0,1),type="n",xlab="variables",ylab="Value",main="Profile Plot")

# Use a loop to generate a profile line for each observation.

for (k in (1:4020))

{

points(1:19,baseball.matrix[k,],type="l")

}

points(1:19, means, type ="b", col="red" )

points(1:19, baseball.matrix[376,], type = "b", col = "green")

points(1:19, baseball.matrix[377,], type = 'b',col = "green")

points(1:19, baseball.matrix[378,], type = 'b', col = "green")

# The hclust computes the cluster object.

baseball.clus <- hclust(dist(baseball2), method="complete")

# specifed (k) number of clusters.

baseball.kclusters <- cutree(baseball.clus,k=7)

table(baseball.kclusters)

which(baseball.kclusters==7, arr.ind = TRUE)

baseball.zscores[c(376,377,378),]

baseball[c(376,377,378),]

clusplot(baseball,baseball.kclusters,stand=TRUE,labels=3,main="Baseball complete link")

# K means cluster method.

# Determine number of clusters

wss <- (nrow(baseball)-1)\*sum(apply(baseball,2,var))

for (i in 2:20) wss[i] <- sum(kmeans(baseball,centers=i)$withinss)

plot(1:20, wss, type="b", xlab="Number of Clusters",

ylab="Within groups sum of squares")

baseball.pkcl <- kmeans(baseball.pca$scores,7,20)

plot(baseball.pca$scores[,1:7], col = baseball.pkcl$cluster)

title ("K Means Cluster on first two PCs - Baseball Data")

clusplot(baseball,baseball.pkcl$cluster,stand=TRUE,labels=2,main="Baseball-kmeans, 7 clusters")

#Determine means of z-scores for each component for each cluster

cluster.vector = baseball.kcl$cluster

clus\_1 =which(baseball.kcl$cluster==1, arr.ind = TRUE)

clus\_1\_z.scores = baseball.zscores[c(clus\_1),]

cluster1\_mean = apply(clus\_1\_z.scores[,c(2:8)], 2, mean)

cluster\_z\_scores = rbind(cluster1\_mean,cluster2\_mean,cluster3\_mean, cluster4\_mean, cluster5\_mean, cluster6\_mean, cluster7\_mean)

baseball.with.clusters = cbind(names,baseball2,cluster.vector)

#Decision tree on the clusters

library(tree)

observation\_cluster = as.factor(cluster.vector)

baseball.tree = tree(observation\_cluster ~ O.Swing. + Z.Swing. + Swing. + O.Contact. + Z.Contact. + Contact. + Zone. + F.Strike. + SwStr. + BB. + K. + ISO + BABIP + LD. + GB. + FB. + IFFB. + HR.FB + IFH., baseball.with.clusters)

summary(baseball.tree)

baseball.tree

plot(baseball.tree)

text(baseball.tree,srt=0,adj=1)

title("Baseball Tree")

cv.baseball.tree <- cv.tree(baseball.tree, ,prune.tree)

plot(cv.baseball.tree,type="b")

pruned.baseball.tree <- prune.tree(baseball.tree,best=10)

plot(pruned.baseball.tree)

text(pruned.baseball.tree, srt = 0, adj = 1)

title("Final Baseball Tree")

summary(pruned.baseball.tree)