Understanding Principal Component Analysis

Leo Sun

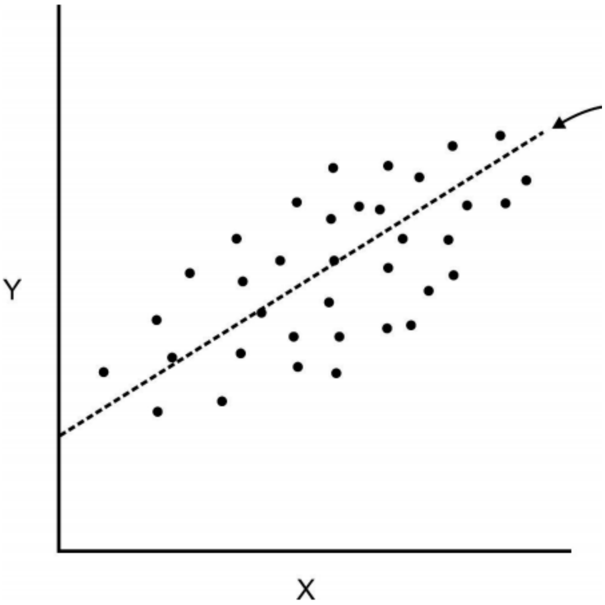
October 31, 2017

Introduction: Teapot

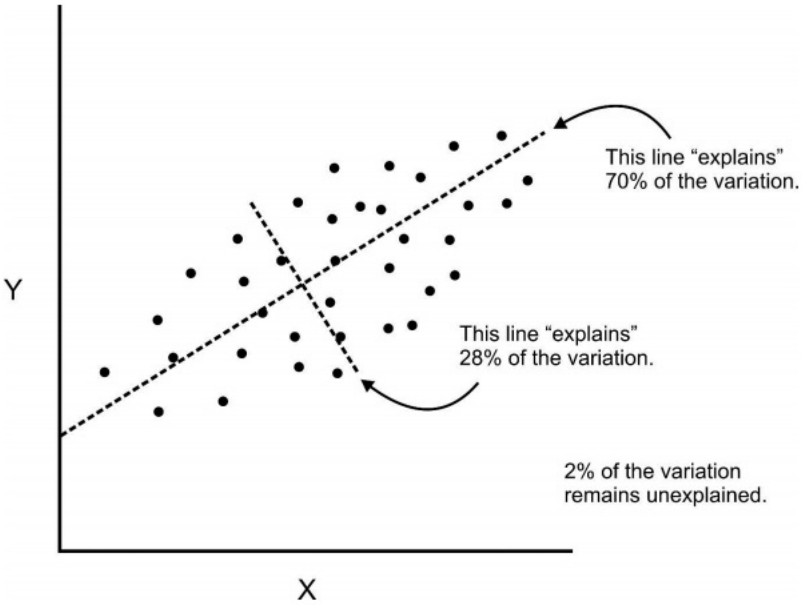
Picture a teapot in an empty 3-D space. Now, think take a mental 2-D image of that teapot in different rotations. Think about which angle captures the most information about that teapot. Does an image of the front capture a lot of information? Or the bottom? Or the top? No, but a picture of the side of the teapot contains the most information. You can see the lid, handle, spout, and the body of the teapot all in one picture. Finding the best angle is essentially using Principal Component Analysis (PCA) to find the most variation of a 3-D teapot. In this post, I will delve into the mysterious waters of PCA and explain how it works and what the results mean. But as a side note, take this information with a grain of salt. I am not a mathematician, so my analysis will be based off of how I interpret my research

PCA

In essence, PCA is dimensionality reduction. For example, you can extract information from a `n` dimension matrix that can be represented as a 2D or 3D matrix. PCA uses correlation to find the which dimensions contain the most variance, and captures that into **principal components**. Each component has its correlation values based off of a specific feature. The first component captures the most amount of variation between the data.



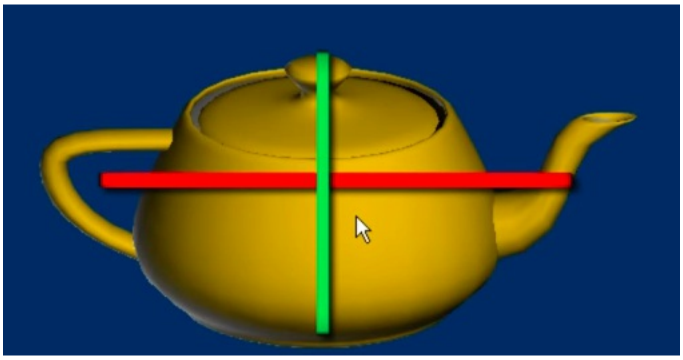
From the image above, you can see that the first component is a linear regression (best fit) line of the graph. The line captures the most variation because the difference between the most bottom left and top right points are the greatest.



The second component will capture the second most variation of the graph, seen above. The third component will capture the third most variation, and so on.

So you might be wondering, how many components do I need really need to capture all the necessary information? While there are differing thoughts on how many components are needed, the most common is to get enough components such that it captures at least 85% of the variation. From the image above, you can see that the first two components capture 98% of the variation, which is more than enough.

Now, picture the teapot again and try to imagine it with the principal components. It should look something like this:



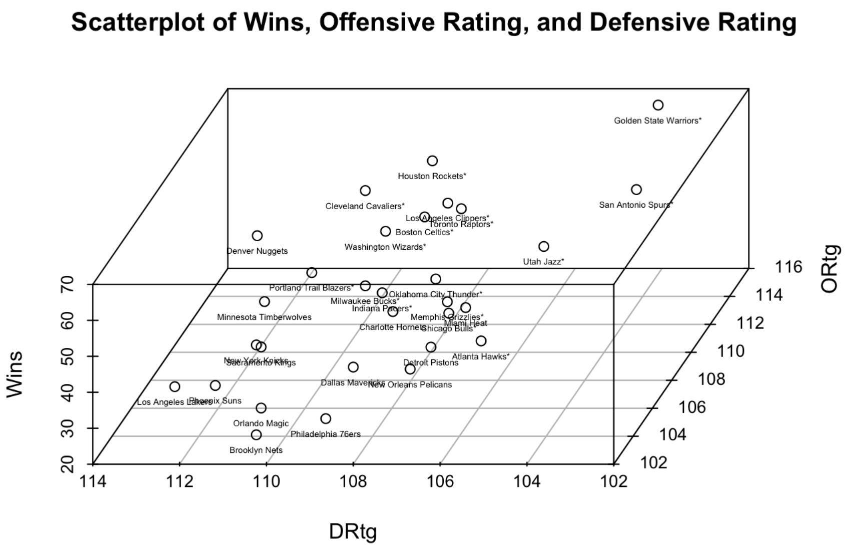
PCA Example

PCA can be confusing for the very first time, so here is a more concrete example.

In basketball (NBA), there are a couple of advance statistics that measures a team’s performance. We measure the correlation between offensive rating and defensive rating, which is an estimation of points produced per 100 possessions and points allowed per 100 possessions, respectively.

```
library(scatterplot3d)
library(dplyr)
library(ggplot2)
# First import that data
dat = read.csv('stats.csv')
# Select the features we want
dat = select(dat, Team, W, ORtg, DRtg)

# 3D graph of the features
s3d = scatterplot3d(dat$ORtg, dat$DRtg, dat$W, main = 'Scatterplot of Wins, Offensive Rating, and Defensive Rating', xlab = 'ORtg', ylab = 'DRtg', zlab = 'Wins', angle = 250)
text(s3d$xyz.convert(dat$ORtg, dat$DRtg, dat$W), labels=dat$Team, pos=1, cex = .40, )
```



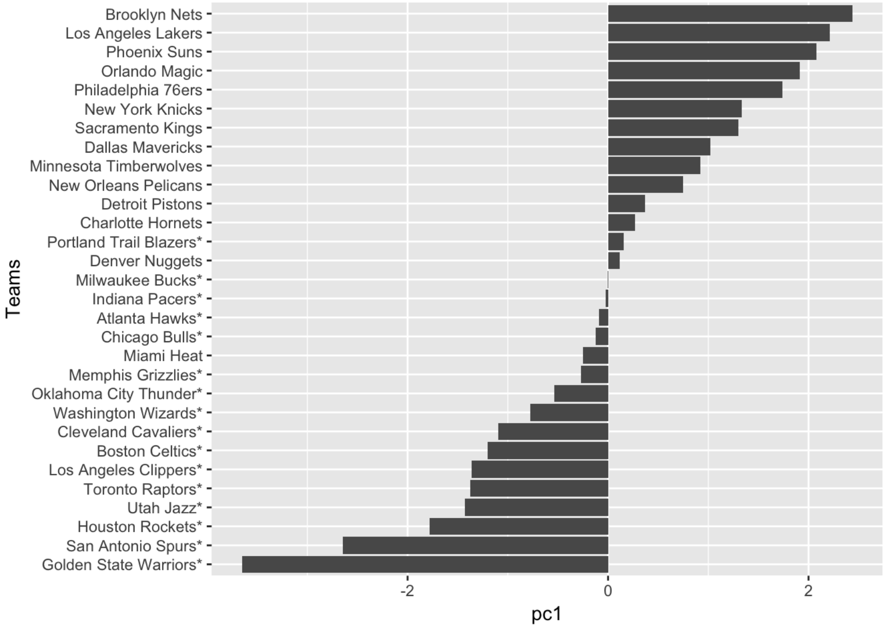
At a glance, we see that the most variation is from the bottom right to the top left.

```
# Performing PCA on each of the features
pca = prcomp(dat[-1], scale. = T)

pc1 = pca$x[, 1]
pc2 = pca$x[, 2]
pc3 = pca$x[, 3]

pcs = data.frame(pc1, pc2, pc3, teams=dat$Team)

# Graphing each team's contribution to the first principal component
ggplot(pcs, aes(y= pc1, x = reorder(teams, pc1))) + geom_bar(stat = 'identity') + coord_flip() + xlab(aes(x = 'Teams'))
```

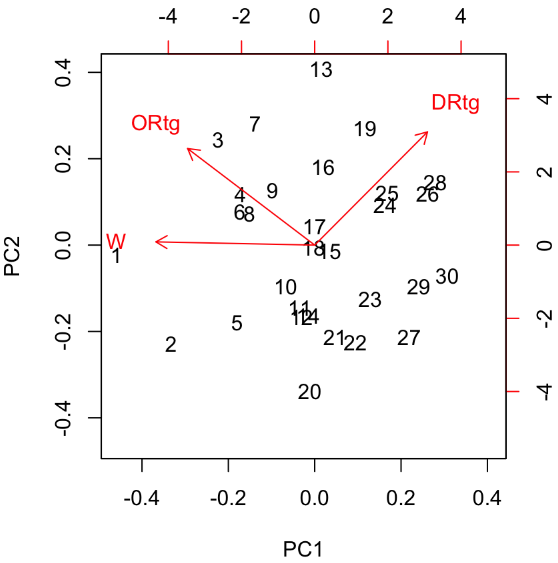


Plotting the first component reveals that the most variation is the axis created by the difference from the bottom right to the top left.

```
pca$rotation
```

```
##              PC1        PC2        PC3
## W    -0.6827376 0.02164314  0.7303430
## ORtg -0.5463541 0.64856638 -0.5299612
## DRtg  0.4851460 0.76085033  0.4309758
```

```
biplot(pca)
```



Here, we can see how each the correlations of features and the principal components. Wins has a very low correlation with PC2, and a strong and negative correlation with PC1. ORtg has a positive and moderate correlation for PC2 and a moderate and negative correlation with PC1. DRtg has a moderate/strong and positive correlation with PC1 and PC2.

Since wins and offensive rating is closer to each other in space, they are more closely related. If we examine our data, we can see why.

```
print(dat$Team[dat$ORtg < dat$DRtg])
```

```
##  [1] Portland Trail Blazers* Milwaukee Bucks*
##  [3] Indiana Pacers*         Minnesota Timberwolves
##  [5] Atlanta Hawks*          Detroit Pistons
##  [7] New Orleans Pelicans    Dallas Mavericks
##  [9] Sacramento Kings        New York Knicks
## [11] Phoenix Suns            Philadelphia 76ers
## [13] Los Angeles Lakers      Orlando Magic
## [15] Brooklyn Nets
## 30 Levels: Atlanta Hawks* Boston Celtics* ... Washington Wizards*
```

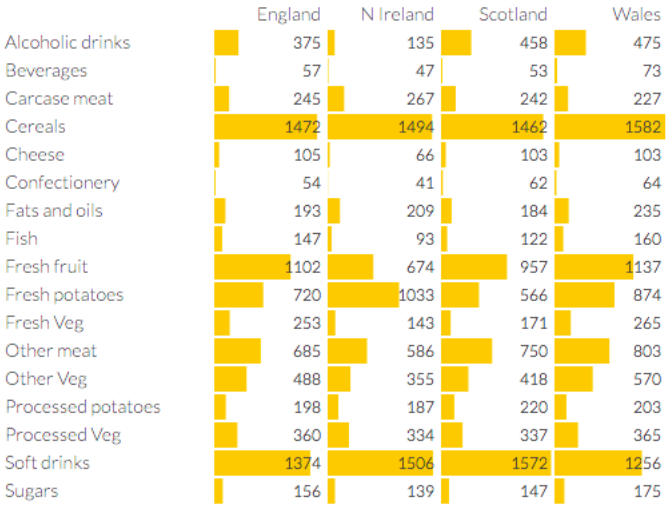
```
print(dat$W[dat$ORtg < dat$DRtg])
```

```
##  [1] 41 42 42 31 43 37 34 33 32 31 24 28 26 29 20
```

Teams with a greater defensive rating than their offensive ratings have the least amount of wins. This makes logical sense because if they have a high DRtg, teams score more on them than they do on the opposing team.

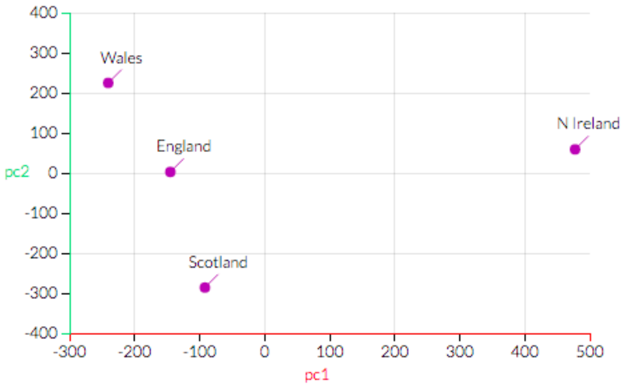
PCA with many Dimensions

In the previous example, we only explored three dimensions. Some data sets will contain a lot more, so it will be hard to visualize the information. This is where the effectiveness of PCA can truly be seen. Referenced below, here is a data set of foods consumed in the British Isles.



	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

Looking at the graph, there doesn’t seem to be a lot of differences. Each place looks like they eat the same types of food. But, if we use PCA, we can use the first and second components to graph the differences.



We can see that Wales, England, and Scotland are relatively close to each other. This implies that their food consumption is rather similar. The only major differences are in the second principal component, so there is some food variation there. But what really stands out is Northern Ireland. This makes it easier for us to find what is causing that variation. Looking at the data set again reveals that the Irish eat more potatoes than fresh fruit. This is what created the huge gap in PC1. It is also possible to find what causes variation in PC2 for the others. For example, the difference in PC2 for Scotland and Wales might be because of the food consumption differences in vegetables and fats.

By using PCA, we could spot the differences in a 17-D data set and isolate the cause of the variation.

Closing Thoughts

Principal component analysis is a strong tool that helps determine the correlation between many features. The first and second principal components are usually enough to explain the data set because it contains most of the variation. Essentially, PCA reduces the many dimensions of a data set into a couple of components that allows us to easily visualize the intricacies of a data set.

References

https://www.youtube.com/watch?v=BfTMmoDFXyE

http://webspace.ship.edu/pgmarr/Geo441/Lectures/Lec%2017%20-%20Principal%20Component%20Analysis.pdf

https://stats.stackexchange.com/questions/76906/how-can-i-interpret-what-i-get-out-of-pca

https://onlinecourses.science.psu.edu/stat505/node/54

https://en.wikipedia.org/wiki/Principal_component_analysis

http://setosa.io/ev/principal-component-analysis/

http://www.basketball-reference.com/

https://www.youtube.com/watch?v=_UVHneBUBW0&t=528s