PCA\_Pizza

Edward Guevel, Kayla Strunk, Teresa Vail

6/01/2022

## PCA Background

Principal Component Analysis (PCA) is a dimension reduction technique. Analysis can help identify patterns in a data set and distill the variables down to their most prominent features so that the data is simplified without losing much of the data’s variability. PCA allows us to find a low-dimensional representation of our data set that contains as much of the variation as possible. PCA reduces the overall dimensional scope of the data set by only highlighting the interesting (high variance) components.

## Data Requirements

-Data for the predictor variables must be numeric

-Data must have no outcome variable

-Must be some correlation (linear relationship) between predictor variables

-Three or more independent predictor variables is encouraged

## Pizza Dataset Information

We will be using “pizza.csv” dataset that contains nutrient analysis of pizzas. Credit goes to Dhilip Subramanian on data.world (<https://data.world/sdhilip/pizza-datasets>)

The variables in the data set are:

brand - Pizza brand (class label)

id - Sample analyzed

mois - Amount of water per 100 grams in the sample

prot - Amount of protein per 100 grams in the sample

fat - Amount of fat per 100 grams in the sample

ash- Amount of ash per 100 grams in the sample (think burnt crust)

sodium - Amount of sodium per 100 grams in the sample

carb - Amount of carbohydrates per 100 grams in the sample

cal - Amount of calories per 100 grams in the sample

## Data Loading and Manipulation

#Code to load in the full dataset:  
pizza.complete<-read.csv("pizza.csv")  
head(pizza.complete)

## brand id mois prot fat ash sodium carb cal  
## 1 A 14069 27.82 21.43 44.87 5.11 1.77 0.77 4.93  
## 2 A 14053 28.49 21.26 43.89 5.34 1.79 1.02 4.84  
## 3 A 14025 28.35 19.99 45.78 5.08 1.63 0.80 4.95  
## 4 A 14016 30.55 20.15 43.13 4.79 1.61 1.38 4.74  
## 5 A 14005 30.49 21.28 41.65 4.82 1.64 1.76 4.67  
## 6 A 14075 31.14 20.23 42.31 4.92 1.65 1.40 4.67

#Check the variable types in your data set:  
str(pizza.complete)

## 'data.frame': 300 obs. of 9 variables:  
## $ brand : chr "A" "A" "A" "A" ...  
## $ id : int 14069 14053 14025 14016 14005 14075 14082 14097 14117 14133 ...  
## $ mois : num 27.8 28.5 28.4 30.6 30.5 ...  
## $ prot : num 21.4 21.3 20 20.1 21.3 ...  
## $ fat : num 44.9 43.9 45.8 43.1 41.6 ...  
## $ ash : num 5.11 5.34 5.08 4.79 4.82 4.92 4.71 5.28 5.02 5.16 ...  
## $ sodium: num 1.77 1.79 1.63 1.61 1.64 1.65 1.58 1.75 1.71 1.66 ...  
## $ carb : num 0.77 1.02 0.8 1.38 1.76 1.4 1.77 2.95 1.18 0.64 ...  
## $ cal : num 4.93 4.84 4.95 4.74 4.67 4.67 4.63 4.72 4.93 4.95 ...

For PCA, we need our variables to be in numeric format.

When running the str() function above, we see that the brand name is “chr” for character. We do not care about the specific brands for running PCA in this example. We are looking for the overall trends of the various qualities of a pizza (moisture, protein, fat, etc.), so we can remove the brand names.

We can also remove the ID name, since it does not carry any useful information either.

#Code to remove brand name (row "brand") and ID (column "Id"):  
  
pizza <- subset(pizza.complete, select = -c(1,2))  
head(pizza)

## mois prot fat ash sodium carb cal  
## 1 27.82 21.43 44.87 5.11 1.77 0.77 4.93  
## 2 28.49 21.26 43.89 5.34 1.79 1.02 4.84  
## 3 28.35 19.99 45.78 5.08 1.63 0.80 4.95  
## 4 30.55 20.15 43.13 4.79 1.61 1.38 4.74  
## 5 30.49 21.28 41.65 4.82 1.64 1.76 4.67  
## 6 31.14 20.23 42.31 4.92 1.65 1.40 4.67

Compare the pizza and the pizza.complete data sets above to ensure the brands and IDs were removed.

## Running PCA

We are going to use the prcomp() from the tidyverse package. This function will take in the data set and by setting scale = TRUE we scale the data to have a mean = 0 and sd = 1. This allows us to standardize the data.

#set scale = TRUE so that all data are scaled to have a mean of 0 and sd=1 before calculating principle components  
pizza.pca <- prcomp(pizza, scale = TRUE)  
  
#Look at the different results given:  
names(pizza.pca)

## [1] "sdev" "rotation" "center" "scale" "x"

#Center and Scale components display the means and standard deviations of the variables that were used for scaling prior to implementing PCA.   
pizza.pca$center

## mois prot fat ash sodium carb cal   
## 40.903067 13.373567 20.229533 2.633233 0.669400 22.864767 3.271000

pizza.pca$scale

## mois prot fat ash sodium carb cal   
## 9.5529866 6.4343924 8.9756583 1.2697235 0.3703577 18.0297225 0.6200343

#PCA scores each principal component under column "x"  
# if you want the entire data set...  
# pizza.pca$x  
# for just the top section of the data set  
head(pizza.pca$x)

## PC1 PC2 PC3 PC4 PC5 PC6  
## 1 5.001985 2.674746 0.0392612138 -0.1438652 0.28354158 -0.002334931  
## 2 5.015375 2.525076 0.0968945341 -0.3533599 0.21490980 0.002946986  
## 3 4.797424 2.669240 0.0753491747 0.1076181 -0.03497911 0.005414628  
## 4 4.462088 2.281218 0.1200327815 0.0539380 0.17420945 0.005617206  
## 5 4.464433 2.155551 0.0007362777 -0.1169702 0.31250437 0.001685765  
## 6 4.497286 2.164357 0.1747249127 -0.1152763 0.20046580 0.005184239  
## PC7  
## 1 0.0009586381  
## 2 0.0010203355  
## 3 0.0008303696  
## 4 0.0009448638  
## 5 0.0009668578  
## 6 0.0009775808

#The rotation matrix provides the principal components. Each column is a principle component, and each row is the variable. The more a number is further away from 0, the more that variable contributes to the principle component.  
pizza.pca$rotation

## PC1 PC2 PC3 PC4 PC5 PC6  
## mois 0.06470937 -0.6282759 0.42166894 0.2207216 0.006470293 -0.4464499018  
## prot 0.37876090 -0.2697067 -0.74602744 0.0105932 0.387982788 0.0001715203  
## fat 0.44666592 0.2343791 0.19930871 0.5070422 -0.173367634 0.5254028685  
## ash 0.47188953 -0.1109904 -0.05627269 -0.5523985 -0.670885701 -0.0588609281  
## sodium 0.43570289 0.2016617 0.45516887 -0.4462769 0.602614079 -0.0031309852  
## carb -0.42491371 0.3203121 -0.05223651 -0.3343395 -0.007436899 0.0005088535  
## cal 0.24448730 0.5674576 -0.11331559 0.2792632 -0.078003175 -0.7219138527  
## PC7  
## mois -0.4185690354  
## prot -0.2767646428  
## fat -0.3776715255  
## ash -0.0560214003  
## sodium 0.0005243238  
## carb -0.7760679112  
## cal -0.0120598098

#Note that eigenvectors in R point to the negative direction by default. Therefore we must multiply our results by -1.  
pizza.pca$rotation <- -(pizza.pca$rotation)  
  
#We also need to reverse the scores  
pizza.pca$x <- -(pizza.pca$x)

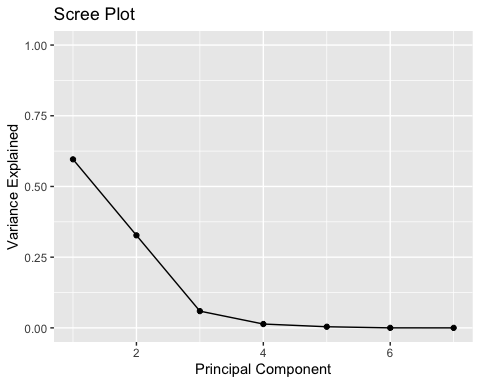
By choosing a threshold, we can determine what variables are significant in the contribution to the principal component. The threshold is arbitrary; it’s based more on the data set at hand and there is no rule of thumb.

## Calculate the Variance

var\_explained<- pizza.pca$sdev^2 / sum(pizza.pca$sdev^2)  
var\_explained

## [1] 5.959688e-01 3.272082e-01 5.922319e-02 1.359632e-02 3.953860e-03  
## [6] 4.822991e-05 1.359826e-06

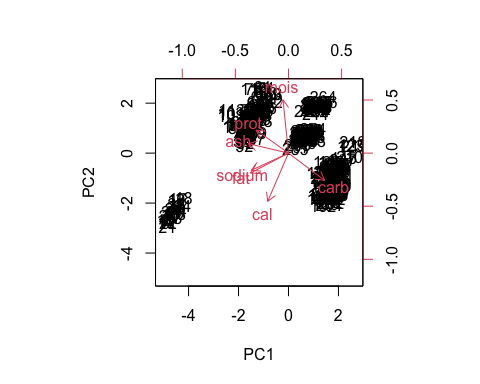
qplot(c(1:7), var\_explained) +   
 geom\_line() +   
 xlab("Principal Component") +   
 ylab("Variance Explained") +  
 ggtitle("Scree Plot") +  
 ylim(0, 1)



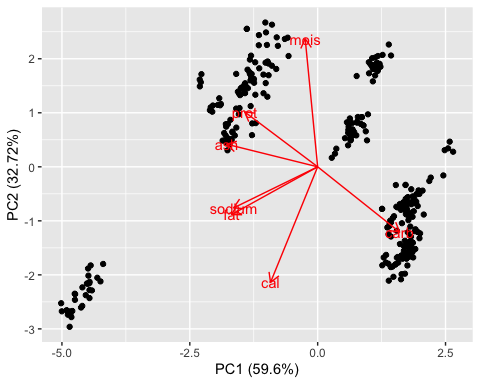
Each variance above is associated with each principal component calculated. The first number above (.596) means that 59.6% of the total variance in the data set is captured in the first principal component. As you can see the variances of principal components 4 to 7 are very close to zero. From this we can gather that including these components in analysis would not provide much further insight into the variability of the overall data set. The first 3 principal components account for most of the variability of the data set.

Now we can plot the first two principle components using two ways: biplot and autoplot.

#Code to plot data:  
biplot(pizza.pca, scale = 0)



#for a better view of plot. Set scale = 0 so the scale is the same for both biplot and autoplot.   
autoplot(pizza.pca, data = pizza.complete, color = 'brand', loadings = TRUE, loadings.label = TRUE, scale = 0 )



# Pizza.pca rotation matrix  
(pizza.pca$rotation)

## PC1 PC2 PC3 PC4 PC5 PC6  
## mois -0.06470937 0.6282759 -0.42166894 -0.2207216 -0.006470293 0.4464499018  
## prot -0.37876090 0.2697067 0.74602744 -0.0105932 -0.387982788 -0.0001715203  
## fat -0.44666592 -0.2343791 -0.19930871 -0.5070422 0.173367634 -0.5254028685  
## ash -0.47188953 0.1109904 0.05627269 0.5523985 0.670885701 0.0588609281  
## sodium -0.43570289 -0.2016617 -0.45516887 0.4462769 -0.602614079 0.0031309852  
## carb 0.42491371 -0.3203121 0.05223651 0.3343395 0.007436899 -0.0005088535  
## cal -0.24448730 -0.5674576 0.11331559 -0.2792632 0.078003175 0.7219138527  
## PC7  
## mois 0.4185690354  
## prot 0.2767646428  
## fat 0.3776715255  
## ash 0.0560214003  
## sodium -0.0005243238  
## carb 0.7760679112  
## cal 0.0120598098

In the graphs above, each of the different pizzas are shown on a 2D space (black numbers). The pizzas (black numbers) that are close to each other in groups in the plot have similar data patterns. We can also see that some pizzas tend to be more associated with a pizza trait. We can see this in clusters of pizza (black numbers) around the pizza trait (red arrow).

As stated in the plot, the red arrows represent the pizza traits. There are three important characteristics of these arrows to note. The first is the direction of the vector in respect to the PC (x and y axes). The more parallel the vector is to a PC axis, the more it contributes to just that PC. The second is the length of the vector. The longer the vector is, the more variability it represents. The third/final characteristic is the angle between vectors. The smaller the angles between two vectors, the higher the positive correlation between the two variables there is. If the angle is right, then there is no correlation. Opposite angles mean high negative correlation.

## Results Explanation

Looking at the first principal component (PC1), in the results above, we select a comparative threshold of .4 to assess the trends across the predictors that experience the most variance. Since fat, ash, sodium, and carb are all above the absolute threshold value, we can say that fat, ash, and sodium all move together (increase or decrease), and with their movement, carb will go in the opposite direction.

92.3% of variability is accounted by the first two principal components. Looking at the graph above, we can see that the fat and sodium are highly correlated since the angle between them is super small. There is no correlation between calories and protein since the angle between is a right angle (90 degrees). Moisture is almost parallel to PC2 (the y-axis), so it is the highest contributor to PC2.