Lab 5 Cluster Analysis Part I

The goal of this lab is to become familiar with the application the non-hierarchical clustering procedure K-means clustering.

## Data

Call in the data set “snail\_data.csv” from your working directory and name it **snails**. This data set consists of three morphological measurements and four environmental measurements for the snail *Littoraria angulifera* at 9 geographic locations ranging from western Africa to the Caribbean and South America. **For this analysis we are only going to use the first three variables** which are the shell morphology measurements. Remember to denote that the first row and column are row names and column names, respectively, when you read in the csv file. The matrix notation with square brackets at the end of this code specifies that we want the snails data object to only contain the first three columns of the dataset contained in the csv file.

snails <- read.csv("G:/Shared drives/MultivariateStatistics/Data/LabData/Lab5/snail\_data.csv",  
 row = 1, header = TRUE)[, 1:3]

## Download packages

You will be using new packages cluster and raster.

library(raster)  
library(cluster)  
library(mvnormtest)  
library(MVN)

# K-means clustering

**K-means** clustering is a **non-hierarchical** clustering method that seeks to find groups that maximize within-group homogeneity:

`?`(kmeans)

## Checking model assumptions

K-means clustering does have the assumption of multivariate normality. Check the histograms of each variable (as discussed in lab 2) and determine if you need to transform them.

Let’s also test for multivariate and univariate normality (even though these tests are conservative).

mshapiro.test(t(snails))  
  
mvn(snails, mvnTest = "mardia")

## Question 1: Compare test results to the histograms you plotted (show all code). Are you able to acchieve univariate and multivariate normality? We will proceed with the analysis either way, but consider how this might change youre interpretation of the results. 10pts

If variables (columns) are measured on different scales or have large differences in variance, you must scale the variables (columns). Let’s check this using the coefficient of variation (cv) on column totals:

snail.tot <- apply(snails, 2, sum)  
cv(snail.tot)

## [1] 39.58346

## Question 2: What do you think? Should you scale the snail data set? (10 pts)

If so, *z standardize* (scale function) the variables.

snails <- scale(snails)

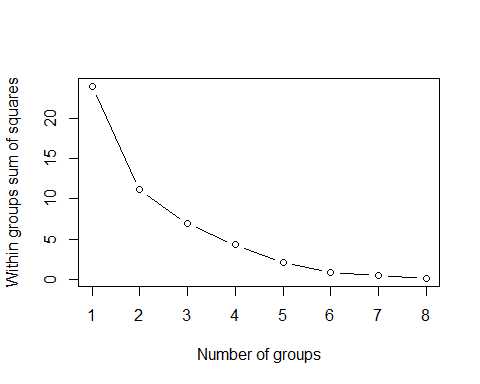
If you remember from lecture, one way to determine the number of clusters, *k*, for the k-means cluster analysis is to look at a scree plot for the within-group sum of squares. Let’s do a loop that will find the within group sum of squares for the full range of possible cluster solutions (nrows-1). The snails data set has nine samples (rows), so you will explore solutions from 1-8 clusters.

#Set a vector for the loop to fill.   
wss <- rep(0, 8)  
  
#Run a loop for 1 to 8 clusters:  
for (i in 1:8) # sets the number of times the loop will be run i.e., the number of clusters in this case)  
   
wss[i] <- sum(kmeans(snails, centers = i,nstart=25)$withinss) # run the kmeans function for each number of clusters (i) and extract the within sum of squares for each.  
  
#Check out you vector of within group sum of squares for one to eight groups:  
wss

## [1] 24.0000000 11.1259137 6.9344101 4.2727792 2.0731364 0.8488141 0.4801028  
## [8] 0.1127559

Make the scree plot:

plot(1:8, wss, type = "b", xlab = "Number of groups", ylab = "Within groups sum of squares")



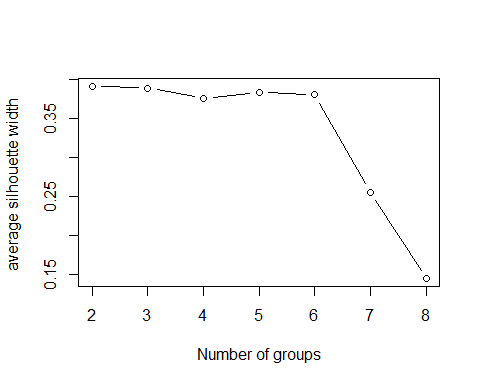
## Question 3: How many clusters are present in the best K-means solution? 10 pts

Silhouette Width is another indicator for determining how many clusters to use.

`?`(silhouette)

We can run a similar loop to look at the average silhouette width:

sil <- rep(0, 8)  
for (i in 2:8) sil[i] <- summary(silhouette(kmeans(snails, centers = i, iter.max = 100,  
 nstart = 25)$cluster, dist(snails)))$avg.width  
plot(2:8, sil[2:8], type = "b", xlab = "Number of groups", ylab = "average silhouette width ")



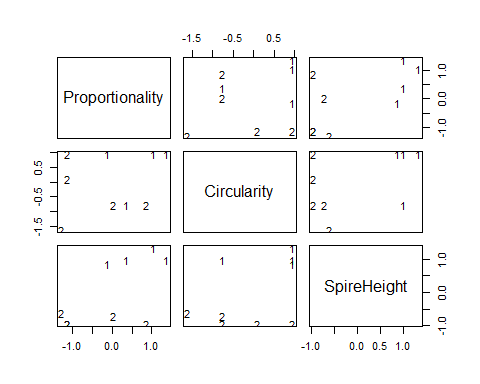
# Question 4: Does the plot for the average silhouette width concur with the scree plot for the within-group sum of squares? 10 pts

Let’s go ahead and plot out the result for the optimal cluster solution that you just found. Fill in the code for the number of centers (clusters) you found above:

snails.kop <- kmeans(snails, centers = 2, iter.max = 10, nstart = 25)

Plot a scatter plot showing cluster designations:

pairs(snails, panel = function(x, y, z) text(x, y, snails.kop$cluster))



# Question 5: Which morphological metric really separates the two groups? 10 pts

Now, let’s plot them against principal components 1 and 2:

Run, plot and interpret the PCA:

snail.pc <- princomp(snails, cor = T)  
summary(snail.pc)

## Importance of components:  
## Comp.1 Comp.2 Comp.3  
## Standard deviation 1.3796772 0.8700575 0.5826583  
## Proportion of Variance 0.6345031 0.2523333 0.1131636  
## Cumulative Proportion 0.6345031 0.8868364 1.0000000

snail.pc$loadings

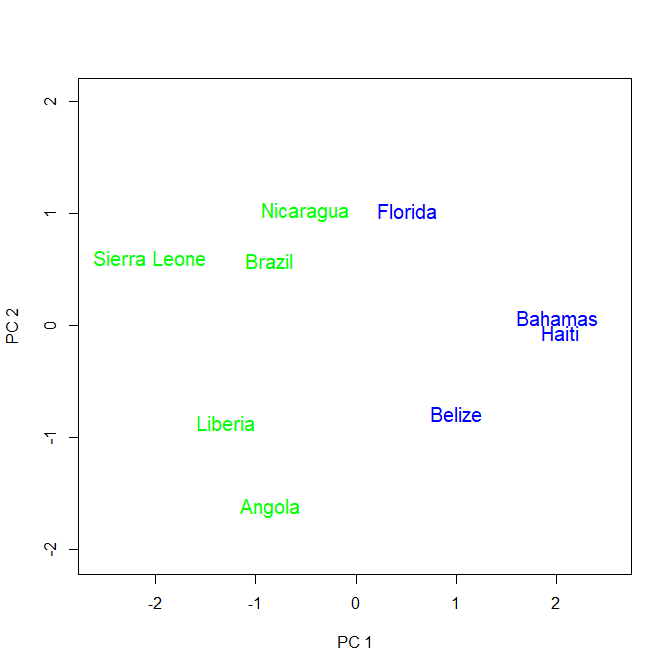
##   
## Loadings:  
## Comp.1 Comp.2 Comp.3  
## Proportionality 0.589 0.533 0.608  
## Circularity 0.486 -0.834 0.260  
## SpireHeight 0.646 0.143 -0.750  
##   
## Comp.1 Comp.2 Comp.3  
## SS loadings 1.000 1.000 1.000  
## Proportion Var 0.333 0.333 0.333  
## Cumulative Var 0.333 0.667 1.000

Set up colors for each cluster:

my.color.vector <- rep("green", times = nrow(snails))  
my.color.vector[snails.kop$cluster == 1] <- "blue"  
my.color.vector[snails.kop$cluster == 2] <- "green"

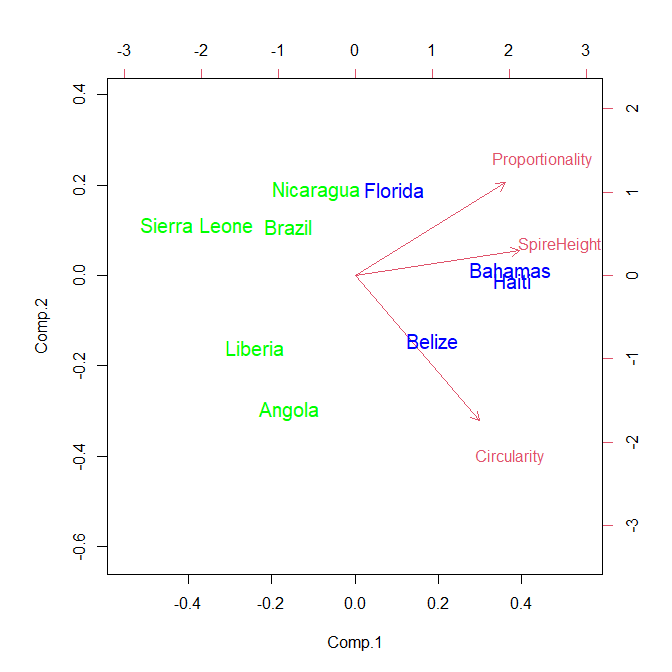
Plot clusters:

plot(snail.pc$scores[, 1], snail.pc$scores[, 2], ylim = range(snail.pc$scores[, 1]),  
 xlim = range(snail.pc$scores[, 1] \* 1.25), xlab = "PC 1", ylab = "PC 2", type = "n",  
 lwd = 2)  
text(snail.pc$scores[, 1], snail.pc$scores[, 2], labels = rownames(snails), cex = 1.25,  
 lwd = 2, col = my.color.vector)



Plot clusters onto biplot:

biplot(snail.pc, xlabs = rep("", 9), xlim = range(-0.55, 0.55))  
text(snail.pc$scores[, 1], snail.pc$scores[, 2], labels = rownames(snails), cex = 1.25,  
 lwd = 2, col = my.color.vector)



## Question 6: How does plotting your clusters on a bi-plot add to your interpretation of the cluster analysis and the PCA? 10 pts

## Question 7: Would it make sense to perform k-means clustering on your own dataset? Consider potential research questions on all or a subset of your data and model assumptions. 40 pts.

**This was potentially a short lab. If you have extra time, you might want to make sure you finalize your project 1 submission.**