**Dealing with multi-dimensional data using R:**

**PCA plots**

Scientists need to be careful about what dimensions are being plotted and how the graph is interpreted. It is tempting to include a considerable amount of information on a graph, however the importance of the parameter of interest quickly becomes lost as the reader becomes overwhelmed with the plot. While simple line graphs and boxplots are useful for a data set with two or three-dimensions, different approaches should be taken when a response variable is regressed against a large number of covariates. One analysis technique commonly used to deal with large and/or multi-dimensional data is the use of a Principal Component Analysis (PCA). The main purpose of a PCA is to determine patterns within a data set by reducing the dimensions of the set, yet also ensuring the minimal loss of information. In other words, it identifies a smaller number of uncorrelated variables, or principal components, from a large or multi-dimensional dataset in order to expose any redundancy in the data. If there are many pairwise correlations between the variables to consider examining numerous scatterplots might not be the preferred method of data analysis, therefore PCA plots can be very useful.

In this tutorial, two examples will be used to determine if any patterns arise in a smaller data set and a larger data set. There are many packages that can be used to construct PCA plots; here ‘ggplot2’ and ‘rgl’ will be used.

Some pointers on how to read this tutorial:

* Comments can be made almost anywhere starting with **#**, and everything to the end of the line is a comment; in this tutorial ## will be used to make comments within the syntax (green)
* Commands are separated using either by a semi-colon (;) or by simply starting a new line
* For more information on a prompt use ? before the command. (i.e. ?matrix)

Lets begin:

Begin by setting the working directory, this is where you want R to save all your work and is the location of the data set you wish to work with. Follow this with reading in the data and storing it, here ‘data’ is used to interact with the data set.

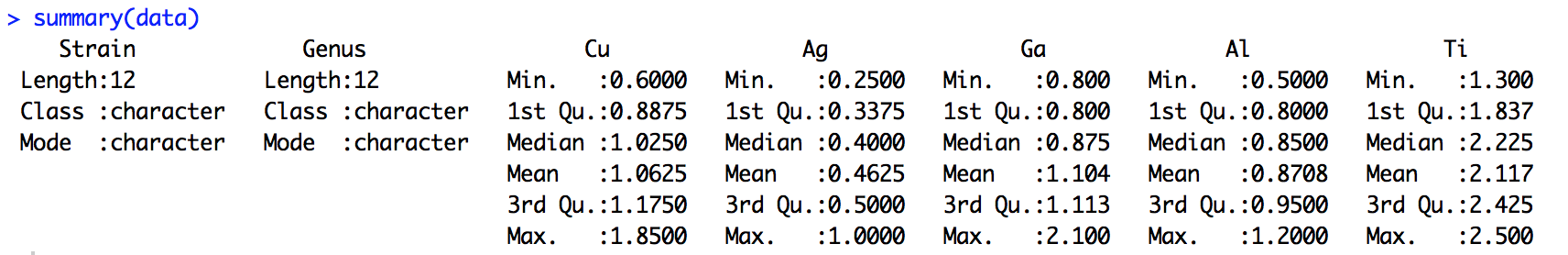
setwd("/Users/NatalieGugala/Desktop/R-course assignmemt") ## location of the ## file

data <- read.csv("Disk Diffusion Data.csv", header = T,stringsAsFactors=F)

head(data) ## this returns the titles of your columns

In this data set Ag is silver, Cu is copper, Ga is gallium, Al is aluminum and Ti is titanium. The values given under each metal are given in cm and depicte the zones of growth inhibition for each strain upon exposure to the metal challenges. In this work we are examining the relationships between the metal antimicrobials and their efficacy against several clinical isolates.

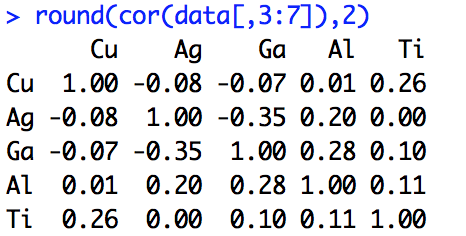
summary(data) ## reports several parameters of interest



In general Ga appeared to be more toxic based on the larger zone of growth inhibition when compared to the other metals.

Lets take a look at how correlated our data is:

round(cor(data[ , 3:7]), 2) ## Select all the rows and columns 3 to 7 and round to 2 ## decimal places



There are two ways of preforming the principal component analysis in R, using princomp and prcomp. By using princomp the data is centered ad scaled before the analysis and the coordinates on each principal component are calculated when both cor and scores are TRUE.

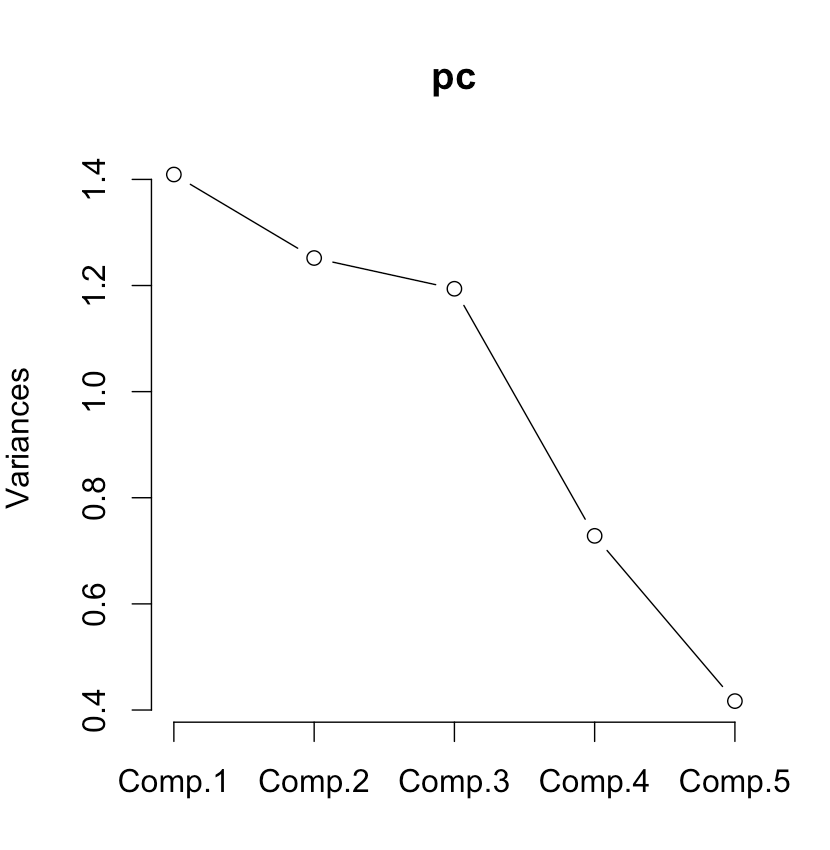
pc <- princomp(data[ , 3:7], cor=TRUE, scores=TRUE)

## store the results

summary(pc) ## Examination of the components

Using the plot function we can examine the variance in the data.

plot(pc, type="lines")



According to the data set, it is not till the last component where the variance decreases, indicating the data set has little correlation.

biplot(pc) ## use this to examine the data in two

## dimensions

Lets plot in three-dimensions. Start by downloading the rgl package from the Internet. This package is easy to work with and has some great features such as the generation of three-dimensional interactive plots of different kinds. Note that application XQuartz must be installed.

install.packages(rgl)

library(rgl) ## load the package into the working directory

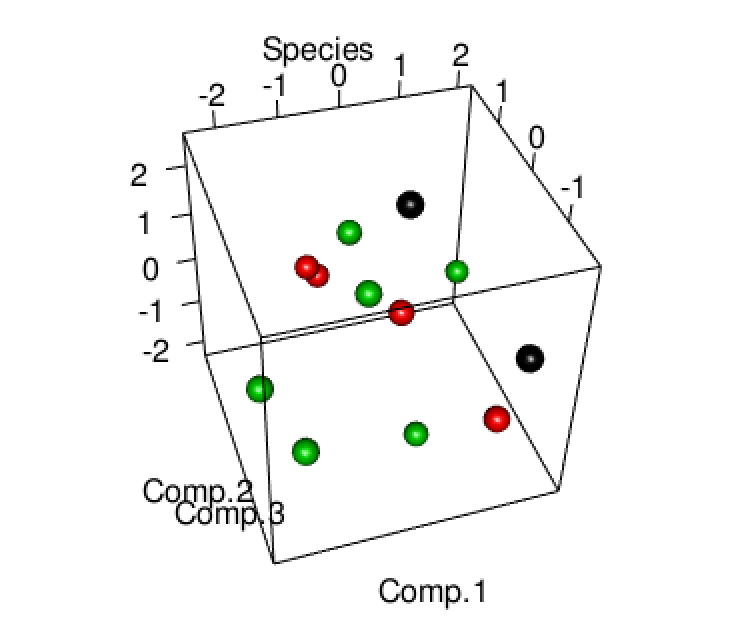
data$Genus <- as.factor(data[ , 2]) ## factor the data so ## it can be changed to numeric data

data$Genus <- as.numeric(data[ , 2]) ## this will allow for ## each genus ([ ,2]) in the data set to have its own

## colour

Next plot the data in three dimensions. Use ?plot3d for more options such as changing the x and y labels, changing the colours and the size of the spheres.

plot3d(pc$scores[,2:4], col=data$Genus, type="s", main = "Species")



This plot is a three dimensional interactive figure. Note there may be a correlation between the red balls (Staphylococcus strains).

Now lets use ggplot2 to produce an aesthetically pleasing two-dimensional PCA plot.

Download the files as above and upload them into the working directory

library(ggplot2)

library(ggfortify)

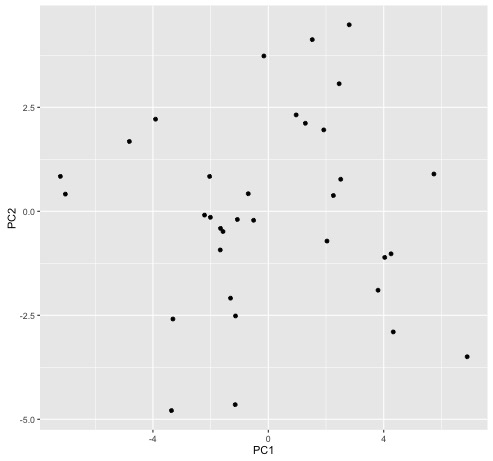
download.file("https://vincentarelbundock.github.io/Rdatasets/csv/car/WeightLoss.csv", destfile = "weight.csv", method = "auto") ## downloaded directly from the

## internet, destfile is the name of the file you wish ## to save it under

weight <- read.csv("weight.csv", header = T, stringsAsFactors = F)

df <- weight[c(3:8)] ## select the columns to be analyzed ## (this means omitting those that are not numeric, such as ## “group” data)

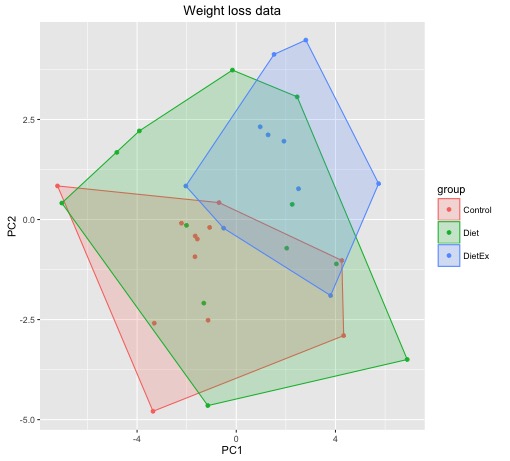
autoplot(prcomp(df)) ## All your data points with no colour ## label



autoplot(prcomp(df), data = weight, colour = "group",

frame = TRUE, main = "Weight loss data") ## sort the

## data points based on the group and frame the groups



Note that one can also use princomp as completed above.

There are some key points one must consider when using PCA plots, which include the correlation between the data and the size of the data. In the first example a small number of data points were used. Little correlation was found between the data. However, this does not mean that relationships cannot be exposed. Note that the advantage of using a three-dimensional plot is the use of the interactive window. Further manipulation identified a correlation between the Staphylococcus strains, however this was difficult to observe since the correlation was examined between all strains at once. In other words, choosing to conduct a PCA plot and how it will be preformed is grounded in the question one wishes to answer.