

## Lab 4

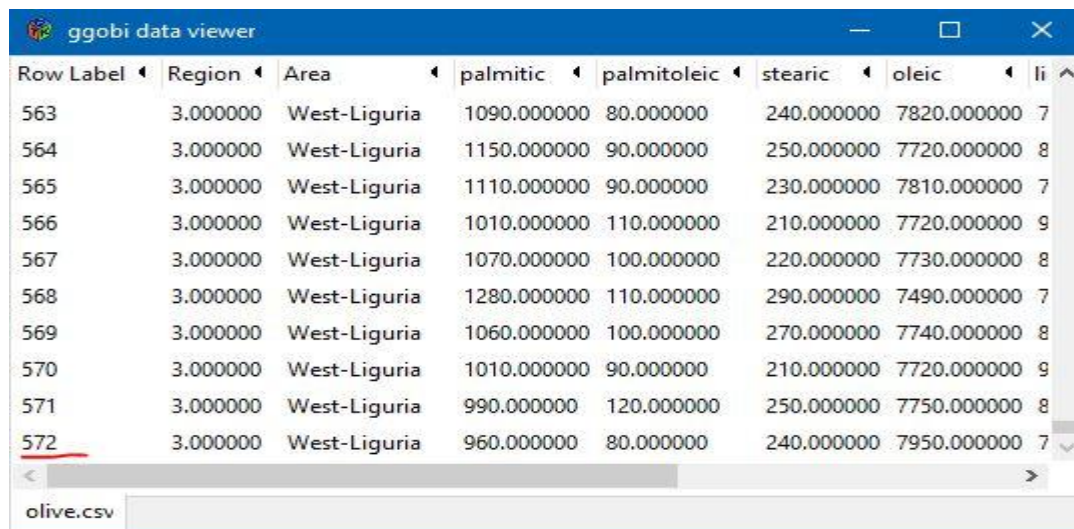
Gustav Sternelöv

September 29, 2016

### Assignment 1

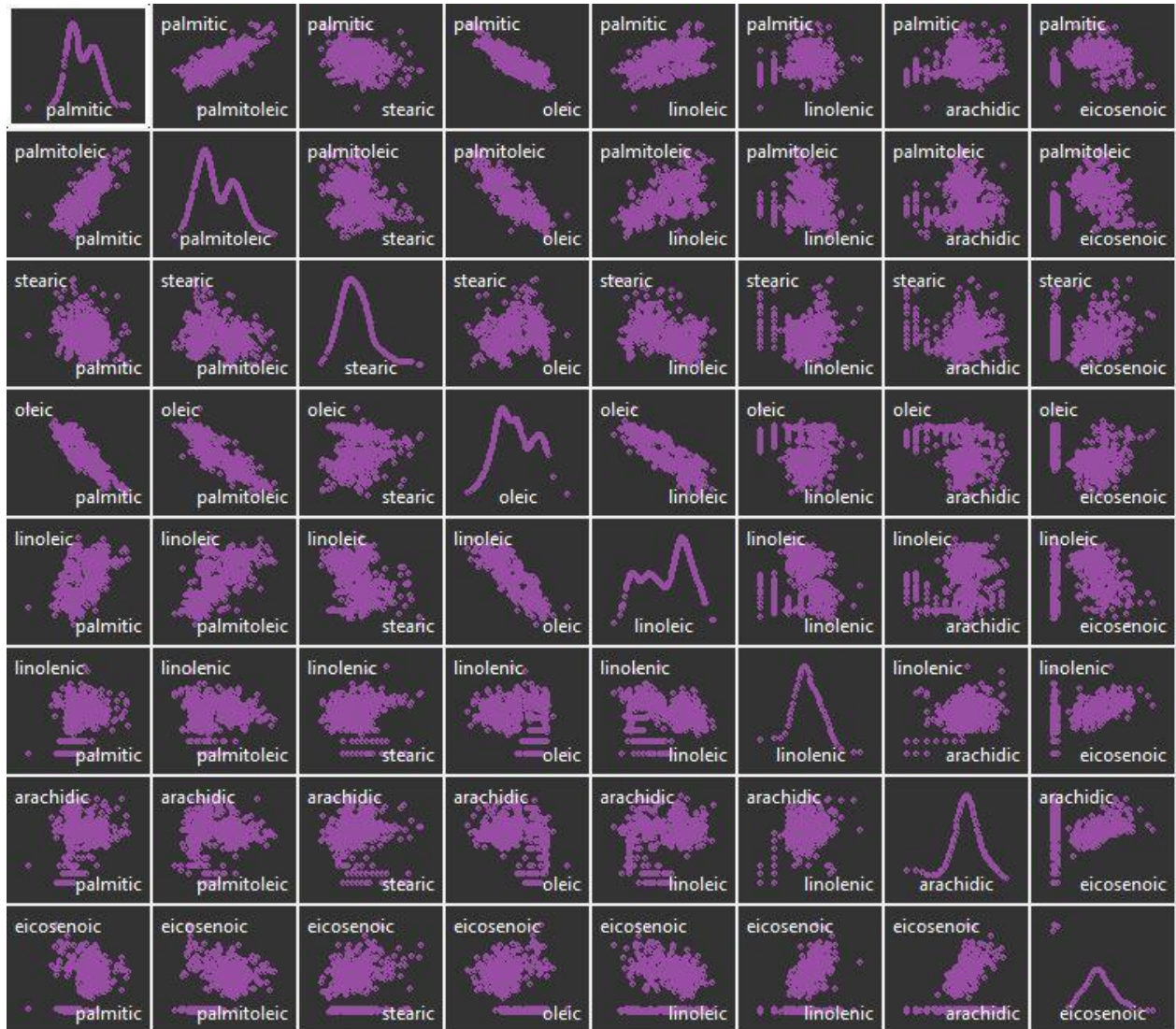
#### 1. Open olive.csv in GGobi and open Data Viewer. How many observations are present in the data?

By using the tool *Data viewer* in GGobi it is easy to look up that there are 572 observations in the data set.



Row Label	Region	Area	palmitic	palmitoleic	stearic	oleic	li
563	3.000000	West-Liguria	1090.000000	80.000000	240.000000	7820.000000	7
564	3.000000	West-Liguria	1150.000000	90.000000	250.000000	7720.000000	8
565	3.000000	West-Liguria	1110.000000	90.000000	230.000000	7810.000000	7
566	3.000000	West-Liguria	1010.000000	110.000000	210.000000	7720.000000	9
567	3.000000	West-Liguria	1070.000000	100.000000	220.000000	7730.000000	8
568	3.000000	West-Liguria	1280.000000	110.000000	290.000000	7490.000000	7
569	3.000000	West-Liguria	1060.000000	100.000000	270.000000	7740.000000	8
570	3.000000	West-Liguria	1010.000000	90.000000	210.000000	7720.000000	9
571	3.000000	West-Liguria	990.000000	120.000000	250.000000	7750.000000	8
572	3.000000	West-Liguria	960.000000	80.000000	240.000000	7950.000000	7

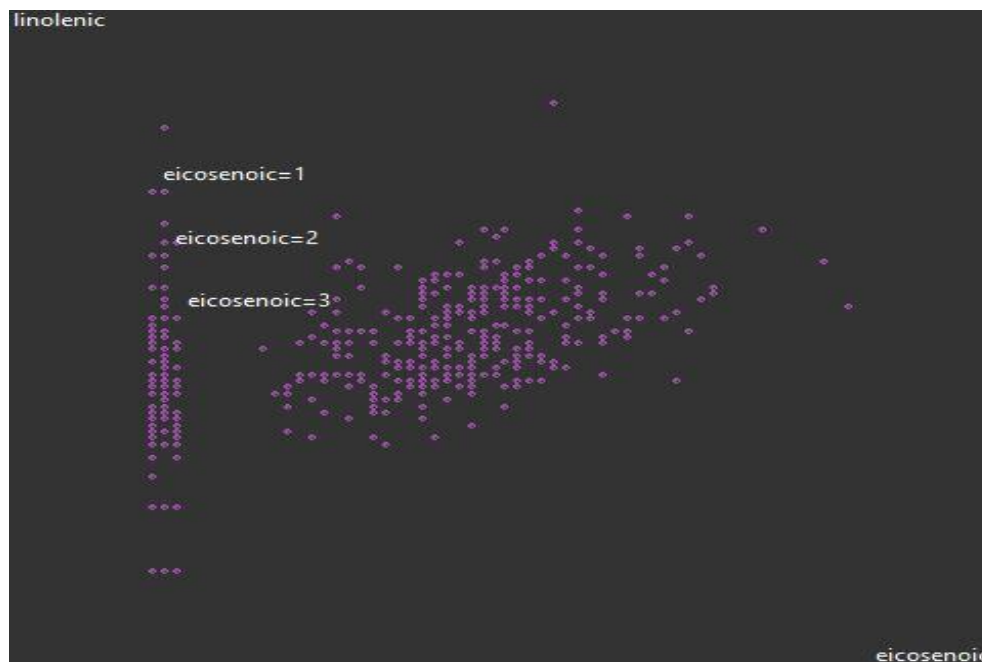
2. Create a scatter plot matrix that shows how the contents of different acids are related to each other. Investigate the matrix to find plots where the clusters are present. Close the plot.



Can see clusters for the variable *eicosenoic* together with all the other variables. In all cases is there one cluster with low values of *eicosenoic* and one with higher values of the acid. There might very well be other clusters, but they are harder to see clearly in this scatter plot matrix.

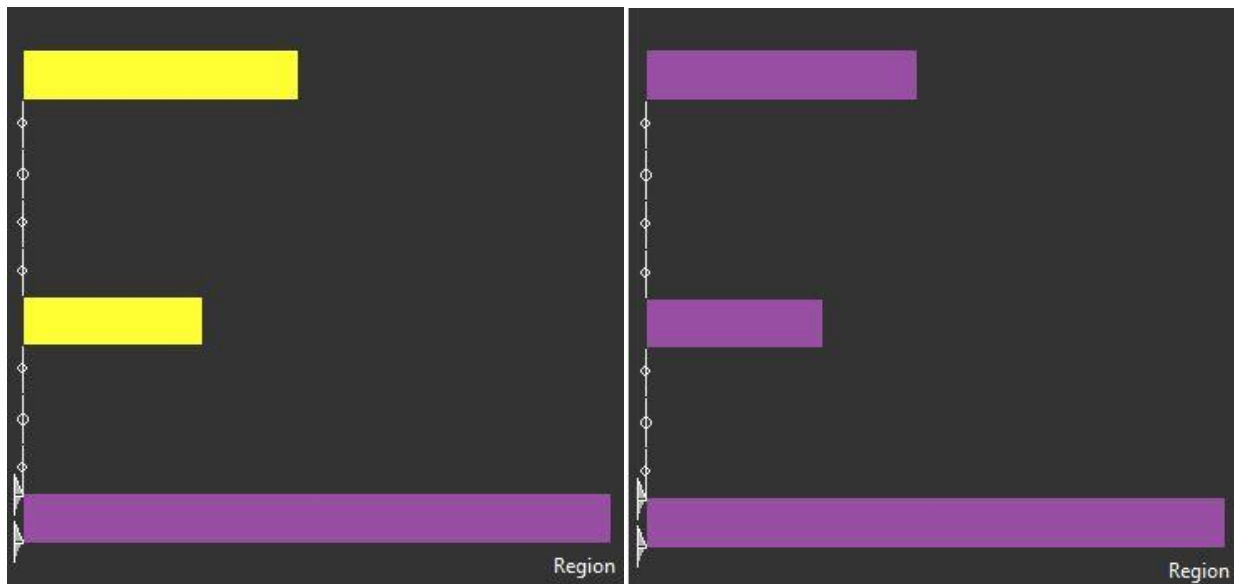
3. Create a scatter plot of the eicosenoic against linoleic. Based on section 2, comment why it can be interesting to investigate this pair of variables. You have probably found a group of observations having unusually low values of eicosenoic. Use identification tool to find out the exact values of eicosenoic for these observations.

The pair of variables plotted below are interesting since it seem to be two different clusters, at least based on the analysis in section 2. This analysis is confirmed by the plot and the cluster with observations which has low values of *eicosenoic* takes the values 1, 2 or 3.



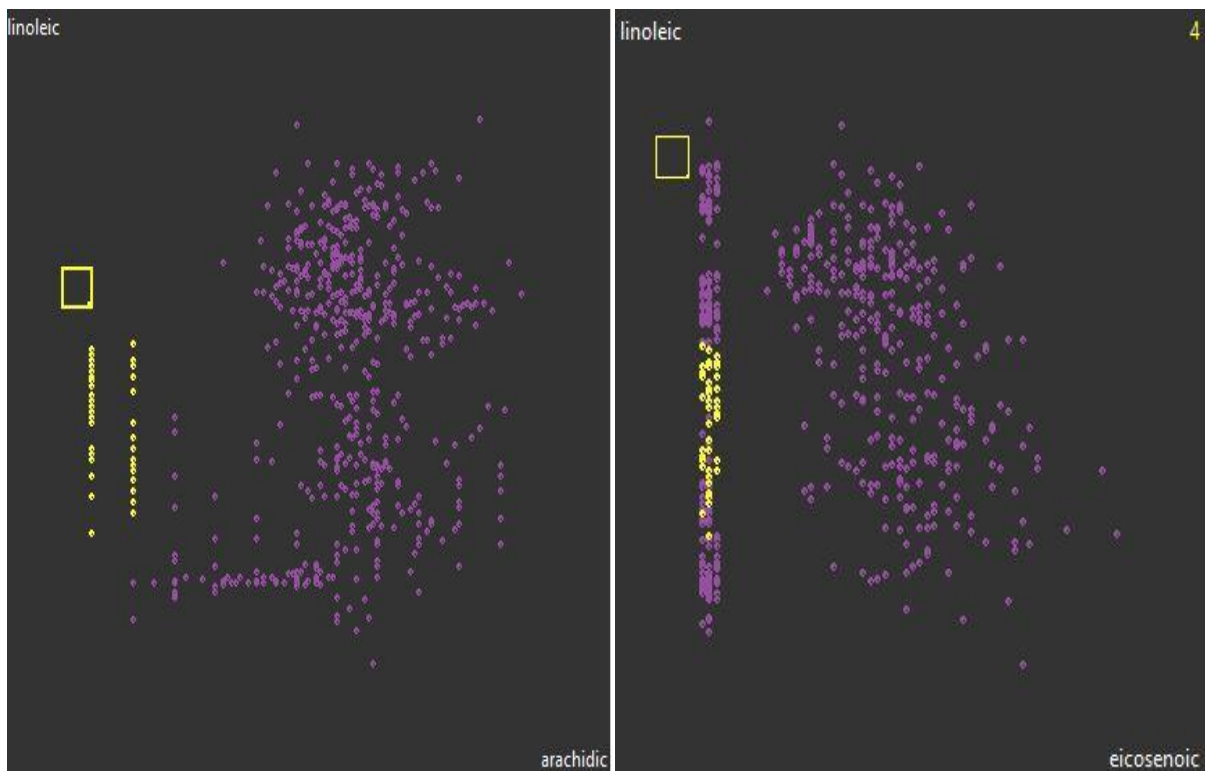
4. Create a histogram that shows how many observations fall within any given region. Use persistent brushing to identify the regions that correspond unusually low values of eicosenoic. Include the plots into your report and then remove the brushing (one way is to restart GGobi)

All the observations with low values of *eicosenoic* are colored in yellow. The histograms on the next page shows that all observations from two of the regions takes low values for the variable and that all observations from the third regions takes higher values.



5. Create scatter plots eicosenoic against linoleic and arachidic against linolenic. Which outliers in (arachidic, linolenic) are also outliers in (eicosenoic, linoleic)? Are outliers grouped in some way?

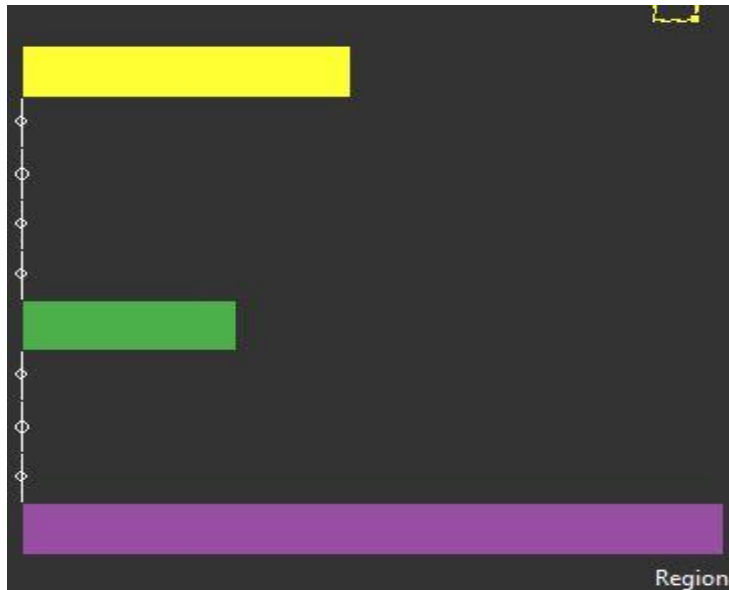
Hard to say exactly what values are outliers. Guessing that the values coloured in yellow in the arachidic versus linolenic scatter plot are the outliers. Common for all the outliers is that they have low values for *eicosenoic*.





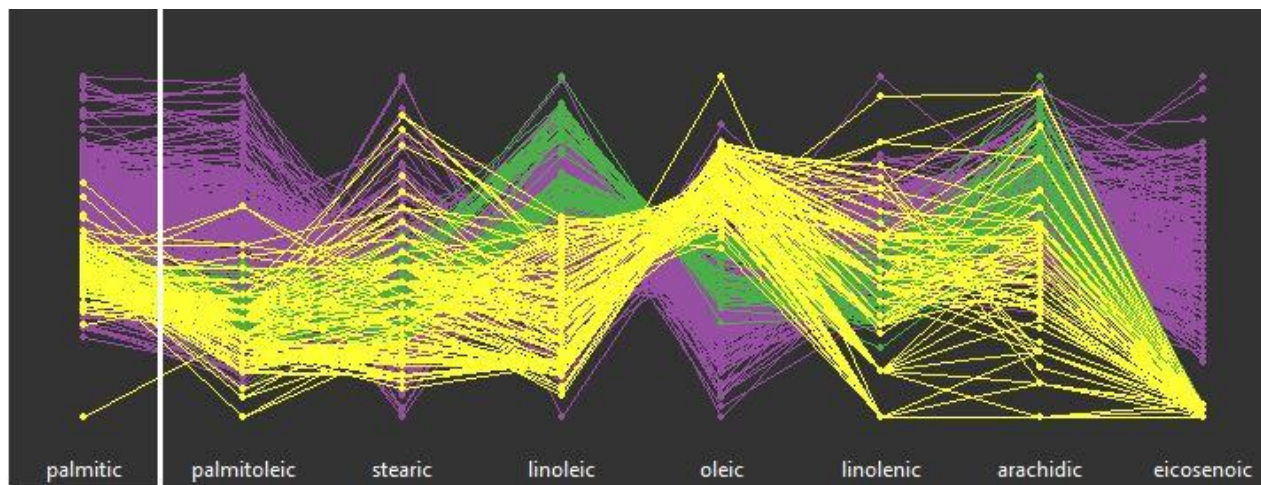
6. Use persistent brushing to paint by different colors the observations that fall into different regions. Keep these coloring during steps 7-9.

The color given for each region is shown with the graph below.



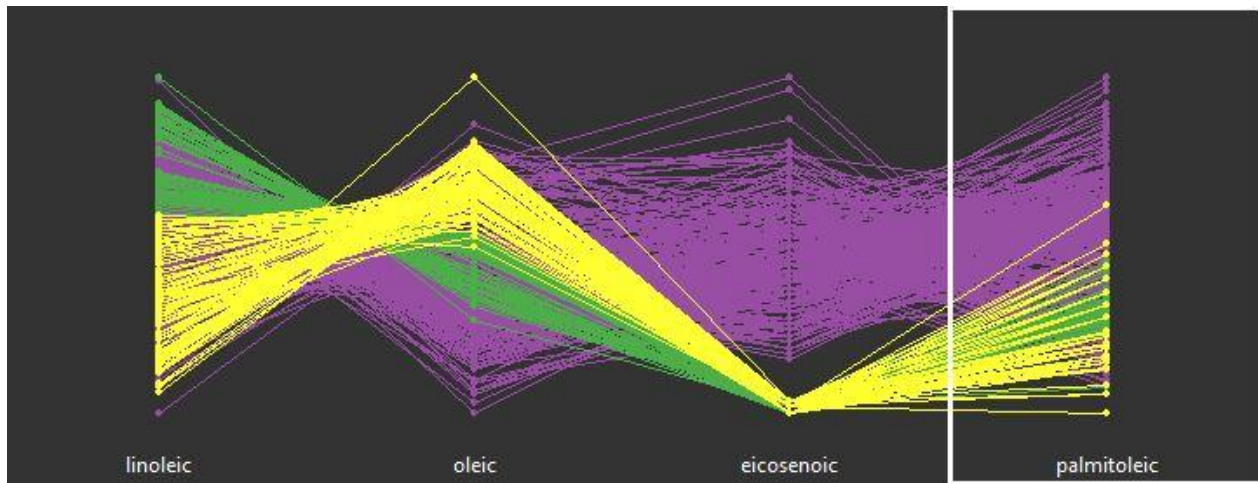
7. Create a parallel coordinate plot for the available eight acids. Select some proper subset of variables and define their order on the plot. Which variables can be taken for identifying clusters? (suggest at least three variables)

*Oleic* and *Eicosenoic* looks to be good for finding clusters. Perhaps also *Linoleic* and *Palmitoleic* can be used for this purpose.



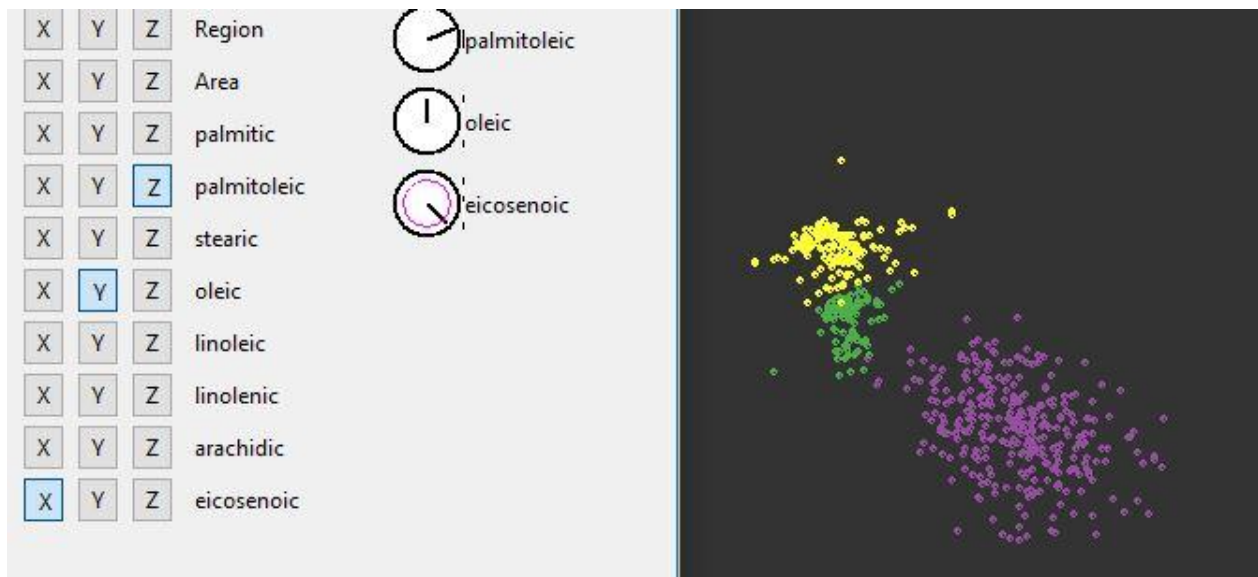
The next parallel plot includes only the four suggested variables from the first parallel plot. With this four variables the observations from the region colored in purple easily are

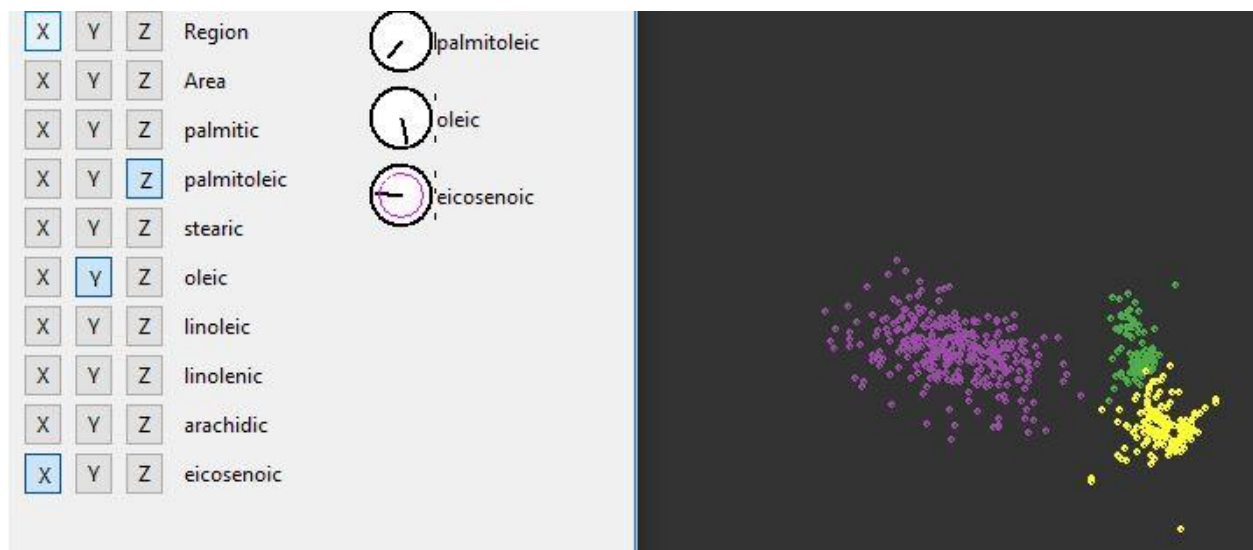
separated from the other observations. The observations from the green and yellow region are mainly separated given the two first variables in the plot.



**8. Create a 3D-rotation plot by using the variables found in step 7. Can you see clusters? Include proper screenshots motivating your answer.**

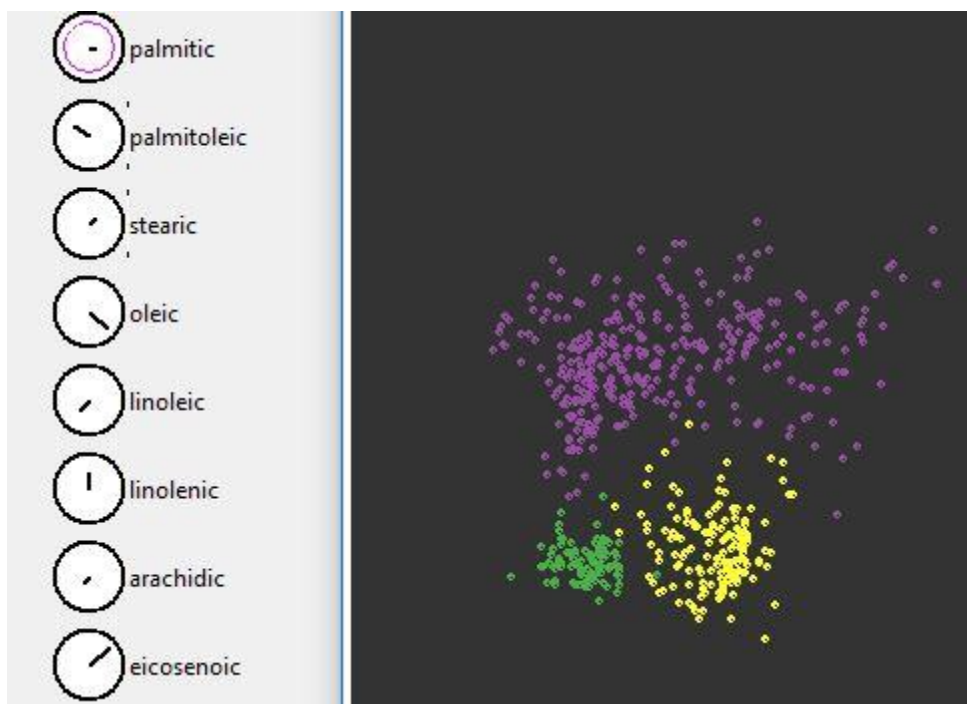
Two screenshots are used for answering this question. From step 7 are the variables *palmitoleic*, *oleic* and *eicosenoic* chosen. In both cases are *eicosenoic* an important variable and in the first case *palmitoleic* more than *oleic* and in the second *oleic* more than *palmitoleic*. It is also very clear that clusters can be seen.





**9. Use all 8 acids and examine a 2D-tour. Try to find a projection with the best separation of the data into clusters. How the clusters detected are related to the regions the oils come from?**

In this projection are the clusters detected mainly related to the acids *palmitoleic*, *oleic* and *eicosenoic*. Each cluster correspond to a region.



## 10. Based on the analysis above, try to suggest a strategy (or, maybe, several strategies) that would use information about the level of acids to discover which region the oil comes from.

Could look at the observed values for some of the variables which have proven to be interesting. For example the two first regions are separated from the third by low values of *eicosenoic*. Then, the two first regions could be separated by the value of the acid *oleic*. Some more variables could be used, but the general principle of the strategy is the same.

## Assignment 2

### 1. Load the file to R and answer whether it is reasonable to scale these data in order to perform a multidimensional scaling (MDS).

```
## Country Car MPG Weight Drive_Ratio Horsepower
## 1 U.S. Buick Estate Wagon 16.9 4.360 2.73 155
## 2 U.S. Ford Country Squire Wagon 15.5 4.054 2.26 142
## 3 U.S. Chevy Malibu Wagon 19.2 3.605 2.56 125
## 4 U.S. Chrysler LeBaron Wagon 18.5 3.940 2.45 150
## 5 U.S. Chevette 30.0 2.155 3.70 68
## 6 Japan Toyota Corona 27.5 2.560 3.05 95
## Displacement Cylinders
## 1 350 8
## 2 351 8
## 3 267 8
## 4 360 8
## 5 98 4
## 6 134 4

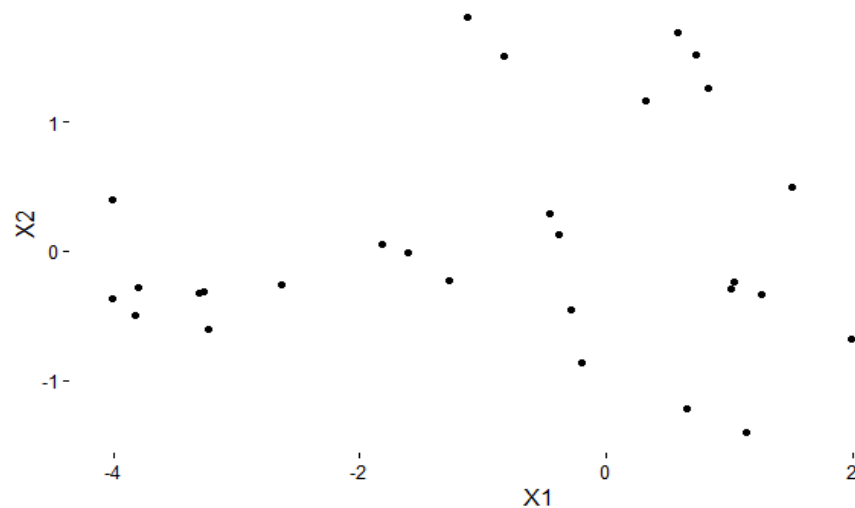
## MPG Weight Drive_Ratio Horsepower Displacement
## 24.760526 2.862895 3.093421 101.736842 177.289474
## Cylinders
## 5.394737

## MPG Weight Drive_Ratio Horsepower Displacement
## 42.8673186 0.4996658 0.2679691 699.3342817 7899.0761024
## Cylinders
## 2.5697013
```

The numeric variables in data have different scales in terms of mean and variance, so yes it is reasonable to scale this data set.

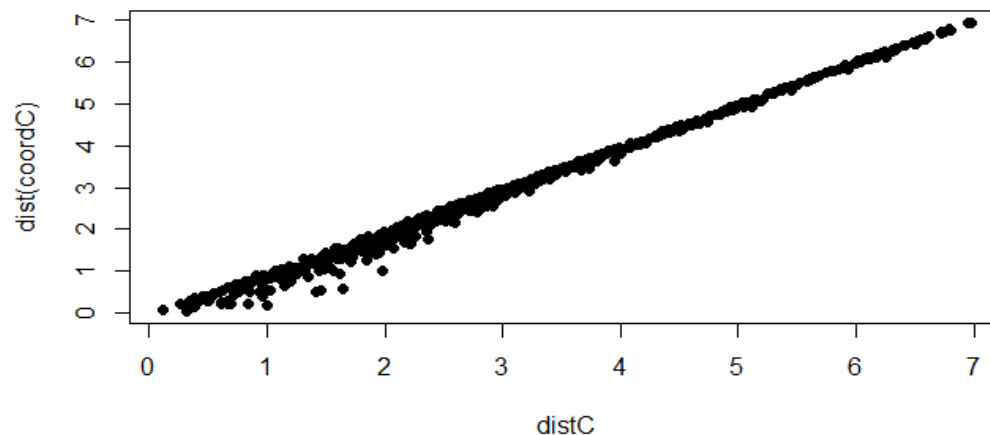


2. Write an R code that performs a metric MDS of the data (numerical columns) into two dimensions. Visualize the resulting observations in GGobi and analyze the plot.



The observations seem to be relatively spread over the graph. Perhaps some small groups can be seen in the graph. One to the left, on the right and one at the top but there are also quite many points that not are close to any of the groups.

3. Create the Shepard plot for the MDS performed and comment about how successful the MDS was.

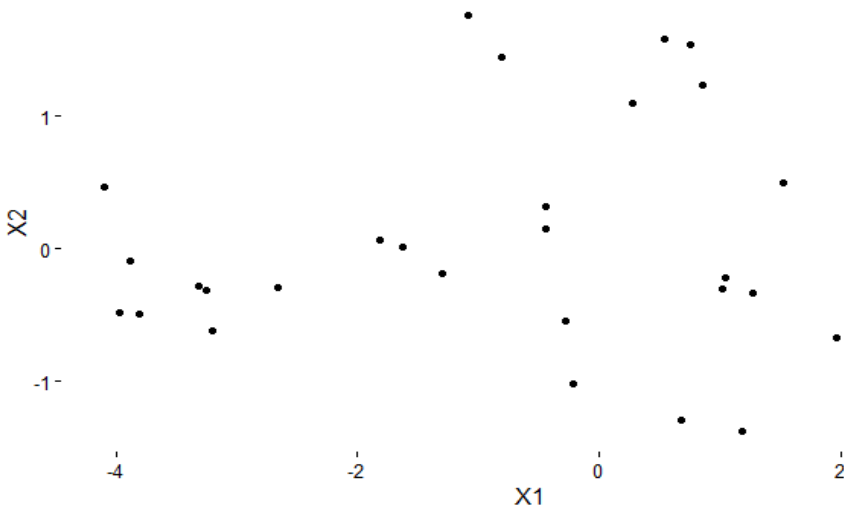


By looking at the Shepard plot the MDS seem to have been rather successful. The majority of the points lies along the diagonal line as the increase of the values is monotonic.

#### 4. Repeat steps 2-3 for the nonmetric MDS and Minkowski distance=2. Report the stress value and whether the MDS was converged.

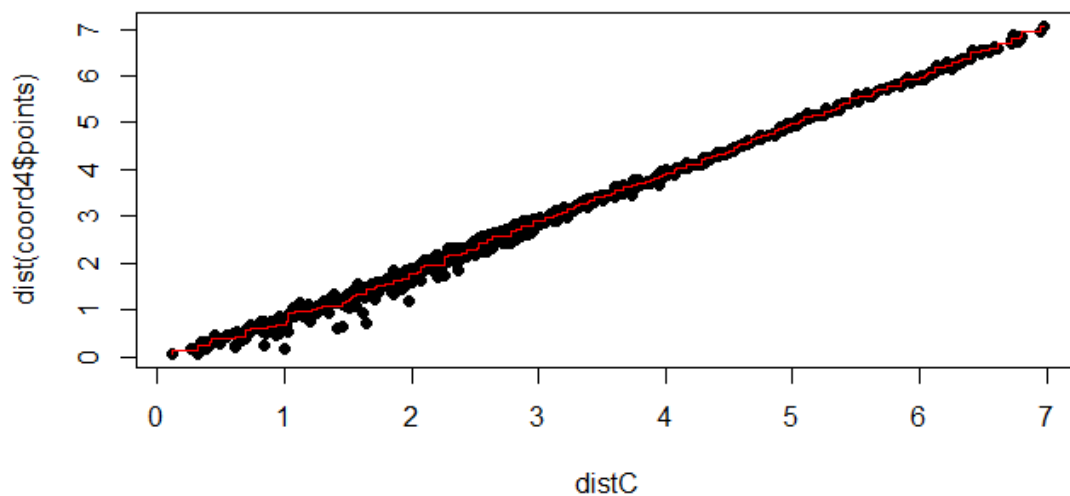
The output is presented below and it can be seen that the stress value is 2.60 and that the MDS has converged

```
## initial value 2.981155
## iter 5 value 2.607683
## final value 2.600426
## converged
```



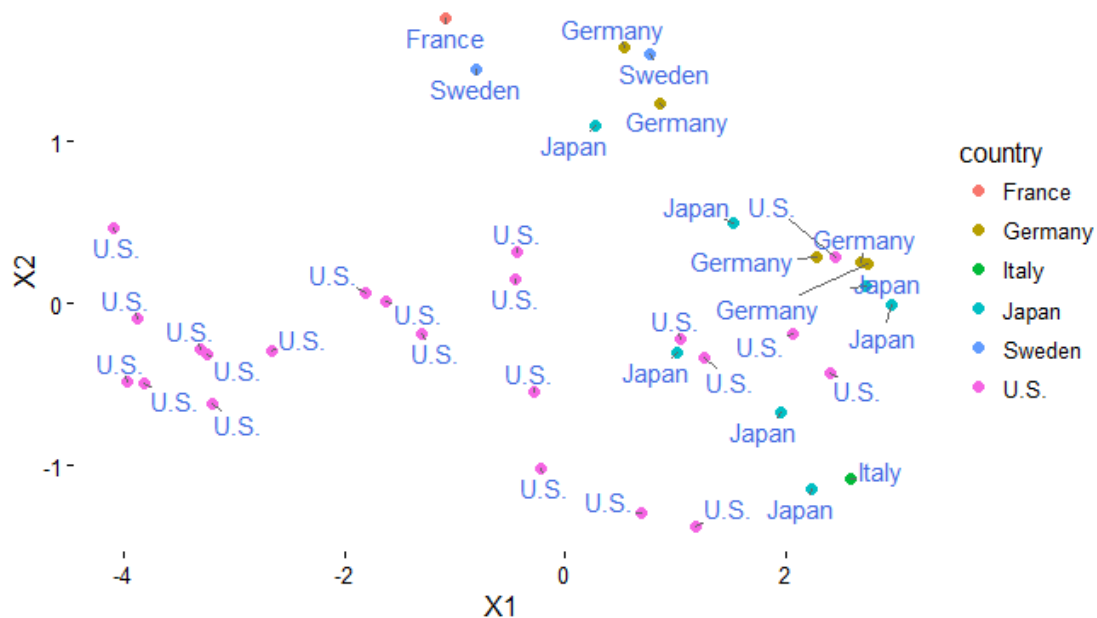
The plot with the observations is very similar to the plot presented in 2.2. It is hard to see any larger differences at all.

The Shepard plot is similar to the one in 2.3, so the non-metric MSD does also seem to have been successful.



## 5. Brush the MDS plot by the value of the Country. Draw conclusions.

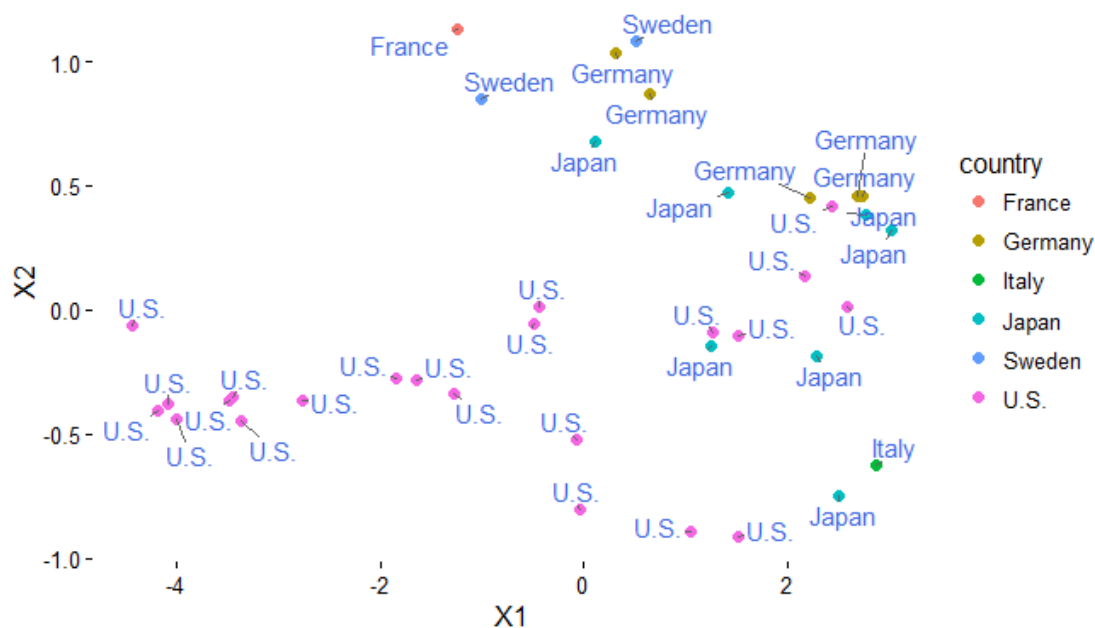
The plot in 2.4 is modified so that the country for each observation is added. It is noted that all the observations to the left in the graph are cars from the U.S. In the group to the right the origins of the cars is more mixed as there are American, European and Asian countries there. At the top of the graph there are five cars from Europe and one from Japan. In some sense the cars are clustered after country or region, but not entirely as there is no distinct grouping really.



## 6. Perform a nonmetric MDS for Minkowski distance $p=1$ . How have this change affected the clustering? Provide a Shepard plot and comment on the quality of the fit.

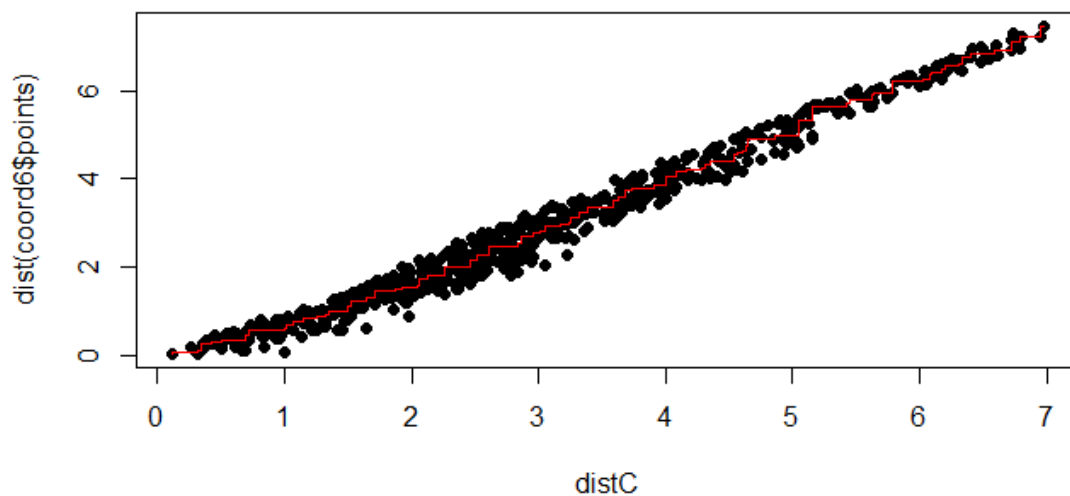
The stress value for this non-metric MDS is 4,47, a little higher than for the earlier non-metric MSD.

```
## initial value 8.814428
## iter 5 value 5.022452
## iter 10 value 4.587726
## iter 15 value 4.488405
## final value 4.472602
## converged
```



The clustering does not seem to have changed that much. In general, the pattern is the same as before.

The Shepard plot is shown on the following page and compared to the earlier Shepard plot this one looks more jittered. The increase is not as clearly monotonic increasing as before. However, it still appears to be a decent fit.

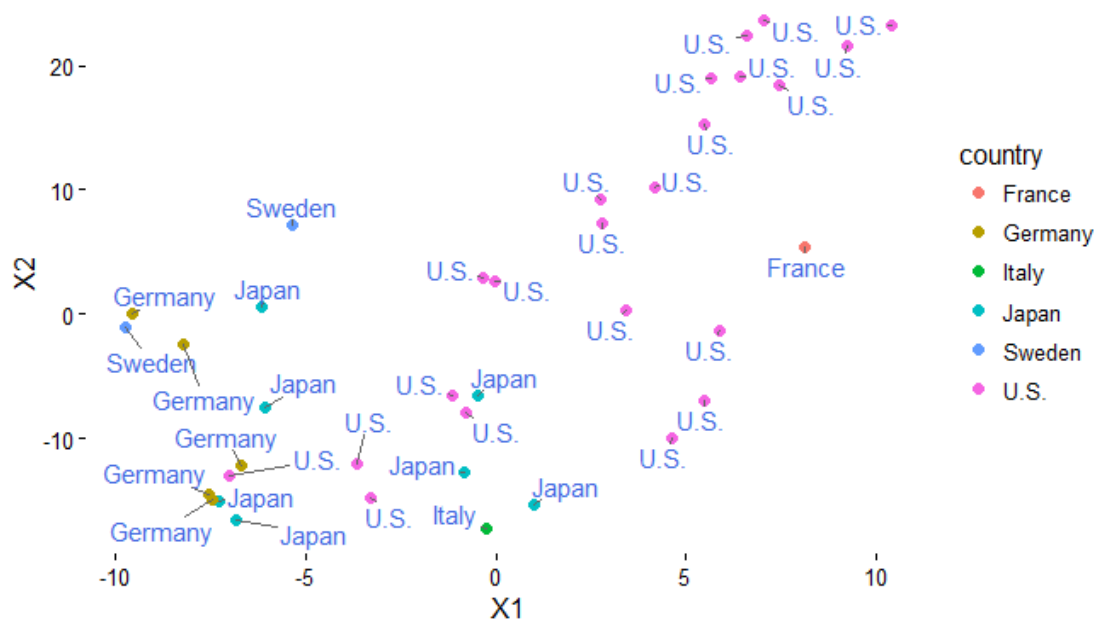


**7. Perform a nonmetric MDS for Minkowski distance  $p=2$ , randomly chosen starting points (uniformly from -1 to 1 in each dimension), and the maximum number of iterations equal to 500. How have this change affected the clustering? Have you got a better stress value? Provide a Shepard plot and comment on the quality of the fit.**

The number of iterations before the MDS converges has increased and the stress value is the highest of all stress values.

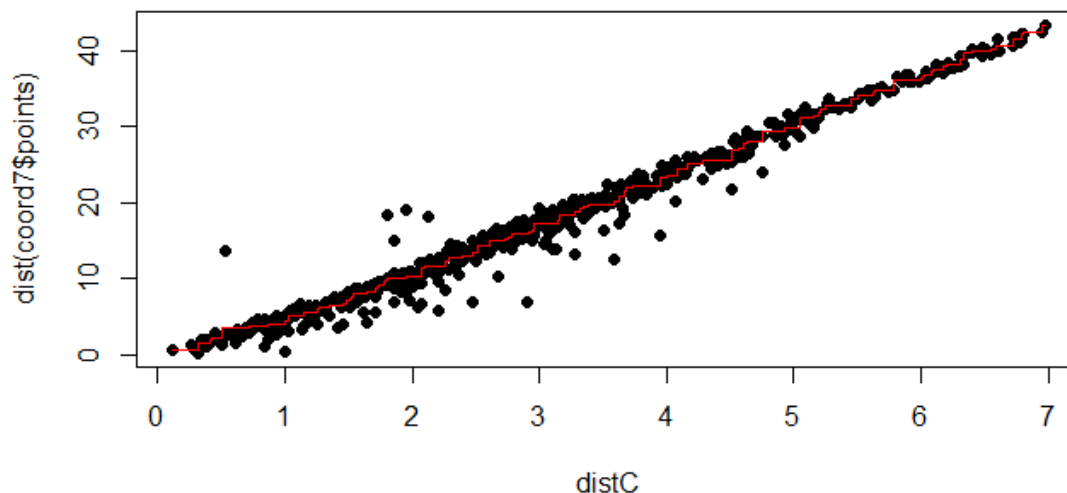
```
## initial value 42.437581
## iter 5 value 38.947935
## iter 10 value 34.941176
## iter 15 value 14.201296
## iter 20 value 9.984102
## iter 25 value 8.348556
## iter 30 value 6.948284
## iter 35 value 6.434648
## final value 6.377942
## converged
```





The clustering has changed a bit. At the top right corner a cluster of cars from the U.S. can be seen. In the left bottom corner there is mainly European and Japanese cars and between the two clusters it is more mixed.

The Shepard plot is the worst so far as the values are more spread than before. The general pattern with monotonic increasing values is still there but there are also some values that lies a bit away from the others. I think that this is an decent fit, but that the ones presented earlier in the report was better than this fit.



## 8. Which of the methods do you think was the best here?

Non-metric MDS with Minkowski distance equal to 1 returned the best result. It had the lowest stress value and a Shephard plot that indicated that the method returned a good fit.

### R-code

```
## ---- echo=FALSE, message=FALSE, warning=FALSE-----
library(ggplot2)
library(ggrepel)
library(gridExtra)
library(XLConnect)
wb =
loadWorkbook("C:\\Users\\Gustav\\Documents\\Visualization\\Lab4\\cars.xlsx")
cars = readWorksheet(wb, sheet = "Blad1", header = TRUE)

head(cars)
colMeans(cars[, 3:8])
diag(var(cars[, 3:8]))

## ---- echo=FALSE, fig.height=4, fig.width=7-----
car.numeric <- scale(cars[, 3:8])
distC <- dist(car.numeric)
coordC <- cmdscale(distC, k = 2)
coordCf <- data.frame(coordC, country = cars$Country, cylinders =
cars$cylinders)

p1 <- ggplot(coordCf, aes(X1, X2)) + geom_point(col = "black") +
theme_classic()
p1

## ---- echo=FALSE, fig.height=4, fig.width=7-----
plot(distC, dist(coordC), pch = 21, bg = "black")

## ---- echo=FALSE, warning=FALSE, message=FALSE, fig.height=4,
## fig.width=7----
library(MASS)
coord4 <- isoMDS(distC, k = 2, p = 2)
coords4 <- data.frame(coord4$points, country = cars$Country, cylinders =
cars$cylinders)

p2 <- ggplot(coords4, aes(X1, X2)) + geom_point(col = "black") +
theme_classic()
p2

## ---- echo=FALSE, fig.height=4, fig.width=7-----
```

```

sh <- Shepard(distC, coord4$points)
plot(distC, dist(coord4$points), pch = 21, bg = "black")
lines(sh$x, sh$yf, type = "S", col = "red")

## ---- echo=FALSE, fig.height=4, fig.width=7-----
ggplot(coords4, aes(X1, X2, col = country)) + geom_point(size = 2) +
theme_classic() +
  geom_text_repel(aes(label = country), col = "royalblue")

## ---- echo=FALSE, fig.height=4, fig.width=7-----
coord6 <- isoMDS(distC, k = 2, p = 1)
coords6 <- data.frame(coord6$points, country = cars$Country, cylinders =
cars$cylinders)

ggplot(coords6, aes(X1, X2, col = country)) + geom_point(size = 2) +
theme_classic() +
  geom_text_repel(aes(label = country), col = "royalblue")

## ---- echo=FALSE, fig.height=4, fig.width=7-----
sh <- Shepard(distC, coord6$points)
plot(distC, dist(coord6$points), pch = 21, bg = "black")
lines(sh$x, sh$yf, type = "S", col = "red")

## ---- echo=FALSE, fig.height=4, fig.width=7-----
set.seed(1897)
initVal <- matrix(runif(76, min = -1, max = 1), ncol = 2)
coord7 <- isoMDS(distC, k = 2, p = 2, maxit = 500, y = initVal)
coords7 <- data.frame(coord7$points, country = cars$Country, cylinders =
cars$cylinders)

ggplot(coords7, aes(X1, X2, col = country)) + geom_point(size = 2) +
theme_classic() +
  geom_text_repel(aes(label = country), col = "royalblue")

## ---- echo=FALSE, fig.height=4, fig.width=7-----
sh <- Shepard(distC, coord7$points)
plot(distC, dist(coord7$points), pch = 21, bg = "black")
lines(sh$x, sh$yf, type = "S", col = "red")

## ----
code=readLines(knitr::purl('C:\\Users\\Gustav\\Documents\\Visualization\\Lab4
\\Lab4.Rmd',documentation
## = 1)), eval = FALSE, tidy=TRUE---- NA

```