Lab Report 4

Thomas Zhang

2016 M10 4

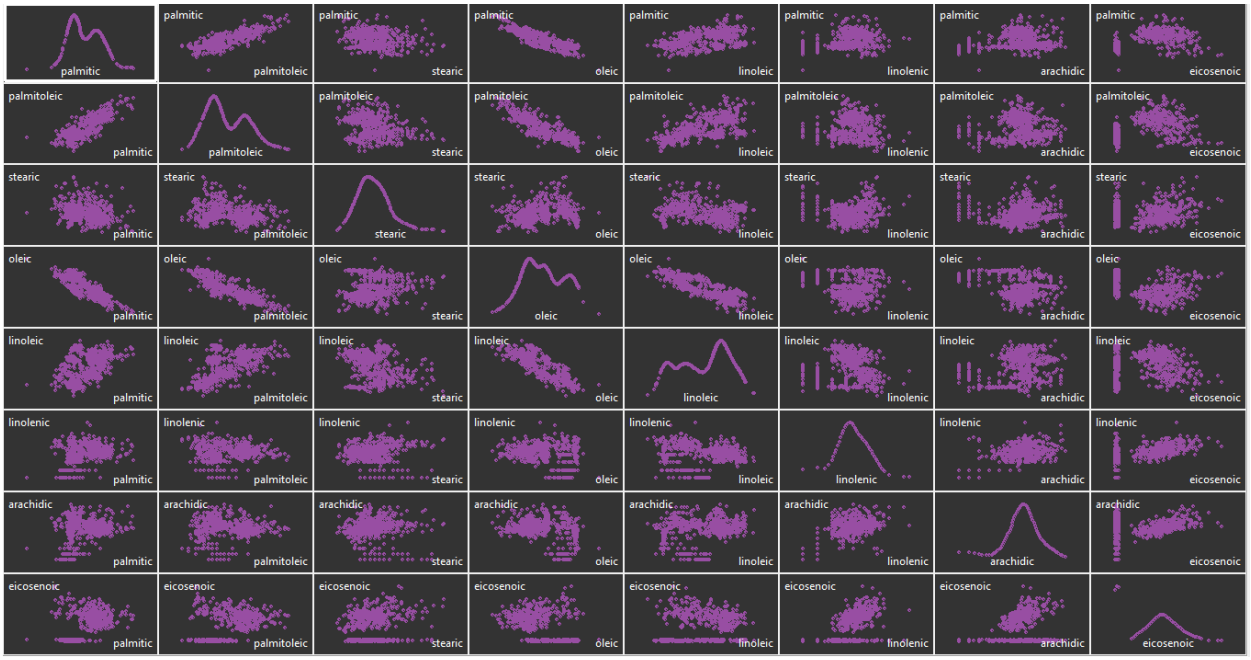
Assignment 1

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

We have 572 observations in the data set `olive.csv`.

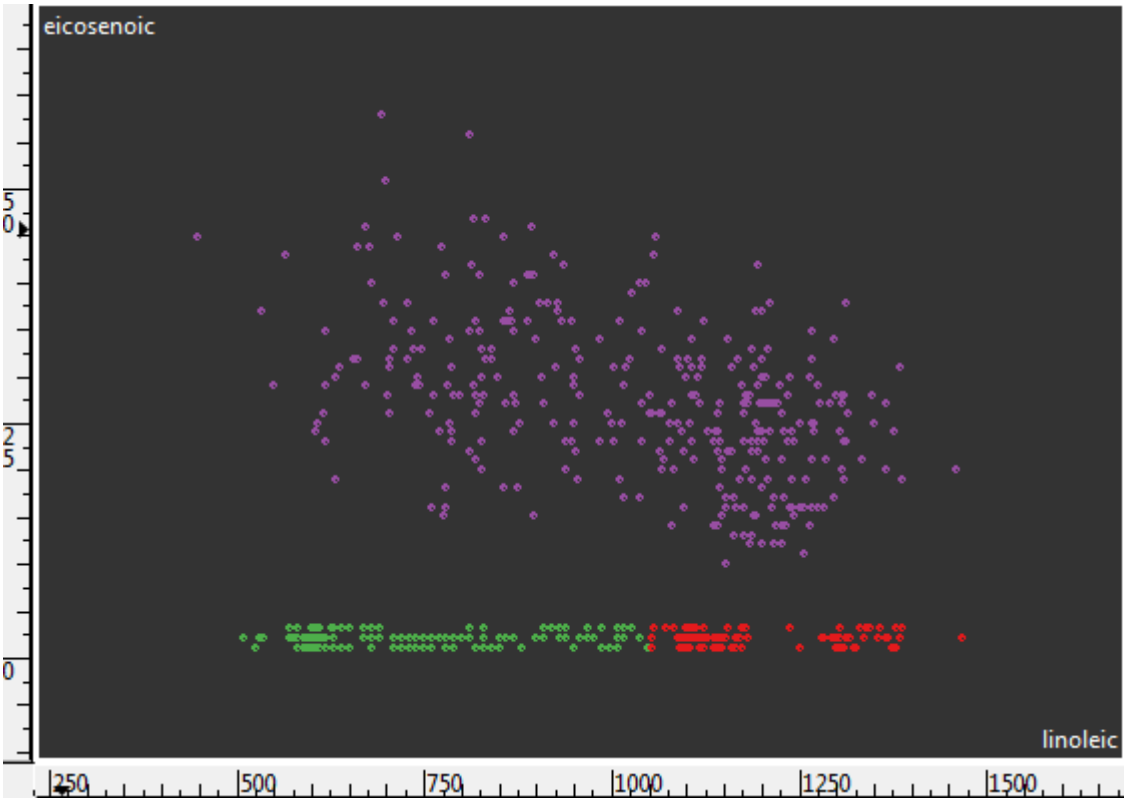
2. Create a scatter plot matrix that shows how the contents of different acids are related to each other. Investigate the matrix to find plots where the clusters are present. Close the plot.

Ok, a scatter matrix looks like this:



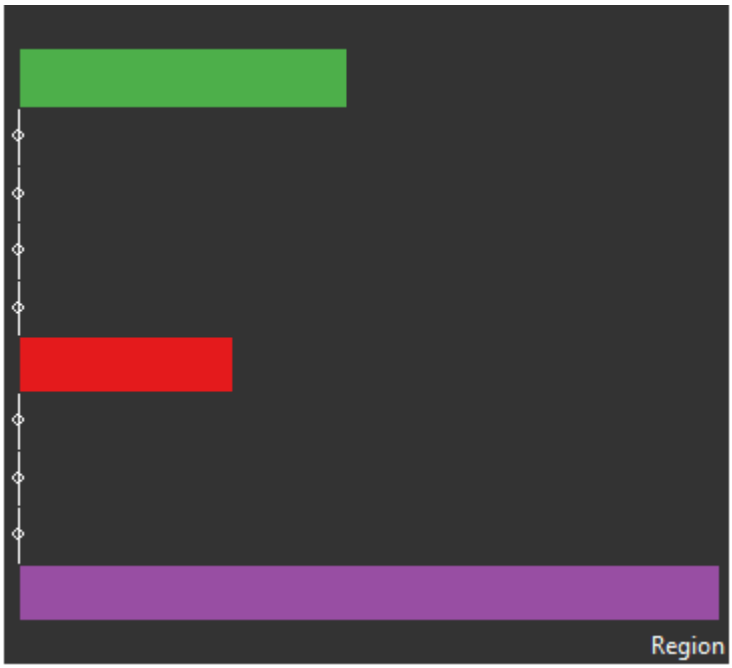
I think there are clearly some positively correlated acids, such as palmitic and palmitoleic, and there are clusters in the variables eicosenoic, linolenic and arachidic acids.

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.



It is fruitful to investigate this particular pair because of the clear partitioning of oils with very low eicosenoic acid content on one hand and all other oils on the other. The values of eicosenoic acids for the low eicosenoic acid clusters are either one, two or three.

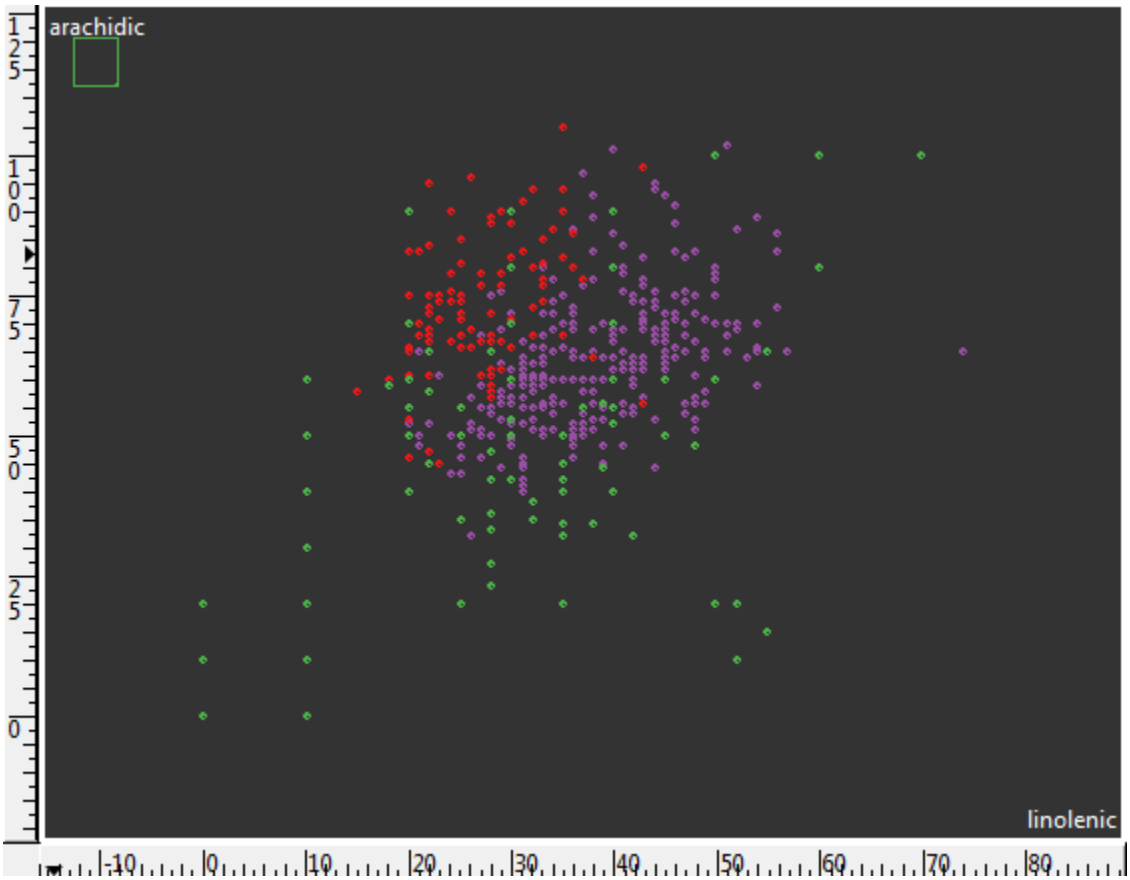
4. Create a histogram that shows how many observations fall within any given region. Use persistent brushing to identify the regions that correspond unusually low values of eicosenoic. Include the plots into your report and then remove the brushing (one way is to restart GGobi)



From the histogram, we see that over half of the observations belong to purple Region 1 (South Italy), while the fewest observations go to red Region 2 (Sardinia) and the rest of the observations go to green Region 3 (North Italy). We have already seen that the low values of eicosenoic acid correspond to Region 2 or Region 3.

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?

We have already seen the eicosenoic vs linoleic scatter plot. The arachidic vs linolenic looks like this:

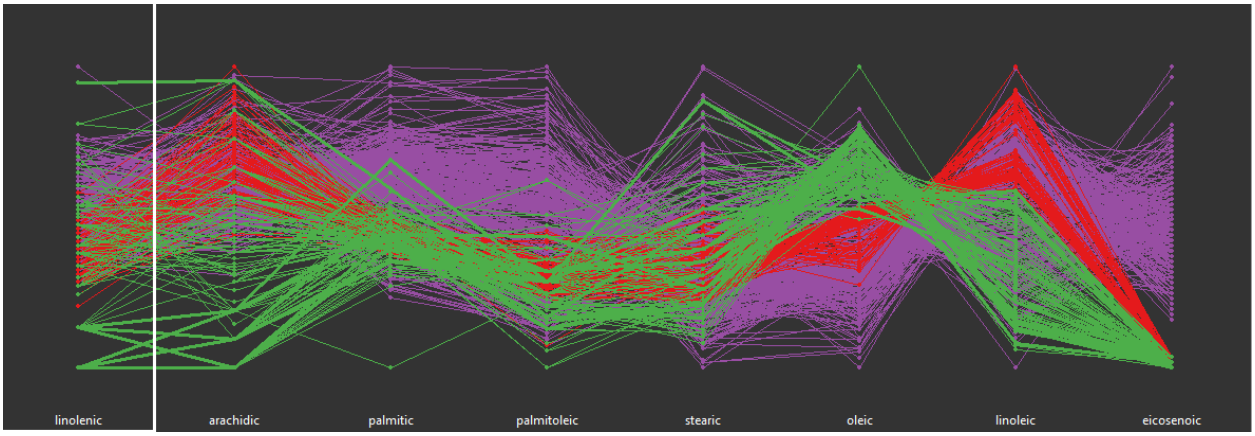


All the outliers from this plot are concentrated in the low eicosenoic acid cluster in the eicosenoic vs linoleic scatter plot.

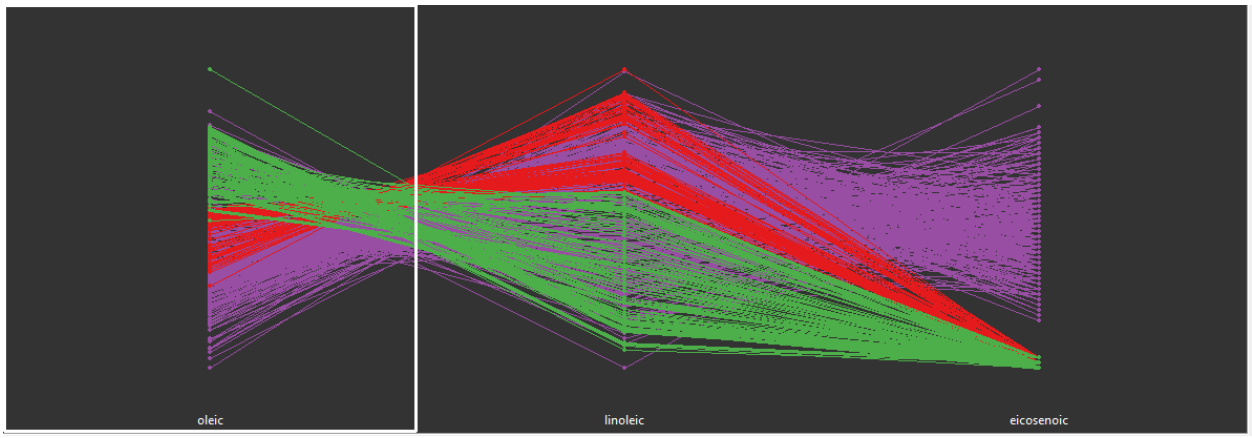
6. Use persistent brushing to paint by different colors the observations that fall into different regions. Keep these coloring during steps 7-9.

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)

A parallel coordinate plot for all eight acids look like this:

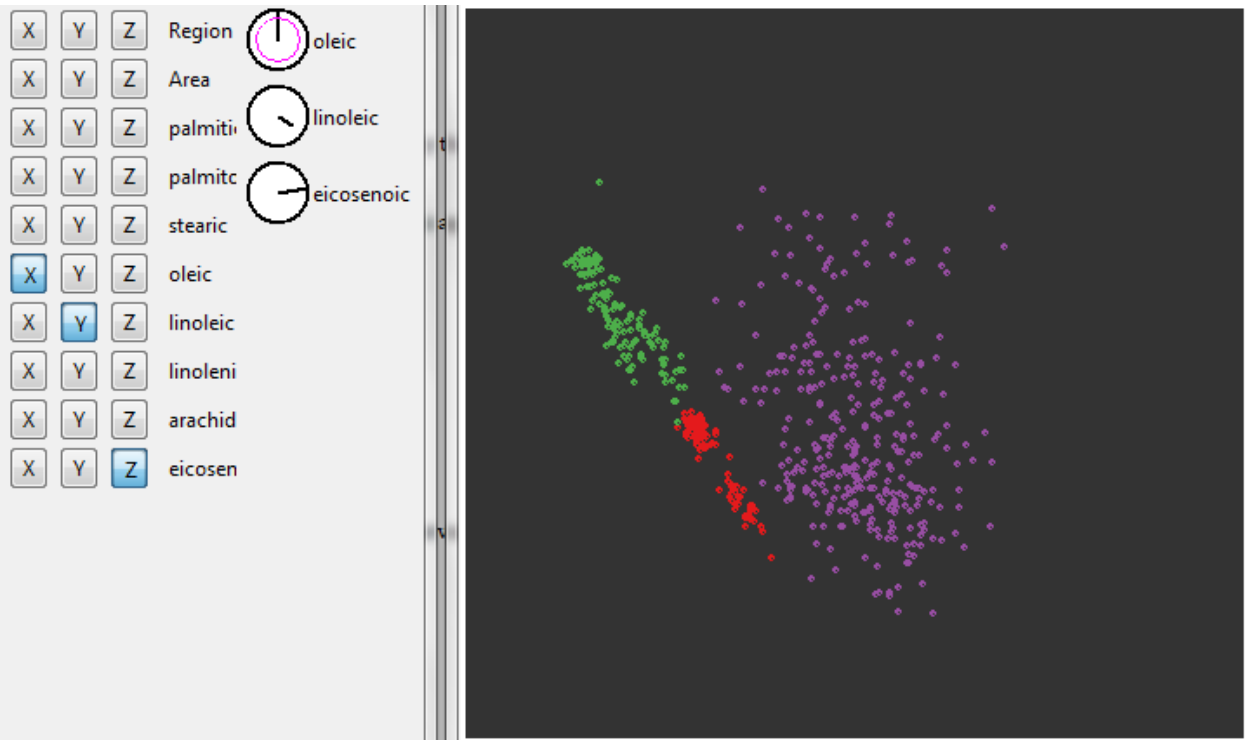


The acids I think are the best for separating the Regions from eachother are the last three, oleic, linoleic and eicosenoic acids.



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

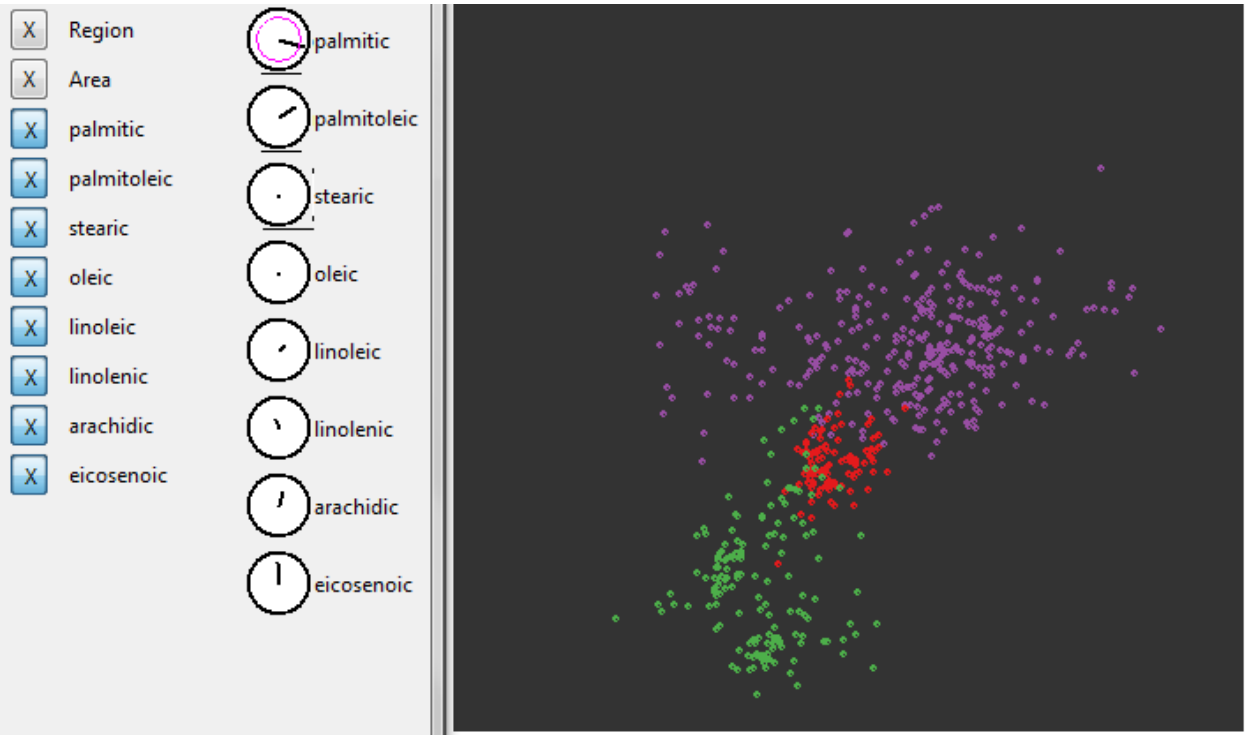
Here is the best snapshot I could make:



I do not think it is terribly informative compared to earlier plots.

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?

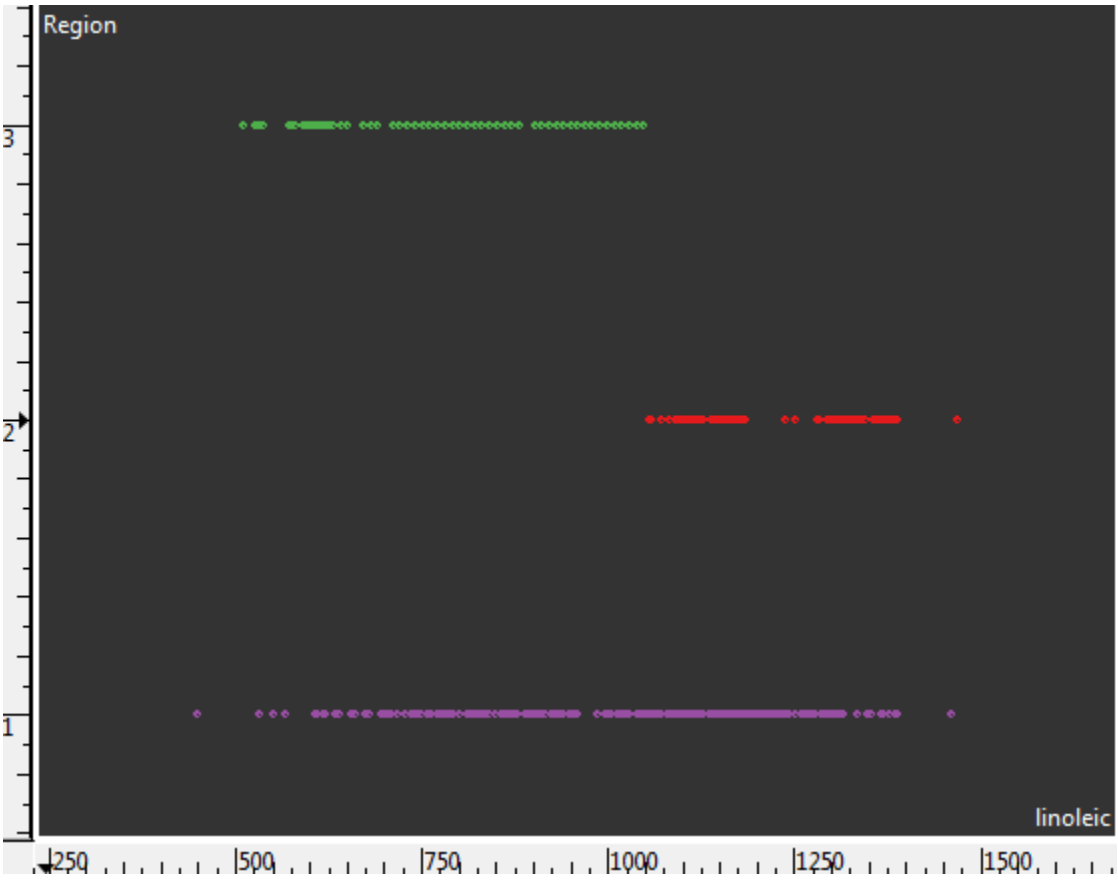
Here is the best snapshot I could make:



I do not think it is terribly informative compared to earlier plots.

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.

I would say that the best strategy is to first find out whether the oil has a eicosenoic acid level of one, two or three. If it does not, then it is probably from Region 1. For the oils that do , we ask how high are their linoleic acid levels.



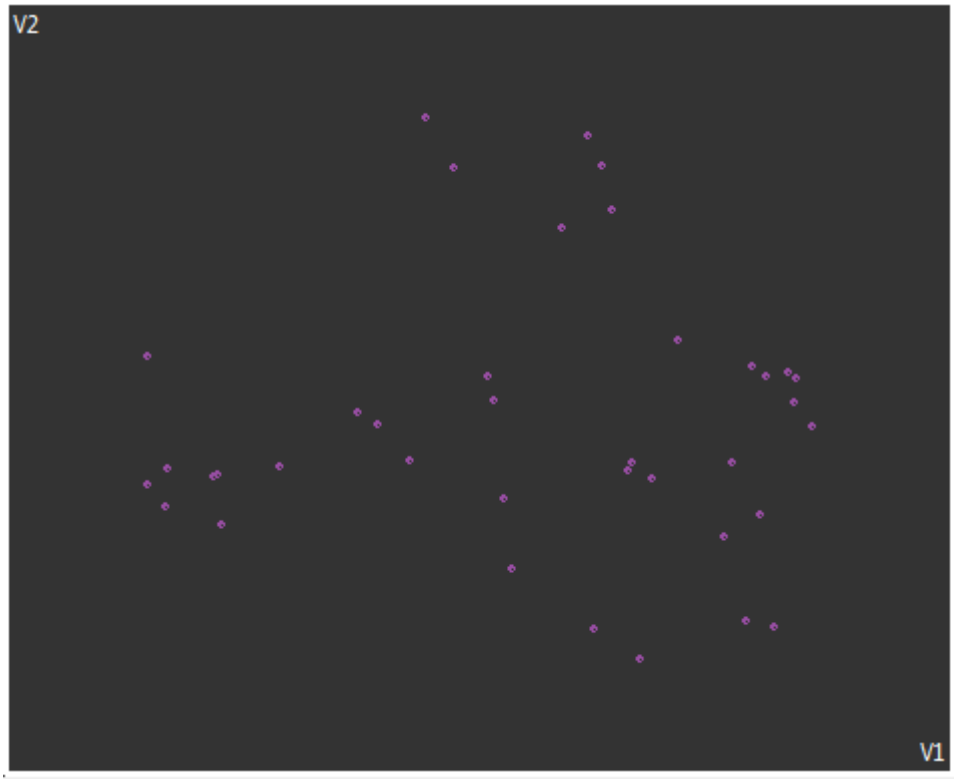
As you can see, the dividing point between oils from regions 2 and 3 is around linoleic acid level 1050. If oil contains lower linoleic acid levels, it is from Region 3, if level is higher then it is from Region 2.

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).

It seems the variables are measuring different things and they are measured in different units. If we do not want only the variable with the numerically largest variation to dominate the analysis we should scale the data.

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.



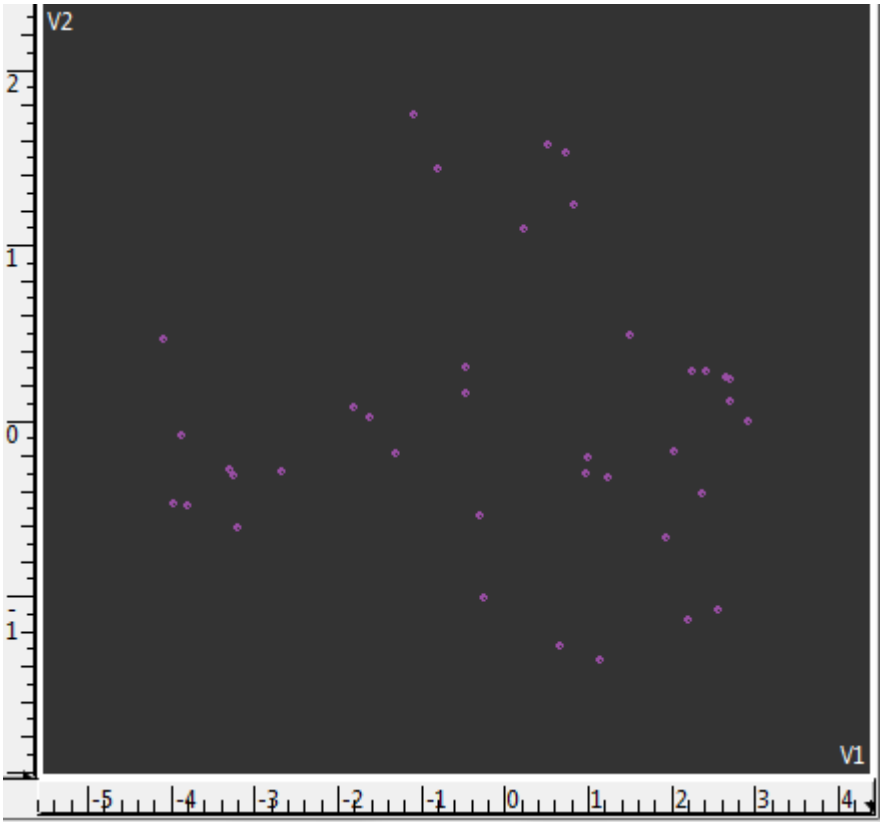
I can't see anything remarkable in this plot.

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

I feel this plot shows that the metric MDS was fairly successful, though a little less so at lower initial distances.

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

The non-metric MDS did converge and the final stress value is about 2.6%.



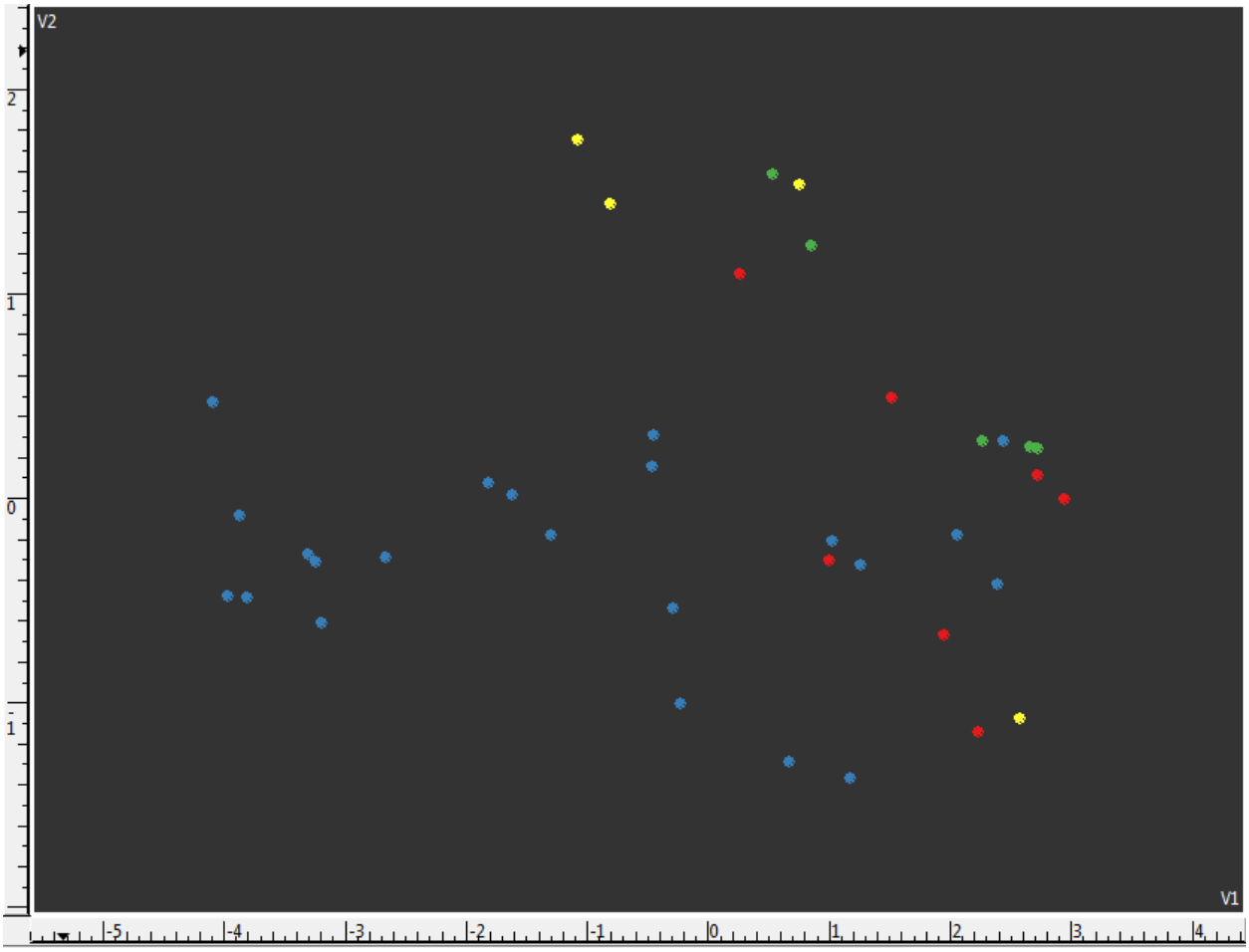
This plot looks almost the same as the metric MDS.

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

There is a slight, almost not noticable, improvement in this shepard plot compared to the shepard plot of the metric MDS.

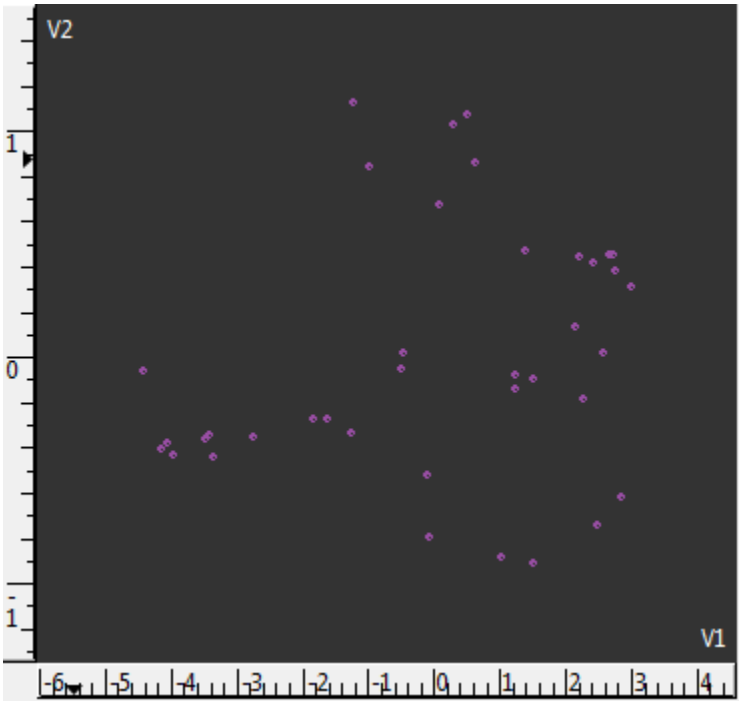
5. Brush the MDS plot by the value of the Country. Draw conclusions.

I have decided to color U.S. cars blue, Japanese cars red, German cars green and yellow signifies any other country.



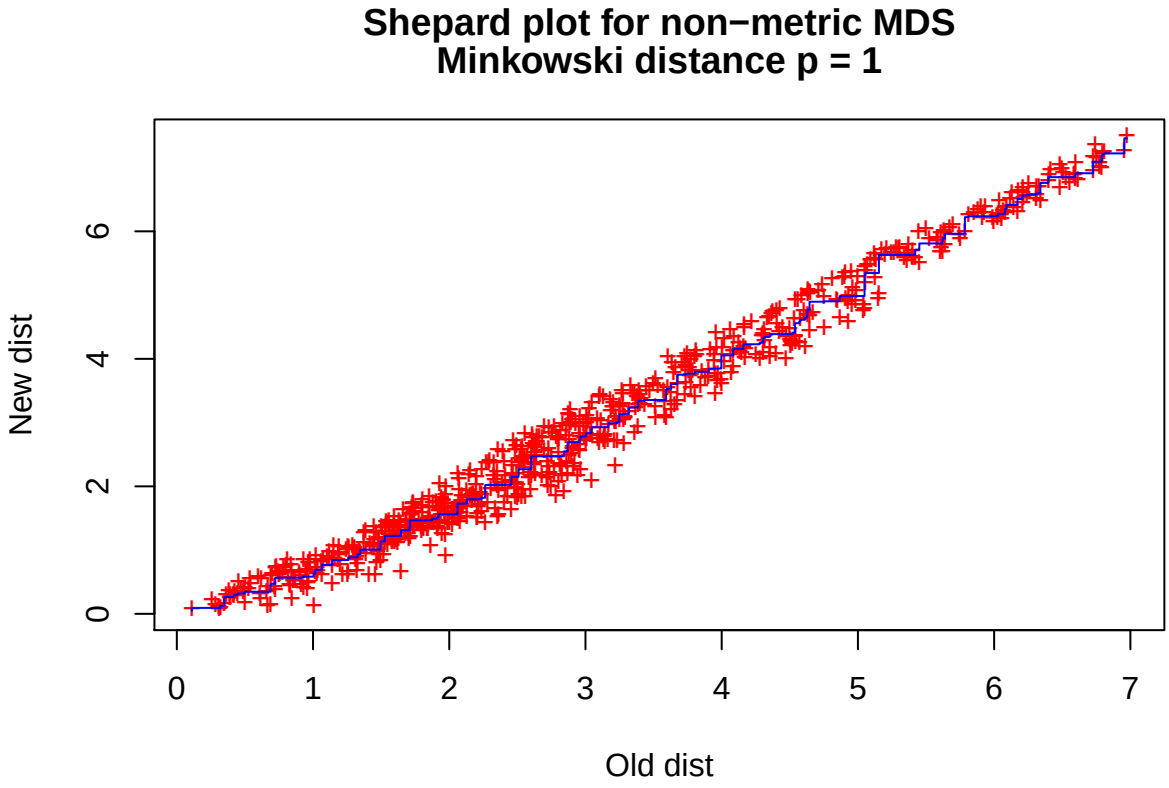
We see that the american cars tend to have quite negative values in the first variable, possibly this is an expression of the larger displacement of U.S. cars compared to other countries, but other than that I can not see any significance.

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.



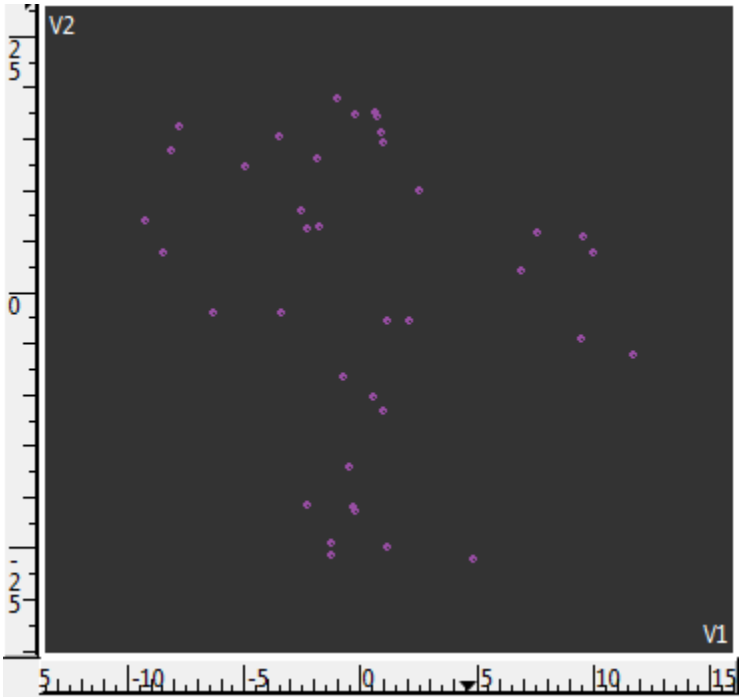
There appears to be very little change in the clustering compared to the two previous attempts. The stress value is worse compared to the first non-metric MDS.

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



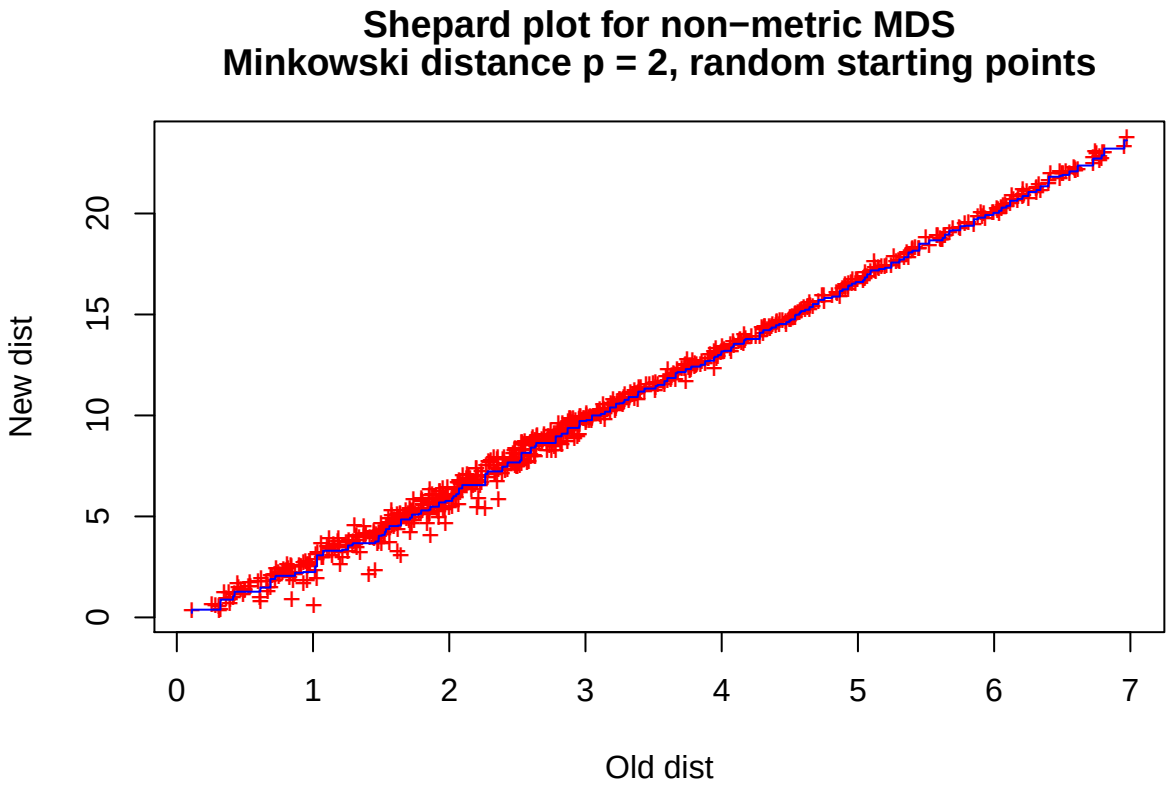
The points are a little more spread out in this Shepard plot.

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.



We see that, due to the random nature of the initial points, the axes have been rotated and mirrored. The overall clustering is very similar to the other cases, however.

```
## initial  value 41.734194
## iter   5 value 34.187351
## iter  10 value 13.929089
## iter  15 value 10.454569
## iter  20 value 9.716491
## iter  25 value 8.106118
## iter  30 value 3.871016
## iter  35 value 2.715610
## iter  40 value 2.605577
## iter  40 value 2.603820
## iter  40 value 2.603419
## final  value 2.603419
## converged
```



This Shepard plot looks fairly good, though I aught to mention that this MDS often fails to start properly. Stress value is ok but not as good as the first non-metric MDS, on average.

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

R code

```
library("MASS")
cars <- read.csv2("cars.csv")

carsN <- scale(cars[,3:8])
d <- dist(carsN)
new <- cmdscale(d, 2)
write.csv(new, file = "carsNew.csv")
plot(d, dist(new),pch ="+",col = "red",
     xlab=c("Old dist"),ylab = c("New dist"),
     main =c("Shepard plot for metric MDS"))
abline(a=0,b=1)
res <- isoMDS(d, k=2, p=2)
coords <- res$points
write.csv(coords, file = "carsNew2.csv")
plot(d,dist(coords),pch ="+",col = "red",
     xlab=c("Old dist"),ylab = c("New dist"),
     main =c("Shepard plot for non-metric MDS"))
sh <- Shepard(d,coords)
```

```r
lines(sh$x, sh$yf, type = "S", col = "blue")
res2 <- isoMDS(d, k=2, p=1)
coords2 <- res2$points
write.csv(coords2, file = "carsNew3.csv")
sh2 <- Shepard(d, coords2)
plot(d, dist(coords2),pch ="+",col = "red",
     xlab=c("Old dist"),ylab = c("New dist"),
     main =c("Shepard plot for non-metric MDS","Minkowski distance p = 1"))
lines(sh2$x, sh2$yf, type = "S", col = "blue")
init <- data.frame(x1 = runif(38, -1,1), x2 = runif(38, -1,1))
res3 <- isoMDS(d, k=2, p=2, maxit = 500, y = as.matrix(init))
coords3 <- res3$points
write.csv(coords3, file = "carsNew4.csv")
sh3 <- Shepard(d, coords3)
plot(d, dist(coords3),pch ="+",col = "red",
     xlab=c("Old dist"),ylab = c("New dist"),
     main =c("Shepard plot for non-metric MDS",
             "Minkowski distance p = 2, random starting points"))
lines(sh3$x, sh3$yf, type = "S", col = "blue")
## NA
```