# 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.

### 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.

### 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.

They take the value 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)

### 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. All the outliers has 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.

### 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 good for finding clusters. Linoleic and Palmitoleic can also be used for that purpose.

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

**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?**

**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.**

## Assignment 2

### 1.

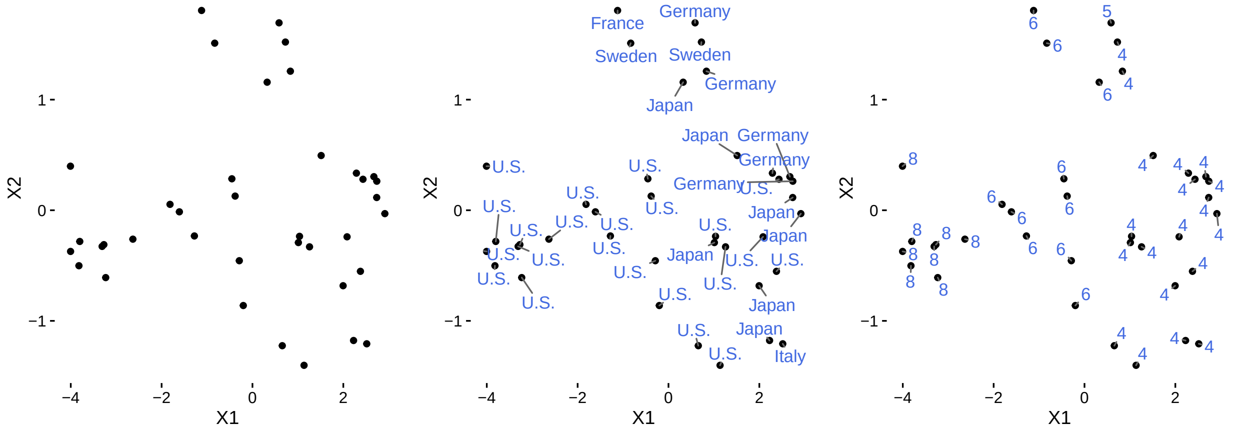
```
##   Country                        Car  MPG Weight Drive_Ratio Horsepower
## 1    U.S.         Buick Estate Wagon 16.9  4.360        2.73        155
## 2    U.S. Ford Country Squire Wagon 15.5  4.054        2.26        142
## 3    U.S.         Chevy Malibu Wagon 19.2  3.605        2.56        125
## 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
```

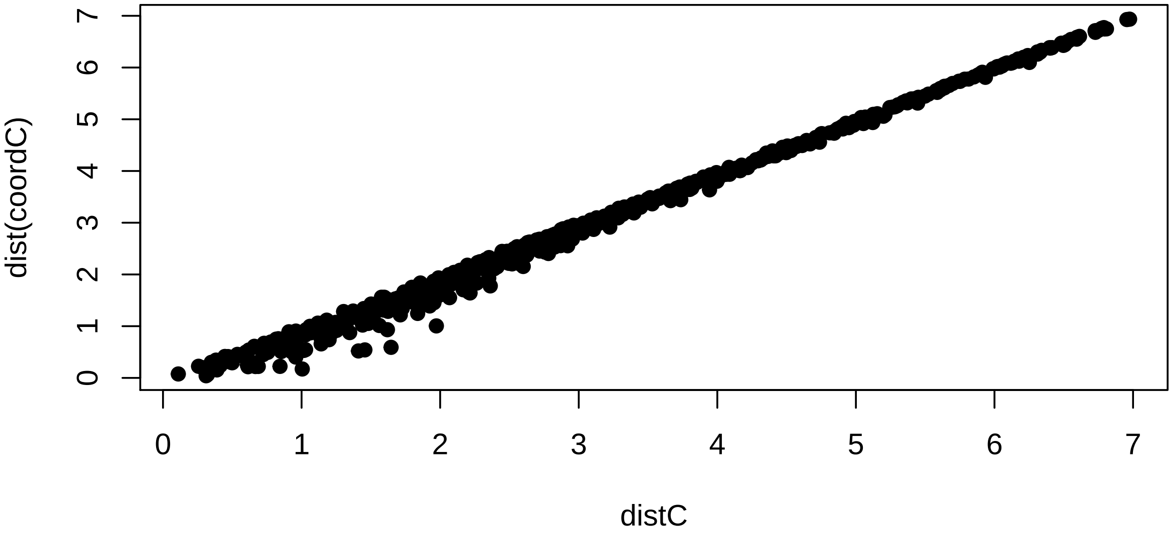
```
##          MPG       Weight  Drive_Ratio   Horsepower Displacement
##   42.8673186    0.4996658    0.2679691  699.3342817 7899.0761024
##    Cylinders
##    2.5697013
```

Data at different scales, so yes might be reasonable to scale this data set. Both the mean and the variance differs clearly so data is scaled.

**2.**

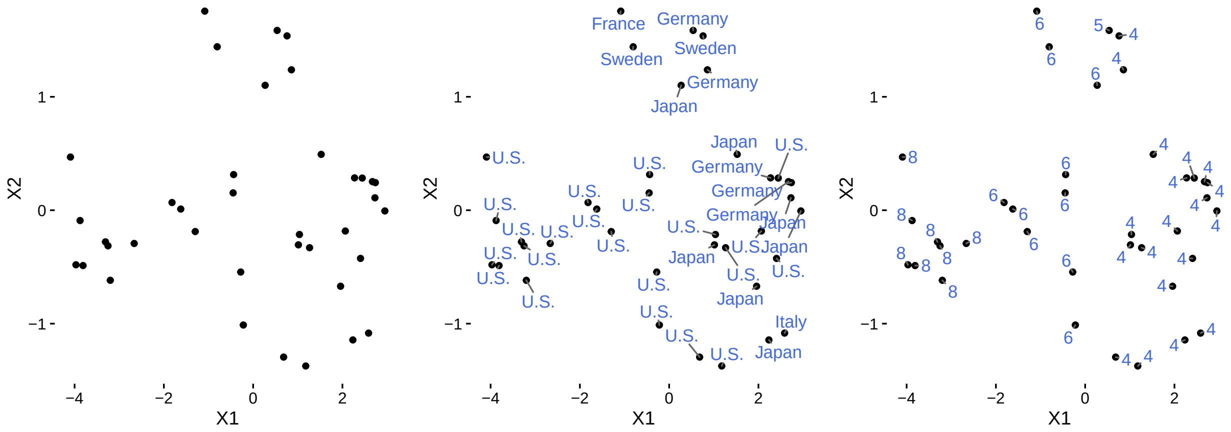


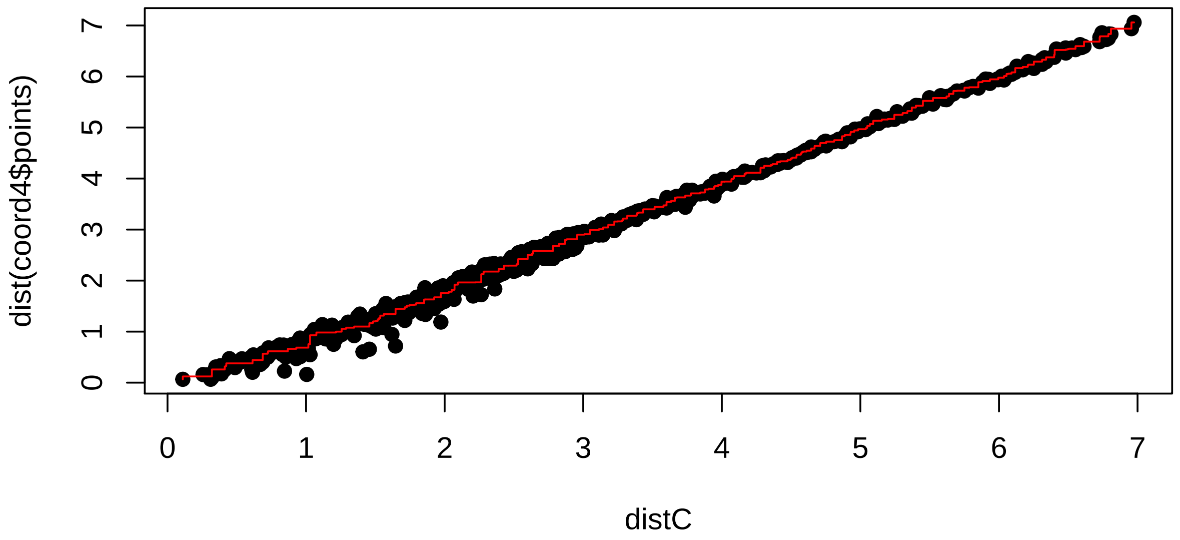
**3.**



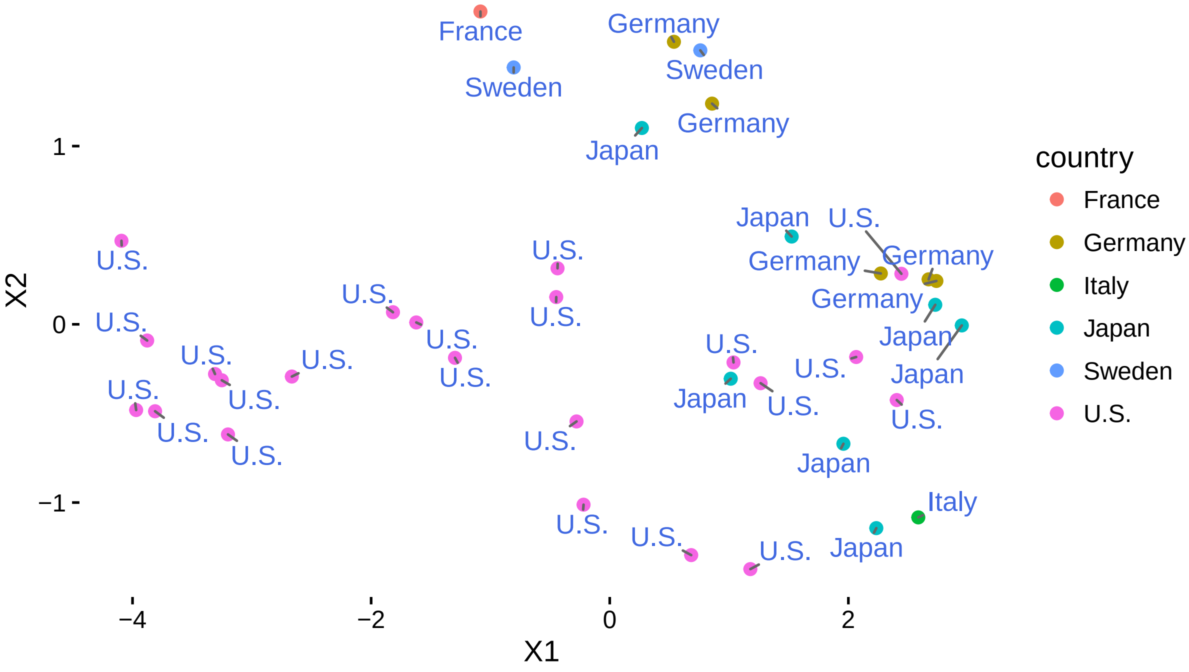
**4.**

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


X1
X2
−4
−2
0
2
−1
0
1
France
Germany
Sweden
Sweden
Germany
Japan
Japan
U.S.
Germany
Germany
Germany
Japan
Japan
U.S.Japan
U.S.
U.S.
Japan
Italy
U.S.
Japan
8
6
5
4

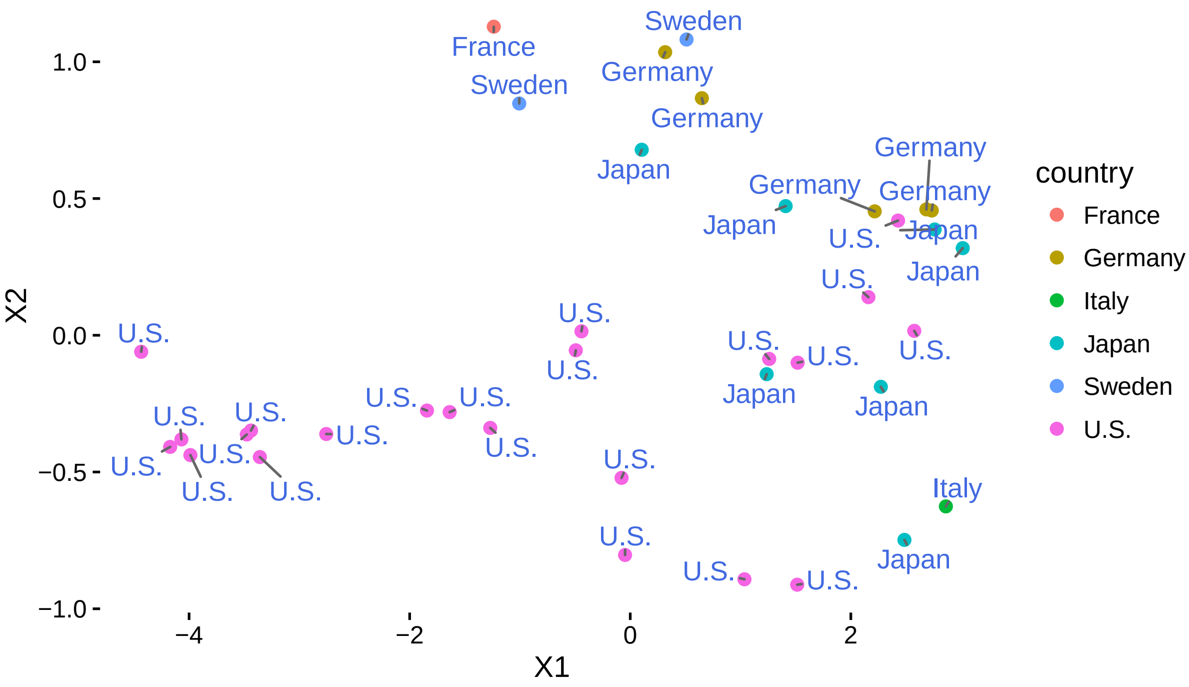

dist(coord4$points)
distC
0
1
2
3
4
5
6
7

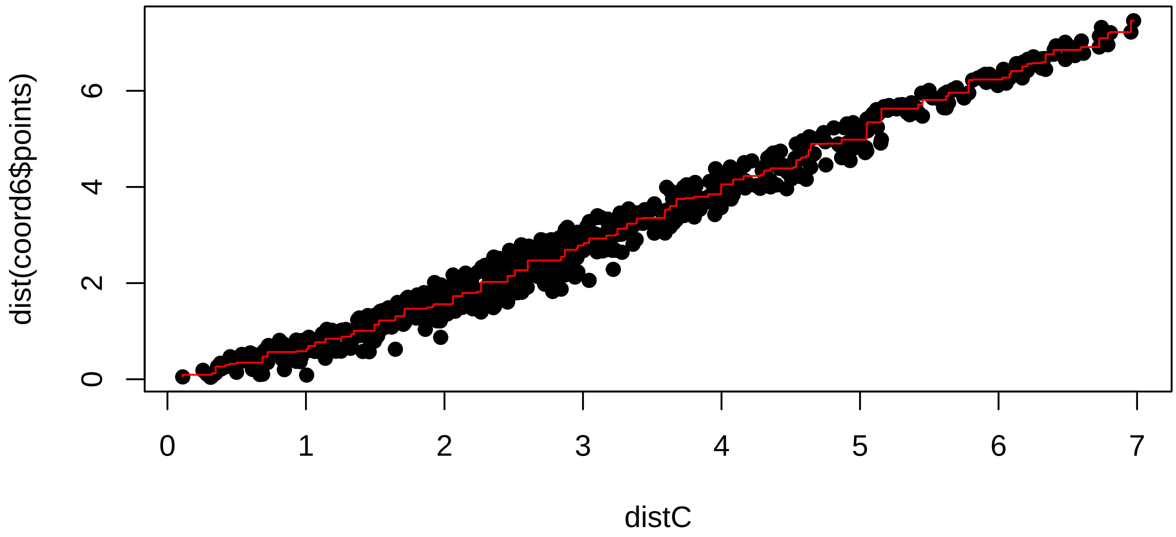
**5.**



**6.**

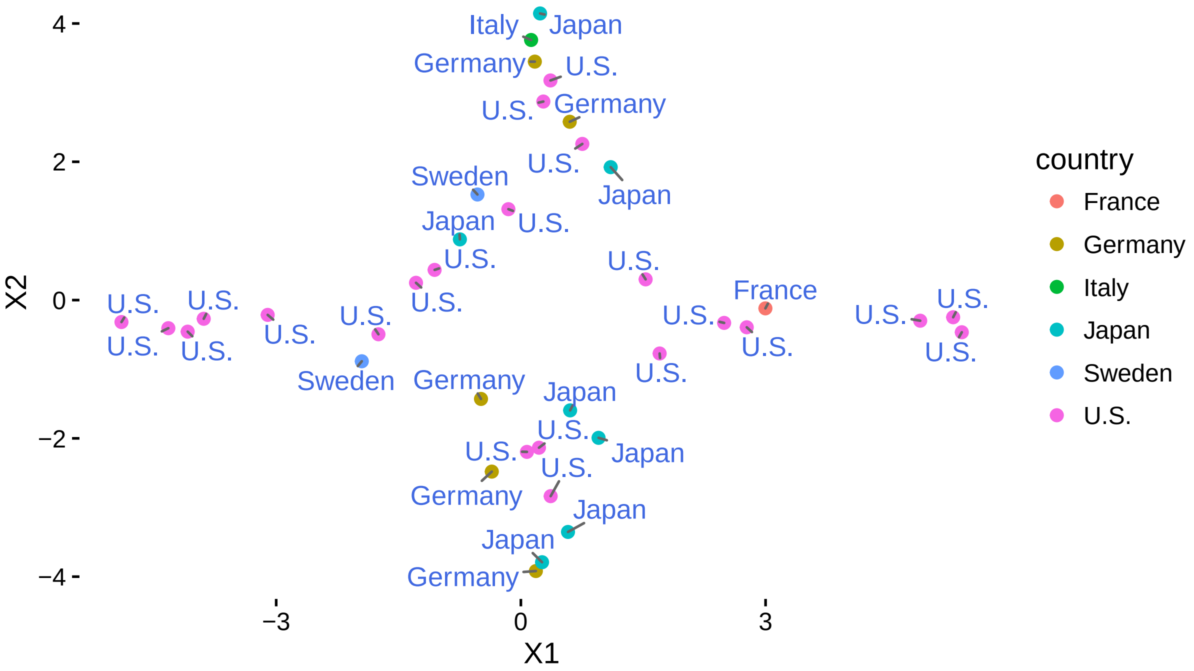
```
## initial  value 8.814428
## iter   5 value 5.022452
## iter  10 value 4.587726
## iter  15 value 4.488405
## final  value 4.472602
## converged
```

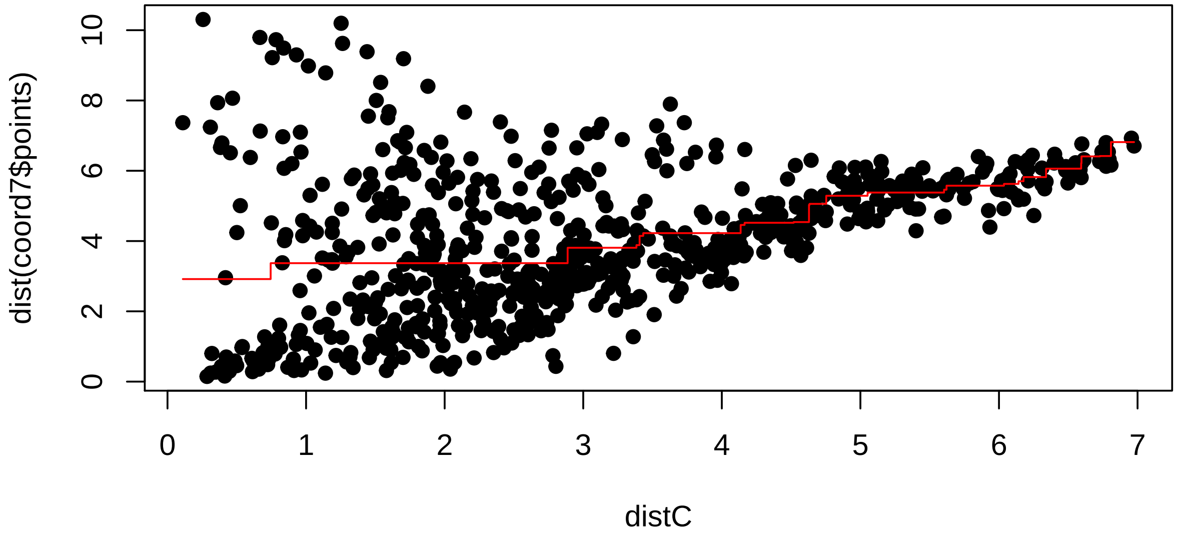




### 7.

```
## initial  value 44.921923
## iter   5 value 39.143612
## iter  10 value 38.002892
## iter  15 value 35.593049
## iter  20 value 34.216651
## iter  25 value 33.000010
## iter  30 value 32.590987
## final  value 32.407134
## converged
```





## 8.

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.