Multivariate Analysis of Oliveoil

Jolene Chen

Contents

Introduction

The dataset *oliveoil* contains 572 rows, each corresponding to a different specimen of olive oil, and 10 columns. The first and the second column correspond to the macro-area (`Centre-North`, `South`, `Sardinia`) and the region of origin of the olive oils, respectively. Columns 3-10 represent the following 8 chemical measurements on the acid components for the oil specimens: `palmitic`, `palmitoleic`, `stearic`, `oleic`, `linoleic`, `linolenic`, `arachidic`, `eicosenoic`. The data set *oliveoil* can be downloaded from:

https://ghuang.stat.nycu.edu.tw/course/multivariate24/files/exam/oliveoil.csv

The data set can also be downloaded from e3 under "midterm".

Import

```r
oil <- read.csv('oliveoil.csv', header = TRUE)
dim(oil)
```

```
## [1] 572  10
```

```r
head(oil)
```

```
##   macro.area       region palmitic palmitoleic stearic oleic linoleic linolenic
## 1      South Apulia.north     1075          75     226  7823      672        36
## 2      South Apulia.north     1088          73     224  7709      781        31
## 3      South Apulia.north      911          54     246  8113      549        31
## 4      South Apulia.north      966          57     240  7952      619        50
## 5      South Apulia.north     1051          67     259  7771      672        50
## 6      South Apulia.north      911          49     268  7924      678        51
##   arachidic eicosenoic
## 1        60         29
## 2        61         29
## 3        63         29
## 4        78         35
## 5        80         46
## 6        70         44
```

```r
oil[c(5,50,100,150,200,250,300,350,400,450,500,550),]
```

```
##       macro.area          region palmitic palmitoleic stearic oleic linoleic
## 5          South    Apulia.north     1051          67     259  7771      672
## 50         South        Calabria     1359          98     351  7262      780
## 100        South    Apulia.south     1286         163     183  7040     1230
## 150        South    Apulia.south     1330         157     228  7055     1108
## 200        South    Apulia.south     1487         246     251  6504     1390
## 250        South    Apulia.south     1434         185     189  6771     1269
## 300        South    Apulia.south     1620         255     166  6628     1212
## 350     Sardinia Sardinia.inland     1106          93     212  7381     1104
## 400     Sardinia Sardinia.inland     1131          78     221  7358     1120
## 450 Centre.North          Umbria     1070          75     188  7980      602
## 500 Centre.North    Liguria.east     1170         110     250  7620      740
## 550 Centre.North    Liguria.west     1040          90     250  7810      810
##     linolenic arachidic eicosenoic
## 5          50        80         46
## 50         41        56         16
```

```
## 100        29         57          12
## 150        42         55          25
## 200        29         53          19
## 250        30         62          25
## 300        29         62          27
## 350        35         68           1
## 400        22         69           2
## 450        22         45           2
## 500        20         90           1
## 550        10         10           2
```

```r
summary(oil)
```

```
##   macro.area           region             palmitic     palmitoleic   
##  Length:572         Length:572         Min.   : 610   Min.   : 15.00  
##  Class :character   Class :character   1st Qu.:1095   1st Qu.: 87.75  
##  Mode  :character   Mode  :character   Median :1201   Median :110.00  
##                                        Mean   :1232   Mean   :126.09  
##                                        3rd Qu.:1360   3rd Qu.:169.25  
##                                        Max.   :1753   Max.   :280.00  
##     stearic          oleic         linoleic        linolenic    
##  Min.   :152.0   Min.   :6300   Min.   : 448.0   Min.   : 0.00  
##  1st Qu.:205.0   1st Qu.:7000   1st Qu.: 770.8   1st Qu.:26.00  
##  Median :223.0   Median :7302   Median :1030.0   Median :33.00  
##  Mean   :228.9   Mean   :7312   Mean   : 980.5   Mean   :31.89  
##  3rd Qu.:249.0   3rd Qu.:7680   3rd Qu.:1180.8   3rd Qu.:40.25  
##  Max.   :375.0   Max.   :8410   Max.   :1470.0   Max.   :74.00  
##    arachidic       eicosenoic   
##  Min.   :  0.0   Min.   : 1.00  
##  1st Qu.: 50.0   1st Qu.: 2.00  
##  Median : 61.0   Median :17.00  
##  Mean   : 58.1   Mean   :16.28  
##  3rd Qu.: 70.0   3rd Qu.:28.00  
##  Max.   :105.0   Max.   :58.00
```

```r
table(oil[, 'macro.area'])
```

```
## 
## Centre.North     Sardinia        South 
##          151           98          323
```

Following, I will perform various multivariate analyses on this data set using the R software.

1. ANOVA

To examine the differences of the 8 acid chemical measurements on the acid components for the oil specimens across three macro-areas, one can do the multivariate mean inferences.

a. MANOVA

Use the one-way MANOVA to examine the overall acid chemical measurement differences among different macro-areas. The model is

$$X_{\ell j} = \mu + \tau_\ell + e_{\ell j},\ \ \ell = 1, 2, 3 \text{ (macro-area: Centre-North, South, Sardinia)},\ \ j = 1, \cdots, n_\ell$$

$$H_0 : \mu_{Centre.North} = \mu_{Sardinia} = \mu_{South} v.s. H_1 : Not\ H_0$$

```
fit <- manova(as.matrix(oil[, 3:10]) ~ as.factor(oil[, 1]), data = oil)
fit
```

```
## Call:
##    manova(as.matrix(oil[, 3:10]) ~ as.factor(oil[, 1]), data = oil)
##
## Terms:
##                 as.factor(oil[, 1]) Residuals
## palmitic                    7517535   8712198
## palmitoleic                  621548    951933
## stearic                        1273    769685
## oleic                      49648363  44385023
## linoleic                   15181903  18479382
## linolenic                     29982     66053
## arachidic                     94004    183120
## eicosenoic                    90444     22808
## Deg. of Freedom                   2       569
##
## Residual standard errors: 123.7393 40.90224 36.77904 279.2943 180.2136 10.77433 17.93958 6.331227
## Estimated effects may be unbalanced
```

```
res <- summary(fit); res
```

```
##                      Df Pillai approx F num Df den Df    Pr(>F)
## as.factor(oil[, 1])   2 1.5937   276.04     16   1126 < 2.2e-16 ***
## Residuals           569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res$SS
```

```
## $`as.factor(oil[, 1])`
##                palmitic  palmitoleic      stearic       oleic    linoleic
## palmitic      7517535.24  2154506.818  -11358.7365 -16313012.3   4413510.5
## palmitoleic   2154506.82   621547.557   -5516.9112  -4916126.6   1491309.0
## stearic        -11358.74    -5516.911    1273.3995    158440.7   -132434.2
## oleic       -16313012.29 -4916126.629  158440.7153  49648363.2 -22971632.2
## linoleic      4413510.54  1491308.999 -132434.1755 -22971632.2  15181902.6
## linolenic      466041.97   135673.062   -1874.3508  -1135935.4    390761.0
## arachidic      409500.04   134446.801  -10109.2121  -1899371.6   1190534.3
## eicosenoic     823641.31   235143.784    -739.1389  -1733475.8    432963.5
##                linolenic   arachidic   eicosenoic
## palmitic      466041.967   409500.04   823641.3147
## palmitoleic   135673.062   134446.80   235143.7841
## stearic        -1874.351   -10109.21    -739.1389
## oleic       -1135935.429 -1899371.59 -1733475.8370
## linoleic      390760.974  1190534.28   432963.5194
```

```
## linolenic      29981.812    34226.86    50590.0723
## arachidic      34226.860    94004.06    41048.1276
## eicosenoic     50590.072    41048.13    90443.6429
##
## $Residuals
##                  palmitic palmitoleic     stearic       oleic    linoleic
## palmitic      8712198.47  2068166.15 -591367.187 -16398164.0   6354230.6
## palmitoleic   2068166.15   951933.35 -239198.820  -5452822.8   3032723.5
## stearic       -591367.19  -239198.82  769685.235    808789.9   -875296.2
## oleic       -16398163.97 -5452822.78  808789.900  44385022.5 -24868011.8
## linoleic      6354230.60  3032723.49 -875296.171 -24868011.8  18479382.0
## linolenic      -67379.53   -99478.02    7021.735    480314.3   -494033.2
## arachidic       74669.45   -78000.09   -8832.249    266028.7   -546173.8
## eicosenoic    -143122.66   -59386.98   42218.812    349340.4   -259104.5
##                linolenic   arachidic  eicosenoic
## palmitic      -67379.527   74669.446 -143122.657
## palmitoleic   -99478.020  -78000.087  -59386.983
## stearic         7021.735   -8832.249   42218.812
## oleic         480314.317  266028.691  349340.368
## linoleic     -494033.183 -546173.842 -259104.523
## linolenic      66053.027   66956.405    9721.942
## arachidic      66956.405  183120.455   17177.110
## eicosenoic      9721.942   17177.110   22808.041
```

For the MANOVA table for comparing population mean vectors, the treatment sum of squares and cross products B is `res$SS$as.factor(macro.area)` with df $= 2$, the residual sum of squares and cross products W is `res$SS$Residuals` with df $= 569$, and the total sum of squares and cross products is $B + W$ with df $= 571$.

- p-value $<$ 2.2e-16, Reject H_0 at $= 0.05$ level.

  At a significance level of 0.05, there is 95% confidence that the overall acid chemical measurement differences among different macro-areas are significant.

b. One-way ANOVA

One-way ANOVA on each each acid measurement (8 variables in total) for its differences over macro-areas Since we need to perform the test for multiple measurements simultaneously in the ANOVA analysis, then according to Bonferroni, we set the cut-off for the p-value $< \frac{0.05}{\text{number of variables}}$ to be significant.

```
alpha <- (0.05/8)
cat('Cutoff for the p-value:', alpha)
```

```
## Cutoff for the p-value: 0.00625
```

Which acid measurement(s) are significantly different over macro-areas?

```
p <- rep(0, 8)
sig <- rep(T, 8)

for (i in 1:8){
  cat('#####################   ANOVA for', colnames(oil)[i+2], '   #####################\n')
  results <- summary(aov(oil[, i+2] ~ as.factor(oil[, 1])), data = oil)
  print(results)
  cat('\n')
```

```r
  p[i] <- results[[1]][["Pr(>F)"]][1]
  sig[i] <- p[i] < alpha
}
```

```
## #####################   ANOVA for palmitic    #####################
##                      Df  Sum Sq Mean Sq F value Pr(>F)    
## as.factor(oil[, 1])   2 7517535 3758768   245.5 <2e-16 ***
## Residuals           569 8712198   15311                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## #####################   ANOVA for palmitoleic    #####################
##                      Df Sum Sq Mean Sq F value Pr(>F)    
## as.factor(oil[, 1])   2 621548  310774   185.8 <2e-16 ***
## Residuals           569 951933    1673                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## #####################   ANOVA for stearic    #####################
##                      Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(oil[, 1])   2   1273   636.7   0.471  0.625
## Residuals           569 769685  1352.7               
## 
## #####################   ANOVA for oleic    #####################
##                      Df   Sum Sq  Mean Sq F value Pr(>F)    
## as.factor(oil[, 1])   2 49648363 24824182   318.2 <2e-16 ***
## Residuals           569 44385023    78005                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## #####################   ANOVA for linoleic    #####################
##                      Df   Sum Sq Mean Sq F value Pr(>F)    
## as.factor(oil[, 1])   2 15181903 7590951   233.7 <2e-16 ***
## Residuals           569 18479382   32477                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## #####################   ANOVA for linolenic    #####################
##                      Df Sum Sq Mean Sq F value Pr(>F)    
## as.factor(oil[, 1])   2  29982   14991   129.1 <2e-16 ***
## Residuals           569  66053     116                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## #####################   ANOVA for arachidic    #####################
##                      Df Sum Sq Mean Sq F value Pr(>F)    
## as.factor(oil[, 1])   2  94004   47002     146 <2e-16 ***
## Residuals           569 183120     322                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## #####################   ANOVA for eicosenoic    #####################
##                      Df Sum Sq Mean Sq F value Pr(>F)
```

```
## as.factor(oil[, 1])   2  90444   45222    1128 <2e-16 ***
## Residuals           569  22808      40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## There are 7 of acid measurement with significant differences.
```

```
## The acid measurement significantly different over macro-areas:
```

```
##
##  - palmitic
##  - palmitoleic
##  - oleic
##  - linoleic
##  - linolenic
##  - arachidic
##  - eicosenoic
```

- Based on above results, only stearic is not significantly different over macro-areas.

  At a significance level of 0.00625, there is confidence that the other 7 acid chemicals show significant differences among the different macro-areas.

Following, we will perform the principal component analysis (PCA), the orthogonal factor analysis (FA) with a proper factor rotation, and the multidimensional scaling (MDS).

2. Principal component analysis

a. PCA

a.1 PCA (original variables)

```
## Eigenvalues:
```

$$\begin{cases} \widehat{\lambda_1} = 230543.82788 \\ \widehat{\lambda_2} = 22789.01058 \\ \widehat{\lambda_3} = 2064.26492 \\ \widehat{\lambda_4} = 758.82269 \\ \widehat{\lambda_5} = 615.20792 \\ \widehat{\lambda_6} = 143.52118 \\ \widehat{\lambda_7} = 51.05564 \\ \widehat{\lambda_8} = 48.74556 \end{cases}$$

```r
cat('Eigenvectors: \n'); eigenvector_s
```

```
## Eigenvectors:
```

```
##                [,1]         [,2]        [,3]        [,4]       [,5]       [,6]
## [1,] -0.284167992  0.637208452 -0.45062836  0.03857271 -0.4524509  0.1462473
## [2,] -0.092012578  0.094554974 -0.16460885  0.57386935  0.6690989  0.3254595
## [3,]  0.011151773  0.014774824  0.72398889  0.39748727 -0.4050560  0.2542609
## [4,]  0.842808624 -0.168763310 -0.33652056  0.09246819 -0.1991783  0.1406756
## [5,] -0.447210266 -0.743751915 -0.30400153  0.06643083 -0.2472581  0.1066613
## [6,] -0.004751237  0.034724051  0.08433954 -0.29315488  0.1110979 -0.2156815
## [7,] -0.013770009  0.009109222  0.14165474 -0.63629076  0.1923104  0.6648742
## [8,] -0.011058482  0.043240557  0.11329544 -0.08614438  0.1827288 -0.5369327
##              [,7]        [,8]
## [1,]  0.2384073 -0.16038084
## [2,]  0.1561341 -0.21948984
## [3,]  0.2132920 -0.20805981
## [4,]  0.2262900 -0.16949136
## [5,]  0.2203788 -0.17007665
## [6,] -0.1583158 -0.90652832
## [7,]  0.3078877  0.03097674
## [8,]  0.8084912  0.04905891
```

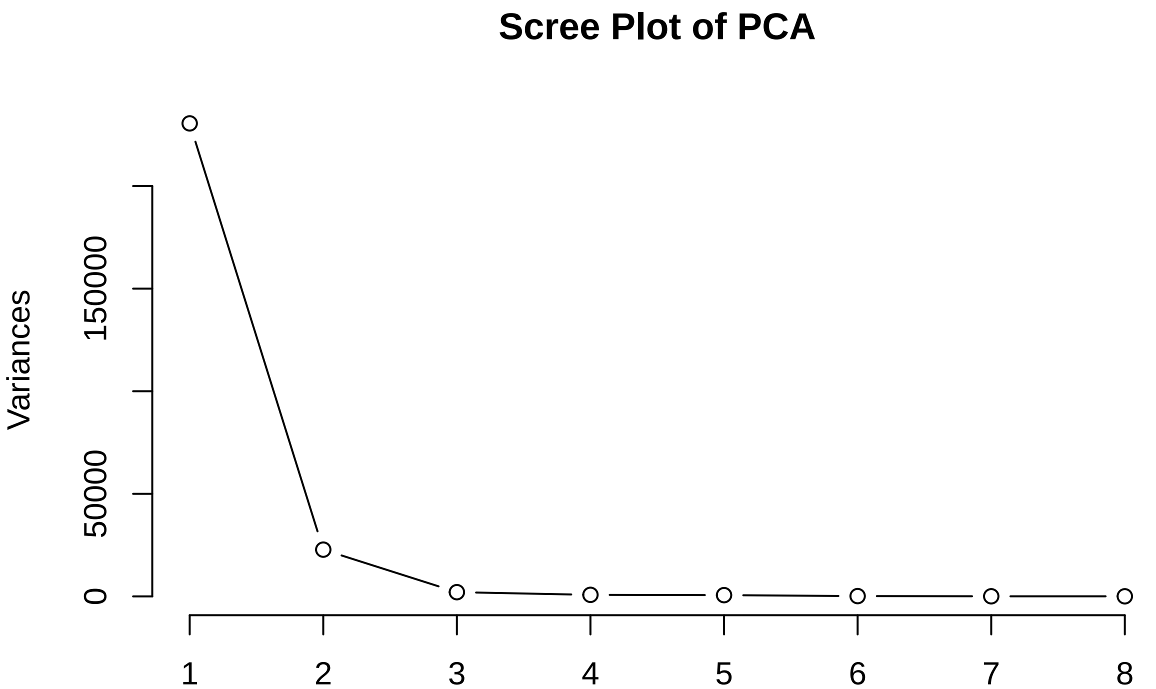
```r
pca = prcomp(oil[, 3:10], center = T)
pca.data = data.frame(pca$x)
pca.variance = pca$sdev^2 / sum(pca$sdev^2)

summary(pca)
```

```
## Importance of components:
##                            PC1       PC2      PC3      PC4      PC5      PC6
## Standard deviation     480.150 150.96029 45.43418 27.54674 24.80339 11.98003
## Proportion of Variance   0.897   0.08867  0.00803  0.00295  0.00239  0.00056
## Cumulative Proportion    0.897   0.98568  0.99371  0.99666  0.99905  0.99961
##                           PC7     PC8
## Standard deviation     7.1453 6.98180
## Proportion of Variance 0.0002 0.00019
## Cumulative Proportion  0.9998 1.00000
```

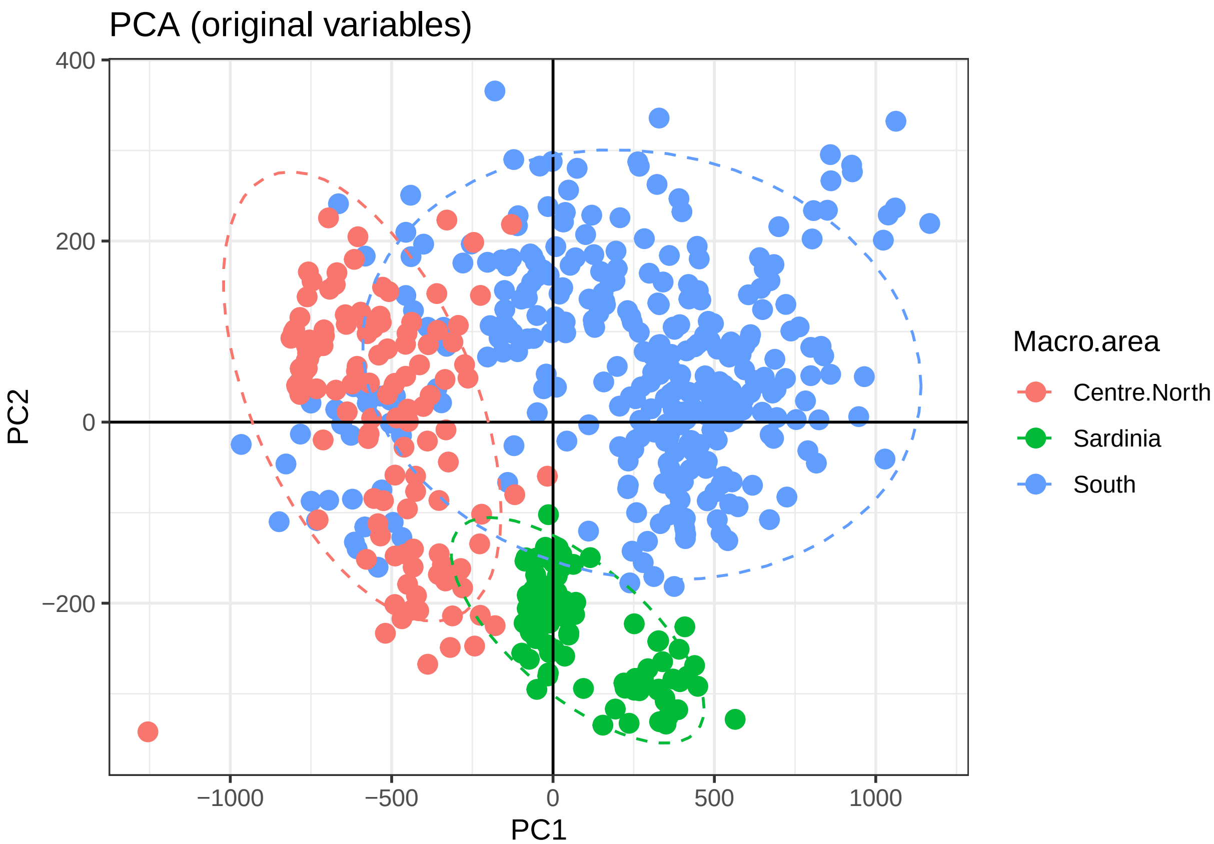
```r
print(pca$rotation)
```

```
##                     PC1          PC2         PC3         PC4        PC5
## palmitic     0.284167992  0.637208452 -0.45062836 -0.03857271 -0.4524509
## palmitoleic  0.092012578  0.094554974 -0.16460885 -0.57386935  0.6690989
## stearic     -0.011151773  0.014774824  0.72398889 -0.39748727 -0.4050560
## oleic       -0.842808624 -0.168763310 -0.33652056 -0.09246819 -0.1991783
## linoleic     0.447210266 -0.743751915 -0.30400153 -0.06643083 -0.2472581
## linolenic    0.004751237  0.034724051  0.08433954  0.29315488  0.1110979
## arachidic    0.013770009  0.009109222  0.14165474  0.63629076  0.1923104
## eicosenoic   0.011058482  0.043240557  0.11329544  0.08614438  0.1827288
##                    PC6        PC7         PC8
## palmitic    -0.1462473  0.2384073  0.16038084
## palmitoleic -0.3254595  0.1561341  0.21948984
## stearic     -0.2542609  0.2132920  0.20805981
## oleic       -0.1406756  0.2262900  0.16949136
## linoleic    -0.1066613  0.2203788  0.17007665
## linolenic    0.2156815 -0.1583158  0.90652832
## arachidic   -0.6648742  0.3078877 -0.03097674
## eicosenoic   0.5369327  0.8084912 -0.04905891
```

```r
library(ggplot2)
screeplot(pca, type = 'lines', main = 'Scree Plot of PCA')
```




```r
ggplot(pca.data, aes(x = PC1, y = PC2, color = oil[, 1])) +
    geom_point(size = 3) +
    geom_hline(yintercept = 0) +
    geom_vline(xintercept = 0) +
    stat_ellipse(aes(x = PC1, y = PC2), linetype = 2, linewidth = 0.5, level = 0.95) +
    guides(colour = guide_legend("Macro.area")) +
    ggtitle('PCA (original variables)') +
    theme_bw()
```



According to the results of the Importance of Components and the Scree Plot, the first principal component can explain 89.7% of the total variance.

- $\hat{\lambda}_1 = 230543.82788$

  The first sample principal component is:

$$\hat{y}_1 = 0.2842 \times \text{palmitic} + \cdots + 0.0111 \times \text{eicosenoic}$$

- `oleic` plays the main role in the first principal.

a.2 PCA using the correlation matrix (standardized variables)

```
## Eigenvalues:
```

$$\begin{cases} \widehat{\lambda_1} = 3.7214 \\ \widehat{\lambda_2} = 1.7658 \\ \widehat{\lambda_3} = 1.0163 \\ \widehat{\lambda_4} = 0.7929 \\ \widehat{\lambda_5} = 0.3338 \\ \widehat{\lambda_6} = 0.2488 \\ \widehat{\lambda_7} = 0.1188 \\ \widehat{\lambda_8} = 0.0021 \end{cases}$$

```r
cat('Eigenvectors: \n'); eigenvector_r
```

```
## Eigenvectors:
```

```
##             [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
## [1,] -0.46074351 -0.04958406  0.11445834 -0.28043124  0.53473943 -0.07699892
## [2,] -0.45022576 -0.24090732  0.14260264 -0.21182252  0.13841908 -0.16728954
## [3,]  0.09864471  0.25837844  0.80215910  0.47082168  0.21340068  0.03064009
## [4,]  0.49417494  0.15866175 -0.08011486 -0.20010742 -0.01552215 -0.11309403
## [5,] -0.36569539 -0.34339930 -0.08747773  0.51249093 -0.40127538  0.30497855
## [6,] -0.21898707  0.60483760 -0.19103316 -0.09881321  0.12507081  0.69784174
## [7,] -0.22830362  0.44719396 -0.42664494  0.48165441  0.14659527 -0.55365142
## [8,] -0.31186781  0.40476916  0.30085585 -0.33222211 -0.67153429 -0.25657629
##             [,7]        [,8]
## [1,]  0.52540418 -0.35438653
## [2,] -0.78680816 -0.08856309
## [3,] -0.07722664 -0.07703841
## [4,] -0.18074878 -0.79903372
## [5,]  0.07768793 -0.46687817
## [6,] -0.19096065 -0.02943890
## [7,] -0.06527504 -0.03996552
## [8,]  0.13959613 -0.04168750
```

```r
pca_std = prcomp(oil[, 3:10], center = T, scale = TRUE)
pca_std.data = data.frame(pca_std$x)
pca_std.variance = pca_std$sdev^2 / sum(pca_std$sdev^2)

summary(pca_std)
```

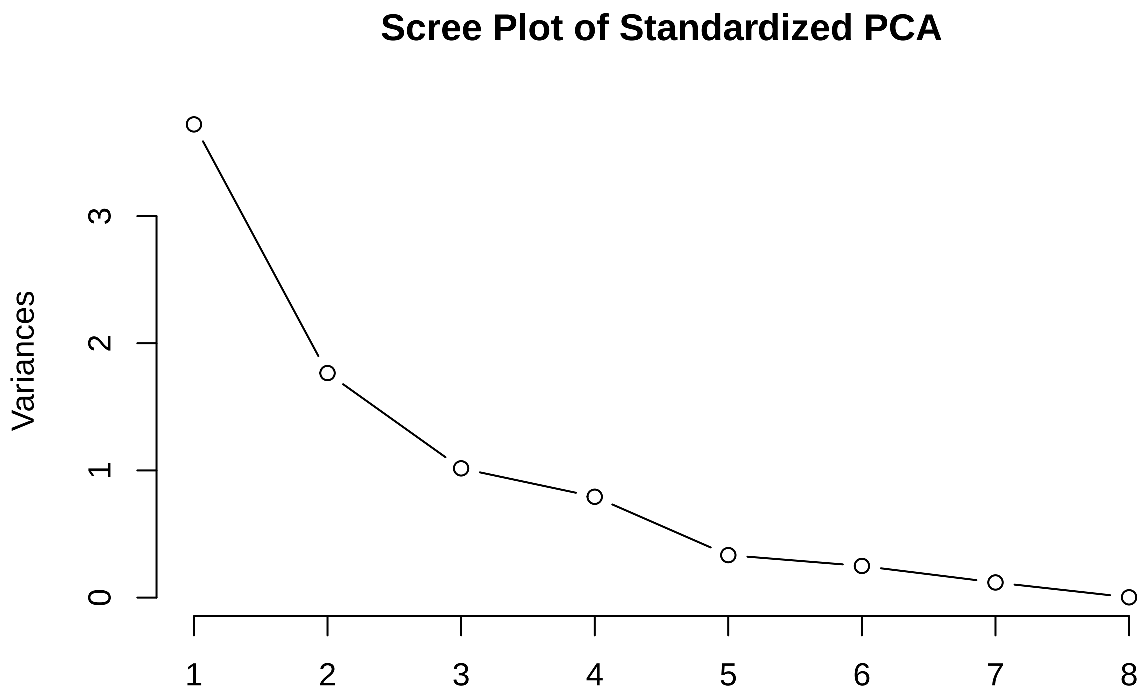
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5    PC6     PC7
## Standard deviation     1.9291 1.3288 1.0081 0.89045 0.57777 0.4988 0.34470
## Proportion of Variance 0.4652 0.2207 0.1270 0.09911 0.04173 0.0311 0.01485
## Cumulative Proportion  0.4652 0.6859 0.8129 0.91206 0.95378 0.9849 0.99974
##                            PC8
## Standard deviation     0.04563
## Proportion of Variance 0.00026
## Cumulative Proportion  1.00000
```

```r
print(pca_std$rotation)
```

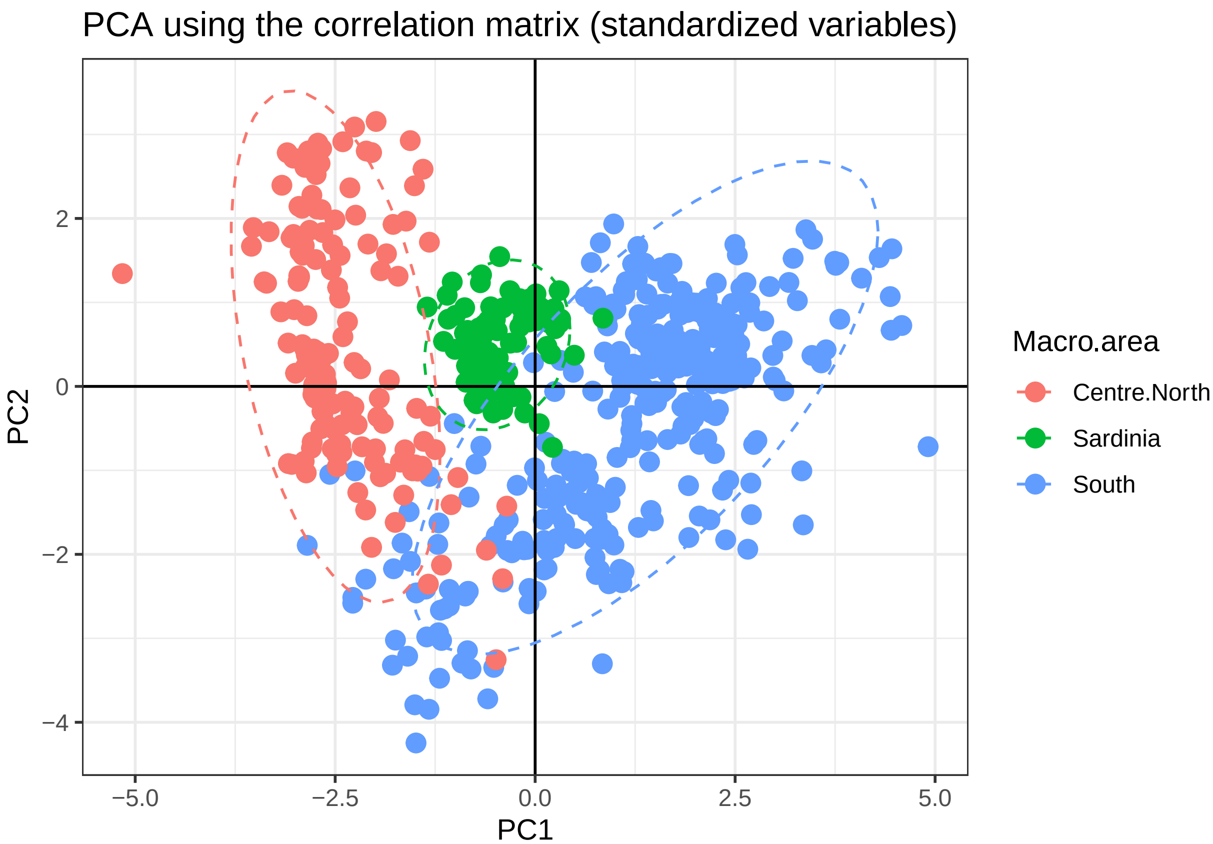
```
##                     PC1         PC2         PC3         PC4         PC5
## palmitic     0.46074351  0.04958406 -0.11445834 -0.28043124  0.53473943
## palmitoleic  0.45022576  0.24090732 -0.14260264 -0.21182252  0.13841908
## stearic     -0.09864471 -0.25837844 -0.80215910  0.47082168  0.21340068
## oleic       -0.49417494 -0.15866175  0.08011486 -0.20010742 -0.01552215
## linoleic     0.36569539  0.34339930  0.08747773  0.51249093 -0.40127538
## linolenic    0.21898707 -0.60483760  0.19103316 -0.09881321  0.12507081
## arachidic    0.22830362 -0.44719396  0.42664494  0.48165441  0.14659527
## eicosenoic   0.31186781 -0.40476916 -0.30085585 -0.33222211 -0.67153429
##                     PC6         PC7        PC8
## palmitic    -0.07699892 -0.52540418 0.35438653
## palmitoleic -0.16728954  0.78680816 0.08856309
## stearic      0.03064009  0.07722664 0.07703841
## oleic       -0.11309403  0.18074878 0.79903372
## linoleic     0.30497855 -0.07768793 0.46687817
```

```
## linolenic    0.69784174  0.19096065 0.02943890
## arachidic   -0.55365142  0.06527504 0.03996552
## eicosenoic  -0.25657629 -0.13959613 0.04168750
```

```r
screeplot(pca_std, type = 'lines', main = 'Scree Plot of Standardized PCA')
```




```r
ggplot(pca_std.data, aes(x = PC1, y = PC2, color = oil[, 1])) +
    geom_point(size = 3) +
    geom_hline(yintercept = 0) +
    geom_vline(xintercept = 0) +
    stat_ellipse(aes(x = PC1, y = PC2), linetype = 2, linewidth = 0.5, level = 0.95) +
    guides(colour = guide_legend("Macro.area")) +
    ggtitle('PCA using the correlation matrix (standardized variables)') +
    theme_bw()
```



- According to the results of the Importance of Components and the Scree Plot, the fourth principal component can explain 91.2% of the total variance.
- $\hat{\lambda}_1 = 3.7214$, $\hat{\lambda}_2 = 1.7658$, $\hat{\lambda}_3 = 1.0164$, $\hat{\lambda}_4 = 0.7929$

  The first sample principal component is:

$$\hat{y}_1 = 0.4607 \times \text{palmitic} + \cdots + 0.3119 \times \text{eicosenoic}$$

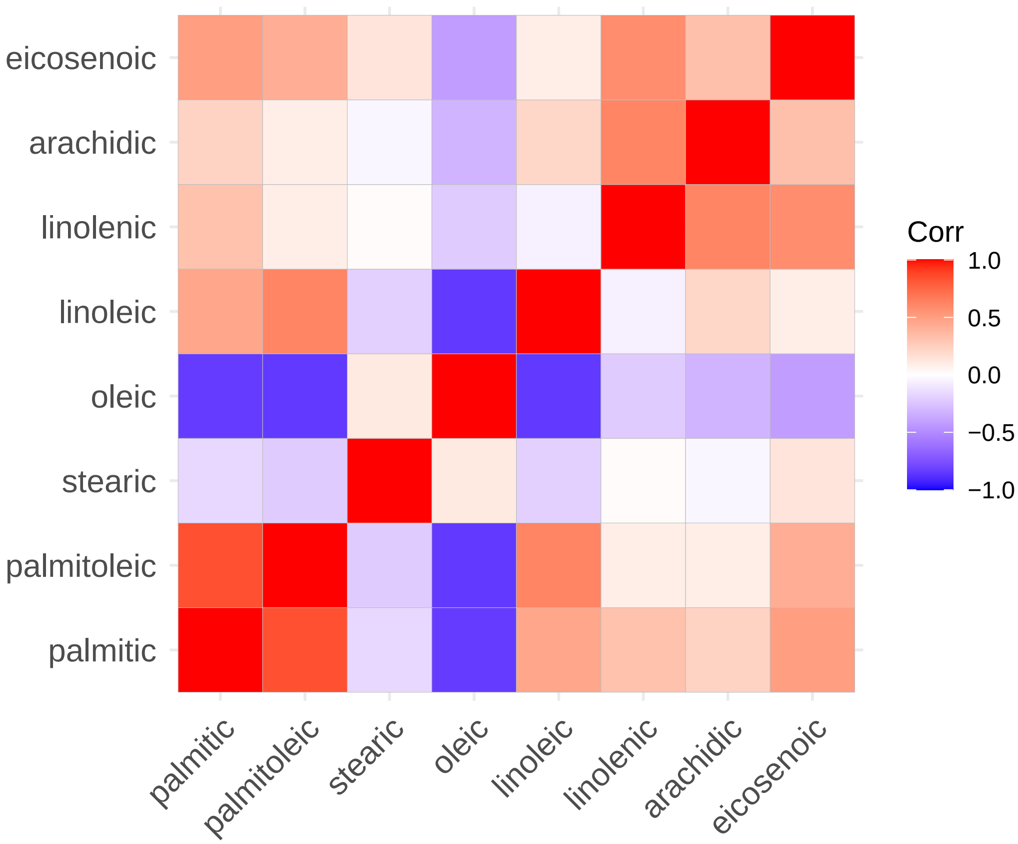
- `oleic` plays the main role in the first principal, `linolenic` plays the main role in the second principal, `stearic` plays the main role in the third principal and `linoleic` plays the main role in the fourth principal.

Conclusion

- In this case, PCA using the covariance matrix requires only one component can explain almost 89.7% of the variance, with two component can explain 98.57% of the variance. Whereas PCA using the correlation matrix needs four components to explain 91.26% of the variance. Furthermore, with only two and three components, PCA using the correlation matrix can explain only 68.59% and 81.29% of the variance. Therefore, for this dataset, PCA using the covariance matrix seems more suitable.

b. Factor Analysis

```
library(ggcorrplot)
ggcorrplot(cor_oil)
```



Principal component solution of the factor model: **Factor loadings is given by**

$$\tilde{\mathbf{L}} = [\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}_1}|\sqrt{\hat{\lambda}_2}\hat{\mathbf{e}_2}|\cdots|\sqrt{\hat{\lambda}_m}\hat{\mathbf{e}_m}]$$

Uniqueness form:

$$\tilde{} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \tilde{\psi}_p \end{bmatrix} \text{with } \tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{\ell}_{ij}^2$$

```
library(psych)
```

b.1 Proportion Variance for m = 1

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
factfit1 <- principal(oil[3:10], nfactors=1, rotate="none",  cor = 'cov'); factfit1
```

```
## Principal Components Analysis
```

```
## Call: principal(r = oil[3:10], nfactors = 1, rotate = "none", cor = "cov")
## Unstandardized loadings (pattern matrix) based upon covariance matrix
##               PC1      h2    u2    H2     U2
## palmitic    -136.4 1.9e+04  9807 0.655 0.3450
## palmitoleic  -44.2 2.0e+03   804 0.708 0.2917
## stearic        5.3 2.9e+01  1322 0.021 0.9788
## oleic        404.7 1.6e+05   921 0.994 0.0056
## linoleic    -214.7 4.6e+04 12843 0.782 0.2179
## linolenic     -2.3 5.2e+00   163 0.031 0.9691
## arachidic     -6.6 4.4e+01   442 0.090 0.9099
## eicosenoic    -5.3 2.8e+01   170 0.142 0.8579
##
##                     PC1
## SS loadings    230543.8
## Proportion Var      0.9
##
##  Standardized loadings (pattern matrix)
##             V   PC1    h2     u2
## palmitic    1 -0.81 0.655 0.3450
## palmitoleic 2 -0.84 0.708 0.2917
## stearic     3  0.15 0.021 0.9788
## oleic       4     1 0.994 0.0056
## linoleic    5 -0.88 0.782 0.2179
## linolenic   6 -0.18 0.031 0.9691
## arachidic   7  -0.3 0.090 0.9099
## eicosenoic  8 -0.38 0.142 0.8579
##
##                 PC1
## SS loadings    7.18
## Proportion Var 0.90
##
## Mean item complexity =  1
## Test of the hypothesis that 1 component is sufficient.
##
## The root mean square of the residuals (RMSR) is  2154.59
##  with the empirical chi square  148701472551  with prob <  0
##
## Fit based upon off diagonal values = 1
```

```r
load_fa1 <- print(factfit1$loadings, digits = 7, cutoff = 1e-7)
```

```
##
## Loadings:
##             PC1
## palmitic    -136.443204
## palmitoleic  -44.179821
## stearic        5.354521
## oleic        404.674390
## linoleic    -214.727919
## linolenic     -2.281306
## arachidic     -6.611667
## eicosenoic    -5.309728
##
##                      PC1
## SS loadings    230543.83
```

```
## Proportion Var  28817.98
```

```r
diag_fa1 <- diag(factfit1$uniquenesses); diag_fa1
```

```
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]    [,8]
## [1,] 9806.604   0.0000    0.000   0.0000     0.00   0.0000   0.0000   0.000
## [2,]    0.000 803.8018    0.000   0.0000     0.00   0.0000   0.0000   0.000
## [3,]    0.000   0.0000 1321.519   0.0000     0.00   0.0000   0.0000   0.000
## [4,]    0.000   0.0000    0.000 920.5747     0.00   0.0000   0.0000   0.000
## [5,]    0.000   0.0000    0.000   0.0000 12843.38   0.0000   0.0000   0.000
## [6,]    0.000   0.0000    0.000   0.0000     0.00 162.9828   0.0000   0.000
## [7,]    0.000   0.0000    0.000   0.0000     0.00   0.0000 441.6178   0.000
## [8,]    0.000   0.0000    0.000   0.0000     0.00   0.0000   0.0000 170.146
```

```r
sum(diag(t(load_fa1) %*% load_fa1)) / tr(cov_oil)
```

```
## [1] 0.8970072
```

```r
factfit2 <- principal(oil[3:10], nfactors=2, rotate="none",  cor = 'cov'); factfit2
```

b.2 Proportion Variance for m = 2

```
## Principal Components Analysis
## Call: principal(r = oil[3:10], nfactors = 2, rotate = "none", cor = "cov")
## Unstandardized loadings (pattern matrix) based upon covariance matrix
##                PC1   PC2     h2   u2    H2     U2
## palmitic    -136.4 -96.2  27870  553 0.981 0.0195
## palmitoleic  -44.2 -14.3   2156  600 0.782 0.2178
## stearic        5.3  -2.2     34 1317 0.025 0.9751
## oleic        404.7  25.5 164410  272 0.998 0.0016
## linoleic    -214.7 112.3  58714  237 0.996 0.0040
## linolenic     -2.3  -5.2     33  136 0.194 0.8057
## arachidic     -6.6  -1.4     46  440 0.094 0.9060
## eicosenoic    -5.3  -6.5     71  128 0.357 0.6430
##
##                             PC1      PC2
## SS loadings           230543.83 22789.01
## Proportion Var             0.90     0.09
## Cumulative Var             0.90     0.99
## Proportion Explained       0.91     0.09
## Cumulative Proportion      0.91     1.00
##
##  Standardized loadings (pattern matrix)
##             item   PC1   PC2    h2     u2
## palmitic       1 -0.81 -0.57 0.981 0.0195
## palmitoleic    2 -0.84 -0.27 0.782 0.2178
## stearic        3  0.15 -0.06 0.025 0.9751
## oleic          4  1.00  0.06 0.998 0.0016
## linoleic       5 -0.88  0.46 0.996 0.0040
## linolenic      6 -0.18 -0.40 0.194 0.8057
## arachidic      7 -0.30 -0.06 0.094 0.9060
## eicosenoic     8 -0.38 -0.46 0.357 0.6430
##
```

```
##                   PC1  PC2
## SS loadings      3.42 1.00
## Proportion Var  0.43 0.13
## Cumulative Var  0.43 0.55
## Cum. factor Var 0.77 1.00
##
## Mean item complexity =  1.4
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  210.58
##  with the empirical chi square  1420489575  with prob <  0
##
## Fit based upon off diagonal values = 1
```

```
load_fa2 <- print(factfit2$loadings, digits = 7, cutoff = 1e-7)
```

```
##
## Loadings:
##             PC1         PC2
## palmitic    -136.443204  -96.193176
## palmitoleic  -44.179821  -14.274047
## stearic        5.354521   -2.230412
## oleic        404.674390   25.476559
## linoleic    -214.727919  112.277008
## linolenic     -2.281306   -5.241953
## arachidic     -6.611667   -1.375131
## eicosenoic    -5.309728   -6.527607
##
##                        PC1       PC2
## SS loadings    230543.83 22789.011
## Proportion Var  28817.98  2848.626
## Cumulative Var  28817.98 31666.605
```

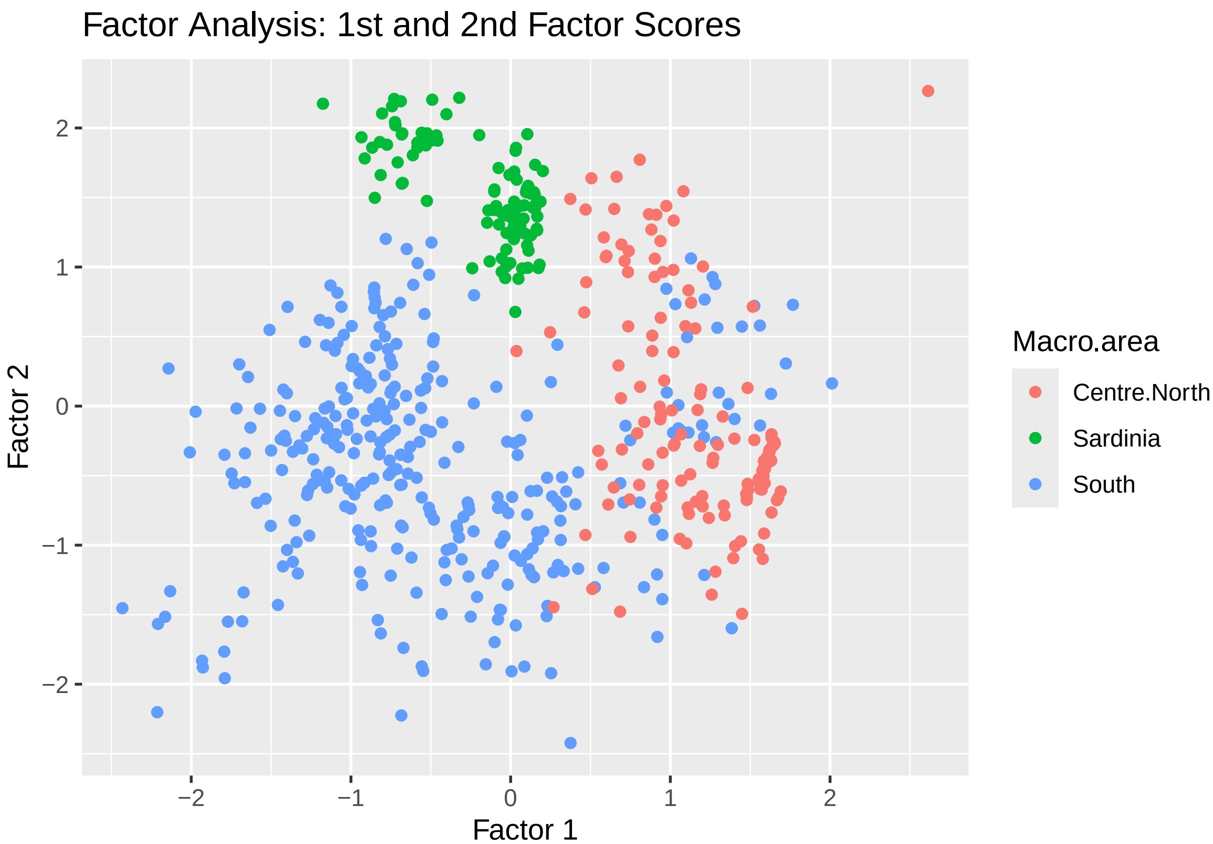
```
diag_fa2 <- diag(factfit2$uniquenesses); diag_fa2
```

```
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
## [1,] 553.4766   0.0000    0.000   0.0000   0.0000   0.0000   0.0000   0.0000
## [2,]   0.0000 600.0534    0.000   0.0000   0.0000   0.0000   0.0000   0.0000
## [3,]   0.0000   0.0000 1316.545   0.0000   0.0000   0.0000   0.0000   0.0000
## [4,]   0.0000   0.0000    0.000 271.5196   0.0000   0.0000   0.0000   0.0000
## [5,]   0.0000   0.0000    0.000   0.0000 237.2559   0.0000   0.0000   0.0000
## [6,]   0.0000   0.0000    0.000   0.0000   0.0000 135.5047   0.0000   0.0000
## [7,]   0.0000   0.0000    0.000   0.0000   0.0000   0.0000 439.7268   0.0000
## [8,]   0.0000   0.0000    0.000   0.0000   0.0000   0.0000   0.0000 127.5363
```

```
sum(diag(t(load_fa2) %*% load_fa2)) / tr(cov_oil)
```

```
## [1] 0.9856754
```

```
factfit2_df <- data.frame(factfit2$scores[, 1], factfit2$scores[, 2], macro.area = oil$macro.area)
ggplot(factfit2_df, aes(factfit2_df[, 1], factfit2_df[, 2], color = factor(macro.area))) +
    geom_point() +
    labs(title = "Factor Analysis: 1st and 2nd Factor Scores",
         x = "Factor 1", y = "Factor 2", color = "Macro.area")
```



- In the factor model with m=1, 89.70% of the total sample variance has been explained by the first factor, moreover, it is clear that oleic plays the main role in the factor.
- In the factor model with m=2, linoleic has the largest loading regarding to the second factor while oleic remains significant in the first factor. In the factor model with m=2, the cumulative proportion of total sample variance explained reaches 98.57%.
- Therefore, 2 factors provide a good fit to the data using a PC solution.

c. MDS

```
library(MASS)
dist_oil <- dist(oil[, 3:10], method = 'euclidean')
mds_oil = isoMDS(dist_oil, k = 3)
```

```
## initial  value 0.695387
## final  value 0.695379
## converged
```

```
head(mds_oil$points)
```

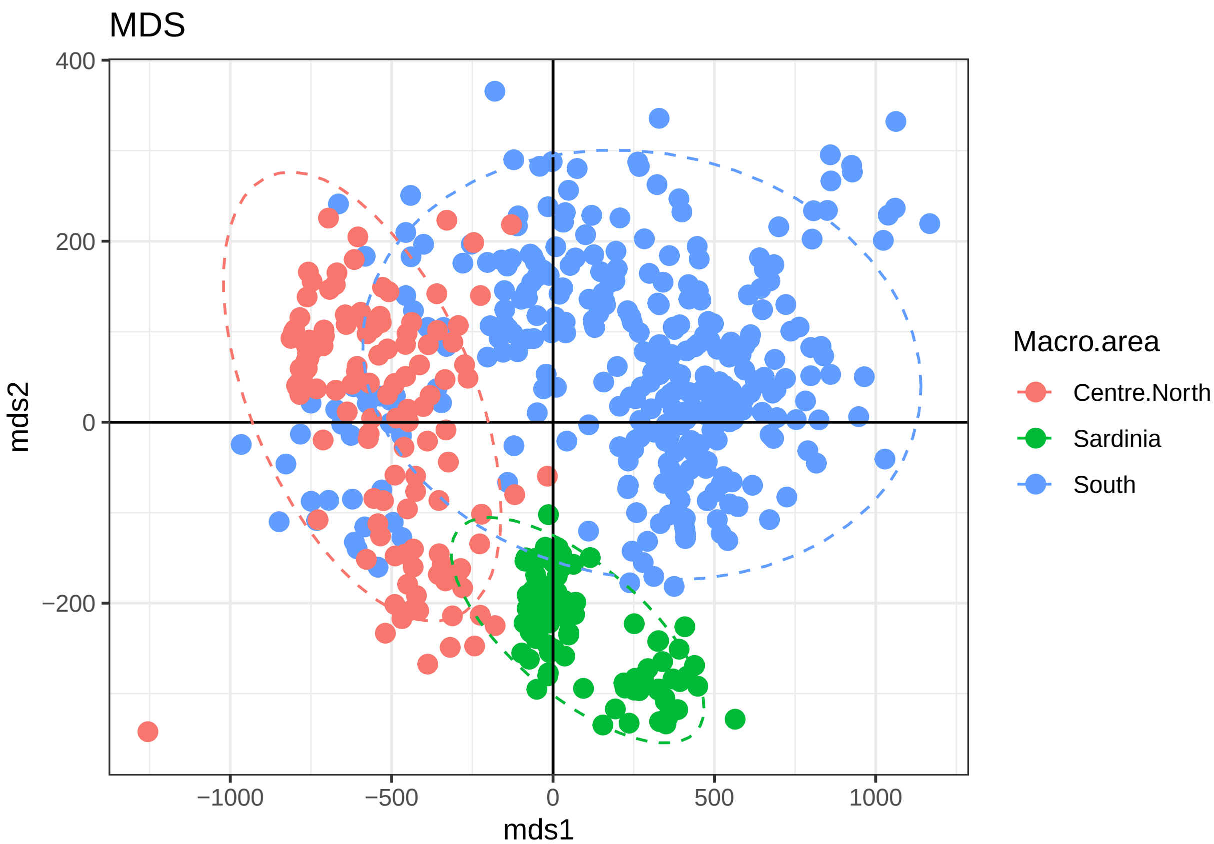
```
##           [,1]      [,2]        [,3]
## [1,] -617.8881  39.14740  -0.7715771
## [2,] -469.5395 -14.78196   1.2582242
## [3,] -966.0508 -24.65113 -32.4156919
## [4,] -782.7186 -13.24558 -40.1001607
```

```
## [5,] -581.4564  33.76464 -60.2341606
## [6,] -749.4180 -87.43976 -77.9305119
```

```r
mds_oil$stress
```

```
## [1] 0.6953795
```

```r
# plot(mds_oil[[1]][,1], mds_oil[[1]][,2])
mds_oil_point <- as.data.frame(mds_oil[[1]])

colnames(mds_oil_point) <- c('mds1', 'mds2', 'mds3')
ggplot(mds_oil_point, aes(x = mds1, y = mds2, color = oil[, 1])) +
    geom_point(size = 3) +
    geom_hline(yintercept = 0) +
    geom_vline(xintercept = 0) +
    stat_ellipse(aes(x = mds1, y = mds2), linetype = 2, linewidth = 0.5, level = 0.95) +
    guides(colour = guide_legend("Macro.area"))+
    ggtitle('MDS')+
    theme_bw()
```



3. Agglomerative hierarchical clustering

Do the agglomerative hierarchical clustering with (1) average linkage, (2) the k-means clustering, and (3) the model-based clustering that adopts the Gaussian mixture model with covariance matrices $\Sigma_1 = \cdots = \Sigma_3 = \Sigma$.

Which approach has the best performance in clustering specimens from the same macro-area together?

a. Average Linkage

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# agglomerative hierarchical with average linkage
oil_avg <- hclust(dist_oil, method = 'average')

oil_avg_cu <- cutree(oil_avg, k = 3)
# table(oil_avg_cu, oil[, 1])

oil_avg_cu_r <- case_when(oil_avg_cu==1 ~ 1,
                          oil_avg_cu==2 ~ 3,
                          oil_avg_cu==3 ~ 2)

table1 <- table(oil_avg_cu_r, oil[, 1]); table1
```

```
##
## oil_avg_cu_r Centre.North Sardinia South
##            1          147        0    45
##            2            1        0     0
##            3            3       98   278
```

```r
cat("\nAccuracy =", sum(diag(table1))/sum(table1), "\n")
```

```
##
## Accuracy = 0.743007
```

b. K-means

```r
set.seed(42)
oil_km <- kmeans(oil[, 3:10], 3)
# table(oil_km$cluster, oil[, 1])

oil_km_r <- case_when(oil_km$cluster == 1 ~ 1,
                      oil_km$cluster == 2 ~ 3,
                      oil_km$cluster == 3 ~ 2
)
```

```r
table2 <- table(oil_km_r, oil[, 1]); table2
```

```
##
## oil_km_r Centre.North Sardinia South
##        1          134        0    42
##        2           17       76    91
##        3            0       22   190
```

```r
cat("\nAccuracy =", sum(diag(table2))/sum(table2), "\n")
```

```
##
## Accuracy = 0.6993007
```

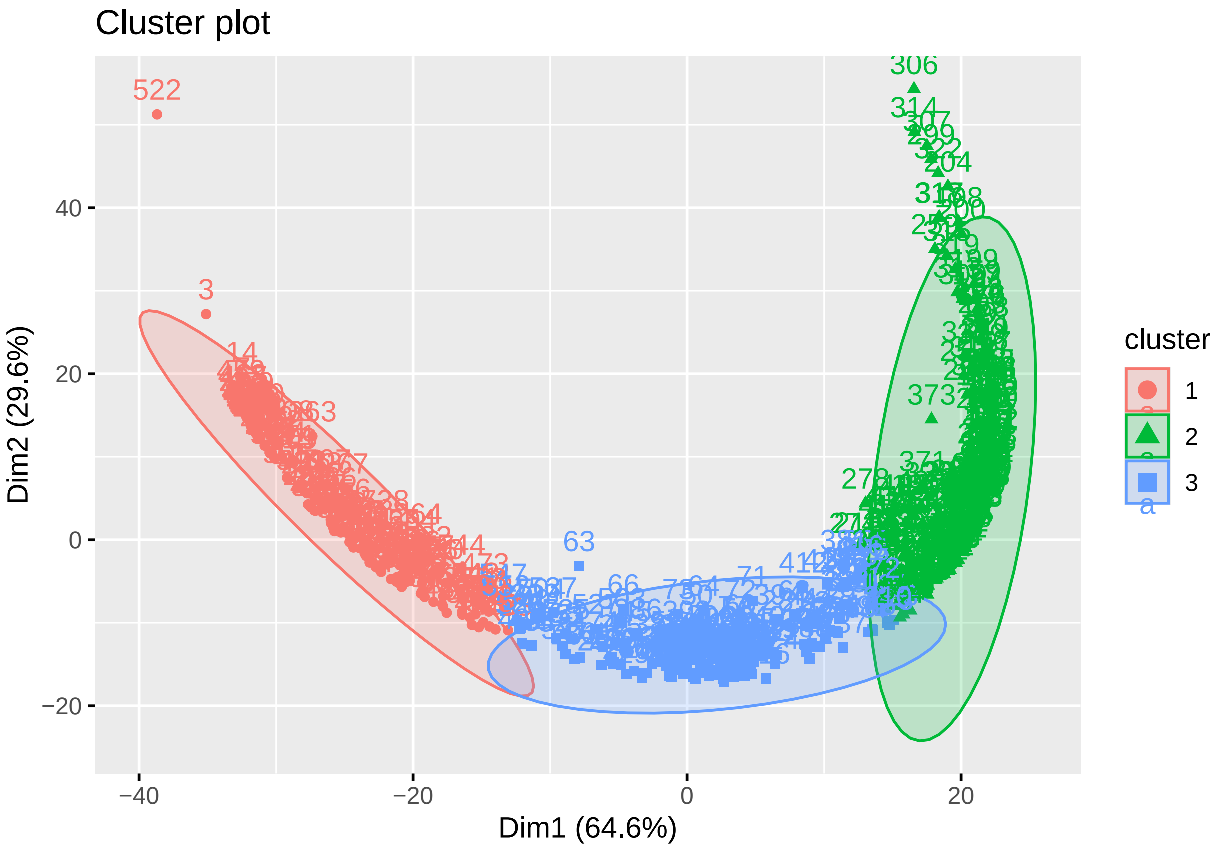
```r
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
fviz_cluster(oil_km,
             data = dist_oil,
             geom = c("point","text"),
             frame.type = "norm")
```

```
## Warning: argument frame is deprecated; please use ellipse instead.
```

```
## Warning: argument frame.type is deprecated; please use ellipse.type instead.
```



c. Model-based clustering

```r
library(mclust)
```

```
## Package 'mclust' version 6.1
## Type 'citation("mclust")' for citing this R package in publications.

##
## Attaching package: 'mclust'

## The following object is masked from 'package:psych':
##
##     sim
```

```r
mbcl <- Mclust(oil[, 3:10], modelNames="EEE", G=3)
summary(mbcl)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEE (ellipsoidal, equal volume, shape and orientation) model with 3
## components:
##
##  log-likelihood   n df       BIC       ICL
##       -21977.31 572 62 -44348.27 -44353.89
##
## Clustering table:
##   1   2   3
## 124 322 126
```

```r
table(mbcl$classification, oil[, 1])
```

```
##
##     Centre.North Sardinia South
##   1          123        0     1
##   2            0        0   322
##   3           28       98     0
```

```r
mbcl5r <- case_when(mbcl$classification==1 ~ 1,
                    mbcl$classification==2 ~ 3,
                    mbcl$classification==3 ~ 2)
table3 <- table(mbcl5r, oil[, 1]); table3
```

```
##
## mbcl5r Centre.North Sardinia South
##      1          123        0     1
##      2           28       98     0
##      3            0        0   322
```

```r
cat("\nAccuracy =", sum(diag(table3))/sum(table3), "\n")
```

```
##
## Accuracy = 0.9493007
```

- By the accuracy rate, model-based clustering has the best performance in clustering specimens from the same macro-area together.

```r
cat("The total sum of squares is:", oil_km$totss)
```

```
## The total sum of squares is: 146755255
```

```r
cat("The (total) within-cluster sum of squares is:", oil_km$tot.withinss)
```

```
## The (total) within-cluster sum of squares is: 30493566
```

```r
cat("The between-cluster sum of squares is:", oil_km$betweenss)
```

```
## The between-cluster sum of squares is: 116261689
```

Gaussian mixture model used for model based clustering In model-based clustering with $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma$, the mixture model is

$$f_{Mix}(x|\mu_1, \mu_2, \mu_3, \Sigma, p_1, p_2, p_3) = \sum_{i=1}^{3} p_i \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^{'}\Sigma^{-1}(x-\mu_i)\right)$$

where $p = 8$ in this case.

Here's the estimated probabilities belonging to each cluster, cluster means and the common covariance matrix.

```r
mbcl$parameters$pro
```

Probabilities

```
## [1] 0.2192619 0.5624473 0.2182907
```

```r
mbcl$parameters$mean
```

Cluster means

```
##                    [,1]       [,2]        [,3]
## palmitic    1084.300658 1333.27151 1118.235271
## palmitoleic   82.506843  155.20298   94.874807
## stearic      227.872435  228.79907  230.033627
## oleic       7816.775836 7096.97403 7357.860409
## linoleic     728.449522 1034.98711 1093.408702
## linolenic     18.700221   38.12399   29.067331
## arachidic     28.036506   63.11635   75.362515
## eicosenoic     2.228855   27.34026    1.902542
```

```r
mbcl$parameters$variance$Sigma
```

Common covaraince matrix

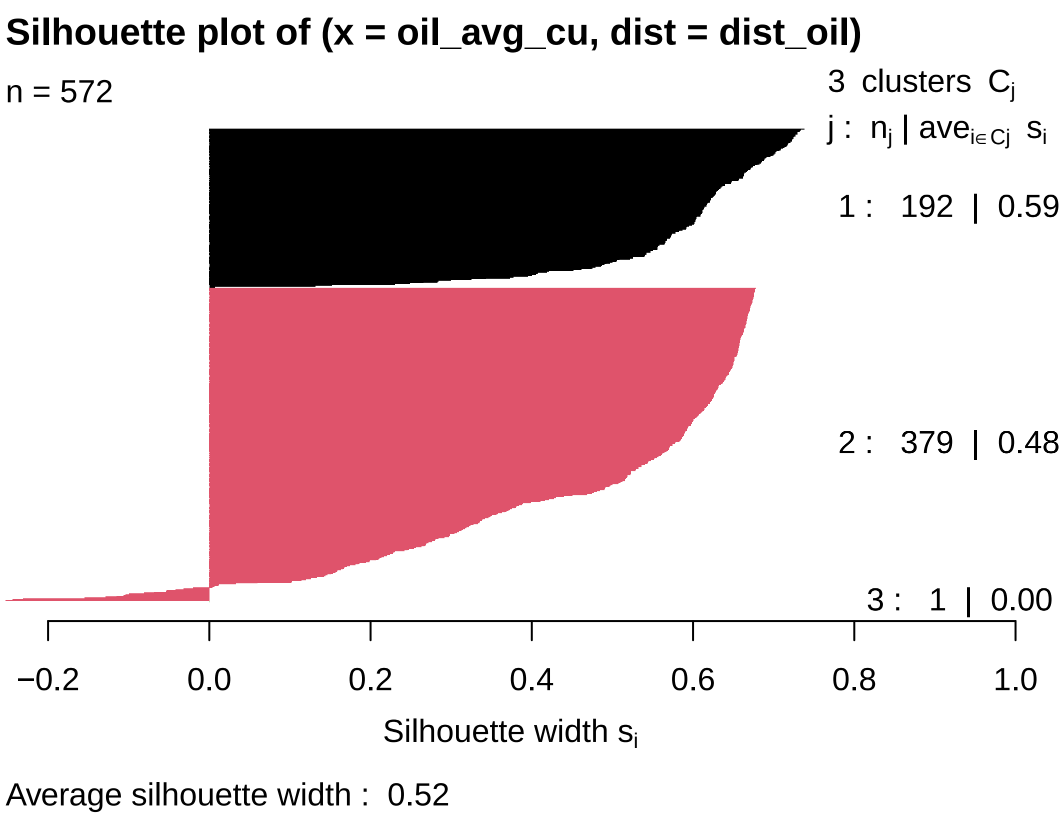
```
##                palmitic palmitoleic      stearic       oleic    linoleic
## palmitic    14996.88446   3537.3968 -1053.083793 -27453.4762  10362.4704
## palmitoleic  3537.39676   1644.9435  -428.267135  -9470.3934   5377.6795
## stearic     -1053.08379   -428.2671  1347.313205   1781.1447  -1843.4030
## oleic      -27453.47622  -9470.3934  1781.144651  82061.9359 -50279.8772
## linoleic    10362.47037   5377.6795 -1843.402951 -50279.8772  40466.1403
## linolenic    -155.37110   -184.0774     7.079657   1095.8309  -1030.9576
## arachidic      15.81169   -153.1214   -43.874861    905.7458  -1114.1569
```

```
## eicosenoic    -252.36466   -106.0817    73.536795    616.9051   -457.1825
##                 linolenic   arachidic eicosenoic
## palmitic     -155.371104    15.81169 -252.36466
## palmitoleic  -184.077408  -153.12138 -106.08171
## stearic         7.079657   -43.87486   73.53679
## oleic        1095.830934   905.74582  616.90506
## linoleic    -1030.957591 -1114.15687 -457.18246
## linolenic     106.150611    82.99742   17.16508
## arachidic      82.997415   207.10874   32.14230
## eicosenoic     17.165082    32.14230   40.77553
```

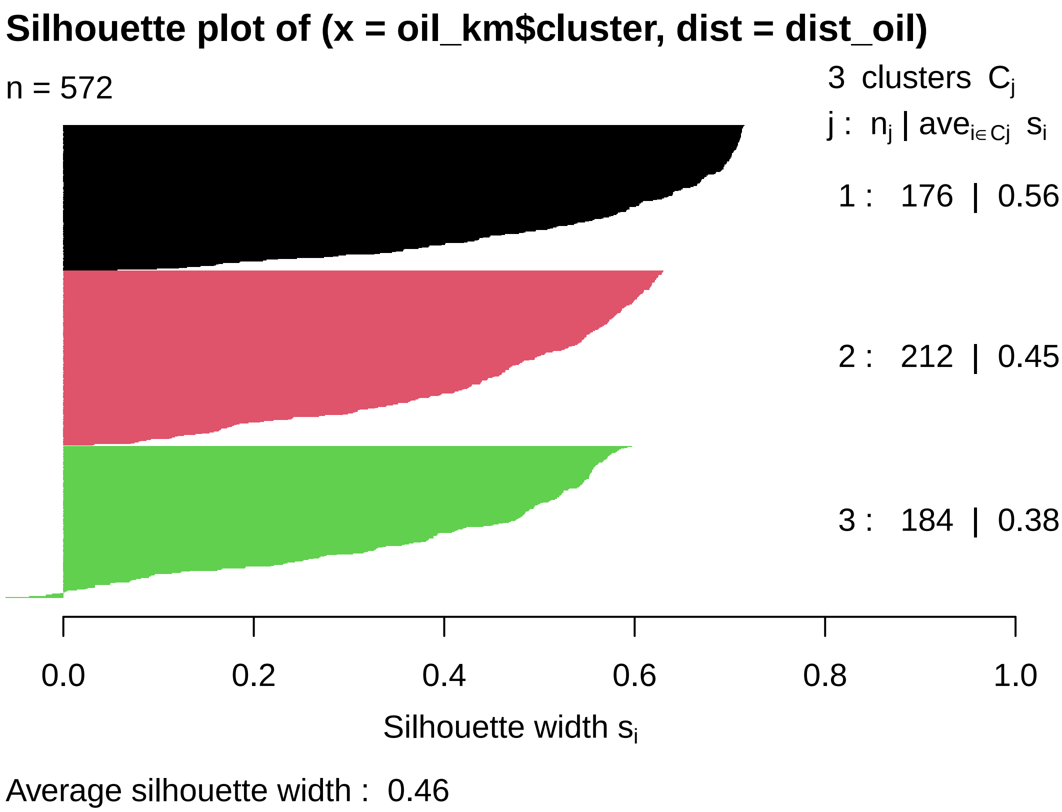
d. Silhouette plot

for each of the clustering approaches.

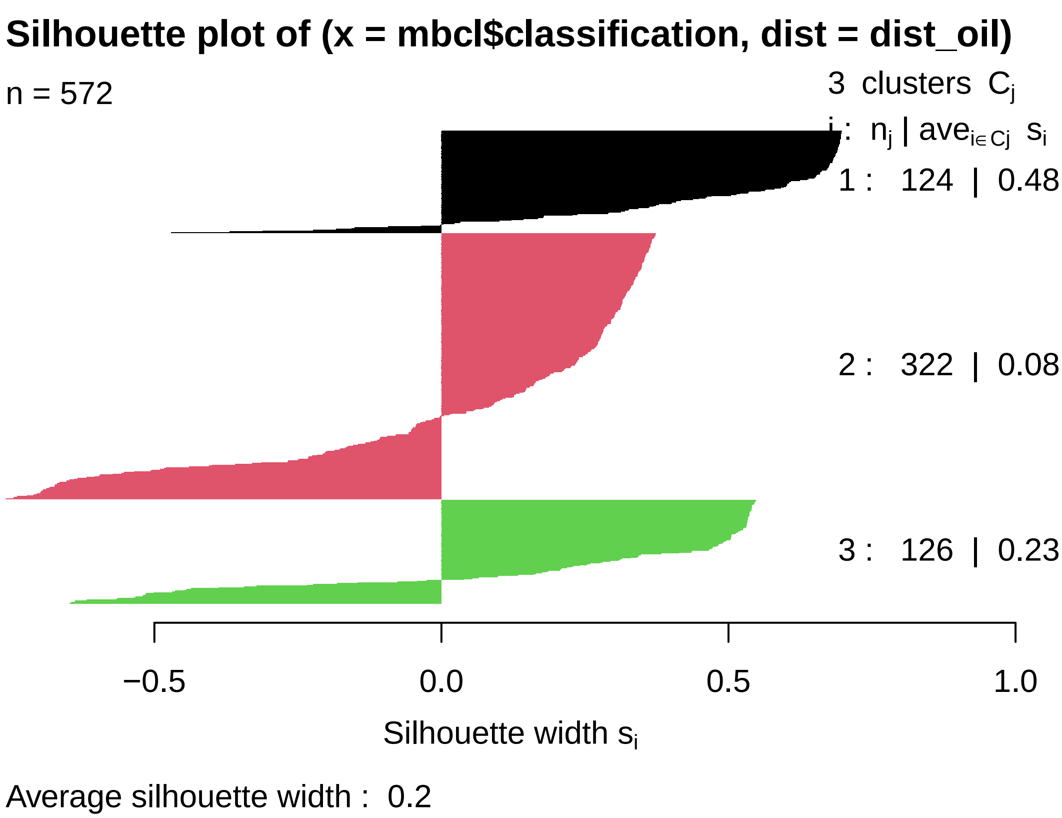
```r
library(cluster)

sia <- silhouette(oil_avg_cu, dist_oil)
plot(sia, col=1:3, border=NA)
```




```r
sik <- silhouette(oil_km$cluster, dist_oil)
plot(sik, col = 1:3, border = NA)
```




```r
sim<-silhouette(mbcl$classification, dist_oil)
plot(sim, col=1:3, border=NA)
```



- By the average ailhoutte width, average linkage has the best cluster fit based on the average silhouette width