APSTA 2011 Project 2

Junyan Yao

January 24, 2017

Data Description A total of 572 olive oil samples were collected from three regions of Italy: the North, the South, and Sardinia. Each region is further divided into areas, with the South having four areas, the North having three, and Sardinia having two. Each olive oil sample was chemically assayed and measured for eight different types of fatty acid. The eight fatty acids are palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, and eisenoic. Each measure represents the percentage of each fatty acid present in each olive oil sample. Using this data, we would like to explore the potential clustering options.

Import and display data

```
require(foreign)
```

```
## Loading required package: foreign
```

```
## Warning: package 'foreign' was built under R version 3.3.2
```

```
olive <- read.dta("C:/Users/jyao/Downloads/olive.dta")
head(olive)
```

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

```
tail(olive)
```

```
##      id palmitic palmitol stearic oleic linoleic linoleni arachidi
## 567 567     1070      100     220  7730      870       10       10
## 568 568     1280      110     290  7490      790       10       10
## 569 569     1060      100     270  7740      810       10       10
## 570 570     1010       90     210  7720      970        0        0
## 571 571      990      120     250  7750      870       10       10
## 572 572      960       80     240  7950      740       10       20
##     eicoseno region         area
## 567        2  North West-Liguria
## 568        2  North West-Liguria
## 569        3  North West-Liguria
## 570        2  North West-Liguria
## 571        2  North West-Liguria
## 572        2  North West-Liguria
```

```
summary(olive)
```

```
##        id         palmitic       palmitol         stearic
##  1      :  1   Min.   : 610   Min.   : 15.00   Min.   :152.0
##  10     :  1   1st Qu.:1095   1st Qu.: 87.75   1st Qu.:205.0
##  100    :  1   Median :1201   Median :110.00   Median :223.0
##  101    :  1   Mean   :1232   Mean   :126.09   Mean   :228.9
##  102    :  1   3rd Qu.:1360   3rd Qu.:169.25   3rd Qu.:249.0
##  103    :  1   Max.   :1753   Max.   :280.00   Max.   :375.0
##  (Other):566
##      oleic         linoleic         linoleni        arachidi
##  Min.   :6300   Min.   : 448.0   Min.   : 0.00   Min.   :  0.0
##  1st Qu.:7000   1st Qu.: 770.8   1st Qu.:26.00   1st Qu.: 50.0
##  Median :7302   Median :1030.0   Median :33.00   Median : 61.0
##  Mean   :7312   Mean   : 980.5   Mean   :31.89   Mean   : 58.1
##  3rd Qu.:7680   3rd Qu.:1180.8   3rd Qu.:40.25   3rd Qu.: 70.0
##  Max.   :8410   Max.   :1470.0   Max.   :74.00   Max.   :105.0
##
##     eicoseno          region                area
##  Min.   : 1.00   South   :323   South-Apulia   :206
##  1st Qu.: 2.00   Sardinia: 98   Inland-Sardinia: 65
##  Median :17.00   North   :151   Calabria       : 56
##  Mean   :16.28                  Umbria         : 51
##  3rd Qu.:28.00                  East-Liguria   : 50
##  Max.   :58.00                  West-Liguria   : 50
##                                 (Other)        : 94
```

```
var(olive$palmitic)
```

```
## [1] 28423.35
```

```
var(olive$palmitol)
```

```
## [1] 2755.658
```

```
var(olive$stearic)
```

```
## [1] 1350.19
```

```
var(olive$oleic)
```

```
## [1] 164681.9
```

```
var(olive$linoleic)
```

```
## [1] 58951.46
```

```
var(olive$linoleni)
```

```
## [1] 168.1871
```

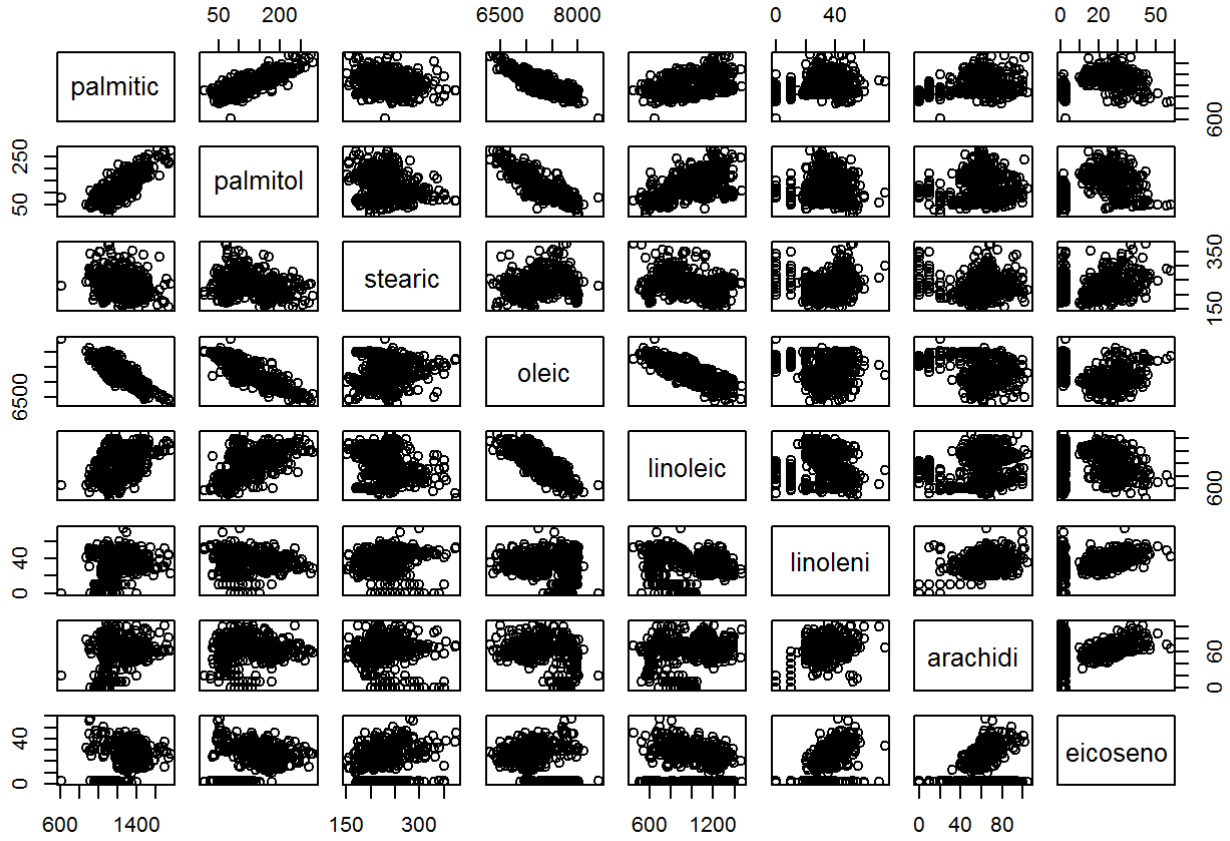
```
var(olive$arachidi)
```

```
## [1] 485.3319
```

```
var(olive$eicoseno)
```

```
## [1] 198.3392
```

```
#All of the variance for each feasure are greater than 150, so there is no reason to drop any va
riables.

#bivariate plot for the original data
pairs(olive[,(2:9)])
```

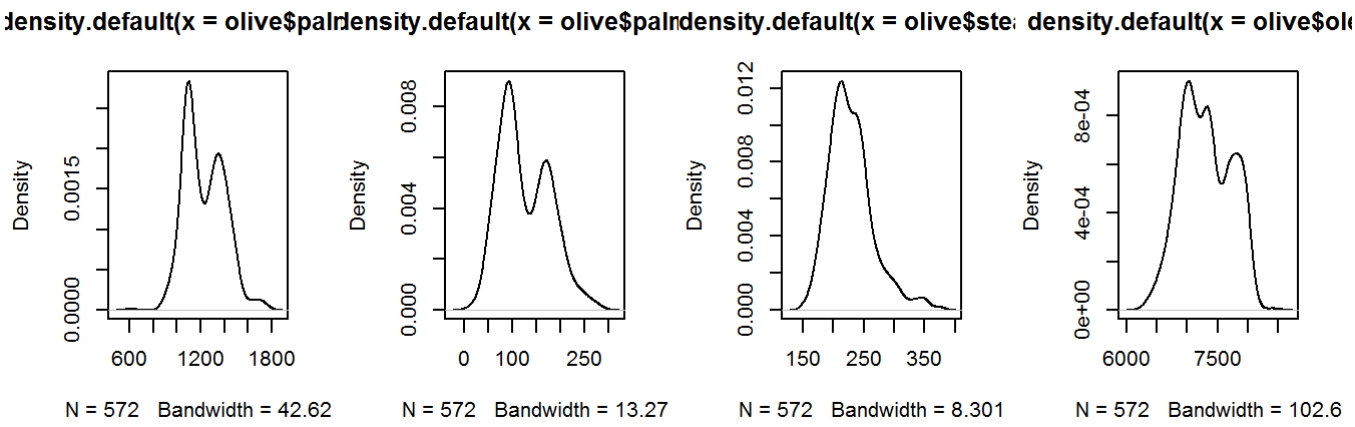


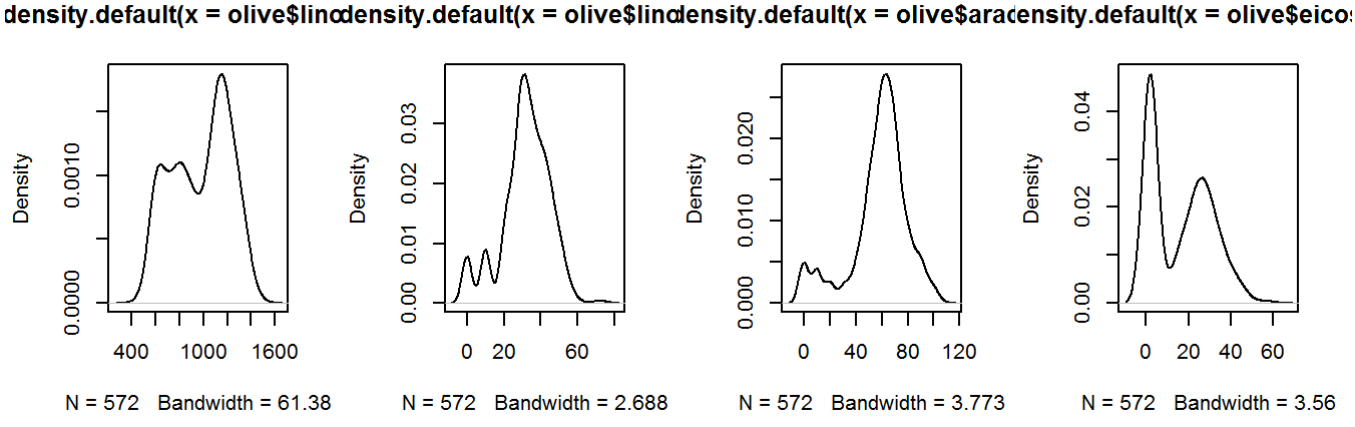

```
#by looking at this bivariate plot, it is hard to tell any potential cluster numbers.
```

Pre-Processing Transformations

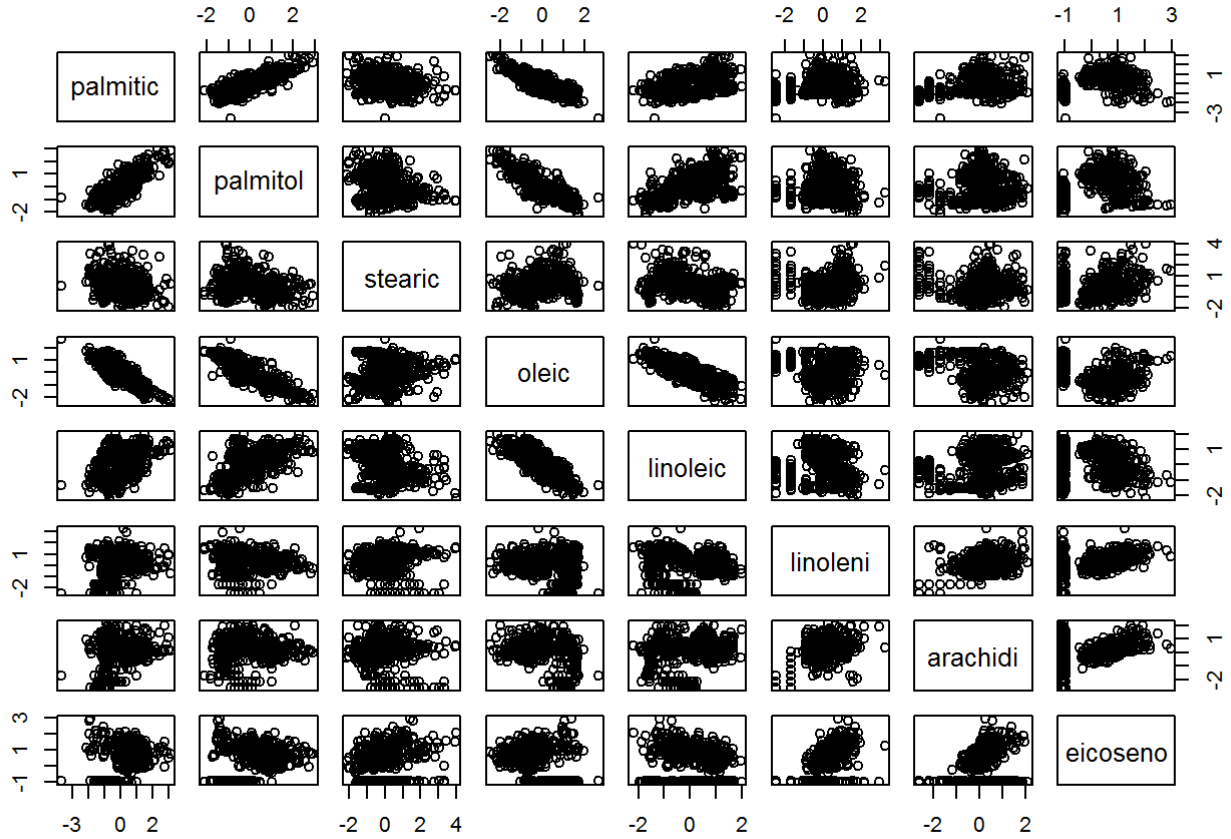
First, I will explore the need to transform or rescale the features. To do this, I plotted the density for each feature measurement.Also, the clustering was done on the original scale, not principal components.

```
#plot the density for each feature
par(mfrow=c(2,4))
plot(density(olive$palmitic))
plot(density(olive$palmitol))
plot(density(olive$stearic))
plot(density(olive$oleic))
plot(density(olive$linoleic))
plot(density(olive$linoleni))
plot(density(olive$arachidi))
plot(density(olive$eicoseno))
```

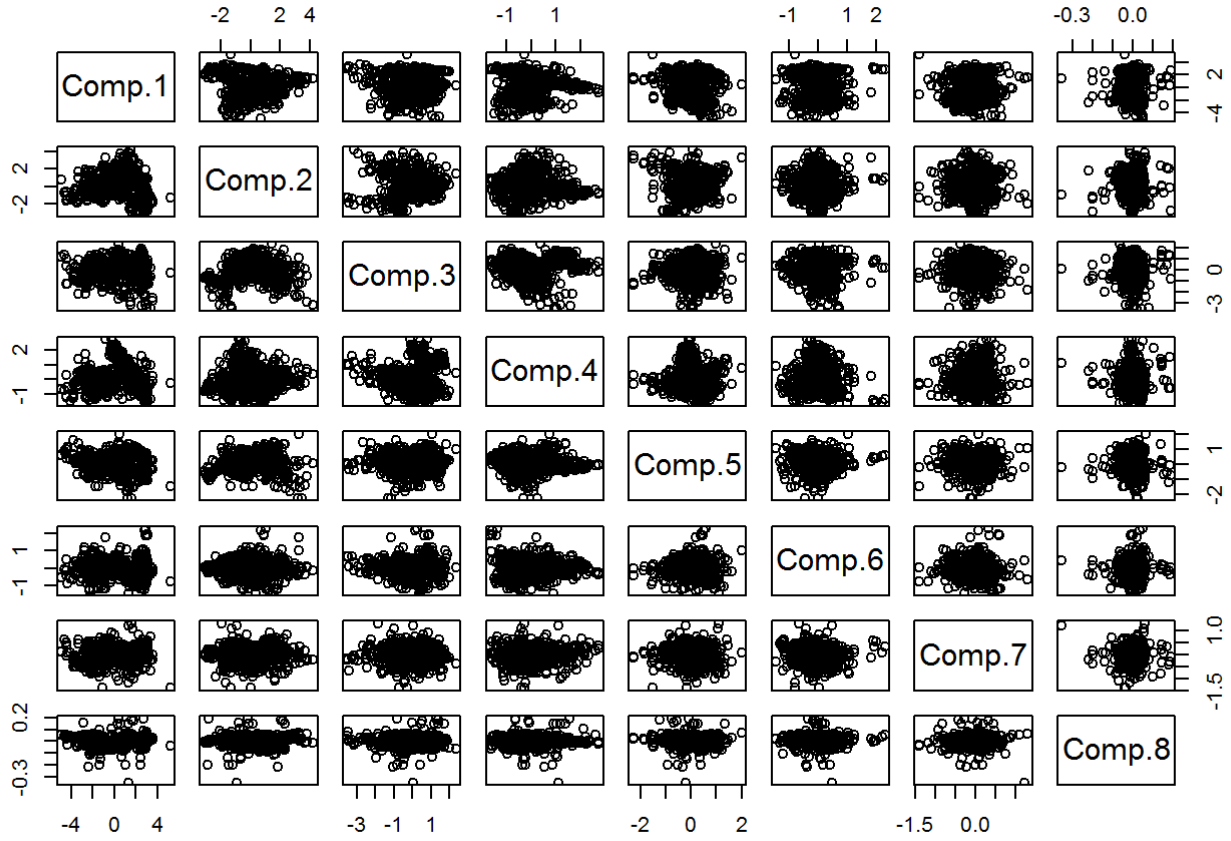





```
#There is some evidence of bimodality in some features; The distributions do not show extreme sk
ew. The log transformation did not really improve a lot. So I will not consider log transformati
on here. It suggests that there are two clusters identified by that one feature.  Transforms can
 confound this.  Only consider transforms in extreme cases in which ONE feature is highly highly
 skewed and the rest are not.  This could be useful for visualizing a bimodality, but likely no
t. Err on the side of not transforming. But we will consider scale the features as the variance
 is very different for some features.

#standardize the data
olive.stdz <- olive;
olive.stdz[,2:9] <- scale(olive[,2:9])

#bivariate plot for standardized data
par(mfrow=c(1,1))
pairs(olive.stdz[,(2:9)])
```




```
#bivariate plots using principal components on standardized data
pairs(princomp(olive.stdz[,2:9])$scores)
```

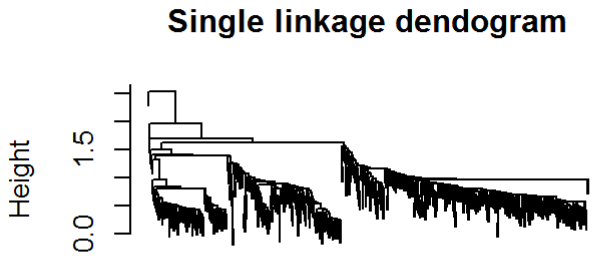


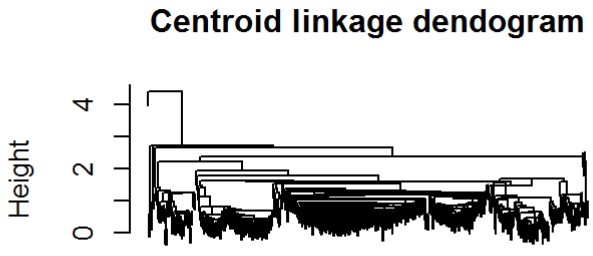
Hierarchical Clustering Analysis

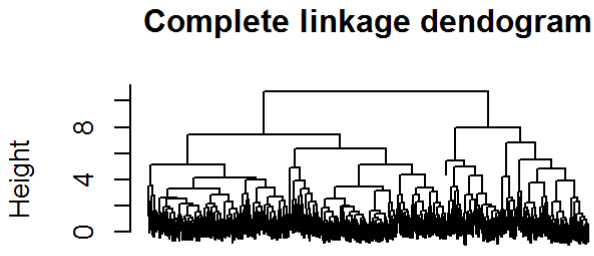
For this analysis, I would like to explore single, centroid, complete, ward and kmeans as potential clustering method for this dataset. Single linkage is the method to find the nearest neighbor. Complete linkage can be refered to furthest neighbor. This is, all possible pairwise distances between elements are evaluated and the largest value is used as the distance between clasters A and B. Using this method, we can ensure all elements contained in one cluster are near all elements in the other cluster. Centroid clustering assigns each newly joined cluster a set of coordinate values based on the mean value for all of the subjects contained within it. In ward's method, we are trying to find two clusters to join such that a total sum of squares associated with the proposed grouping increases by the smallest amount possible,

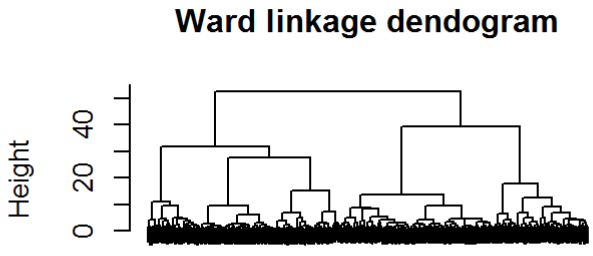
*** Single, Centroid, Complete, and Ward Dendogram***

```
par(mfrow=c(2,2))
#Single linkage hierarchical clustering
hcl.single <- hclust(dist(olive.stdz[,2:9]),meth='single')
plot(hcl.single,labels=F,main='Single linkage dendogram',xlab='',sub="")

#centroid linkage hierarchical clustering
hcl.centroid <- hclust(dist(olive.stdz[,2:9]),meth='centroid')
plot(hcl.centroid,labels=F,main='Centroid linkage dendogram',xlab='',sub="")

#complete linkage hierarchical clustering
hcl.complete <- hclust(dist(olive.stdz[,2:9]),meth='complete')
plot(hcl.complete,labels=F,main='Complete linkage dendogram',xlab='',sub="")


#Ward linkage hierarchical clustering
hcl.ward <- hclust(dist(olive.stdz[,2:9]),meth='ward.D2')
plot(hcl.ward,labels=F,main='Ward linkage dendogram',xlab='',sub="")
```

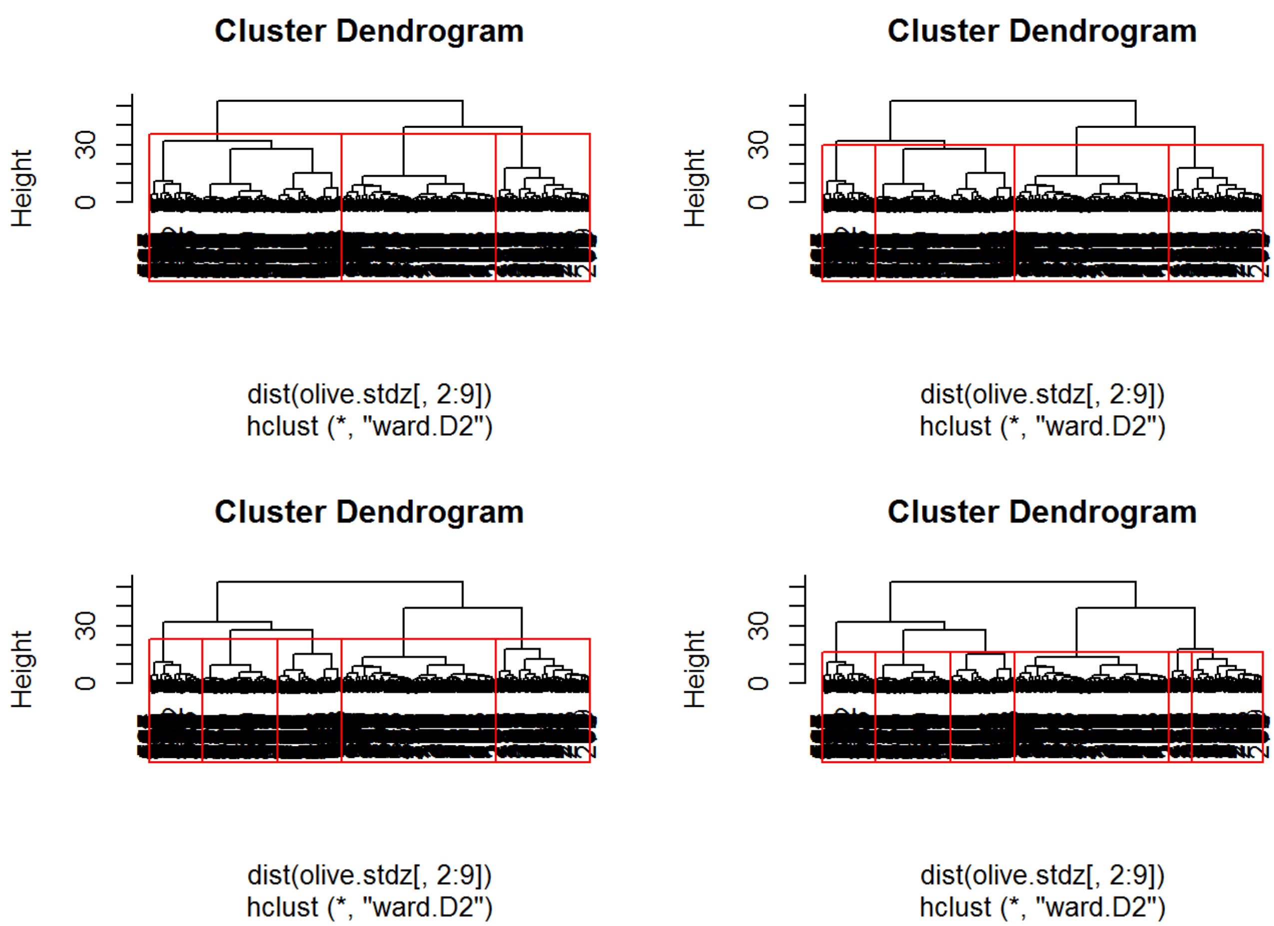








```
#By looking at these four dendogram, we can rule out the single method obviously(it seems very s
tingy);I think ward dendogram looks evenly equal sized clustered (very clean). So I prefere ward
 method.

#choose cluster solution by looking at the dendogram
par(mfrow=c(2,2))
plot(hcl.ward)
rect.hclust(hcl.ward,k=3)
plot(hcl.ward)
rect.hclust(hcl.ward,k=4)
plot(hcl.ward)
rect.hclust(hcl.ward,k=5)
plot(hcl.ward)
rect.hclust(hcl.ward,k=6)
```



I think it is not obvious to find an optimal cluster solution by looking at the dendogram. But we still can see when we choose K=5, the cluster size is a little evenly separated comparing with other cluster solutions.

Choose cluster solution by looking at the Cg maximization In this section, I would like to use two methods-Ward and kmean. The below results shows that when we choose 5 cluster solution, the Cg was maximized for both methods.

```
#############   Ward   #####################
require(NbClust)
```

```
## Loading required package: NbClust
```

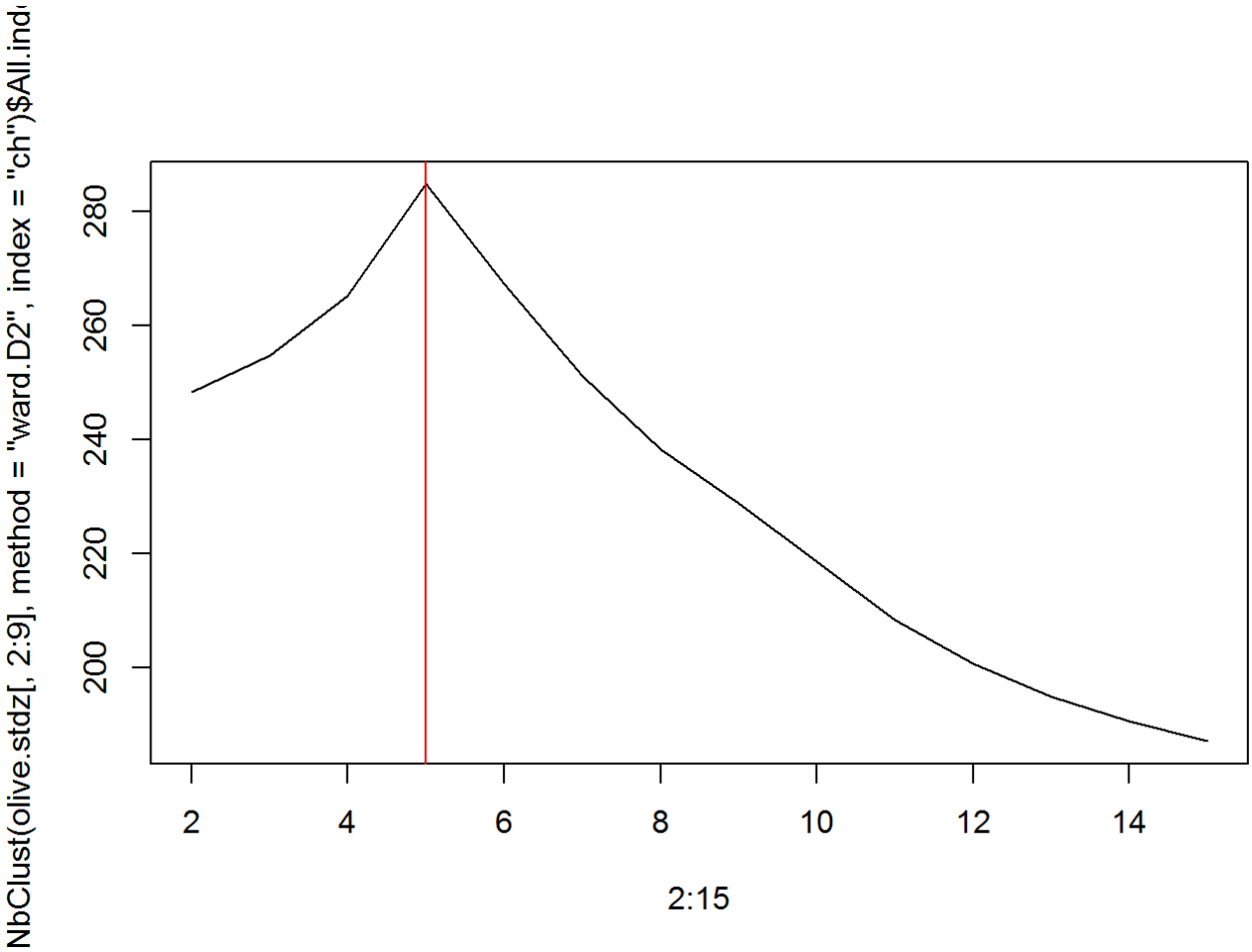
```
## Warning: package 'NbClust' was built under R version 3.3.2
```

```
NbClust(olive.stdz[,2:9],method='ward.D2',index='ch')
```

```
## $All.index
##        2        3        4        5        6        7        8        9
## 248.2752 254.7468 265.1023 284.7418 267.3441 251.1167 238.2705 228.8246
##       10       11       12       13       14       15
## 218.5555 208.3295 200.6838 195.0730 190.6353 187.1614
##
## $Best.nc
## Number_clusters     Value_Index
##          5.0000        284.7418
##
## $Best.partition
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##   1   1   1   1   1   1   1   1   1   2   2   2   2   2   2   2   2   2
##  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
##   2   2   2   2   2   2   2   2   2   2   2   2   1   2   2   2   2   2
## 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234
##   2   2   2   2   2   2   2   2   2   2   1   2   2   1   2   2   1   2
## 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270
##   2   2   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288
##   1   1   1   1   1   1   1   2   2   1   1   1   1   1   1   1   1   1
## 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306
##   1   1   1   1   1   2   2   1   2   2   2   2   2   2   2   2   2   2
## 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   3
## 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342
##   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
## 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360
##   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
## 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378
##   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
```

```
## 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396
##   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
## 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414
##   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
## 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432
##   3   3   3   3   3   3   3   4   4   4   4   4   4   4   4   4   4   4
## 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450
##   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
## 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468
##   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
## 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486
##   4   4   4   4   4   4   4   4   4   5   5   4   4   4   4   4   4   4
## 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504
##   4   4   4   4   4   4   5   4   5   4   4   4   4   4   4   5   5   5
## 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522
##   5   4   5   5   5   4   4   5   5   5   4   4   4   4   5   5   5   5
## 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540
##   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5
## 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558
##   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5   5
## 559 560 561 562 563 564 565 566 567 568 569 570 571 572
##   5   5   5   5   5   5   5   5   5   5   5   5   5   5
```

```
#Show this in graph
par(mfrow=c(1,1))
plot(2:15,NbClust(olive.stdz[,2:9],method='ward.D2',index='ch')$All.index, type="l")
abline(v=5,col="red")
```


NbClust(olive.stdz[, 2:9], method = "ward.D2", index = "ch")$All.ind
200
220
240
260
280
2
4
6
8
10
12
14
2:15

```
#For Ward method, when n=5 Cg is maximized;
############## Kmeans   ###################
c.crit <- function(km.obj) {
    #based on k-means, for convenience due to amt of addl info in the km result object.
    #cd be generalized.
    sizes <- km.obj$size
    n <- sum(sizes)
    g <- length(sizes)
    msW<-sum(km.obj$withinss)/(n-g)
    overall.mean <- apply(km.obj$centers*km.obj$size,2,sum)/sum(km.obj$size)
    msB<-sum(km.obj$size*(t(t(km.obj$centers)-overall.mean))^2)/(g-1)
    list(msB=msB,msW=msW,C.g=msB/msW)
}


set.seed(2011)
km.olive.stdz.2<-kmeans(olive.stdz[,2:9],2,nstart = 100)
km.olive.stdz.3<-kmeans(olive.stdz[,2:9],3,nstart = 100)
km.olive.stdz.4<-kmeans(olive.stdz[,2:9],4,nstart = 100)
km.olive.stdz.5<-kmeans(olive.stdz[,2:9],5,nstart = 100)
km.olive.stdz.6<-kmeans(olive.stdz[,2:9],6,nstart = 100)
km.olive.stdz.7<-kmeans(olive.stdz[,2:9],7,nstart = 100)
km.olive.stdz.8<-kmeans(olive.stdz[,2:9],8,nstart = 100)
km.olive.stdz.9<-kmeans(olive.stdz[,2:9],9,nstart = 100)
km.olive.stdz.10<-kmeans(olive.stdz[,2:9],10,nstart = 100)



c.crit(km.olive.stdz.2) #301.75
```

```
## $msB
## [1] 1581.202
##
## $msW
## [1] 5.239997
##
## $C.g
## [1] 301.7562
```

```
c.crit(km.olive.stdz.3) #275.6649
```

```
## $msB
## [1] 1123.988
##
## $msW
## [1] 4.077371
##
## $C.g
## [1] 275.6649
```

```
c.crit(km.olive.stdz.4) #290.0637
```

```
## $msB
## [1] 921.3038
##
## $msW
## [1] 3.176212
##
## $C.g
## [1] 290.0637
```

```
c.crit(km.olive.stdz.5) #305.2753 This is when Cg is maximized
```

```
## $msB
## [1] 779.8762
##
## $msW
## [1] 2.554665
##
## $C.g
## [1] 305.2753
```

```
c.crit(km.olive.stdz.6) #282.3322
```

```
## $msB
## [1] 652.1307
##
## $msW
## [1] 2.309799
##
## $C.g
## [1] 282.3322
```

```
c.crit(km.olive.stdz.7) #265.1015
```

```
## $msB
## [1] 561.7826
##
## $msW
## [1] 2.119122
##
## $C.g
## [1] 265.1015
```

```
c.crit(km.olive.stdz.8) #256.1535
```

```
## $msB
## [1] 496.4243
##
## $msW
## [1] 1.937996
##
## $C.g
## [1] 256.1535
```

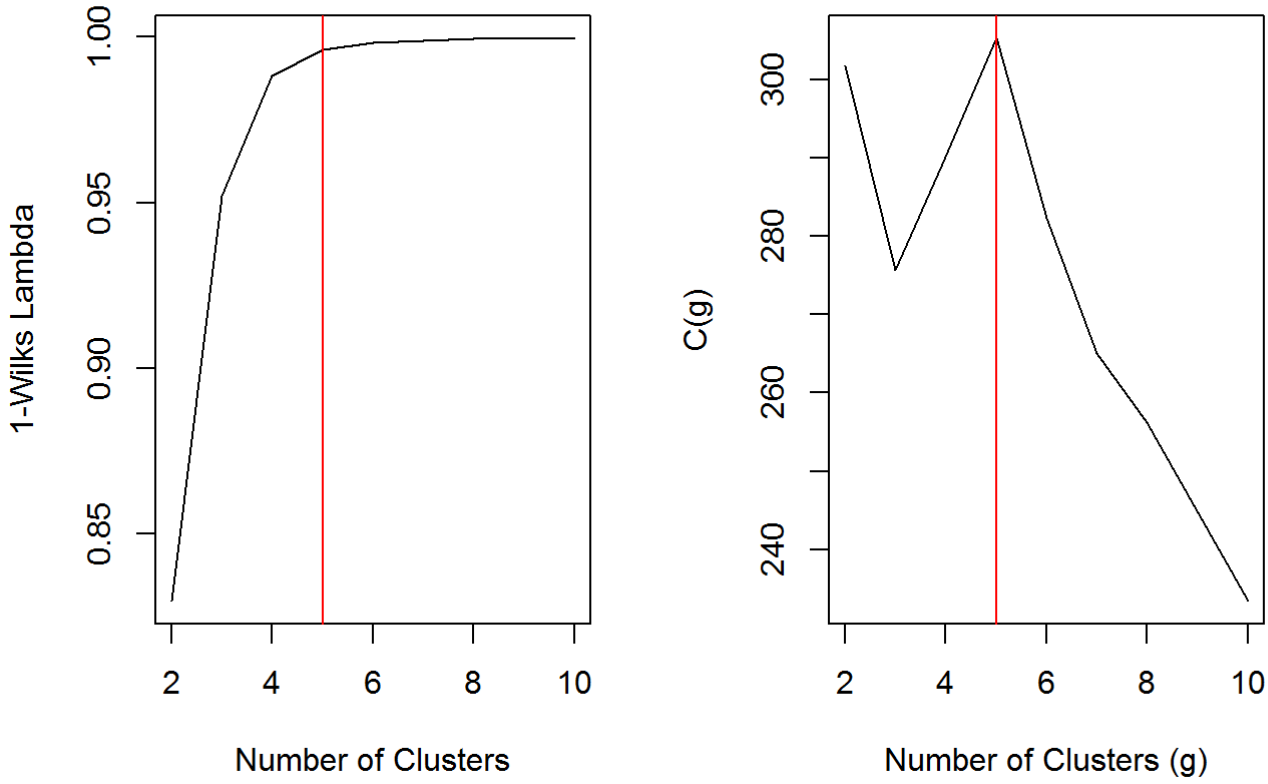
```
c.crit(km.olive.stdz.9) #244.692
```

```
## $msB
## [1] 443.4585
##
## $msW
## [1] 1.812313
##
## $C.g
## [1] 244.692
```

```
c.crit(km.olive.stdz.10) #233.4239
```

```
## $msB
## [1] 400.4335
##
## $msW
## [1] 1.715478
##
## $C.g
## [1] 233.4239
```

```
#show this in graph
numGroupSearch <- function(features,rng=c(2,10),wilks=T,nstart=100) {

    mn <- rng[1]
    mx <- rng[2]
    m.list <- km.list <- vector("list",length=mx-mn+1)
    cFn <- p.rsq <- rep(NA,mx-mn+1)
    i <- 0
    for (k in mn:mx) {
        i <- i+1
        km.list[[i]] <- kmeans(features,k,nstart=nstart)
        U <- as.matrix(features)
        m.list[[i]] <- manova(U~factor(km.list[[i]]$cluster))
        if (wilks) { #avoids some degenerate cases
            p.rsq[i] <- 1-summary(m.list[[i]],test="Wilks")$stats[1,2]
        }
        cFn[i] <- c.crit(km.list[[i]])$C.g
    }
    return(list(km.list=km.list,m.list=m.list,p.rsq=p.rsq,cFn=cFn))
}

ngp.olive <- numGroupSearch(olive.stdz[,2:9])

par(mfrow=c(1,2))
plot(2:10,ngp.olive$p.rsq,type='l',xlab='Number of Clusters',ylab='1-Wilks Lambda')
abline(v=5,col="red")
plot(2:10,ngp.olive$cFn,type='l',xlab='Number of Clusters (g)',ylab='C(g)')
abline(v=5,col="red")
```




```
#Both 1-wilks lambda and Cg yield the same results. They are very consistant in choosing the clu
ster solution here.
#For kmean method, when n=5, Cg is maximized;
```

Comparing ward to kmeans, both with 5 cluster solution

```
set.seed(2011)
lbls.ward.4<- cutree(hcl.ward,k=4)
lbls.ward.5<- cutree(hcl.ward,k=5)
lbls.ward.6<- cutree(hcl.ward,k=6)

xtabs(~km.olive.stdz.5$cluster+lbls.ward.5)
```

```
##                         lbls.ward.5
## km.olive.stdz.5$cluster   1   2   3   4   5
##                       1  17 200   0   0   0
##                       2   0   0   0   0  60
##                       3  99   0   0   5   0
##                       4   0   0  98   5   0
##                       5   7   0   0  73   8
```

```
#write a function to get the maximal agreement
require(gtools)
```

```
## Loading required package: gtools
```

```
## Warning: package 'gtools' was built under R version 3.3.2
```

```
optLabel <- function(src,trg) {
    #input two sets of labels, find permuation that maximizes agreement
    #to be complete search, and handle simpler diag eval, trg must have larger # of labels
    n1 <- length(unique(src))
    n2 <- length(unique(trg))
    tbl <- xtabs(~src+trg)
    best.match <- sum(diag(tbl)) #still works for a non-square matrix.
    best.perm <- 1:n2
    allPerms <- permutations(n2,n2)
    for (i in 1:dim(allPerms)[1]) {
        cur.match <- sum(diag(tbl[,allPerms[i,]]))
        if (cur.match>best.match) {
            best.match <- cur.match
            best.perm <- allPerms[i,]
        }
    }
    list(best.match=best.match,best.perm=best.perm,best.tbl=tbl[,best.perm])
}

optLabel(km.olive.stdz.5$cluster,lbls.ward.5)
```

```
## $best.match
## [1] 530
##
## $best.perm
## [1] 2 5 1 3 4
##
## $best.tbl
##    trg
## src   2   5   1   3   4
##   1 200   0  17   0   0
##   2   0  60   0   0   0
##   3   0   0  99   0   5
##   4   0   0   0  98   5
##   5   0   8   7   0  73
```

```
optLabel(km.olive.stdz.4$cluster,lbls.ward.5)
```

```
## $best.match
## [1] 467
##
## $best.perm
## [1] 2 1 3 5 4
##
## $best.tbl
##    trg
## src   2   1   3   5   4
##   1 200  20   0   0   0
##   2   0 101   0   0  15
##   3   0   0  98   0   8
##   4   0   2   0  68  60
```

```
optLabel(km.olive.stdz.6$cluster,lbls.ward.5)
```

```
## $best.match
## [1] 507
##
## $best.perm
## [1] 3 5 2 1 4
##
## $best.tbl
##    trg
## src   3   5   2   1   4
##   1  98   0   0   0   5
##   2   0  60   0   0   0
##   3   0   0 199   5   0
##   4   0   0   1  76   3
##   5   0   8   0   4  74
##   6   0   0   0  38   1
```

```
optLabel(km.olive.stdz.5$cluster,lbls.ward.4)
```

```
## $best.match
## [1] 462
##
## $best.perm
## [1] 2 4 1 3
##
## $best.tbl
##    trg
## src   2   4   1   3
##   1 200   0  17   0
##   2   0  60   0   0
##   3   0   0  99   5
##   4   0   0   0 103
##   5   0   8   7  73
```

```
optLabel(km.olive.stdz.5$cluster,lbls.ward.6)
```

```
## $best.match
## [1] 506
##
## $best.perm
## [1] 3 6 2 4 5 1
##
## $best.tbl
##    trg
## src   3   6   2   4   5   1
##   1 200   0  17   0   0   0
##   2   0  60   0   0   0   0
##   3   0   0  75   0   5  24
##   4   0   0   0  98   5   0
##   5   0   8   1   0  73   6
```

```
optLabel(km.olive.stdz.4$cluster,lbls.ward.6)
```

```
## $best.match
## [1] 439
##
## $best.perm
## [1] 3 2 4 6 1 5
##
## $best.tbl
##    trg
## src   3   2   4   6   1   5
##   1 200  20   0   0   0   0
##   2   0  73   0   0  28  15
##   3   0   0  98   0   0   8
##   4   0   0   0  68   2  60
```

```
optLabel(km.olive.stdz.6$cluster,lbls.ward.4)
```

```
## $best.match
## [1] 438
##
## $best.perm
## [1] 3 4 2 1
##
## $best.tbl
##    trg
## src   3   4   2   1
##   1 103   0   0   0
##   2   0  60   0   0
##   3   0   0 199   5
##   4   3   0   1  76
##   5  74   8   0   4
##   6   1   0   0  38
```

```
optLabel(km.olive.stdz.6$cluster,lbls.ward.6)
```

```
## $best.match
## [1] 534
##
## $best.perm
## [1] 4 6 3 2 5 1
##
## $best.tbl
##    trg
## src   4   6   3   2   5   1
##   1  98   0   0   0   5   0
##   2   0  60   0   0   0   0
##   3   0   0 199   5   0   0
##   4   0   0   1  76   3   0
##   5   0   8   0   1  74   3
##   6   0   0   0  11   1  27
```

I also tried a range of clusters- one more and one less than Cg suggested, which is 4 clusters and 6 clusters. We find that the maximal agreement is 534, when k= 6 clusters for both Ward and Kmeans. The maximal agreement is 530 using 5 cluster solution for both Ward and Kmeans in this dataset. Acutually choosing 5 or 6 does not make much significant difference. Therefore, I would consider 5 as the cluster solutions for both methods(I will include more details in the Appendix about 6 clusters).

Evaluate the distribution of the known demographics for the kmeans and ward cluster solution Do the clusters seem to divide in a manner consistent with demographic differences? Justify your answer by comparing the frequency distribution of demographics within each cluster

```
xtabs(~km.olive.stdz.5$cluster+olive.stdz$region)
```

```
##                         olive.stdz$region
## km.olive.stdz.5$cluster South Sardinia North
##                       1   217        0     0
##                       2     0        0    60
##                       3    99        0     5
##                       4     0       98     5
##                       5     7        0    81
```

```
xtabs(~km.olive.stdz.5$cluster+olive.stdz$area)
```

```
##                         olive.stdz$area
## km.olive.stdz.5$cluster Calabria Coast-Sardinia East-Liguria
##                       1        3              0            0
##                       2        0              0           10
##                       3       52              0            5
##                       4        0             33            5
##                       5        1              0           30
##                         olive.stdz$area
## km.olive.stdz.5$cluster Inland-Sardinia North-Apulia Sicily South-Apulia
##                       1               0            0     10          204
##                       2               0            0      0            0
##                       3               0           19     26            2
##                       4              65            0      0            0
##                       5               0            6      0            0
##                         olive.stdz$area
## km.olive.stdz.5$cluster Umbria West-Liguria
##                       1      0            0
##                       2      0           50
##                       3      0            0
##                       4      0            0
##                       5     51            0
```

```
xtabs(~lbls.ward.5+olive.stdz$region)
```

```
##            olive.stdz$region
## lbls.ward.5 South Sardinia North
##           1   123        0     0
##           2   200        0     0
##           3     0       98     0
##           4     0        0    83
##           5     0        0    68
```

```
xtabs(~lbls.ward.5+olive.stdz$area)
```

```
##            olive.stdz$area
## lbls.ward.5 Calabria Coast-Sardinia East-Liguria Inland-Sardinia
##           1       56              0            0               0
##           2        0              0            0               0
##           3        0             33            0              65
##           4        0              0           32               0
##           5        0              0           18               0
##            olive.stdz$area
## lbls.ward.5 North-Apulia Sicily South-Apulia Umbria West-Liguria
##           1           25     34            8      0            0
##           2            0      2          198      0            0
##           3            0      0            0      0            0
##           4            0      0            0     51            0
##           5            0      0            0      0           50
```

Kmeans Under kmeans method, The fourth cluster contains 98 samples from Sardinia region and only 5 samples from North region;we can see the second cluster contains all olive oil samples from North region (60 objects); The first cluster contains all samples from south region; In the last cluster, 81 olive oil samples are from North region and only 7 samples are from South region. In the third cluster, 99 samples are from south region and only 5 are from North region. Overall, the kmean clustering separates the samples fairly well. Samples from South region are contained in the first and third clusters. All 98 Sardinia samples are contained in the fourth cluster. Most North samples are fairly split into the second and the last cluster.

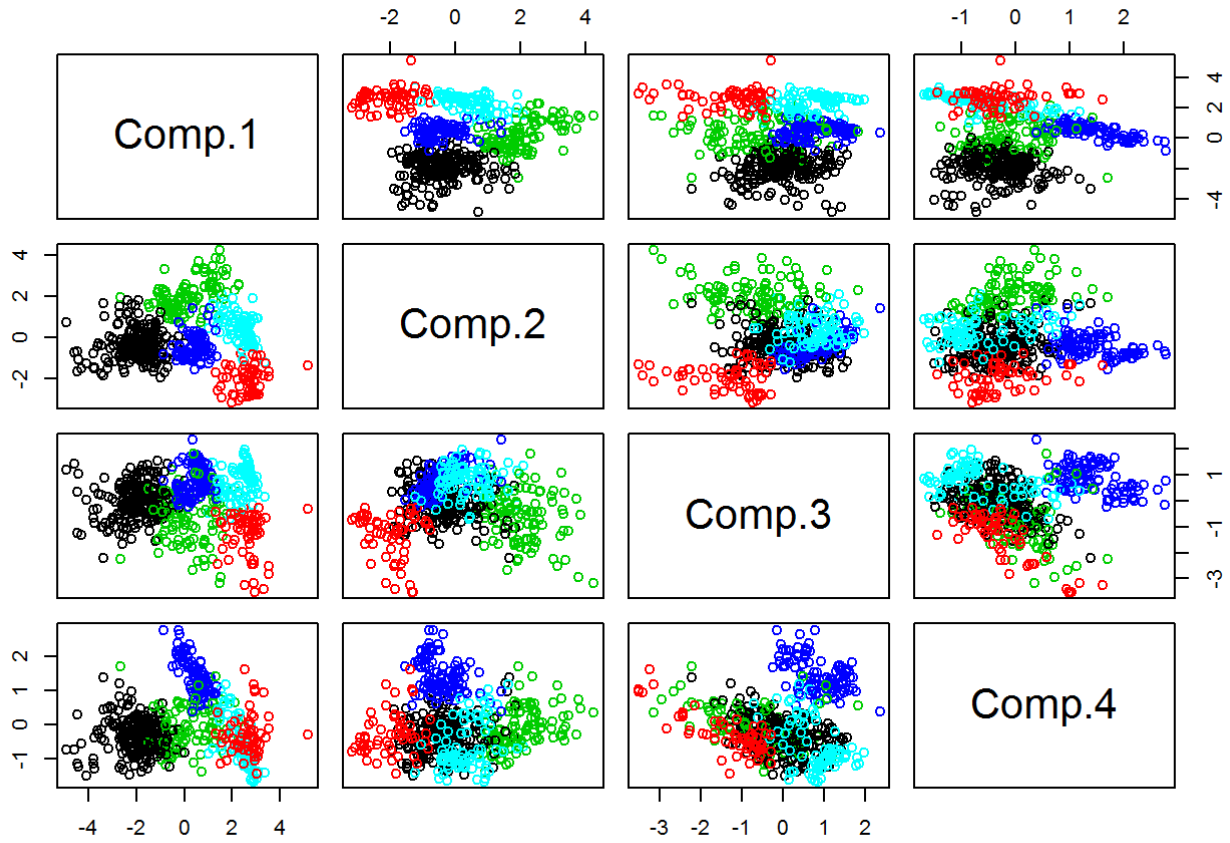
By comparing with area, we can see that 50 olive oil samples are from West-Liguria and 10 samples from East-Liguria in the second cluster; In the fourth cluster, 65 olive oil samples are from Inland-Sardinia area and 33 sample from Coast-Sardinia area. Only 5 samples in the fourth cluster are from East Liguria. In the first cluster, it mostly contains samples from South-Apulia. In the last cluster, it contains 51 samples from Umbria area and 30 samples from East-Liguria area. In the third cluster, it contains 52 olive oil samples from Calabria area, 26 samples from Sicily, 19 samples from North-Apulia and 2 samples from South-Apulia. Overall, this 5 cluster solution does a fairly good job in separating the groups. Samples from Calabria area are almost contained in the third cluster. All samples from Coast-Sardinia area are included in the fourth cluster. All samples from Inland-Sardinia are contained in the fourth cluster. Almost all the samples from South-Apulia area are included in the first cluster. All samples from Umbria are contained in the last cluster. All samples from West-Liguria area are in the second cluster.

Ward By looking at the crosstab result, we would say the ward method 5 cluster solution does a good job in differentiate the regions and area. This reinforces our understanding that olive oil samples are well separated by these features. The first cluster and second cluster contains all samples from South region; The second cluster contains all the samples from Sardinia region; The third and fourth clusters contains all samples from North region. The crosstab comparing the cluster solutions and areas shows that 56 samples from Calabria, 25 samples from North Apulia, 34 samples from Sicily and 8 samples from south Apulia are in the first cluster. In the second cluster, 198 samples are from South Apulia and only 2 samples are from Sicily. 33 Coast-Sardinia and 65 Inland-Sardinia samples are included in the third cluster. 32 East-Liguria and 51 Umbria samples are contained in the fourth cluster. In the last cluster, we can find 18 samples from East-Liguria and 50 samples from West-Liguria. In another way, we can see that all samples from Calabria are in the first cluster; all samples from Coast-Sardinia are in the third cluster; 32 samples from East-Liguria are in the fourth cluster and the rest 18 samples from East-Liguria are in the last cluster; All samples from Inland-Sardinia are in the third cluster. All samples from North-Apulia are in the first cluster. Almost all samples from Sicility are also in the first cluster. Almost all samples from South-Aqualia are in the second cluster, with only 8 samples in the first cluster. All samples from Umbria are in the fourth cluster and all samples from West-Liguria are in the last cluster.

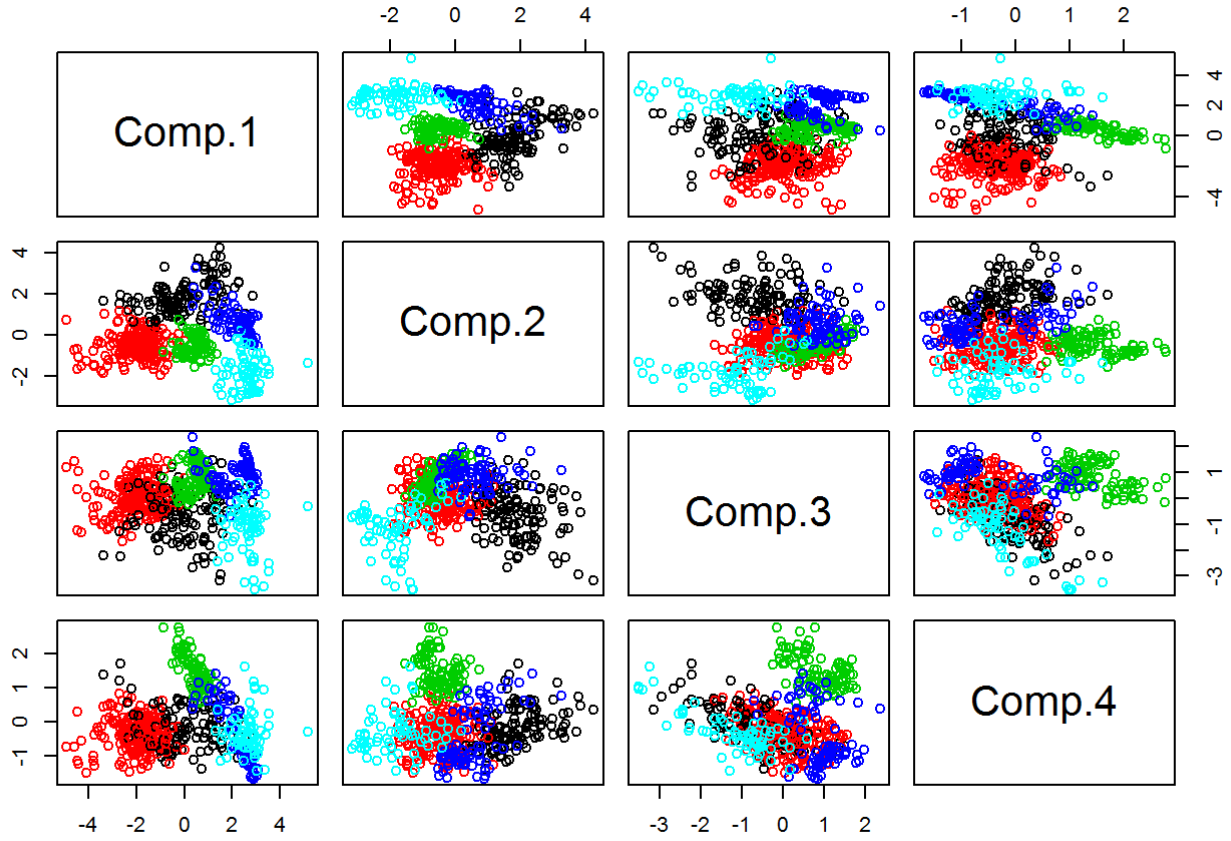
Overall, both kmeans and ward clustering did a good job so that the clusters seem to divide in a manner consistent with demographic differences. I think ward could be better when we are referring the consistency with demographic differences

plot principal component by using kmean and ward Plots (like pairs) of the first 3 or 4 principle components. Colored by cluster solutions (show different approaches to contrast them)

```
pc.olive<- princomp(olive.stdz[,2:9])$scores
#kmeans
pairs(pc.olive[,1:4],col=c(1,2,3,4,5)[km.olive.stdz.5$cluster])
```




```
#wards
pairs(pc.olive[,1:4],col=c(1,2,3,4,5)[lbls.ward.5])
```



By looking at these two above graphs, those two methods split the clusters using different approaches fundamentally, but overall they both looks fairly evenly sized. Take the first component vesus second component for example, there are only a few overlapped points.

Results/Implication

Based on the dendogram and Cg calculation, I choose five clusters using both ward and kmeans. The crosstab comparison suggests the five clusters divide in a manner consistent with the demographic differences.

Appendix

Since choosing 6 can get the maximal agreement. So I would like to evaluate the distribution of the demographics for the kmeans and ward cluster solution using 6 here.

```
xtabs(~km.olive.stdz.6$cluster+olive.stdz$region)
```

```
##                          olive.stdz$region
## km.olive.stdz.6$cluster South Sardinia North
##                       1     0       98     5
##                       2     0        0    60
##                       3   204        0     0
##                       4    77        0     3
##                       5     4        0    82
##                       6    38        0     1
```

```
xtabs(~km.olive.stdz.6$cluster+olive.stdz$area)
```

```
##                         olive.stdz$area
## km.olive.stdz.6$cluster Calabria Coast-Sardinia East-Liguria
##                       1        0             33            5
##                       2        0              0           10
##                       3        1              0            0
##                       4       54              0            3
##                       5        1              0           31
##                       6        0              0            1
##                         olive.stdz$area
## km.olive.stdz.6$cluster Inland-Sardinia North-Apulia Sicily South-Apulia
##                       1              65            0      0            0
##                       2               0            0      0            0
##                       3               0            0      4          199
##                       4               0            1     15            7
##                       5               0            3      0            0
##                       6               0           21     17            0
##                         olive.stdz$area
## km.olive.stdz.6$cluster Umbria West-Liguria
##                       1      0            0
##                       2      0           50
##                       3      0            0
##                       4      0            0
##                       5     51            0
##                       6      0            0
```

```
xtabs(~lbls.ward.6+olive.stdz$region)
```

```
##            olive.stdz$region
## lbls.ward.6 South Sardinia North
##           1    30        0     0
##           2    93        0     0
##           3   200        0     0
##           4     0       98     0
##           5     0        0    83
##           6     0        0    68
```

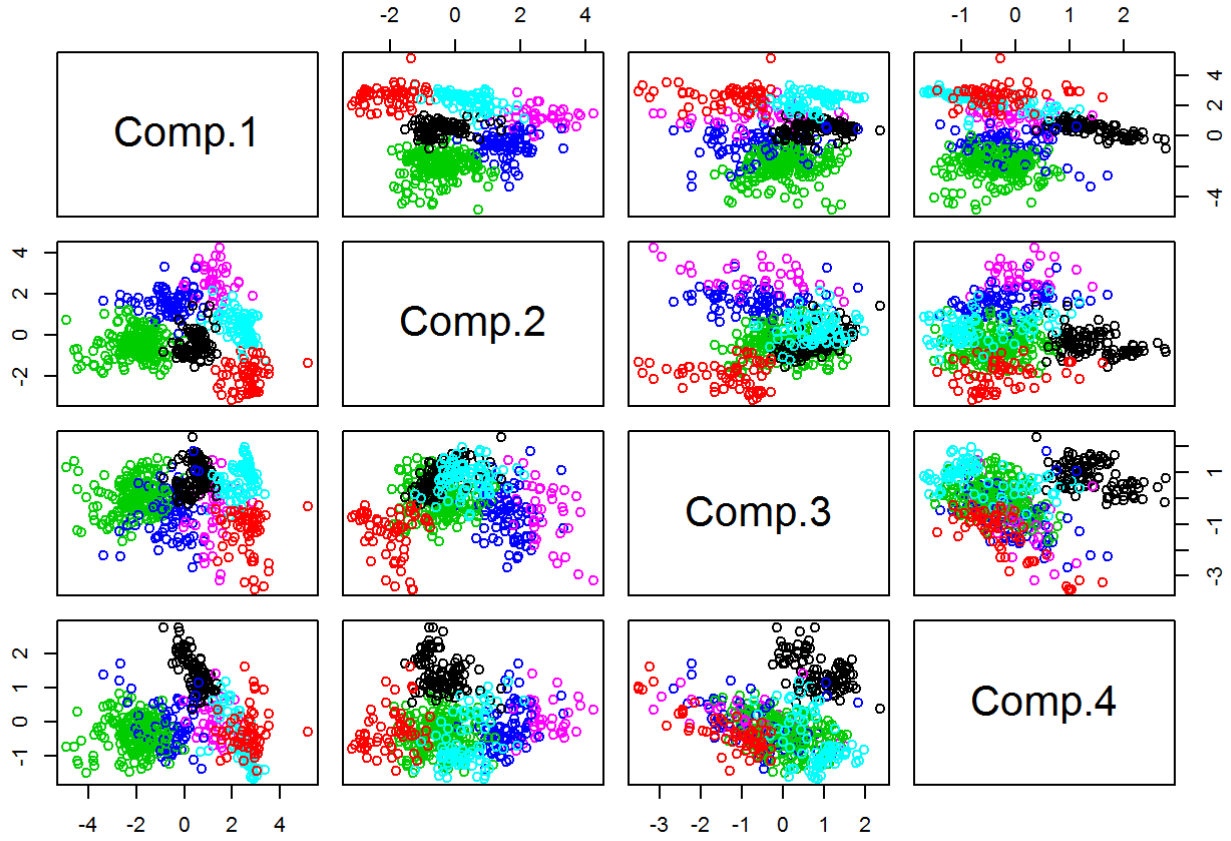
```
xtabs(~lbls.ward.6+olive.stdz$area)
```

```
##            olive.stdz$area
## lbls.ward.6 Calabria Coast-Sardinia East-Liguria Inland-Sardinia
##           1        0              0            0               0
##           2       56              0            0               0
##           3        0              0            0               0
##           4        0             33            0              65
##           5        0              0           32               0
##           6        0              0           18               0
##            olive.stdz$area
## lbls.ward.6 North-Apulia Sicily South-Apulia Umbria West-Liguria
##           1           24      6            0      0            0
##           2            1     28            8      0            0
##           3            0      2          198      0            0
##           4            0      0            0      0            0
##           5            0      0            0     51            0
##           6            0      0            0      0           50
```

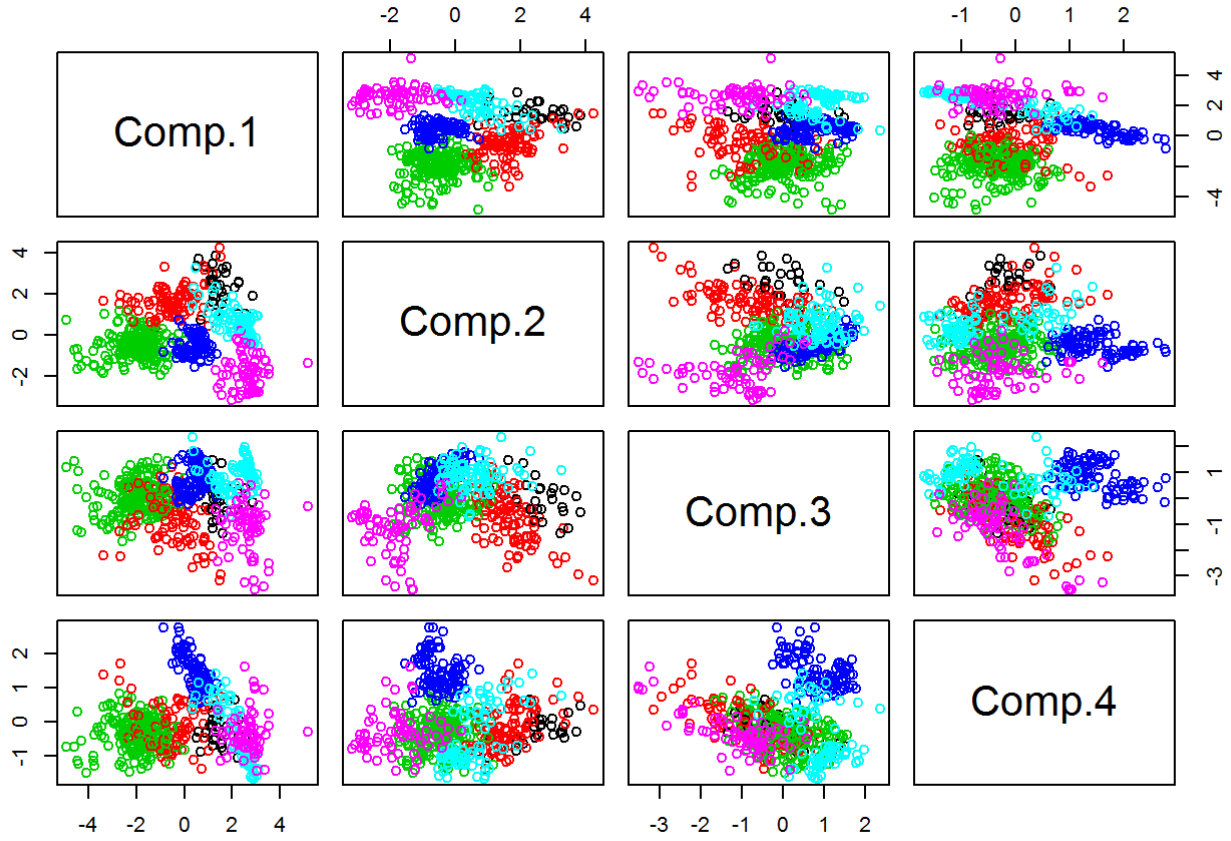
Kmeans The crosstab results shows that almost all samples from South regions are in the third,fourth, and last clusters. Only 4 samples from South are in the fifth cluster. Samples from Sardinia region are all contained in the first cluster. Almost all samples from North are contained in the second and fifth clusters. The crosstab results did not improve a lot comparing the 5 cluster solution.

Ward The crosstab results shows that samples from South are spread among first, second and third clusters. Samples from Sardinia are contained solely in the fourth cluster; Samples from North region are contained in the last two clusters. Again, the results did not improve a lot comparing the 5 cluster solution.

```
#kmeans
pairs(pc.olive[,1:4],col=c(1,2,3,4,5,6)[km.olive.stdz.6$cluster])
```




```
#wards
pairs(pc.olive[,1:4],col=c(1,2,3,4,5,6)[lbls.ward.6])
```



The plots (like pairs) of the first 3 or 4 principle components colored by 6 cluster solutions do not show significant new finding compared with 5 clusters solution.