

# homework8

2022-12-05

```
library(fpc)
library(smacof)

## Caricamento del pacchetto richiesto: plotrix

## Caricamento del pacchetto richiesto: colorspace

## Caricamento del pacchetto richiesto: e1071

##
## Caricamento pacchetto: 'smacof'

## Il seguente oggetto è mascherato da 'package:base':
##      transform

library(cluster)
library(pdfCluster)

## pdfCluster 1.0-3

library(prabclus)

## Warning: il pacchetto 'prabclus' è stato creato con R versione 4.2.2

## Caricamento del pacchetto richiesto: MASS

## Caricamento del pacchetto richiesto: mclust

## Warning: il pacchetto 'mclust' è stato creato con R versione 4.2.2

## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.

##
## Caricamento pacchetto: 'prabclus'

## Il seguente oggetto è mascherato da 'package:fpc':
##      con.comp
```

```

library(mclust)
library(teigen)
library(mixsmsn)

## Caricamento del pacchetto richiesto: mvtnorm

##
## Caricamento pacchetto: 'mvtnorm'

## Il seguente oggetto è mascherato da 'package:mclust':
##
##      dmvnorm

library(fda)

## Warning: il pacchetto 'fda' è stato creato con R versione 4.2.2

## Caricamento del pacchetto richiesto: splines

## Caricamento del pacchetto richiesto: fds

## Warning: il pacchetto 'fds' è stato creato con R versione 4.2.2

## Caricamento del pacchetto richiesto: rainbow

## Warning: il pacchetto 'rainbow' è stato creato con R versione 4.2.2

## Caricamento del pacchetto richiesto: pcaPP

## Warning: il pacchetto 'pcaPP' è stato creato con R versione 4.2.2

## Caricamento del pacchetto richiesto: RCurl

## Caricamento del pacchetto richiesto: deSolve

## Warning: il pacchetto 'deSolve' è stato creato con R versione 4.2.2

##
## Caricamento pacchetto: 'fda'

## Il seguente oggetto è mascherato da 'package:graphics':
##
##      matplot

library(flexmix)

## Caricamento del pacchetto richiesto: lattice

```

```

##  

## Caricamento pacchetto: 'lattice'  

## Il seguente oggetto è mascherato da 'package:fda':  

##  

##      melanoma  

library(funFEM)  

## Warning: il pacchetto 'funFEM' è stato creato con R versione 4.2.2  

## Caricamento del pacchetto richiesto: elasticnet  

## Caricamento del pacchetto richiesto: lars  

## Loaded lars 1.3  

phonemes1000 <- read.table("C:/Users/Utente/OneDrive/Desktop/bigData/datasets/phonemes1000.dat", quote=  

phonemes256 <- as.matrix(phonemes1000[,1:256])  

phonemes257 <- phonemes1000[,257]

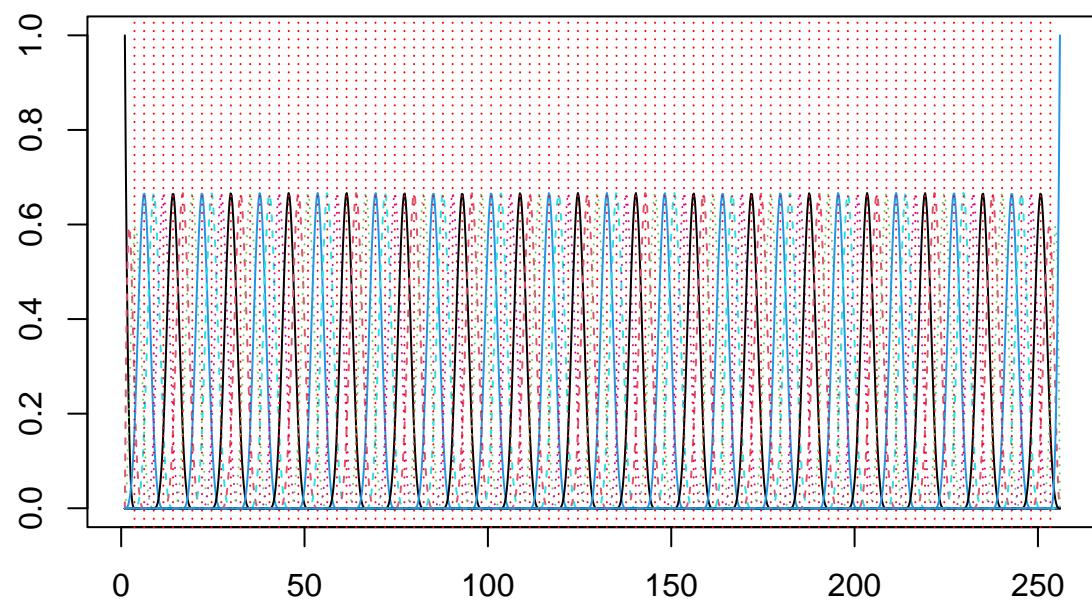
```

1) Represent the data in terms of a suitable B-spline basis. Show how well two exemplary observations are approximated in this way.

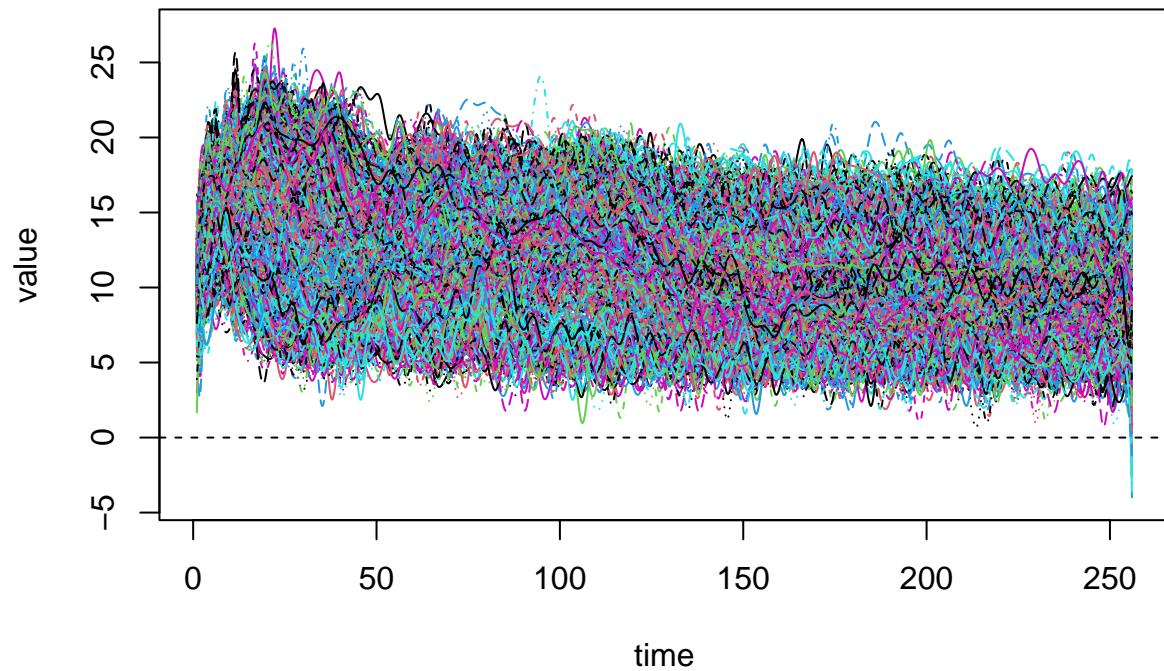
```

basis<-create.bspline.basis(c(1,256), nbasis=100)
fd.phon<-Data2fd(1:256,y=t(phonemes256),basisobj=basis)
plot(basis)

```

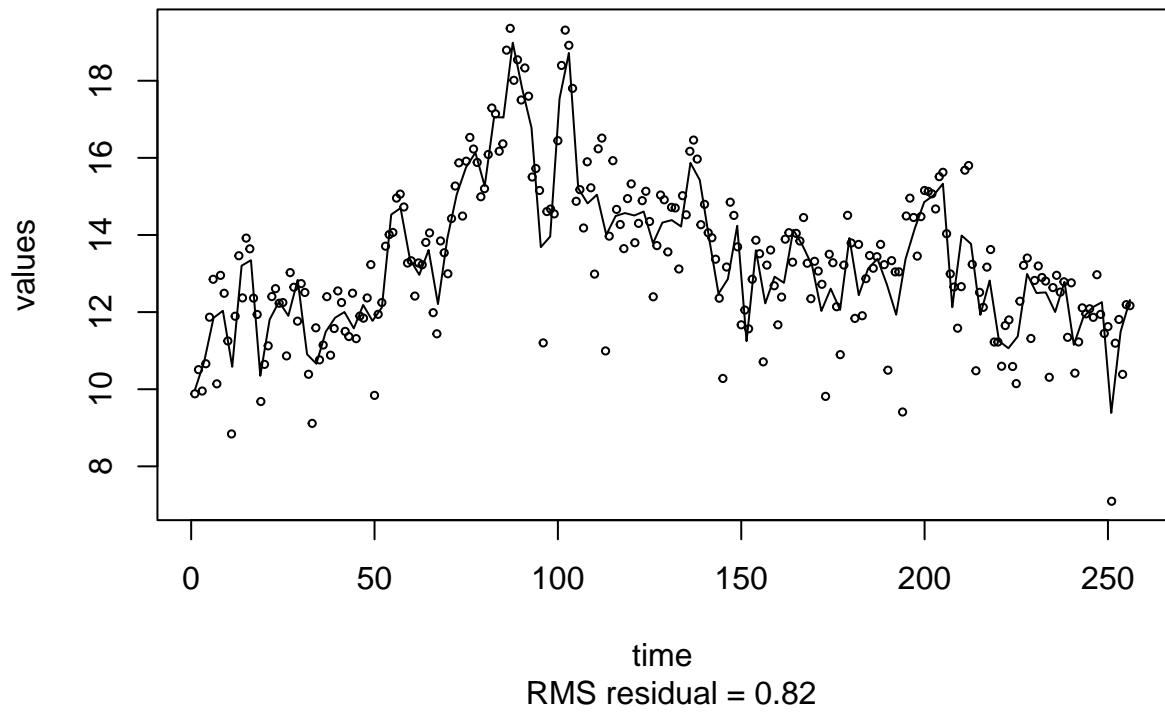


```
plot(fd.phon)
```



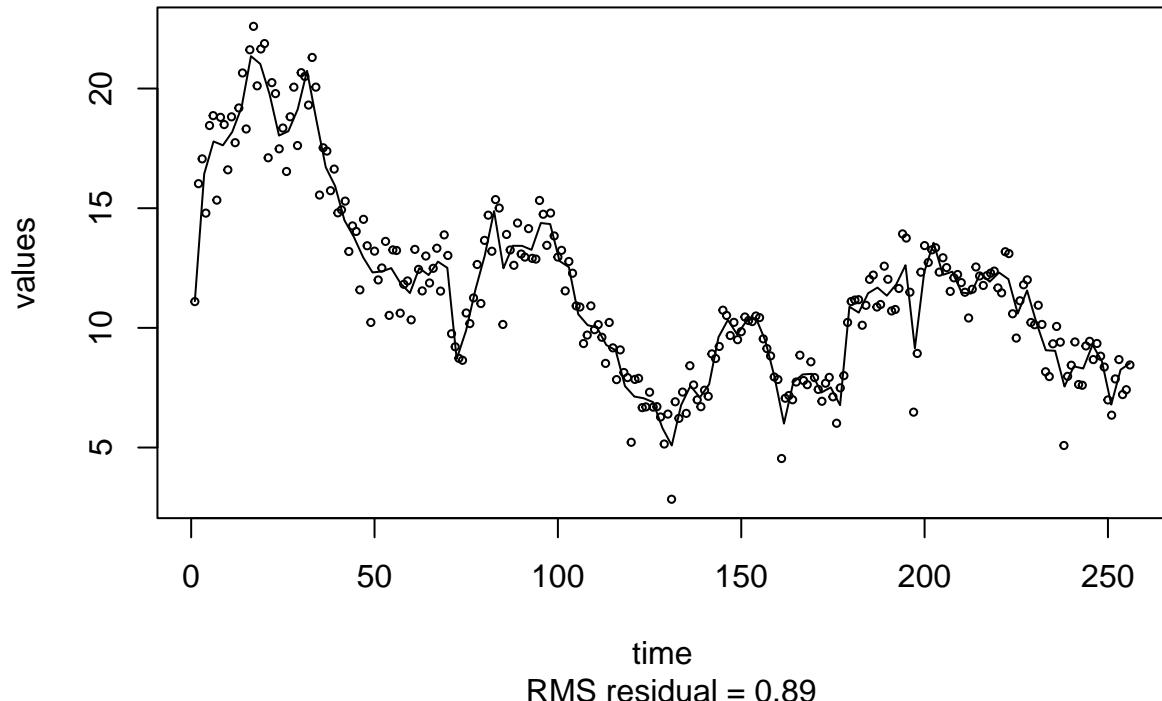
```
## [1] "done"  
plotfit.fd(t(phonemes256), 1:256, fd.phon, index=30, cex.pch=0.5)
```

**rep30**



```
plotfit.fd(t(phonemes256), 1:256, fd.phon, index=130, cex.pch=0.5)
```

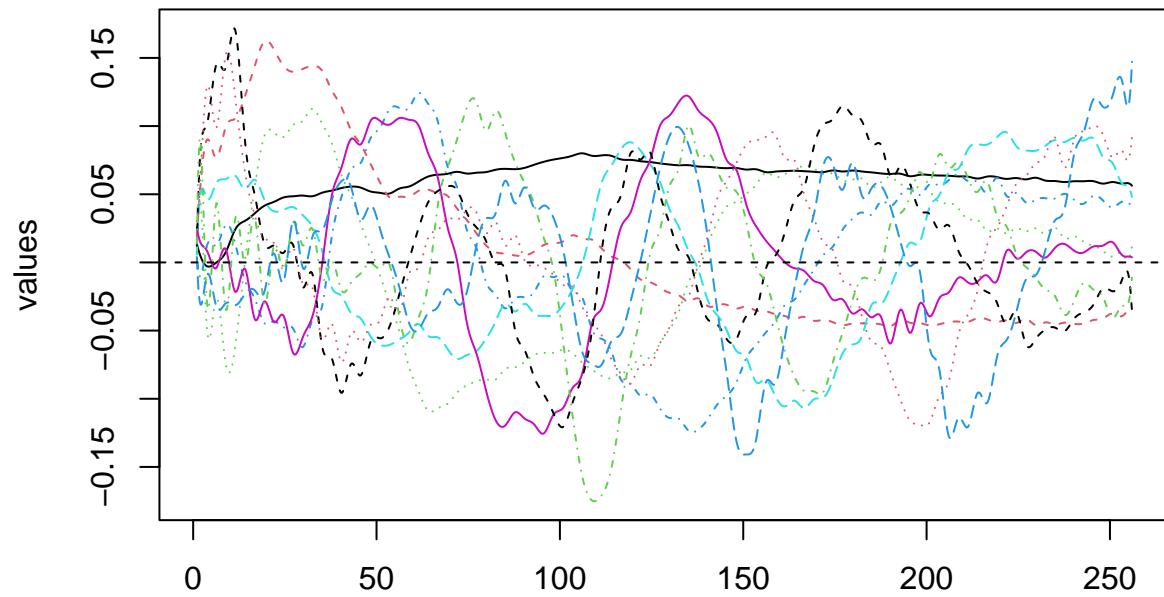
## rep130



If I consider only 10 basis the line of approximations will not follow the true behavior of points. We can note the high variance of the error where points are far from the line, this can be seen especially in the first plot for observation number 30.

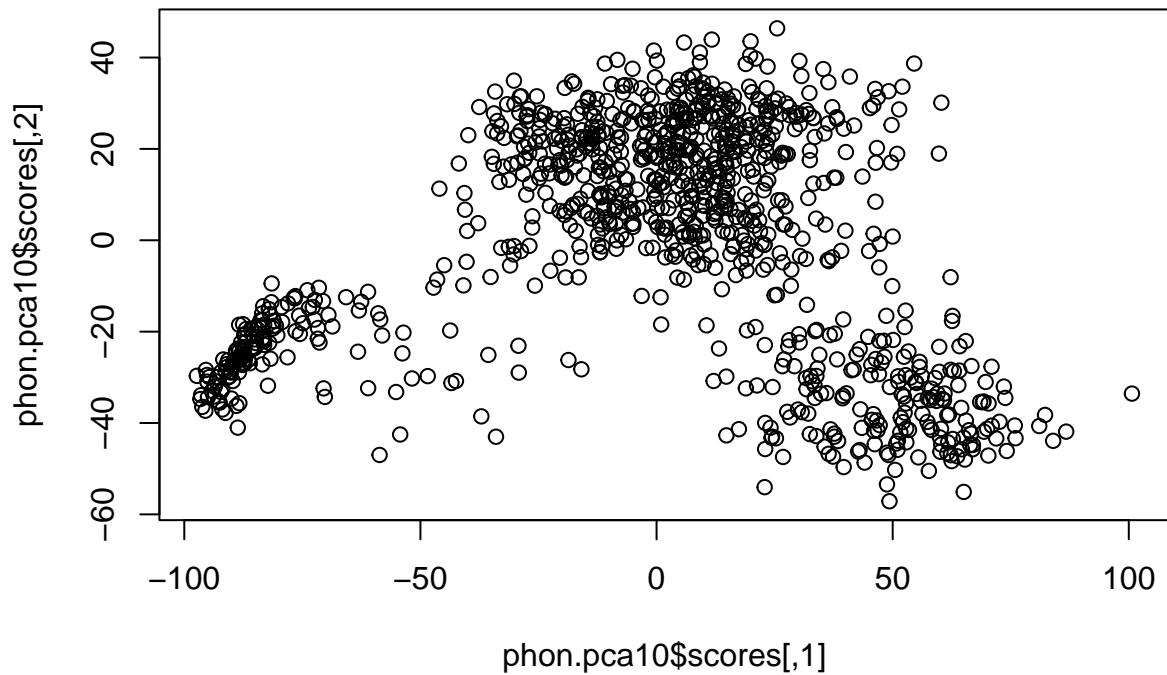
Run a functional principal component analysis,

```
phon.pca10 <- pca.fd(fd.phon, nharm = 10 )
plot(phon.pca10$harmonics) #there is not a lot of variation
```

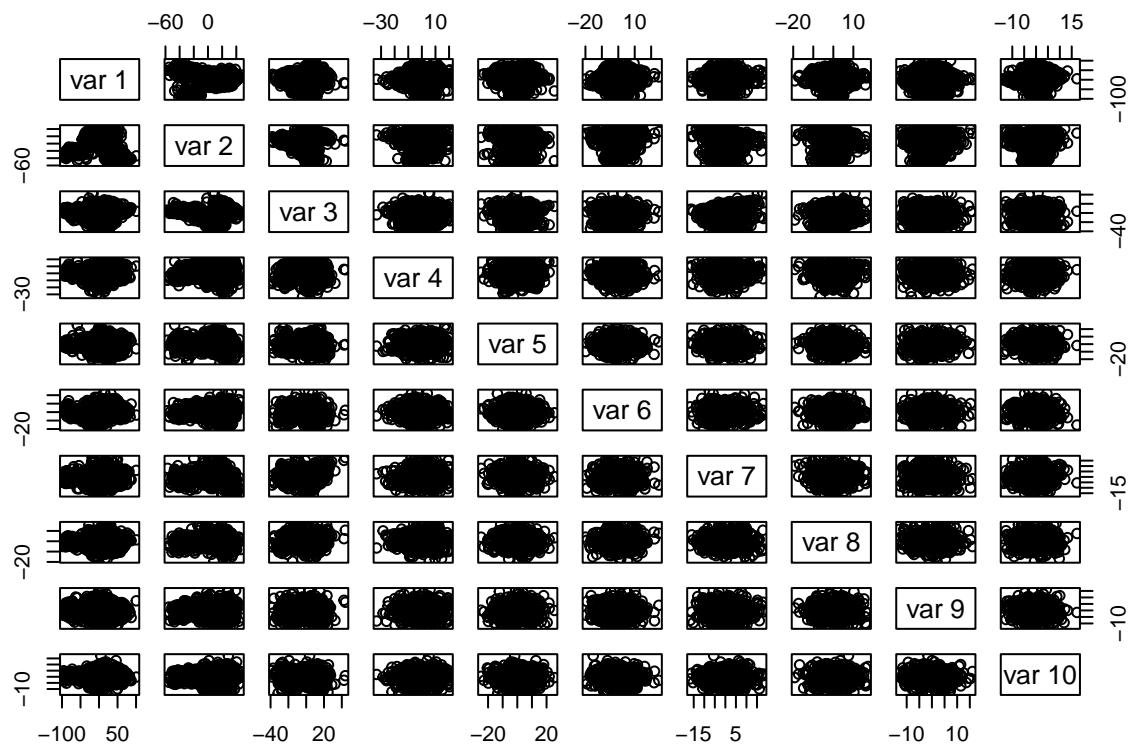


```
## [1] "done"
```

```
plot(phon.pca10$scores)
```



```
pairs(phon.pca10$scores)
```

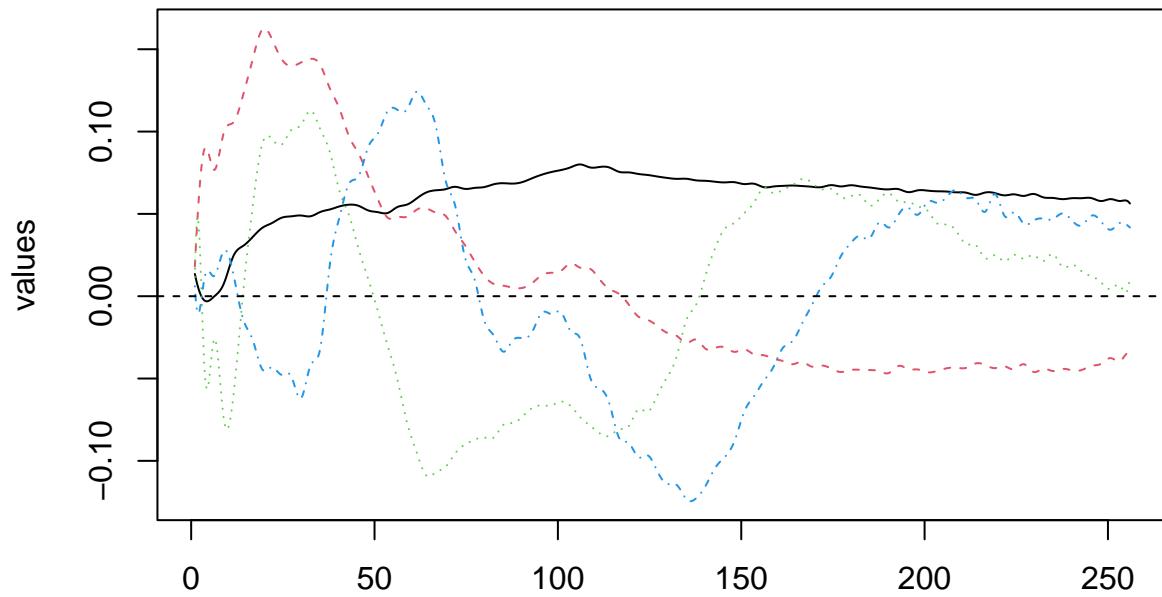


```
cumsum(phon.pca$varprop)
```

```
## [1] 0.5763016 0.7799804 0.8390231 0.8603975 0.8747653 0.8868737 0.8963245
## [8] 0.9047813 0.9114037 0.9165083
```

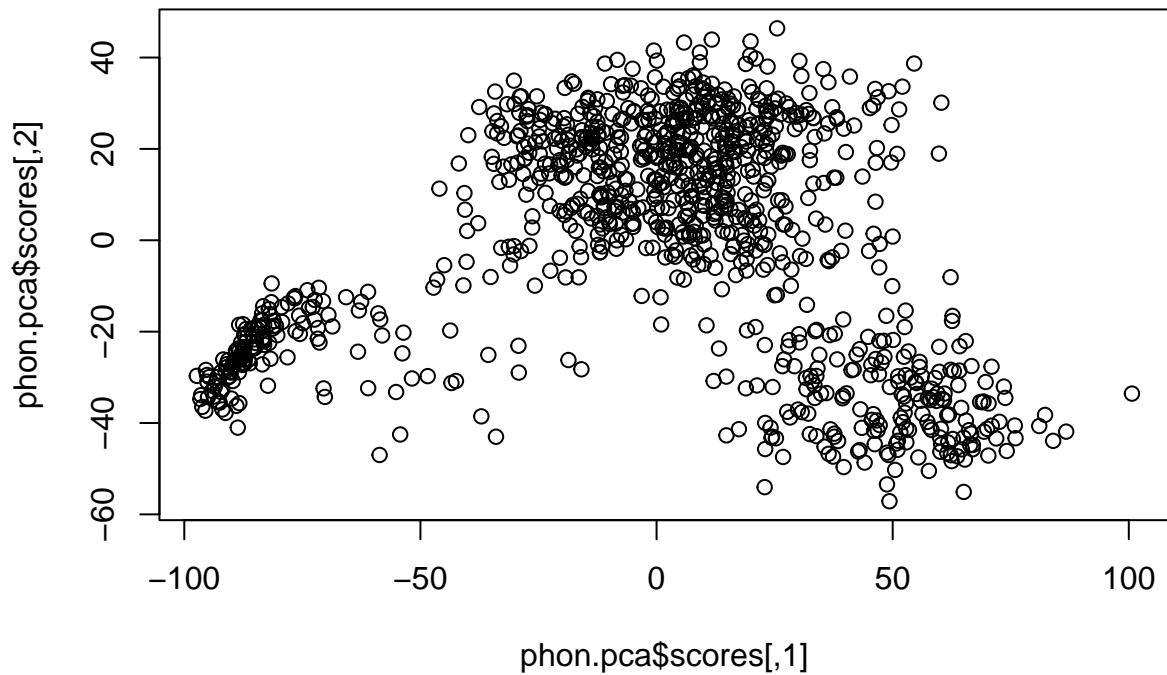
4 principal components are sufficient to explain the 86% of the total variance.

```
phon.pca <- pca.fd(fd.phon, nharm = 4)
plot(phon.pca$harmonics)
```

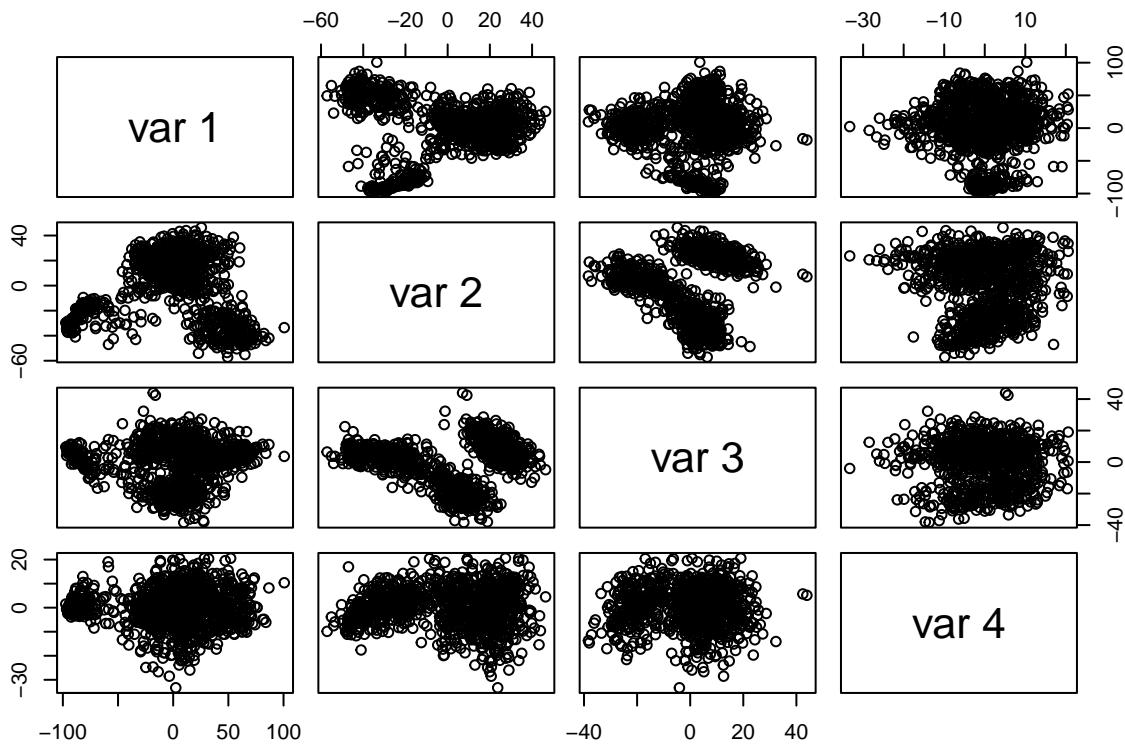


```
## [1] "done"
```

```
plot(phon.pca$scores)
```



```
pairs(phon.pca$scores)
```



```

mean.phon<-mean.fd(fd.phon)

approx<-phon.pca$harmonics
i <- 1
pcacoefi <- phon.pca$harmonics$coefs %*% phon.pca$scores[,]+mean.phon$coefs
approx$coefs <- pcacoefi

for (i in 2:256){
  pcacoefi <- phon.pca$harmonics$coefs %*% phon.pca$scores[i,]+mean.phon$coefs
  approx$coefs <- cbind(approx$coefs, pcacoefi)
}
#dimnames(approx$coefs)[[2]]<-phonemes256[,1]
#plotfit.fd(t(phonemes256),1:ncol(phonemes256),approx,index=30,cex.pch=0.5)
#plotfit.fd(t(phonemes256),1:ncol(phonemes256),approx,index=130,cex.pch=0.5)

```

Run a funFEM-clustering,

```

set.seed(1234)
femmodels <- c("DkBk", "DkB", "DBk",
"DB", "AkjBk", "AkjB", "AkB", "AkBk", "AjBk", "AjB", "ABk",
"AB")
nmodels <- length(femmodels)
femresults <- list()
bestk <- bestbic <- numeric(0)

```











































































```

## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN
## Warning in log(det(D[k, (1:d), (1:d)])): Si è prodotto un NaN

## [1] "DBk"
## [1] "DB"
## [1] "AkjBk"
## [1] "AkjB"
## [1] "AkB"
## [1] "AkBk"
## [1] "AjBk"
## [1] "AjB"
## [1] "ABk"
## [1] "AB"

besti <- which(bestbic==max(bestbic,na.rm=TRUE))
femresults[[11]]$K

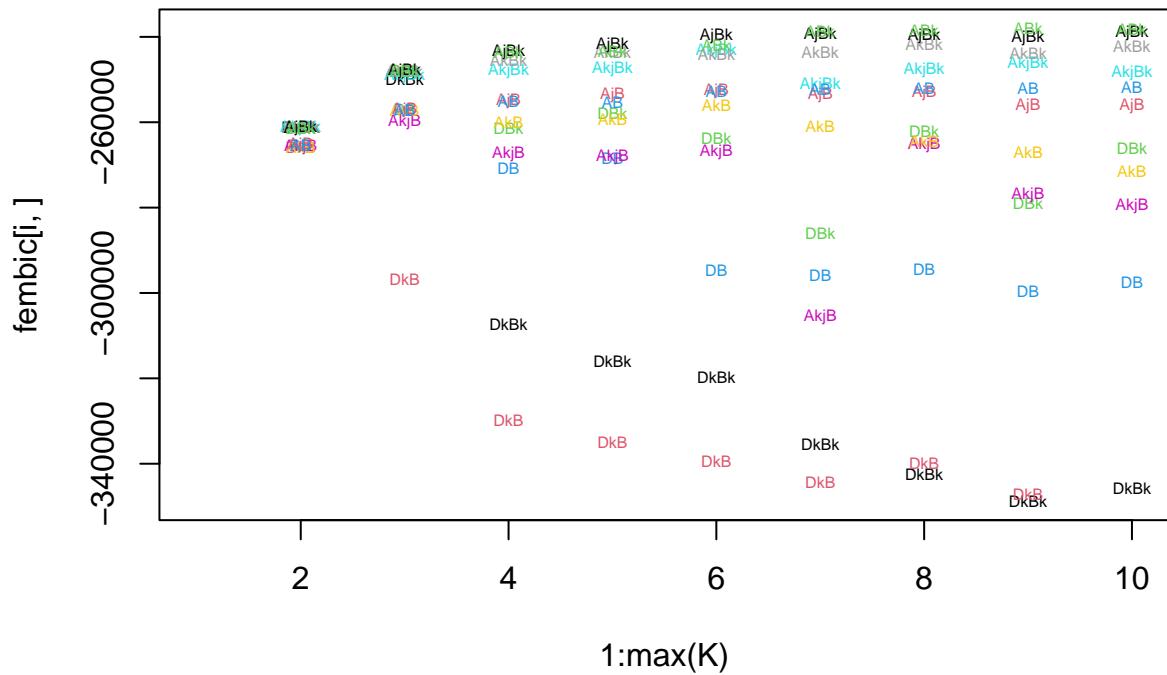
## [1] 9

femresults11 <- femresults[[11]]
table(femresults11$cls)

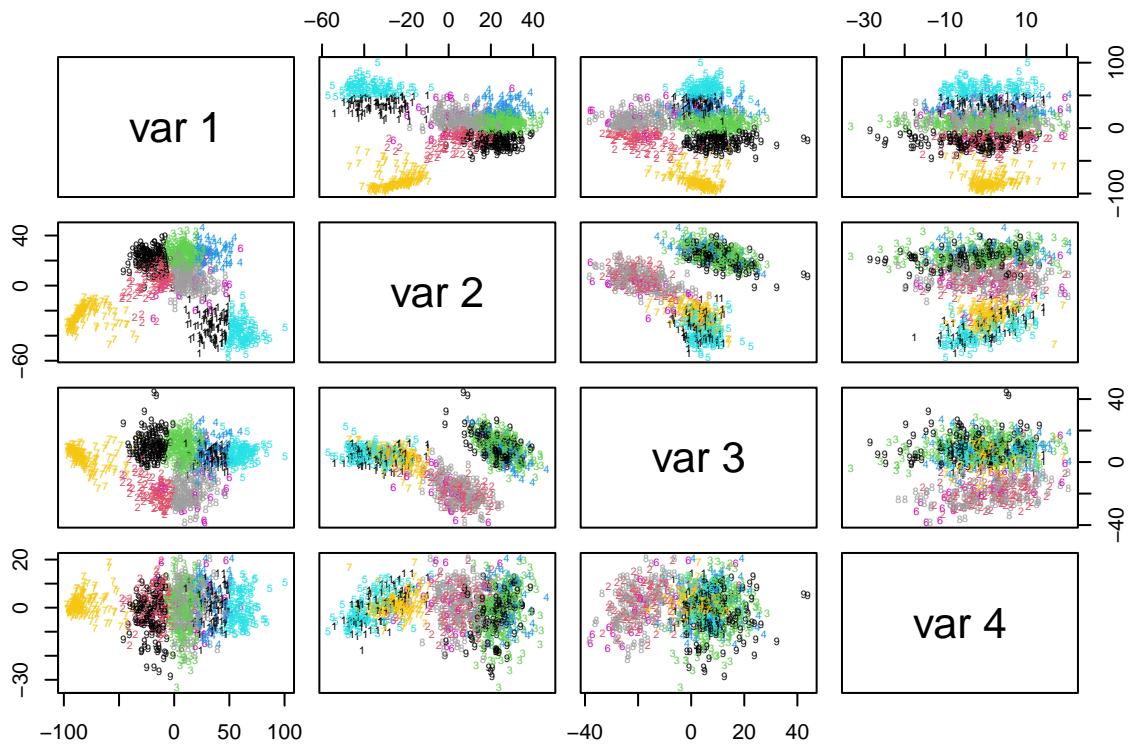
##
##    1    2    3    4    5    6    7    8    9
##   88   95  181   73  109   31  148  150  125

i <- 1
plot(1:max(K),fembic[i,],col=i,pch=i,
      ylim=c(min(fembic,na.rm=TRUE),max(fembic,na.rm=TRUE)),type="n")
for(i in 1:nmodels){
  text(1:max(K),fembic[i,],femmodels[i],col=i, cex=0.5)
}

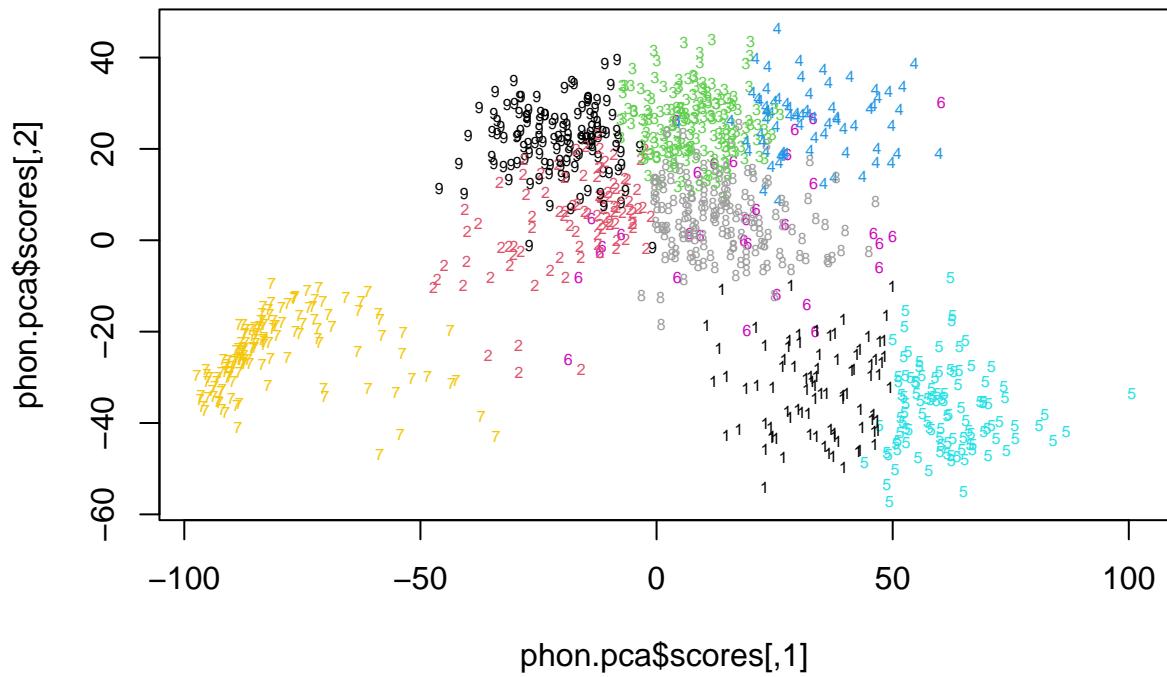
```



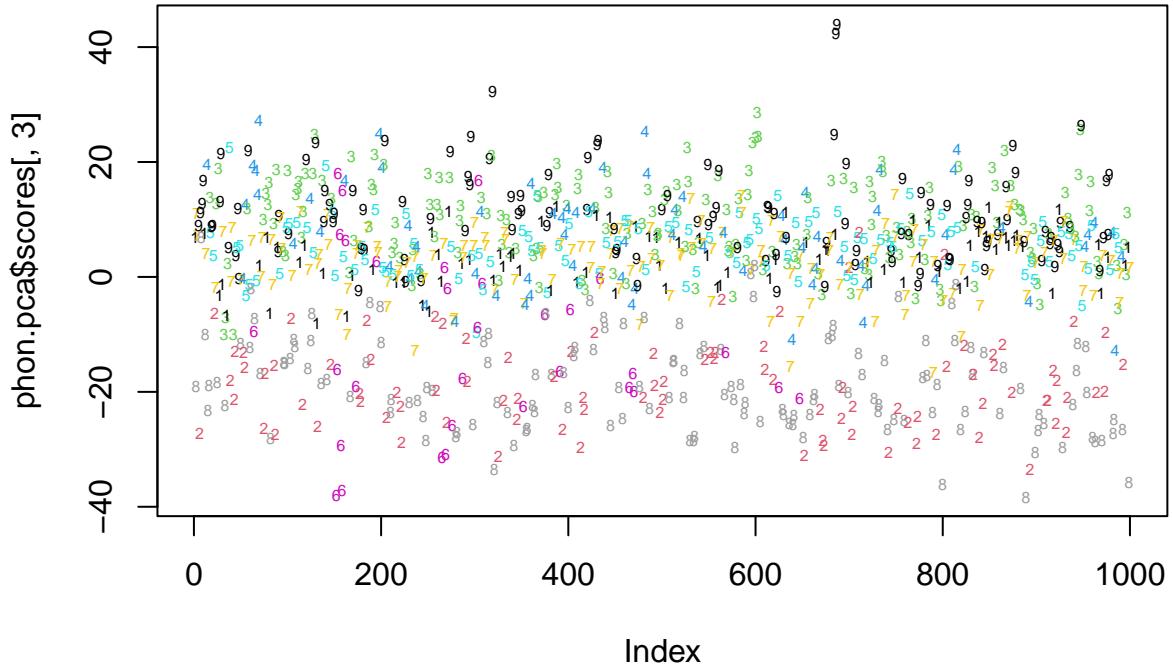
```
pairs(phon.pca$scores, col=femresults11$cls, pch=clusym[femresults11$cls], cex = 0.5)
```



```
plot(phon.pca$scores, col=femresults11$cls, pch=clusym[femresults11$cls], cex = 0.5)
```



```
plot(phon.pca$scores[,3], col=femresults11$cls, pch=clusym[femresults11$cls], cex = 0.5)
```



From the plot we can see that the model which provides the highest BIC is the ABk. The plot of first two principal components is the best in terms of clusters separation while the plot of 3 and 4 principal components is the worst.

and for comparison run a cluster analysis of your choice on the functional principal component scores.

```

gapnc<- function(data,FUNcluster=kmeans,
                  K.max=10, B = 100, d.power = 2,
                  spaceH0 ="scaledPCA",
                  method ="globalSEmax", SE.factor = 2,...){
  gap1 <- clusGap(data,kmeans,K.max, B, d.power,spaceH0,...)
  nc <- maxSE(gap1$Tab[,3],gap1$Tab[,4],method, SE.factor)
  kmopt <- kmeans(data,nc,...)
  out <- list()
  out$gapout <- gap1
  out$nc <- nc
  out$kmopt <- kmopt
  out
}

set.seed(1234)
phon.gap <- gapnc(phon.pca$scores)
table(phon.gap$kmopt$cluster)

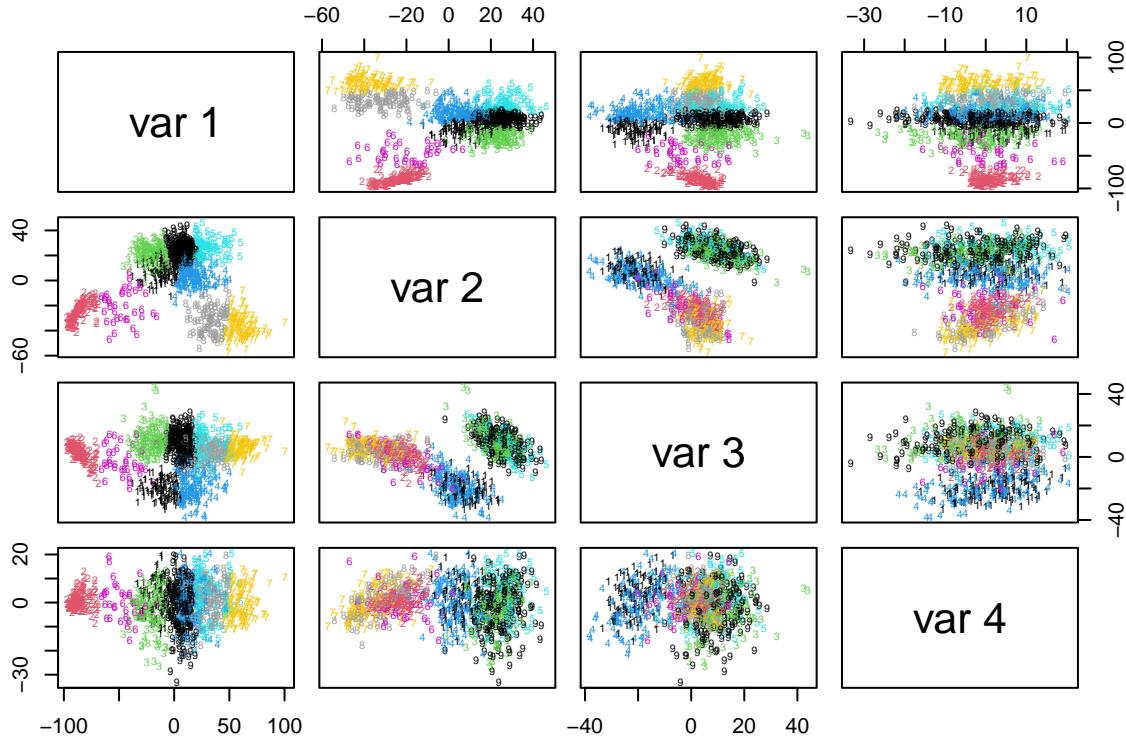
```

```

## 
##   1   2   3   4   5   6   7   8   9
## 111 127 124 138  91  37 115  86 171

```

```
pairs(phon.pca$scores, col=phon.gap$kmopt$cluster, pch=clusym[phon.gap$kmopt$cluster], cex = 0.5)
```



9 clusters are found. Here the majority of units are classified in the first and last cluster, while before in the central clusters. As previous the plot referred to principal components 3 and 4 is so confused and clusters are not well separated.

Visualise your results suitably, and compare the clustering results with the true phonem classes.

```
adjustedRandIndex(femresults11$cls, phonemes257)
```

```
## [1] 0.5304411
```

```
adjustedRandIndex(phon.gap$kmopt$cluster, phonemes257)
```

```
## [1] 0.5201529
```

The ARI values are not so good, the clustering is not very different from the truth but their are not so sufficiently similar.

3)

- (a) Explain in your own words what discriminant coordinates and asymmetric weighted discriminant coordinates are, and how they work.

```
?plotcluster
```

```
## avvio in corso del server httpd per la guida ... fatto
```

1)The Dc method regards projection of high dimensional data with a given grouping to a lower dimensional subspace, it implicitly assumes that all classes have the same covariance matrix. The difference between the classes are measured as differences between class means. The within groups cov matrix must be an adequate measure for the cov structure within all classes, which holds under approximate equality of their cov matrices. Under this assumption the within cov matrix of the projected data is the identity, so projected groups appear spherical. When there is only one Hclass and others are merged in one class, only first DC is informative, so there is an extension of this technique for s=2 that change the covariance structure by choosing further projections orthogonal to the DC in order to maximise the difference in the projected covariance matrices. They are SYMMETRIC so invariant with respect to the numbering of classes.

2)With weighted asymmetric discriminant coordinates, we deal with class treated as homogeneous by an asymmetric projection method while the other may be heterogeneous. It is an improvement of a direct asymmetrisation of DC with respect to objects in the less homogeneous class (Nclass). This class often contains the heaviest outliers. The idea is to weight the points of the Nclass according to their mahalanobis distance to the Hclass in order to deal with the point in Nclass close to Hclass. The ratio between the projection of a between classes separation matrix and the projection of the covariance matrix within the homogeneous class, has to be maximised.

- (b) For the optimal 10-clusters Gaussian mixture clustering of the olive oil data show 2-dimensional discriminant coordinates, and asymmetric weighted discriminant coordinates for all clusters.

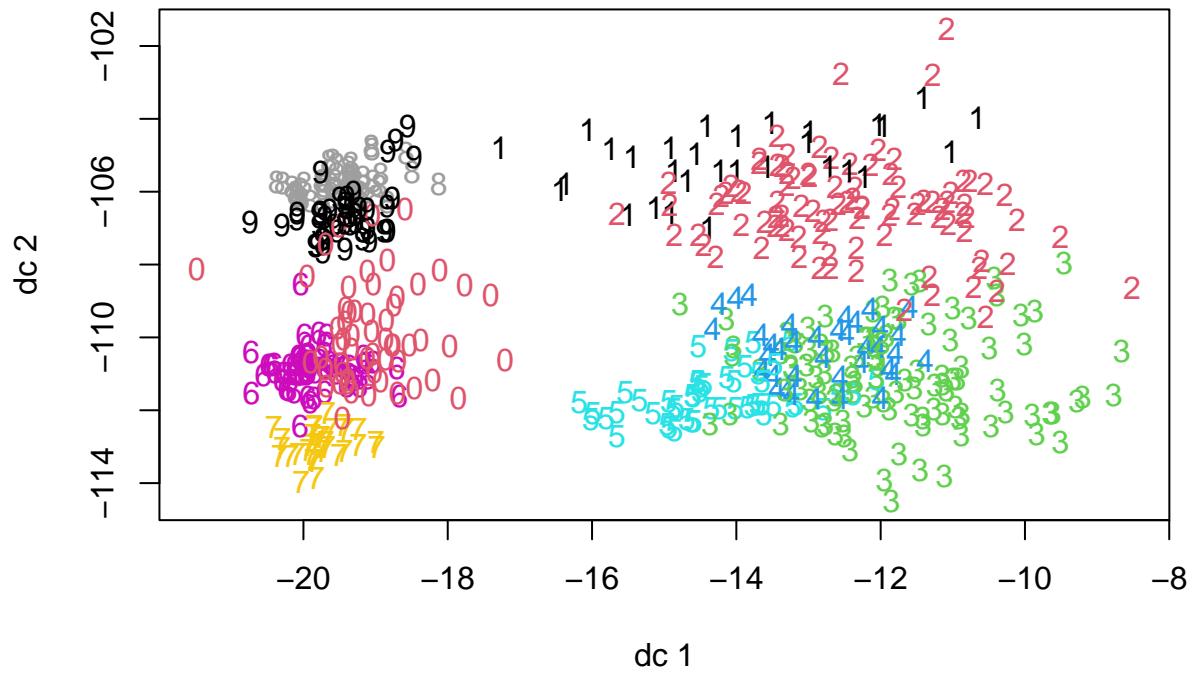
```
data("oliveoil")
olive<-oliveoil[,3:10]
set.seed(1234567)
molute<-Mclust(olive, G=10)
summary(molute)
```

```
## -----
## Gaussian finite mixture model fitted by EM algorithm
## -----
## 
## Mclust VVE (ellipsoidal, equal orientation) model with 10 components:
## 
##   log-likelihood   n   df       BIC       ICL
##             -20448.03 572 197 -42146.84 -42193.07
## 
## Clustering table:
##   1   2   3   4   5   6   7   8   9   10
##  30  96 110  37  50  64  34  49  45  57
```

```
molute$G
```

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

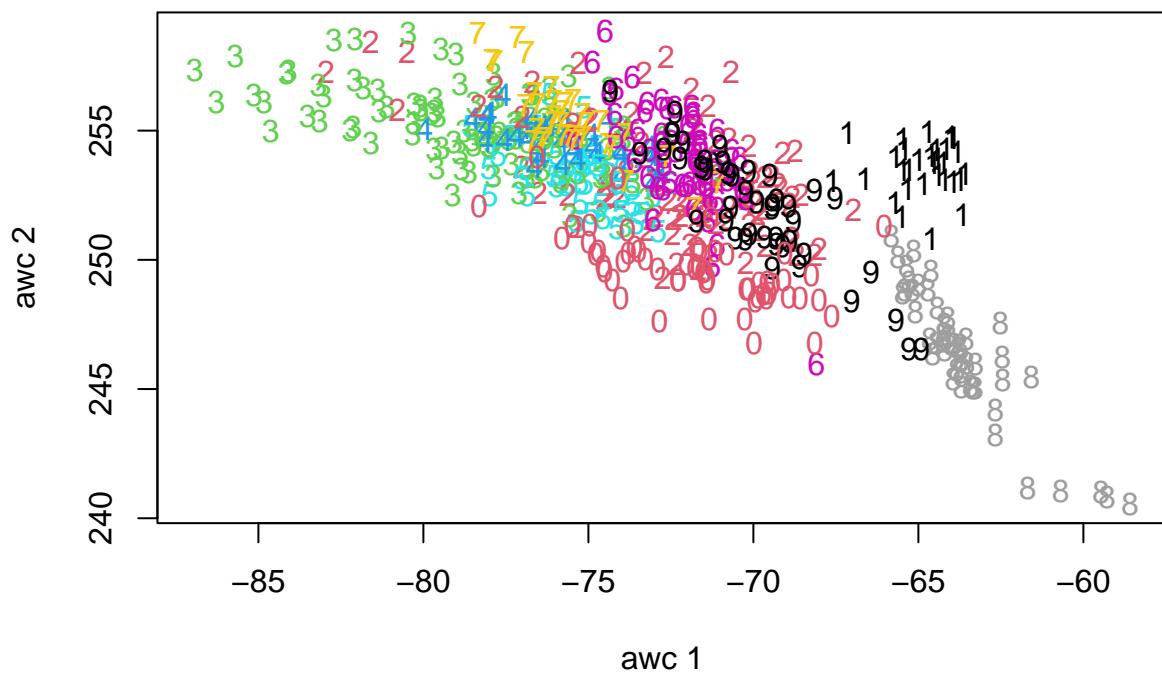
```
clvecd<-molute$classification
plotcluster(olive,clvecd, method = "dc", col = molute$classification, pch=clusym[molute$classification])
```

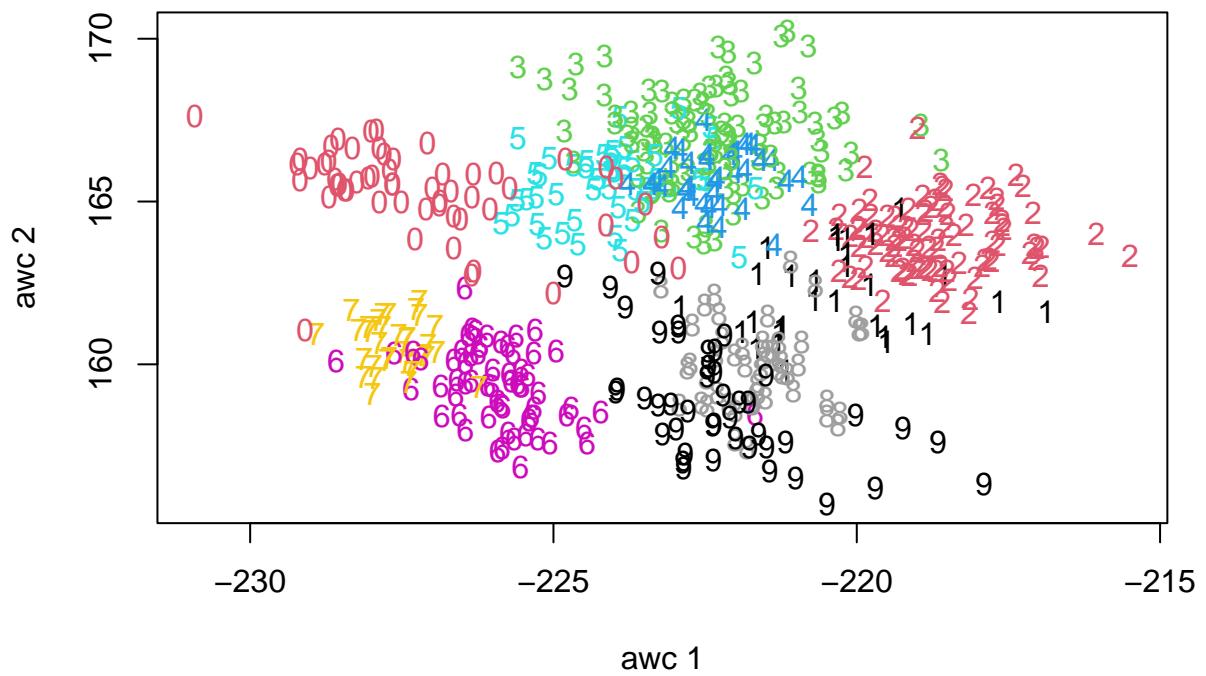


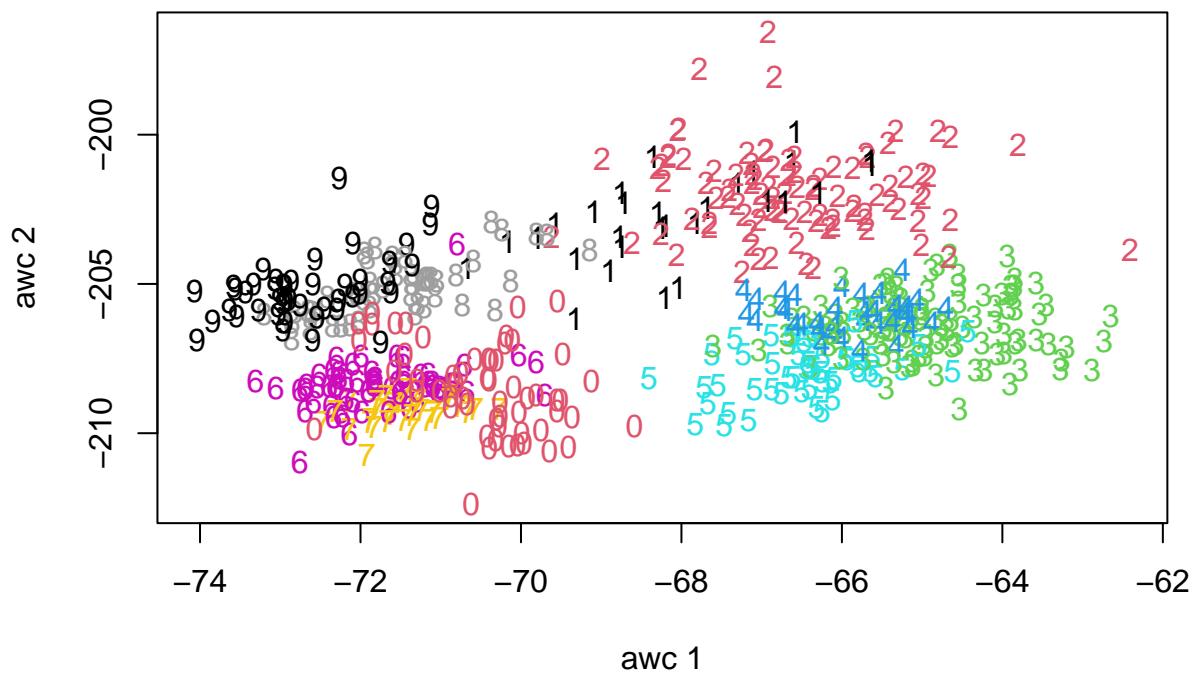
```

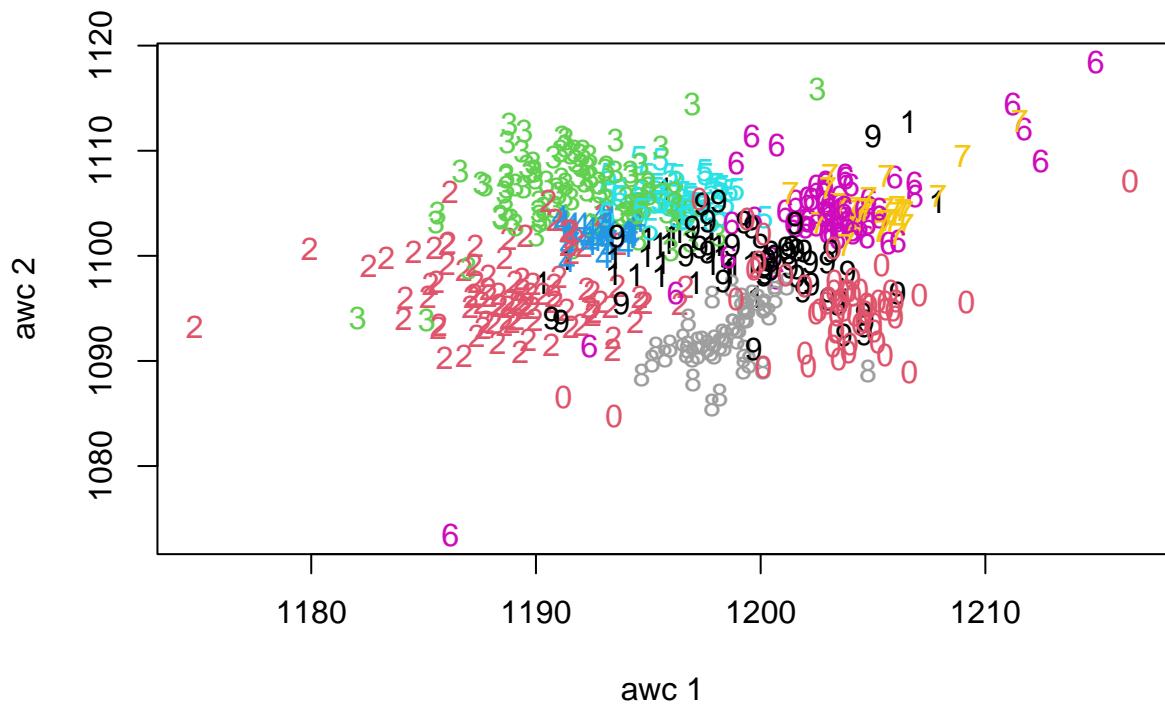
for (i in 1:10) {
  plotcluster(olive, clvecd, clnum = i, method = "awc", col = m olive$classification, pch=clusym[m olive$]
}

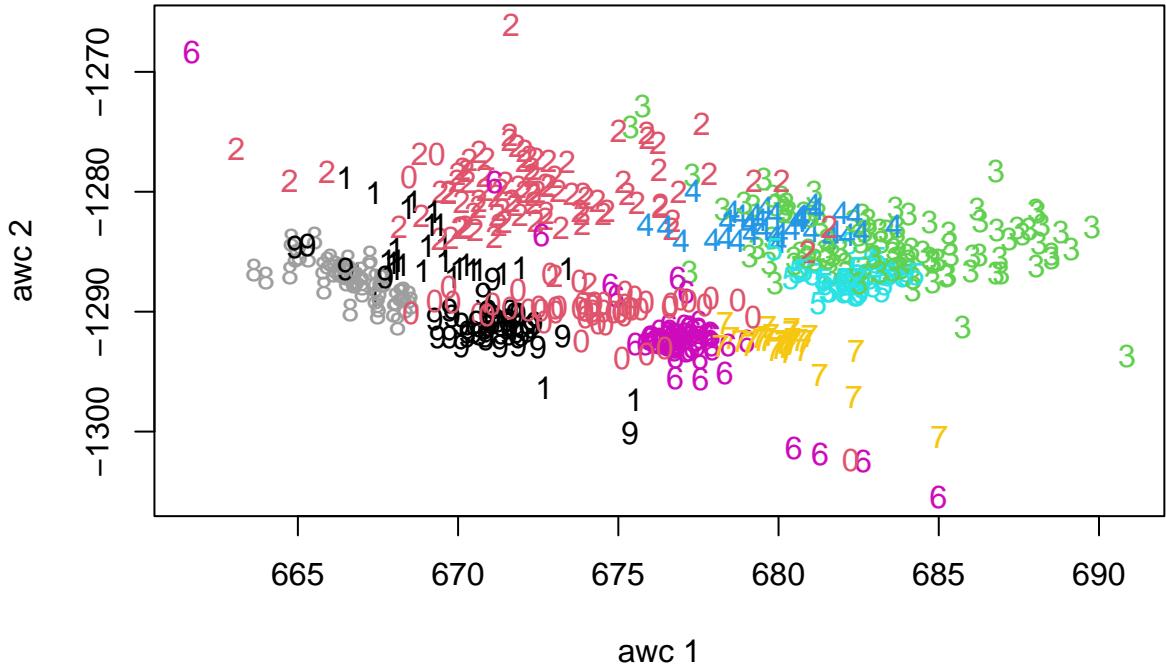
```

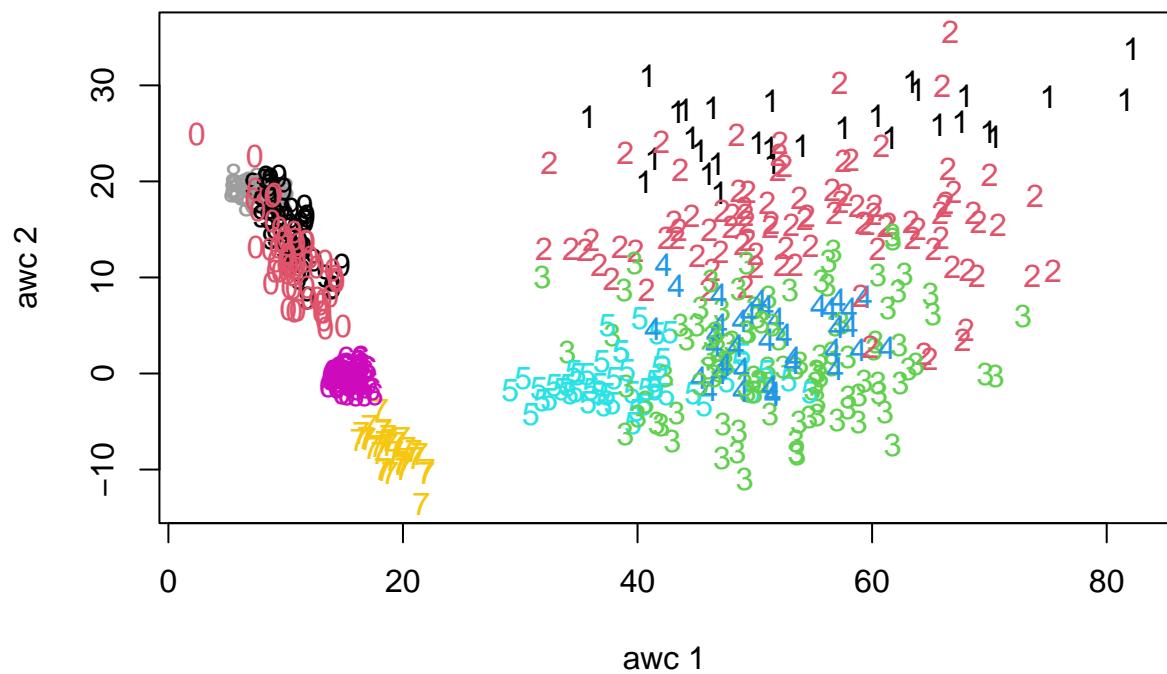


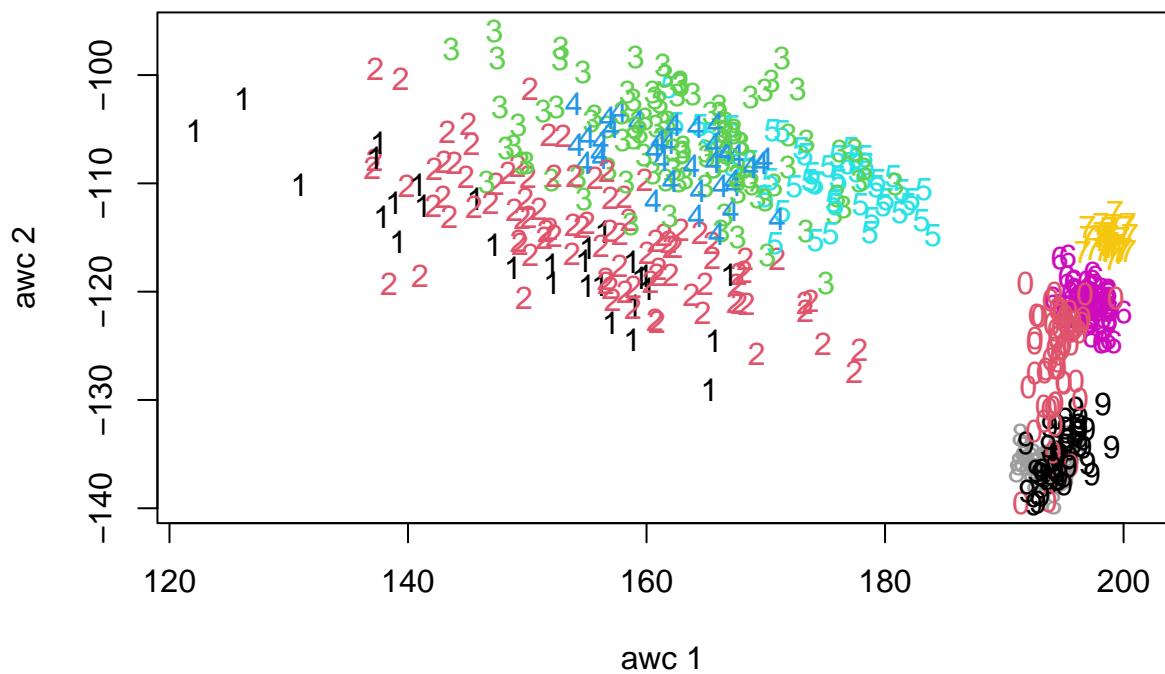


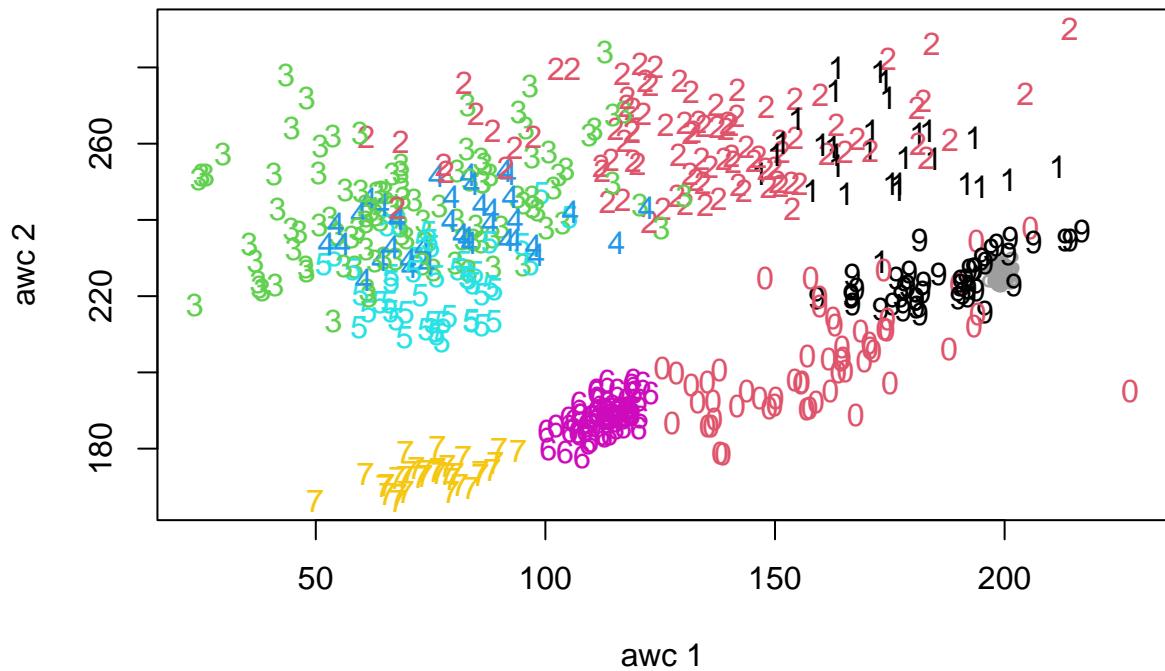


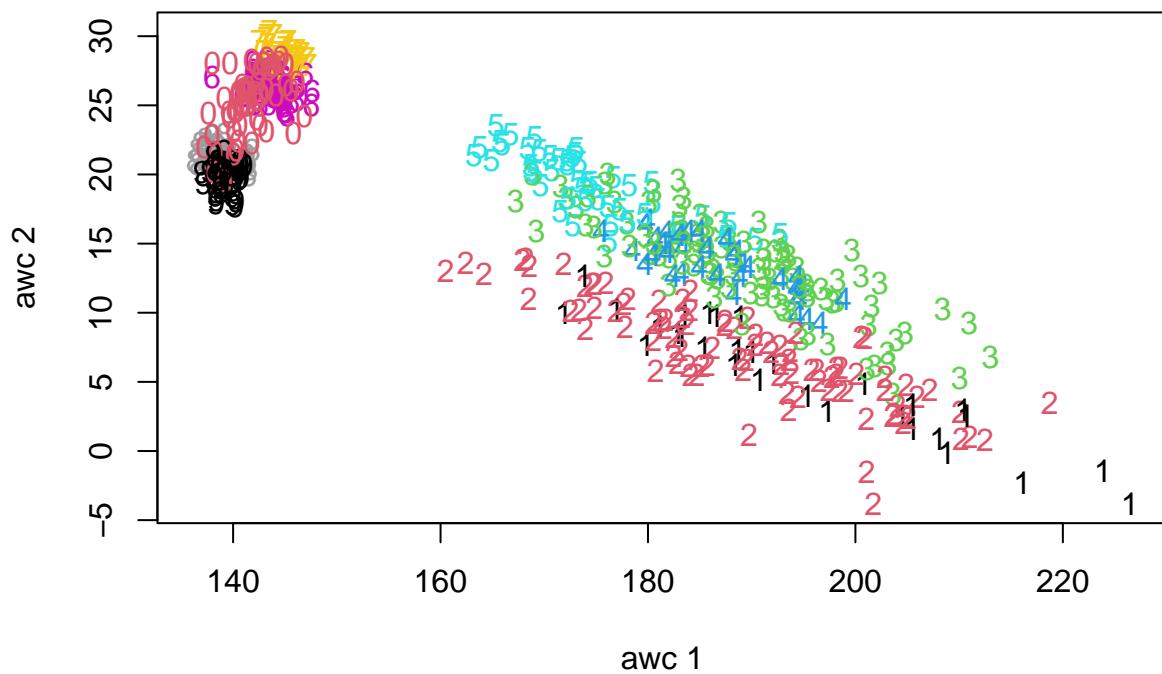


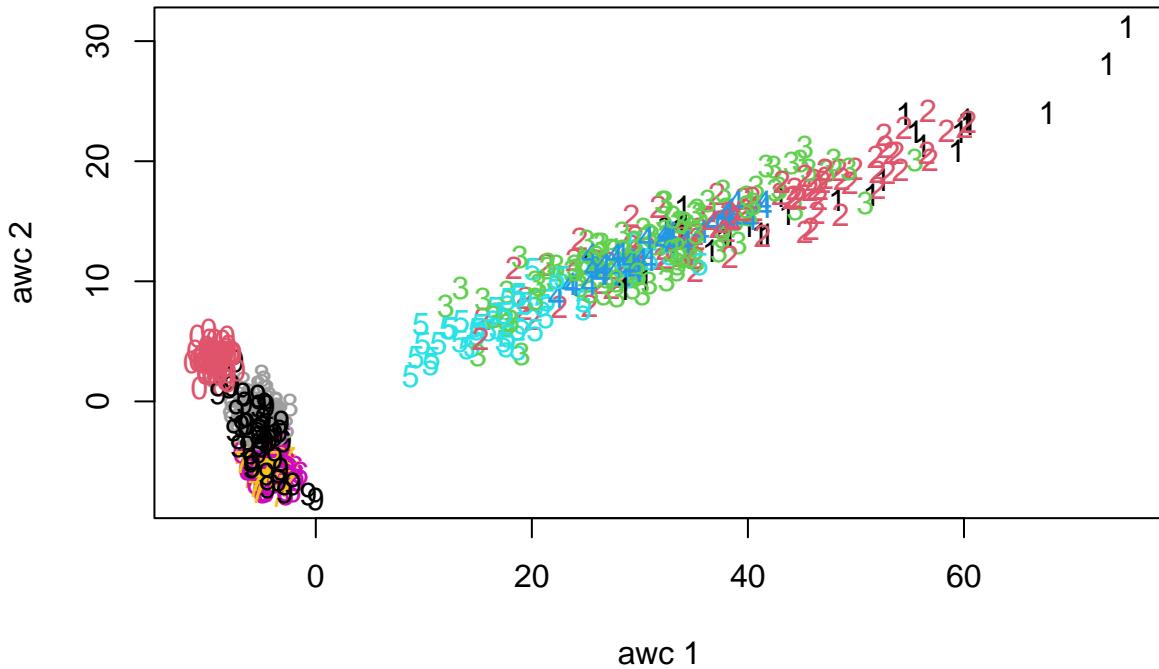












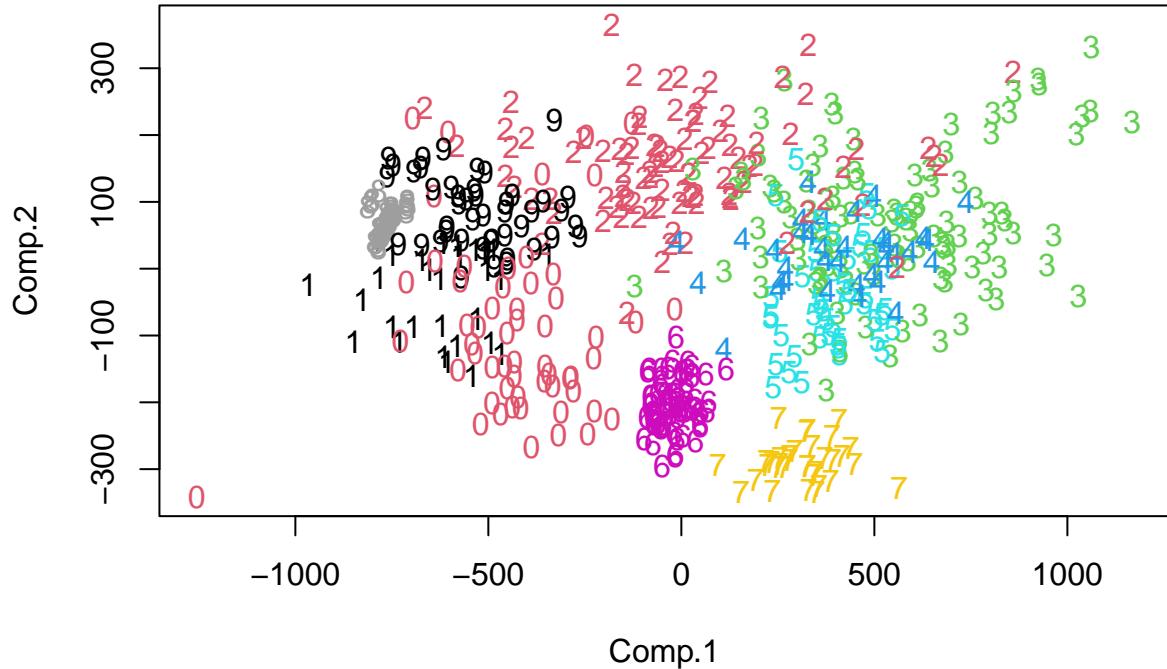
The plot that represents point using the first two discriminant coordinates does not provide a good separation of clusters. The plots of weighted asymmetric discriminant coordinates are drawn considering different numbers of clusters considered as homogeneous. Maybe the best clusters separation is provided by treating 6 or 7 clusters as homogeneous.

Comment on how these plots compare to the principal components plot in terms of showing the separation of the clusters.

```
?princomp
n=nrow(olive)
pcaolive <- princomp(olive)
summary(pcaolive)

## Importance of components:
##              Comp.1      Comp.2      Comp.3      Comp.4
## Standard deviation   479.7299024 150.82827868 45.394449751 27.522646558
## Proportion of Variance  0.8970072  0.08866821  0.008031707  0.002952451
## Cumulative Proportion  0.8970072  0.98567544  0.993707152  0.996659603
##                         Comp.5      Comp.6      Comp.7      Comp.8
## Standard deviation    24.78169442 1.196956e+01 7.1390744088 6.9756965249
## Proportion of Variance  0.00239367 5.584168e-04 0.0001986489 0.0001896608
## Cumulative Proportion  0.99905327 9.996117e-01 0.9998103392 1.00000000000
```

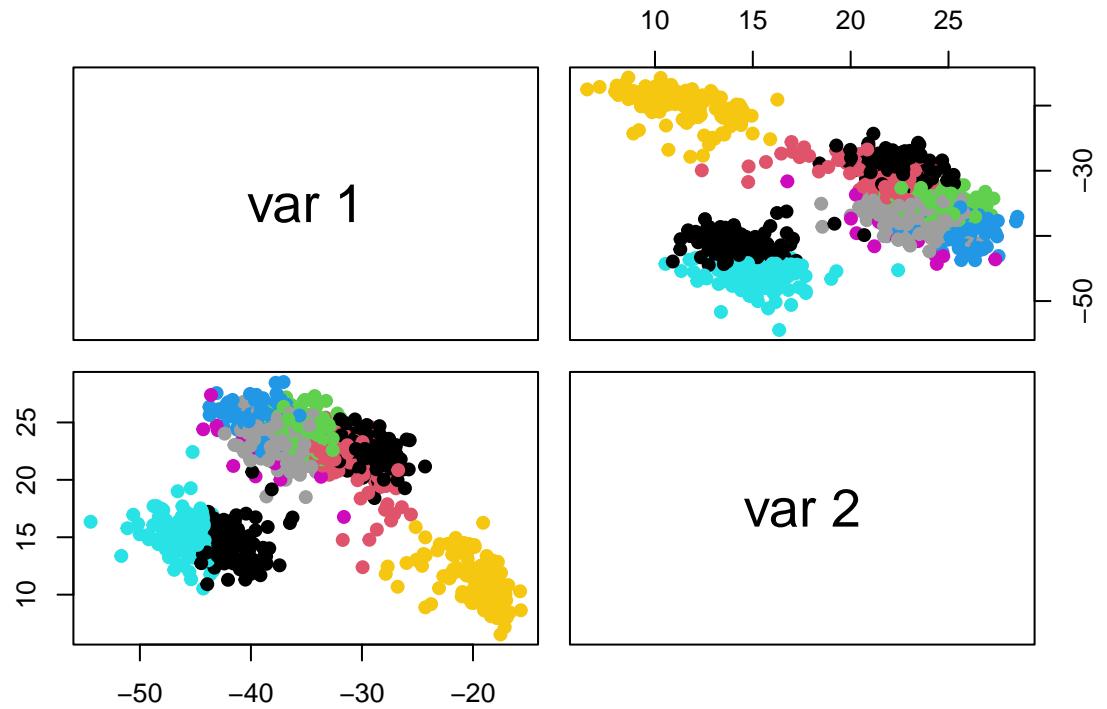
```
plot(pcaolive$scores[,1:2], col=molive$classification, pch=clusym[molive$classification])
```



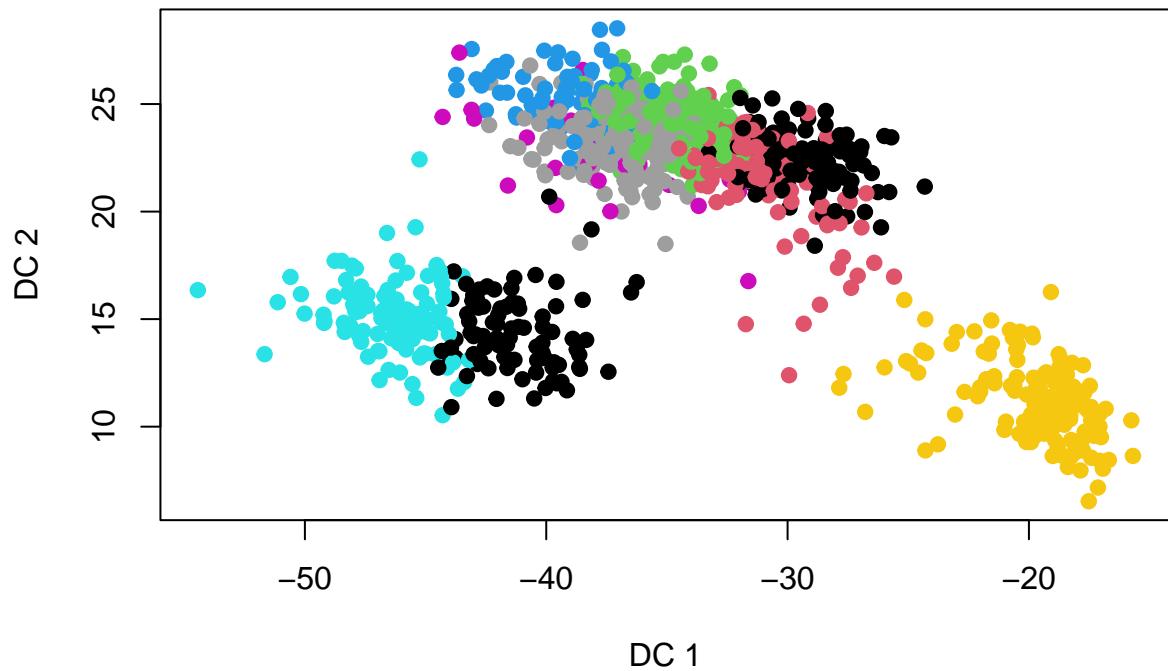
The plot of showing the points projected in the first two principal components is a little bit confused but it seems similar to the plot of awd with two clusters treated as homogeneous. The best method seems to be the one with first 2 discriminant coordinates.

- (c) Considering the phoneme data from question 1 and the funFEM-clustering, compare the plot of the first two dimensions of the Fisher discriminating subspace from funFEM with what you get when applying discriminant coordinates using plotcluster to the data set of the coefficients of the full dimensional B-spline basis and the funFEM-clustering.

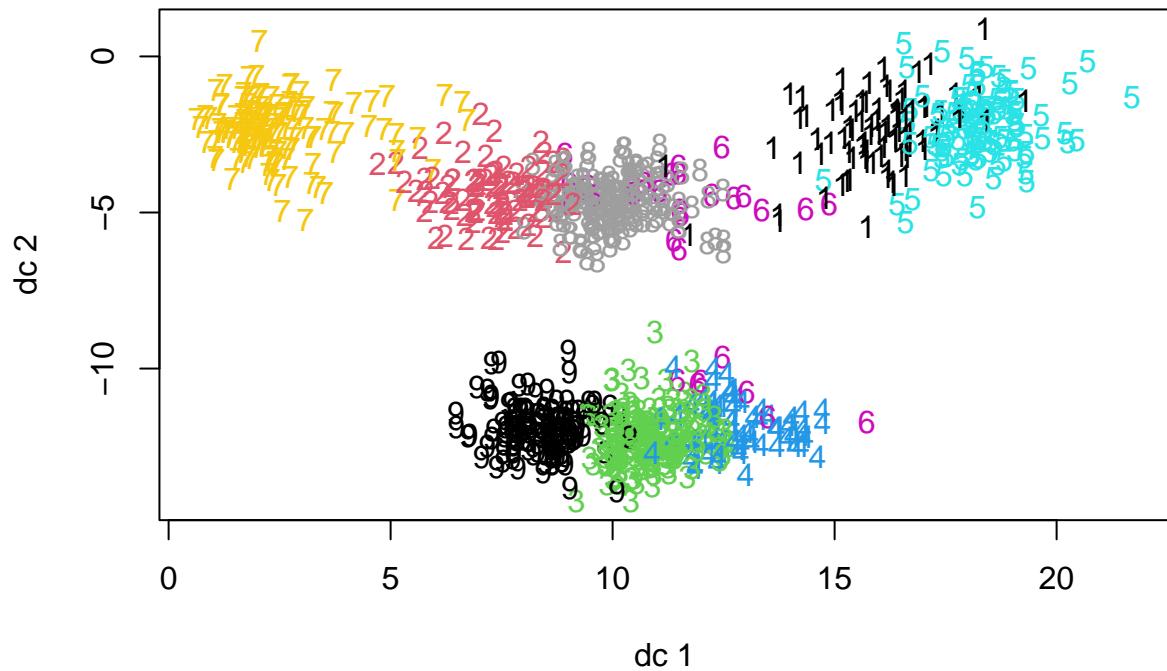
```
fdproj <- t(fd.phon$coefs) %*% femresults11$U[,1:2] # first two dimensions of the fisher discriminating subspace
pairs(fdproj, col=femresults11$cls, pch=19)
```



```
plot(fdproj, col=femresults11$cls, pch=19, xlab="DC 1", ylab="DC 2")
```



```
plotcluster(t(fd.phon$coefs), clvecd = femresults11$cls, method = "dc", col = femresults11$cls, pch=clus
```



The worst specified cluster is the sixth, but the others are quite good separated.