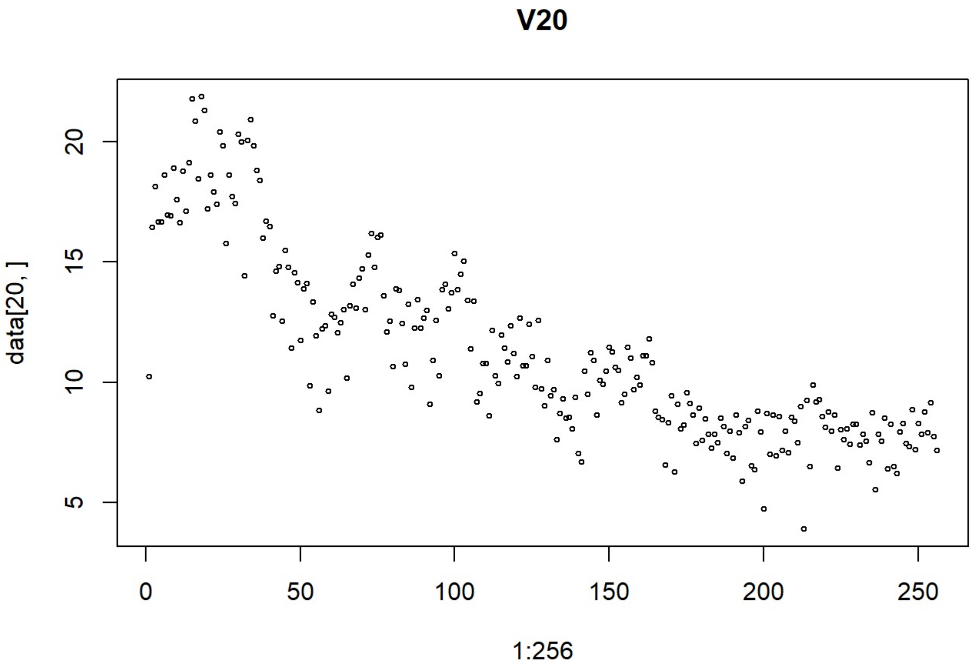
Exercise 8

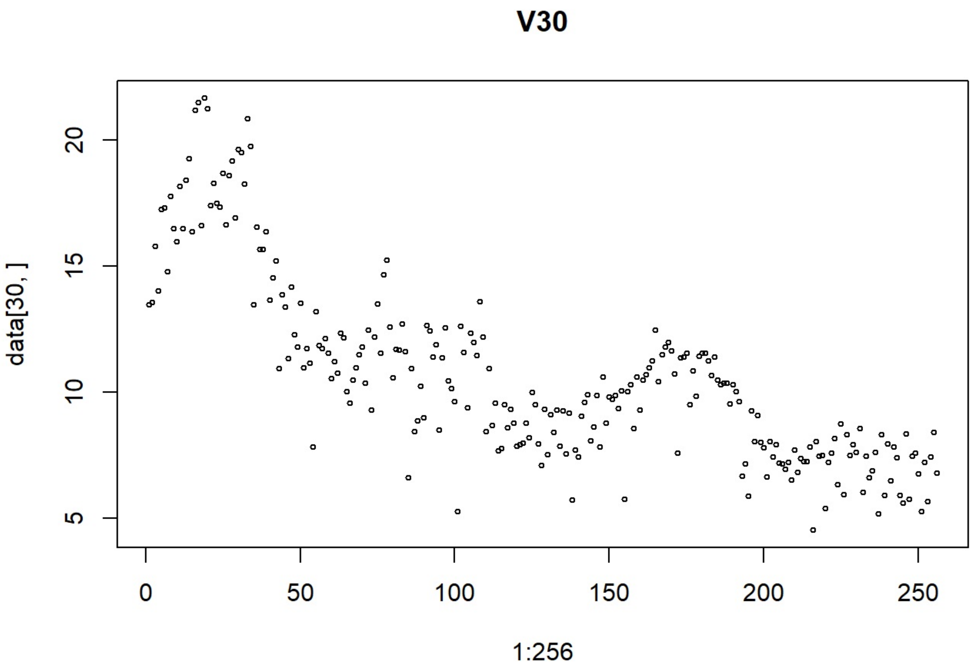
Jacopo Palombarini

2023-12-08

Exercise 1

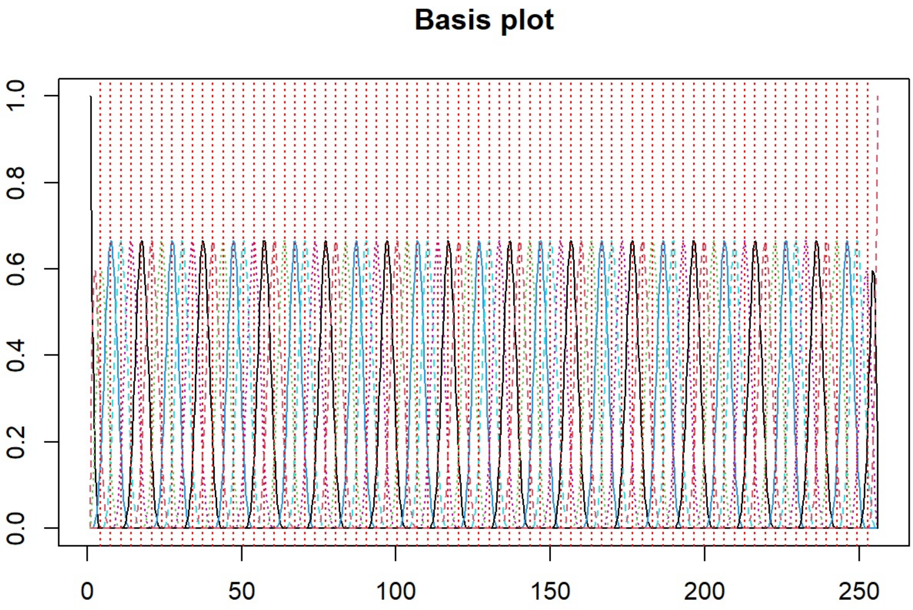
Representing data





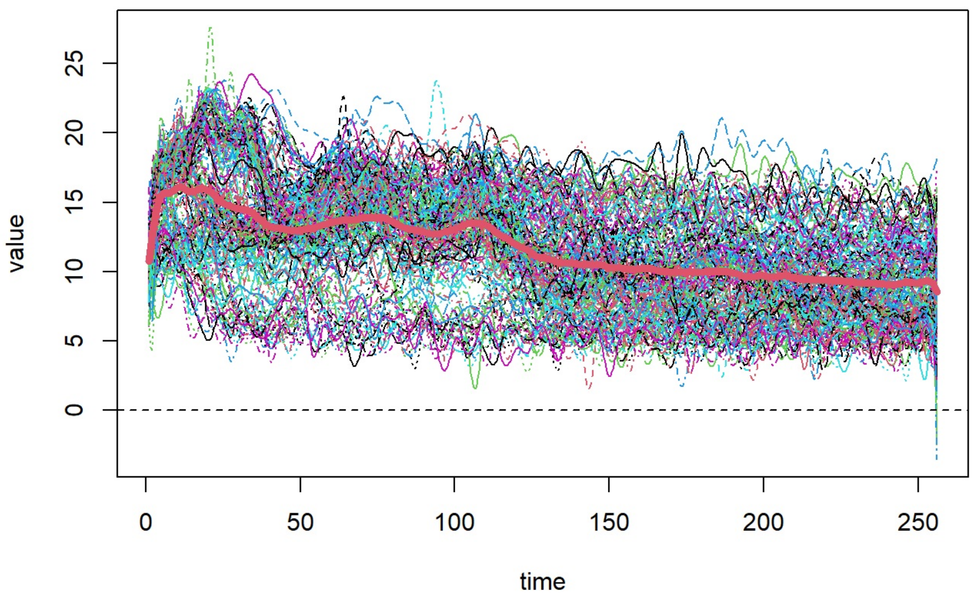
Represent the data in terms of a suitable B-spline basis.

Constructing B-spline basis

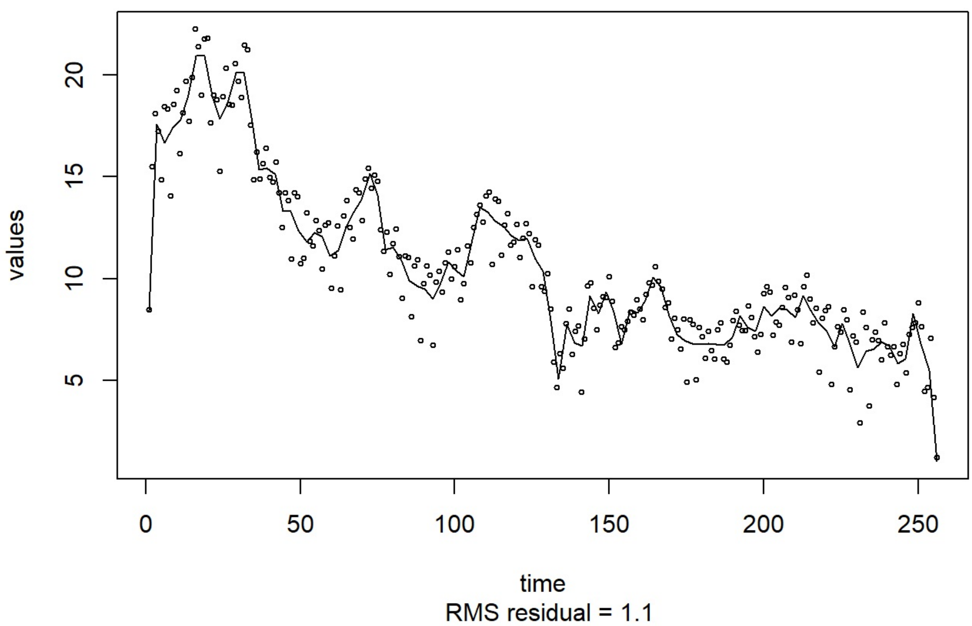



```
## [1] "done"
```

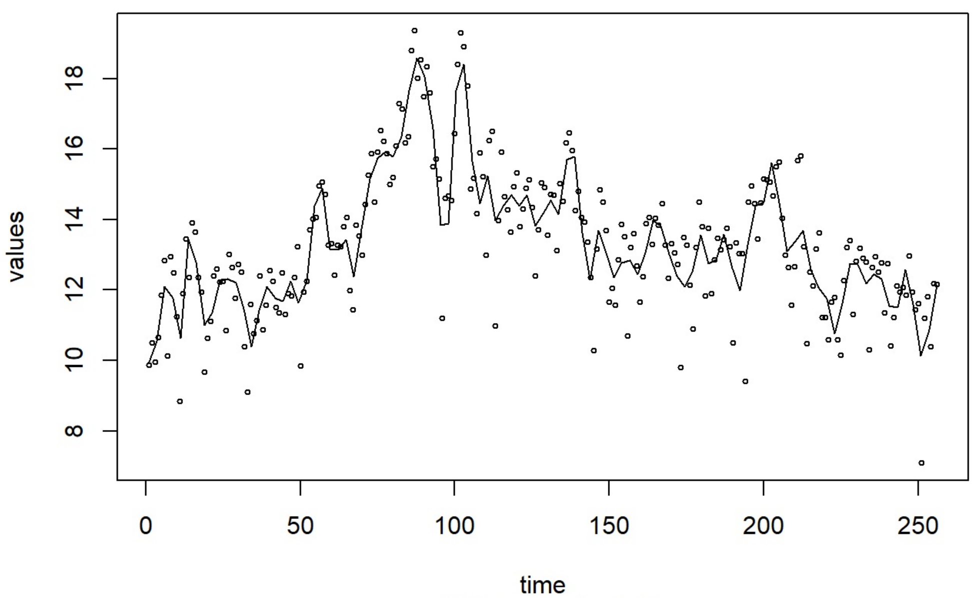
Smooth splines for data with smooth mean function



smooth fit of 20th obs.

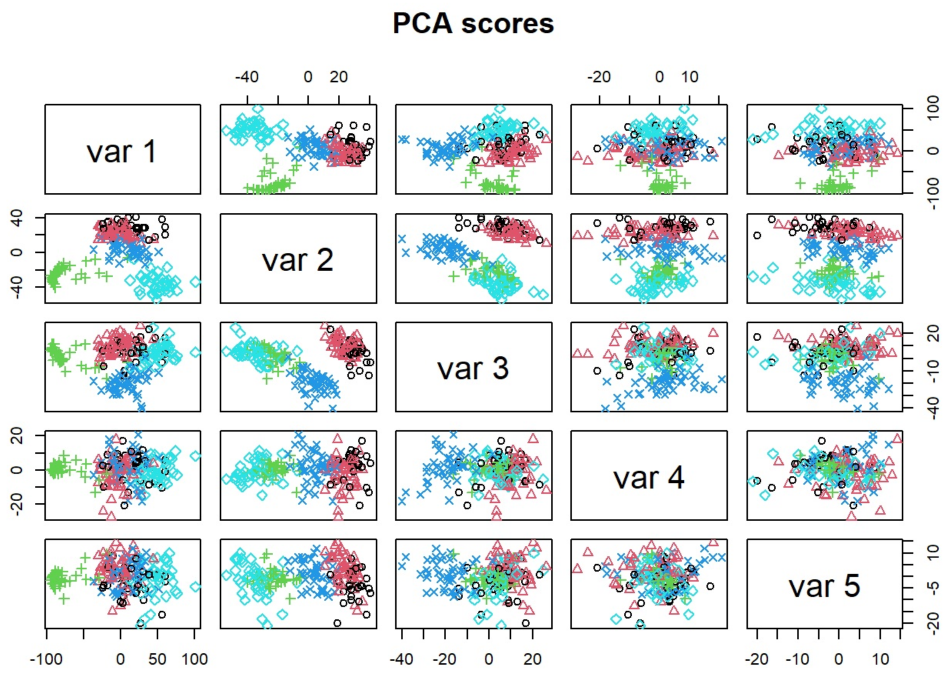


smooth fit of 30th obs.



RMS residual = 0.93

Run a functional principal component analysis



Create functional data object of PCA approximations

Run a funFEM-clustering

```
besti <- which(bestbic==max(bestbic,na.rm=TRUE))
besti
```

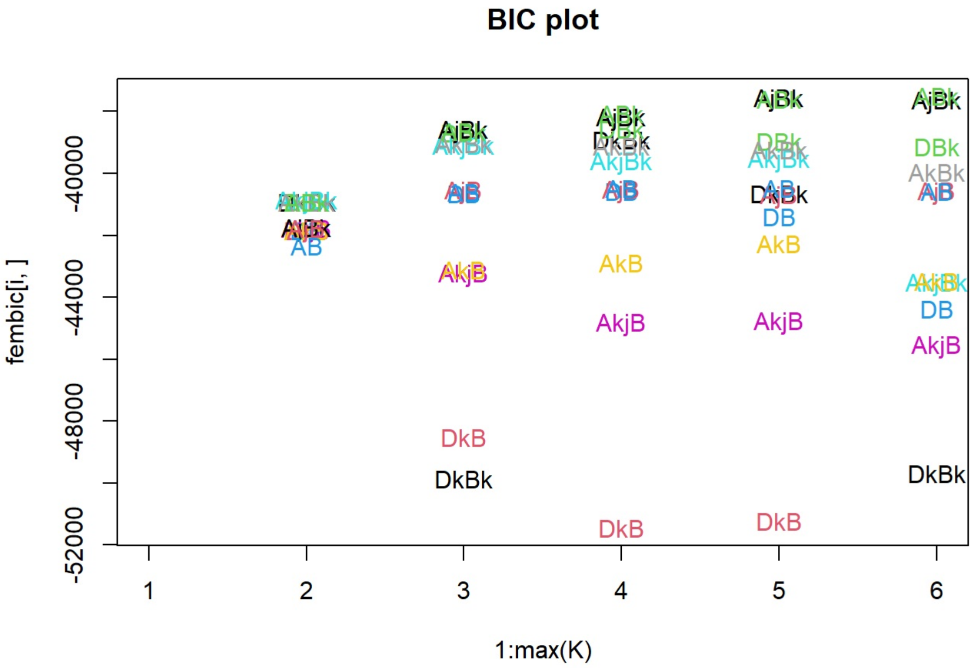
```
## [1] 11
```

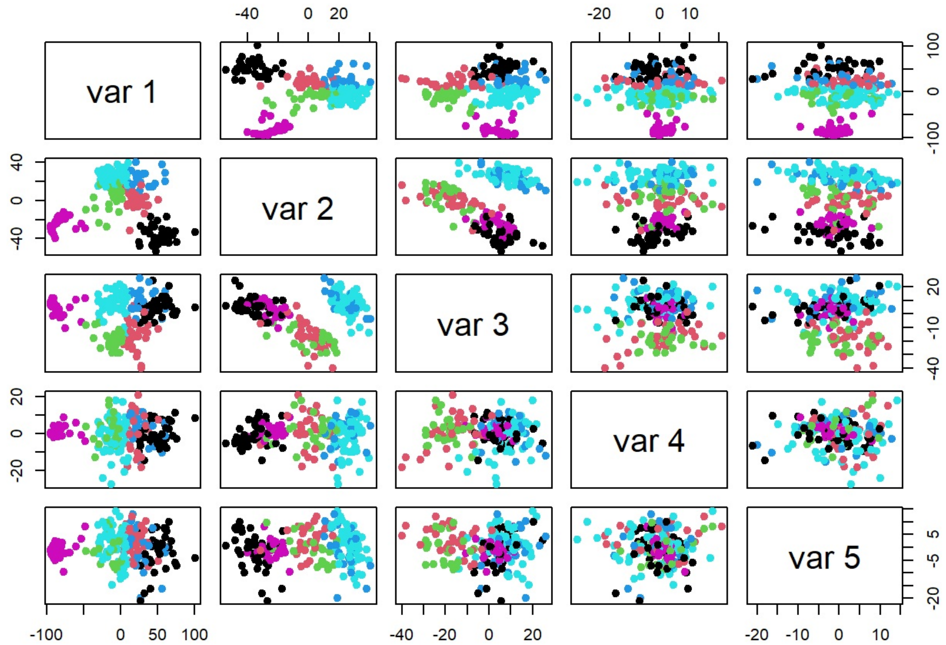
```
femmodels[besti]
```

```
## [1] "ABk"
```

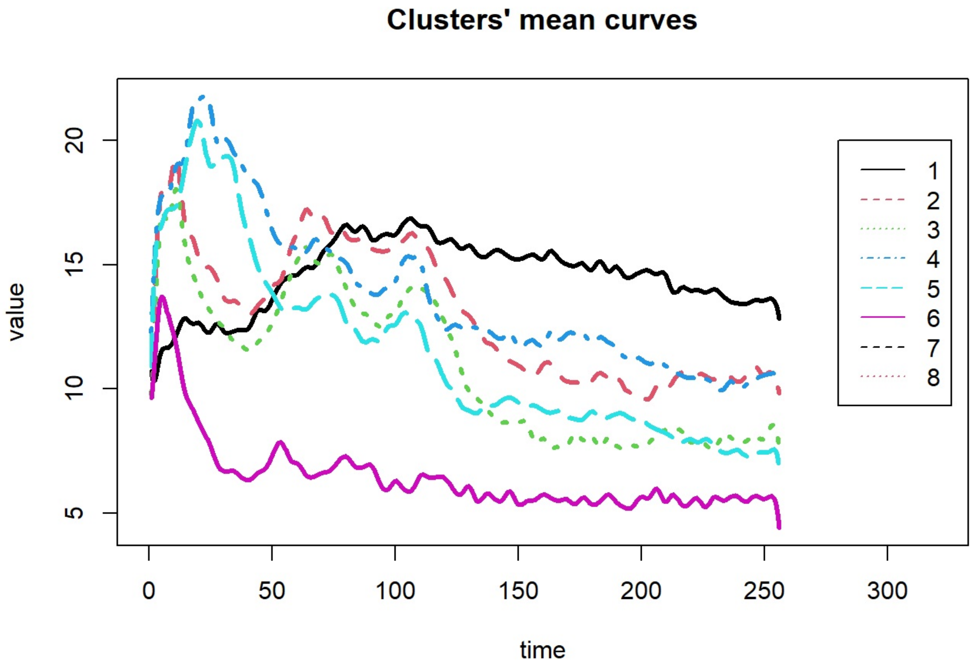
```
bestk # K=6 optimal for model 11 "ABk"
```

```
##  [1] 4 2 4 4 3 2 2 3 5 4 6 4
```

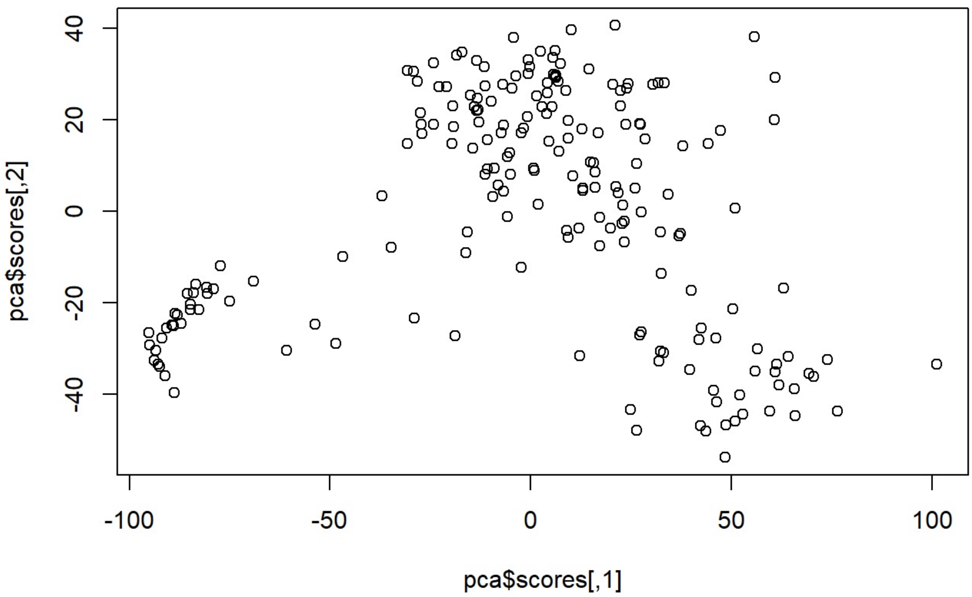

BIC plot
fembic[i,]
1:max(K)
-40000
-44000
-48000
-52000
1
2
3
4
5
6
AjBk
AkjBk
AjB
AB
AkjB
AkB
DkB
DkBk
DBk
DB
AkBk


var 1
var 2
var 3
var 4
var 5

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



Run a cluster analysis of your choice on the functional principal component scores.



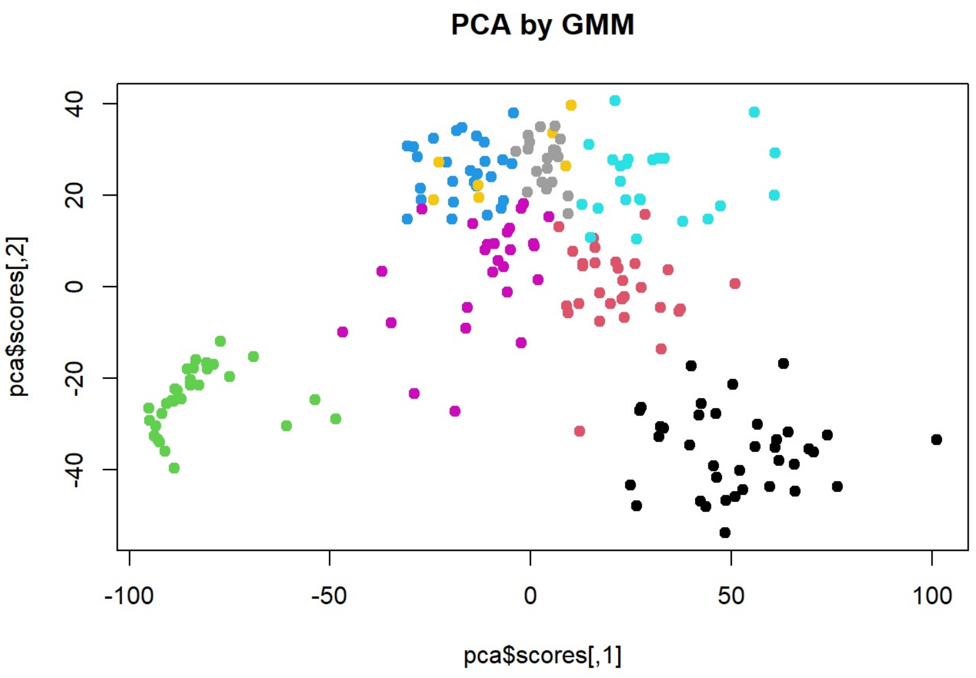
Comparison

GMM

```
adjustedRandIndex(gmm$classification, phonemes1000$g)
```

```
## [1] 0.5803799
```

```
plot(pca$scores,col=gmm$classification,pch = 19, main = "PCA by GMM")
```




```
table(gmm$classification, phonemes1000$g)
```

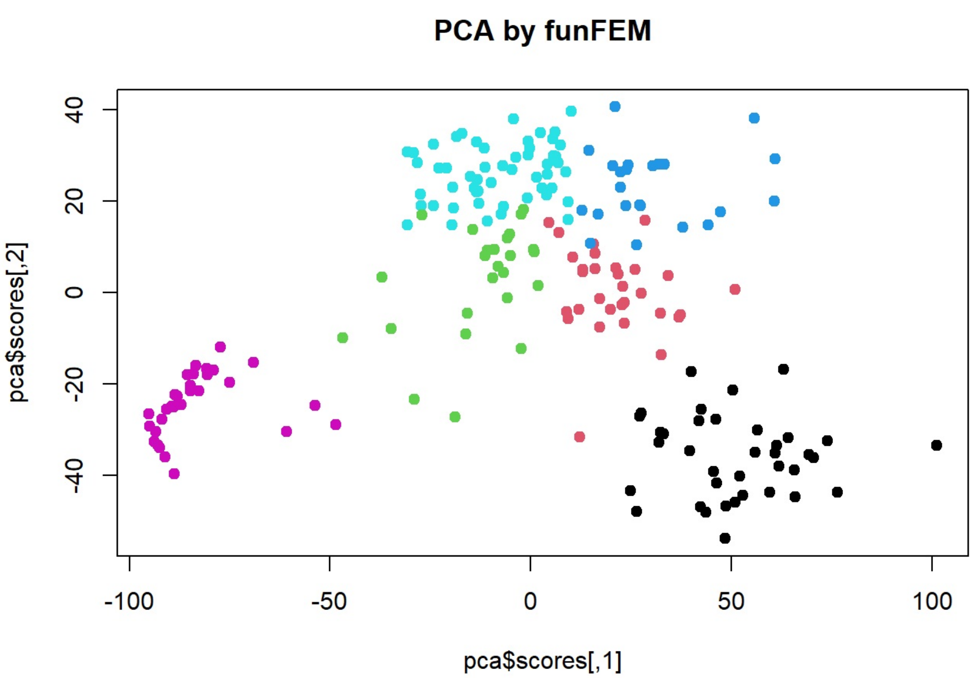
```
##
##     aa ao dcl iy sh
##   1  0  0   0  0 37
##   2  0  0   1 26  2
##   3  0  0  30  0  0
##   4  6 21   0  0  0
##   5 16  7   0  0  0
##   6  0  0   4 22  0
##   7  3  4   0  0  0
##   8  5 16   0  0  0
```

funFEM

```
adjustedRandIndex(femresult11$cls, phonemes1000$g)
```

```
## [1] 0.6492008
```

```
plot(pca$scores,col=femresult11$cls,pch = 19, main = "PCA by funFEM")
```



```r
table(femresult11$cls, phonemes1000$g)
```

```
##
##     aa ao dcl iy sh
##   1  0  0   0  0 37
##   2  0  0   1 27  2
##   3  0  0   4 21  0
##   4 16  7   0  0  0
##   5 14 41   0  0  0
##   6  0  0  30  0  0
```

Exercise 3

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

Discriminant Coordinates (DC):

Discriminant Coordinates are used for projecting high-dimensional data into a lower-dimensional space. The goal is to find a set of projection vectors that maximize the separation between different classes in the projected space. The method relies on the concept of discriminant analysis, which aims to find directions that maximize the ratio of between variance to within variance. In the context of DC, the between variance is captured by the covariance matrix of mean differences between classes (Q), and the within variance is represented by the pooled within covariance matrix (R).

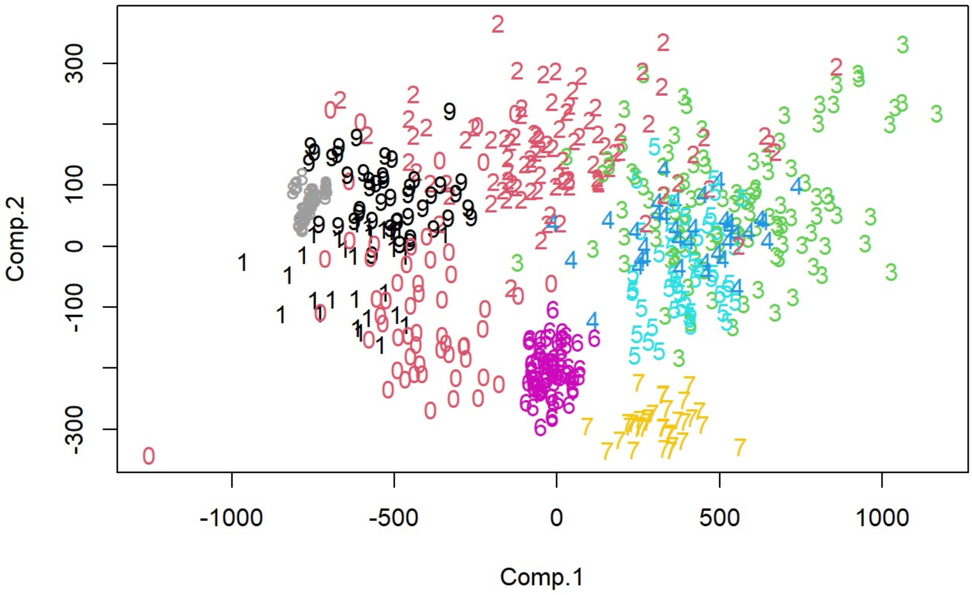
Asymmetric Weighted Discriminant Coordinates (AWC):

AWC is an extension of the DC method that introduces asymmetry in the treatment of classes. It recognizes the potential heterogeneity within classes and addresses situations where one class may be more homogeneous than another. AWC aims to emphasize the separation of a “homogeneous class” from a “non-homogeneous class”. The key innovation in AWC is the introduction of weights for points in the N-class based on their Mahalanobis distance to the H-class. In the AWC method, the within covariance matrix (R) is replaced by a modified between covariance matrix (B∗∗), which considers the Mahalanobis distances of points in the N-class to the H-class. The weights assigned to points in the N-class are determined by their distances, giving more emphasis to points closer to the H-class. This weighting scheme is designed to ensure that extreme points in the N-class, especially those far from the H-class, do not dominate the projection and that the resulting visualization highlights the differences between the H-class and the N-class in a more nuanced manner.

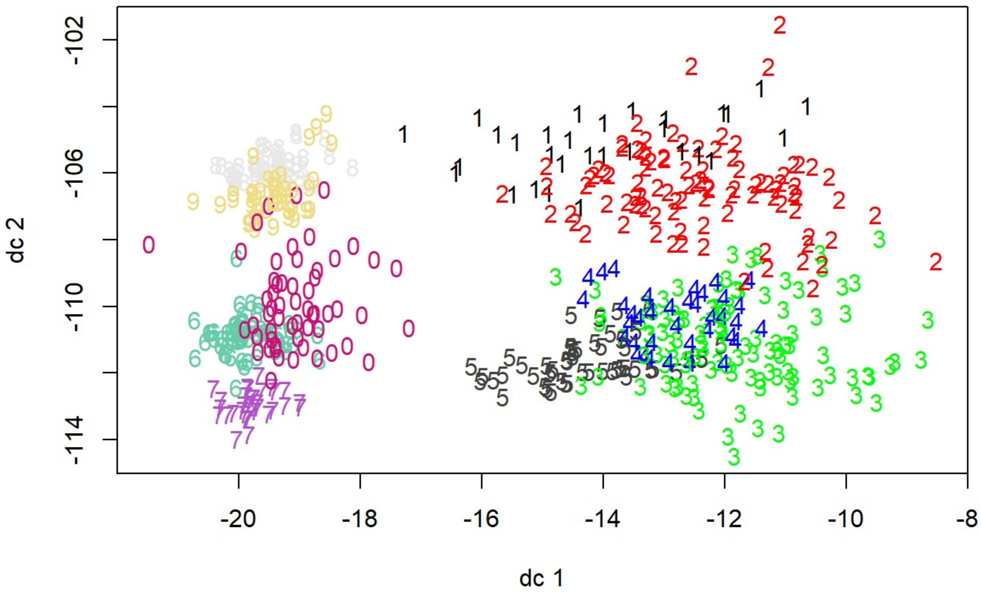
(b) For the optimal 10-clusters Gaussian mixture clustering of the olive oil data (Example 6.5 on the course slides) show 2-dimensional discriminant coordinates, and asymmetric weighted discriminant coordinates for all clusters. Comment on how these plots compare to the principal components plot in terms of showing the separation of the clusters.

```
## ----------------------------------------------------
## 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
```

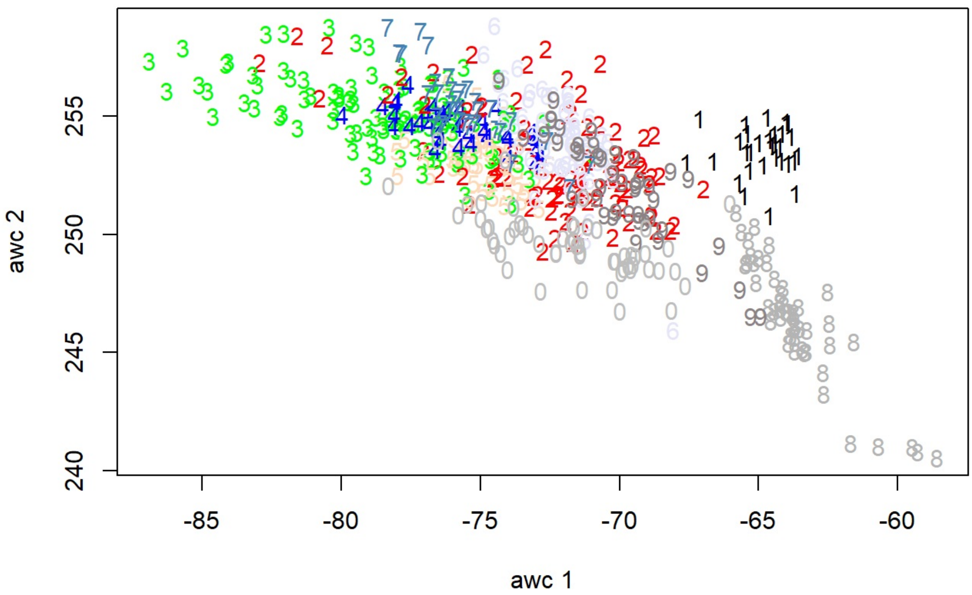
PCA by GMM



DC plot



AWC plot



(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 cordinates using plotcluster to the data set of the coefficients of the full dimensional B-spline basis and the funFEM-clustering.