TACv2

Nazanin Zounemat Kermani

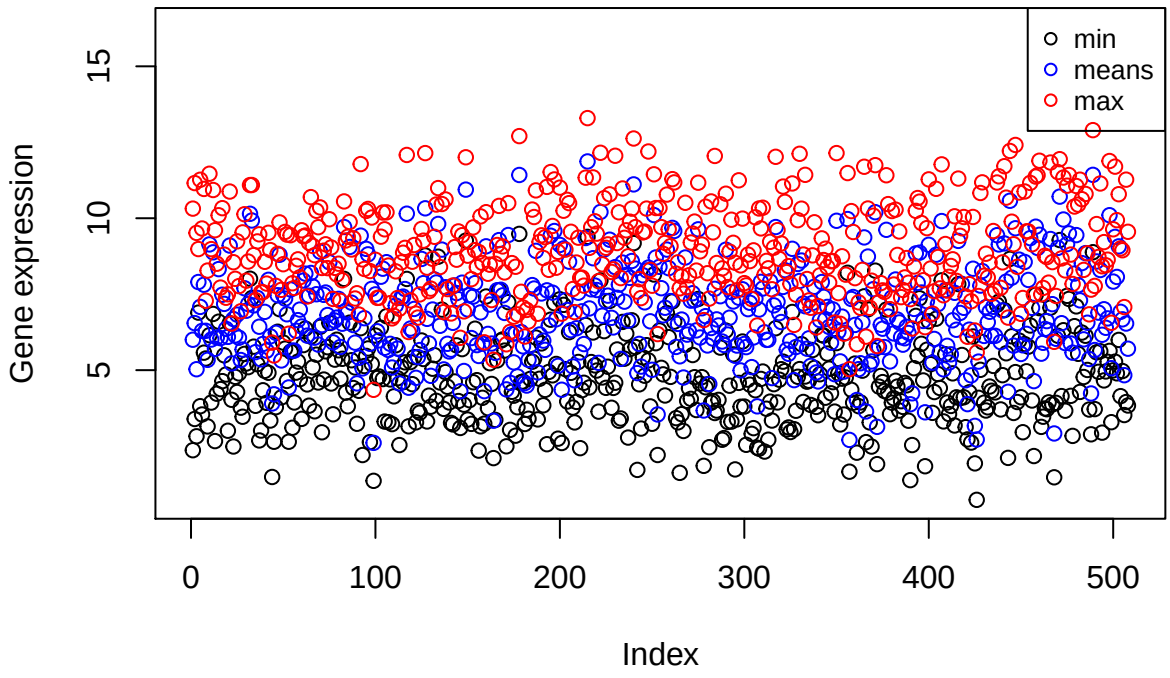
January 24, 2018

Find the optimal number of clusters

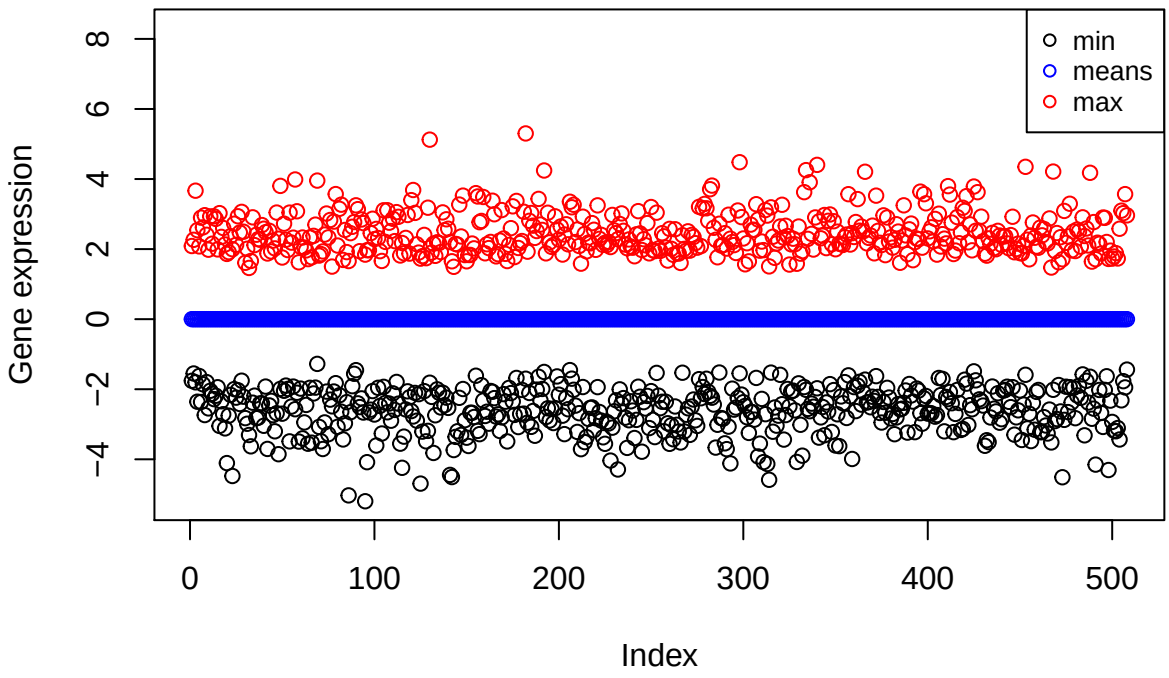
This the R notebook to find the optimal number of clusters in the data set. visual inspection of heatmap of the similarity matrix in the main reference paper suggests that the largest cluster includes subclusters. In this section of the report 3 methods were studied. The underlying rational behind employing several measurements and methods to determine the number of clusters is two fold. firstly, there is not single method that find the optimal number of clusters in the data set. The reason is that the number of clusters is highly subjective and governed by the similarity measurement and the clustering methods' parameters. five methods were used to find the optimal number of clusters. Gap statisitics, (more info if stayed in the analysis), WSS and silhouette, Gaussian model based, progency.

Read the data and check for outliers:

```
data <- t(read.table("C:/Users/nz1413/Desktop/dataTAC/sputum_508genes.txt",                    heade
dataSafety(data)
```




```
dataScaled = scale(data)
dataSafety(dataScaled)
```

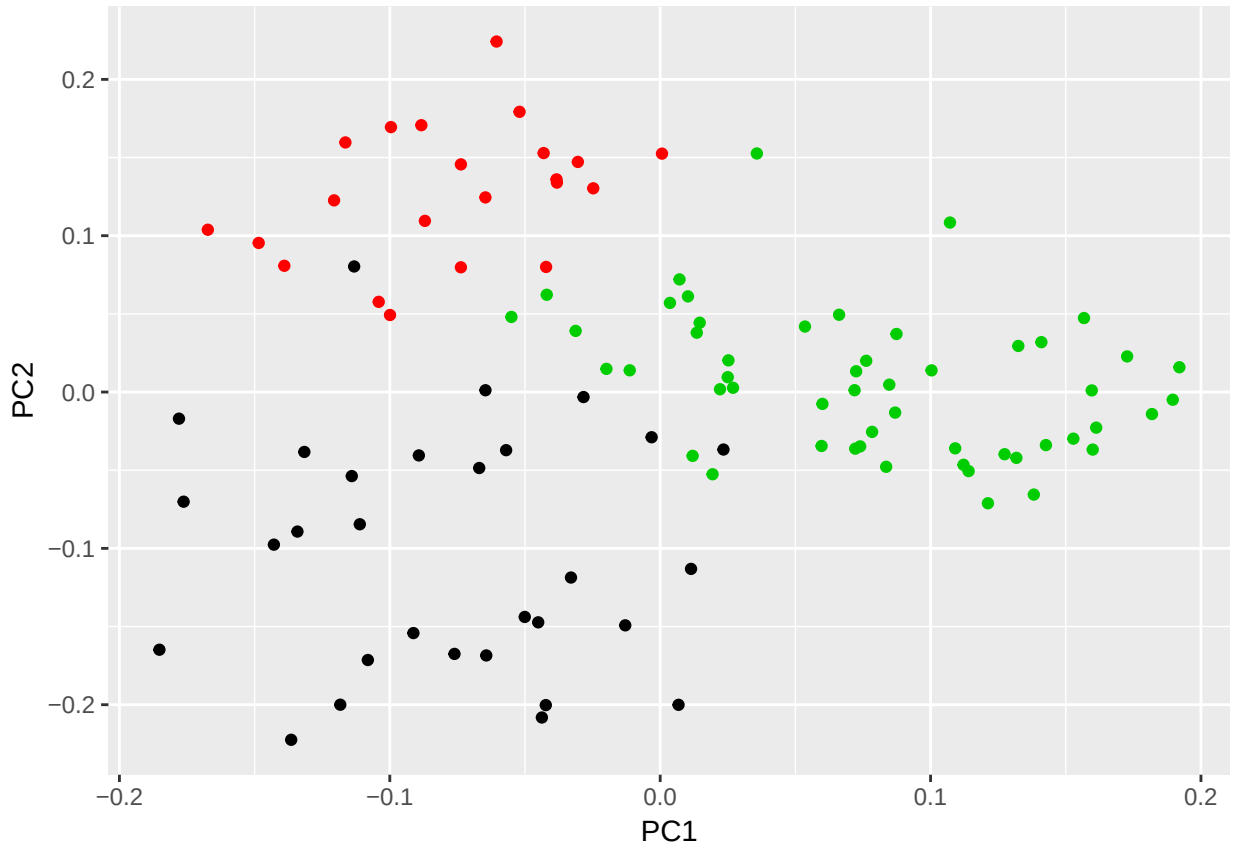


Plot the data before any furthur analysis to inspect the clustering tendency.

```
library(ggfortify)
```

```
## Loading required package: ggplot2
```

```
TAClabels <- t(read.table("C:/Users/nz1413/Desktop/dataTAC/correspondanceTAC.txt",  header = FALSE))
inds = as.integer(sapply(rownames(data) , function(x) which(x == TAClabels[2,])))
autoplot(prcomp(dataScaled), data = dataScaled, colour = TAClabels[1,inds], legendLabs = levels(factor(T
```



WSS(elbow), silhouette and gap

model-based clustering

My personal choise if the EM based model. This model:

```
library(mclust)
```

```
## Package 'mclust' version 5.4
## Type 'citation("mclust")' for citing this R package in publications.
```

```
mbc<-Mclust(dataScaled)
summary(mbc)
```

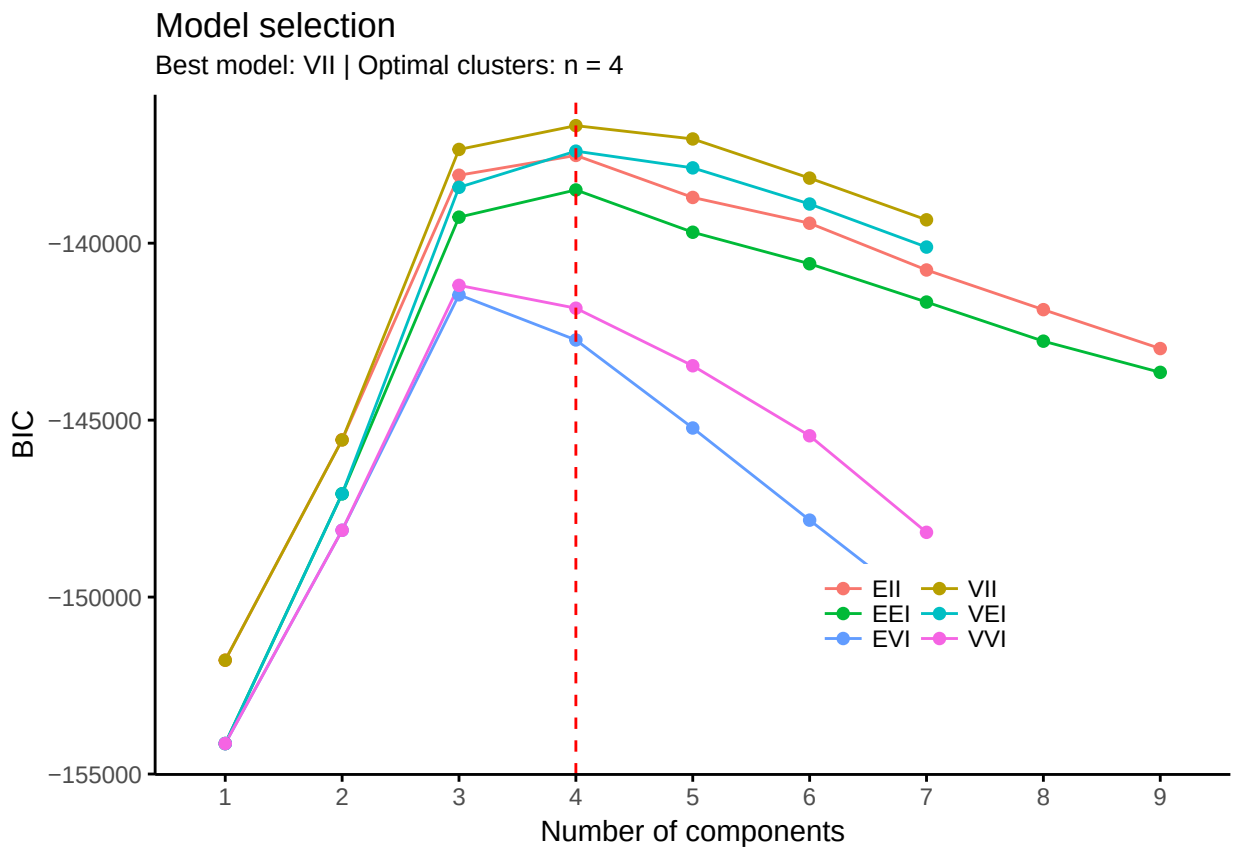
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VII (spherical, varying volume) model with 4 components:
##
##  log.likelihood   n   df       BIC       ICL
##       -63604.42 104 2039 -136678.7 -136678.8
##
## Clustering table:
##  1  2  3  4
## 24 23 28 29
```

visualization

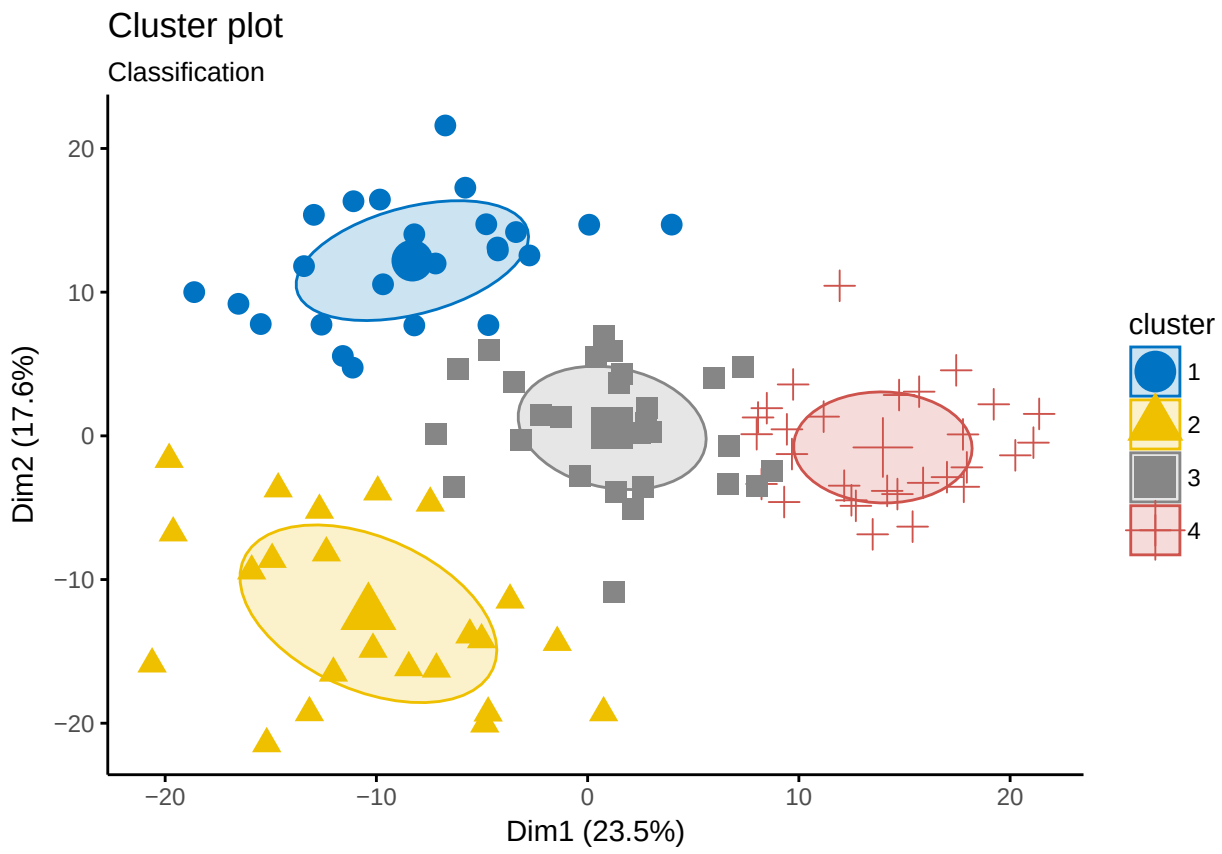
```
library(factoextra)
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

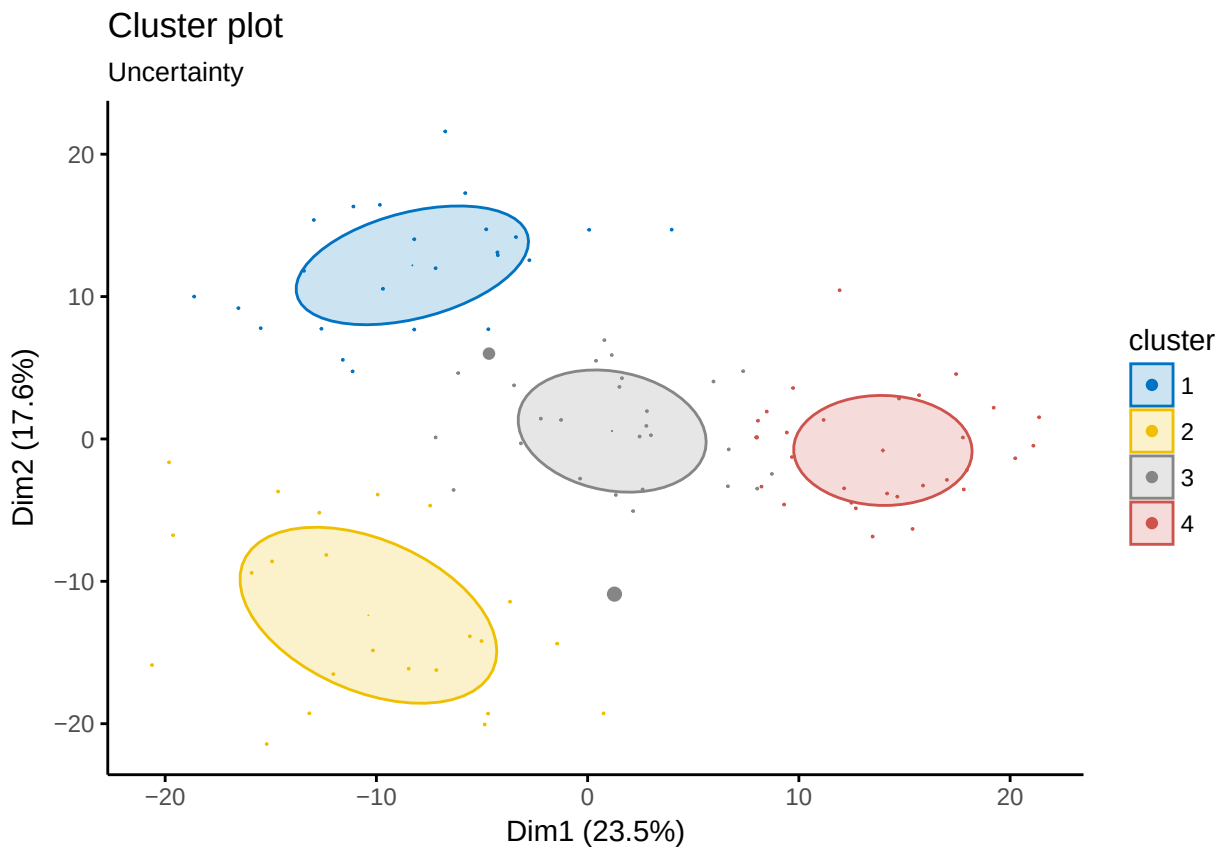
```
# BIC value for choosing the number of clusters
fviz_mclust(mbc, "BIC", pallette = "jco")
```




```
#Classification plot
fviz_mclust(mbc, "classification", geom="point",
            pointsize= 3.5, palette = "jco")
```




```
#classification uncertainity
fviz_mclust(mbc, "uncertainty",
            palette = "jco")
```


Cluster plot
Uncertainty
Dim2 (17.6%)
Dim1 (23.5%)
20
10
0
−10
−20
−20
−10
0
10
20
cluster
1
2
3
4