CS282R - Sepsis Cluster Optimzation

David Wihl

11/3/2017

Using the suggestions from this StackOverflow question, this note will attempt to find the optimal clustering for the Sepsis data set.

```
lots_of_clusters = function(d, max_k) {
  mydata <- d
  wss <- (nrow(mydata) - 1) * sum(apply(mydata, 2, var))
    for (i in 2:15)
      wss[i] <- sum(kmeans(mydata,centers = i)$withinss)
  plot(1:15,wss,type = "b",xlab = "Number of Clusters",ylab = "Within groups sum of squares")
  pamk.best <- pamk(d)
  cat("number of clusters estimated by optimum average silhouette width:",pamk.best$nc,"\n")
  plot(pam(d, pamk.best$nc))
  asw <- numeric(20)
  for (k in 2:20)
    asw[[k]] <- pam(d, k)$silinfo$avg.width
  k.best <- which.max(asw)
  cat("silhouette-optimal number of clusters:", k.best, "\n")
# calinsky

  fit <- cascadeKM(scale(d, center = TRUE,  scale = TRUE), 1, 10, iter = 1000)
  plot(fit, sortg = TRUE, grpmts.plot = TRUE)
  calinski.best <- as.numeric(which.max(fit$results[2,]))
  cat("Calinski criterion optimal number of clusters:", calinski.best, "\n")

  # BIC
# See http://www.jstatsoft.org/v18/i06/paper
# http://www.stat.washington.edu/research/reports/2006/tr504.pdf
#
# Run the function to see how many clusters
# it finds to be optimal, set it to search for
# at least 1 model and up max_k.
d_clust <- Mclust(as.matrix(d), G=1:20)
m.best <- dim(d_clust$z)[2]
cat("model-based optimal number of clusters:", m.best, "\n")
plot(d_clust)

# Affinity Propogation
d.apclus <- apcluster(negDistMat(r=2), d)
cat("affinity propogation optimal number of clusters:", length(d.apclus@clusters), "\n")
# 4
heatmap(d.apclus)
plot(d.apclus, d)

# Gap Statistic


clusGap(d, kmeans, 50, B = 100, verbose = interactive())
```

```
# NbClust
# nb <- NbClust(d, diss="NULL", distance = "euclidean",
#         min.nc=2, max.nc=15, method = "kmeans",
#         index = "alllong", alphaBeale = 0.1)
# hist(nb$Best.nc[1,], breaks = max(na.omit(nb$Best.nc[1,])))

d_dist <- dist(as.matrix(d))   # find distance matrix
plot(hclust(d_dist))           # apply hirarchical clustering and plot

x <- as.matrix(d)
d.bclus <- bclust(x, transformed.par = c(0, -50, log(16), 0, 0, 0))
viplot(imp(d.bclus)$var); plot(d.bclus); ditplot(d.bclus)
dptplot(d.bclus, scale = 20, horizbar.plot = TRUE,varimp = imp(d.bclus)$var, horizbar.distance = 0, den

# For high dimensional data

 d.pv <- pvclust(d)
plot(d.pv)
}
```

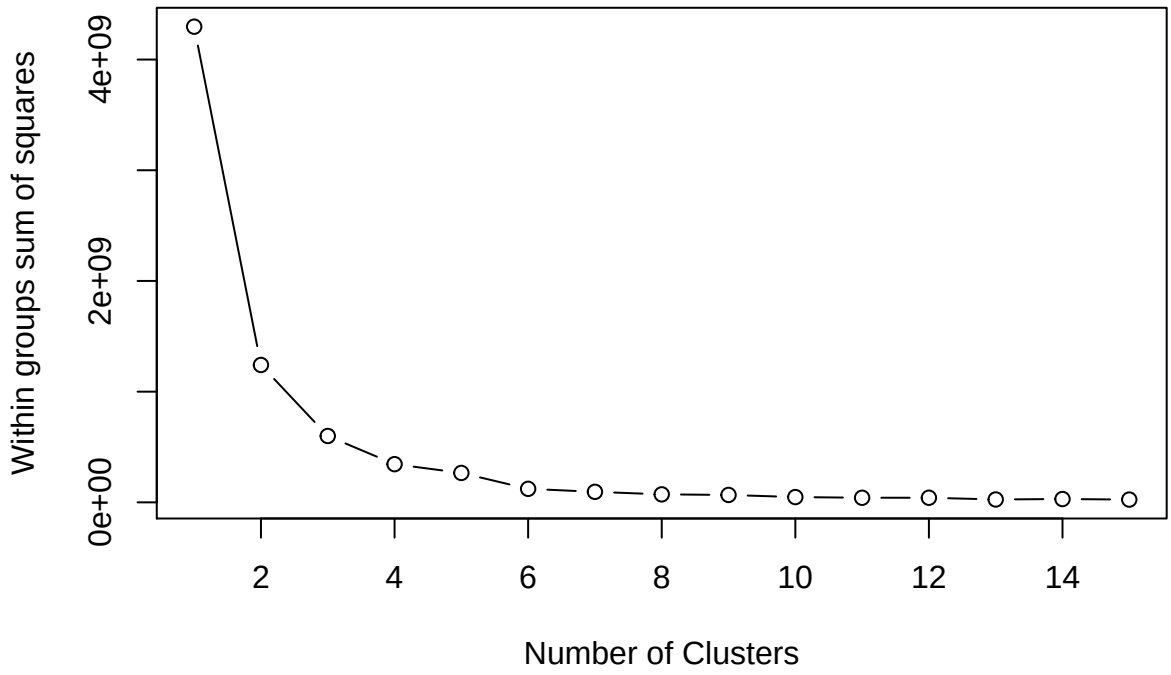
```
# Sample data
# n = 100
# g = 6
# set.seed(g)
# d <- data.frame(x = unlist(lapply(1:g, function(i) rnorm(n/g, runif(1)*i^2))),
#                 y = unlist(lapply(1:g, function(i) rnorm(n/g, runif(1)*i^2))))
# plot(d)

sep = read.csv("../data/Sepsis_imp.csv", header = TRUE)
demographics = c("age","HR","MeanBP","SpO2","Arterial_lactate","Creatinine","SOFA")
df_initial = sep[sep$bloc==1,demographics]
max_k =50
```

```
d = df_initial[sample(nrow(df_initial),100),]
summary(d)
```

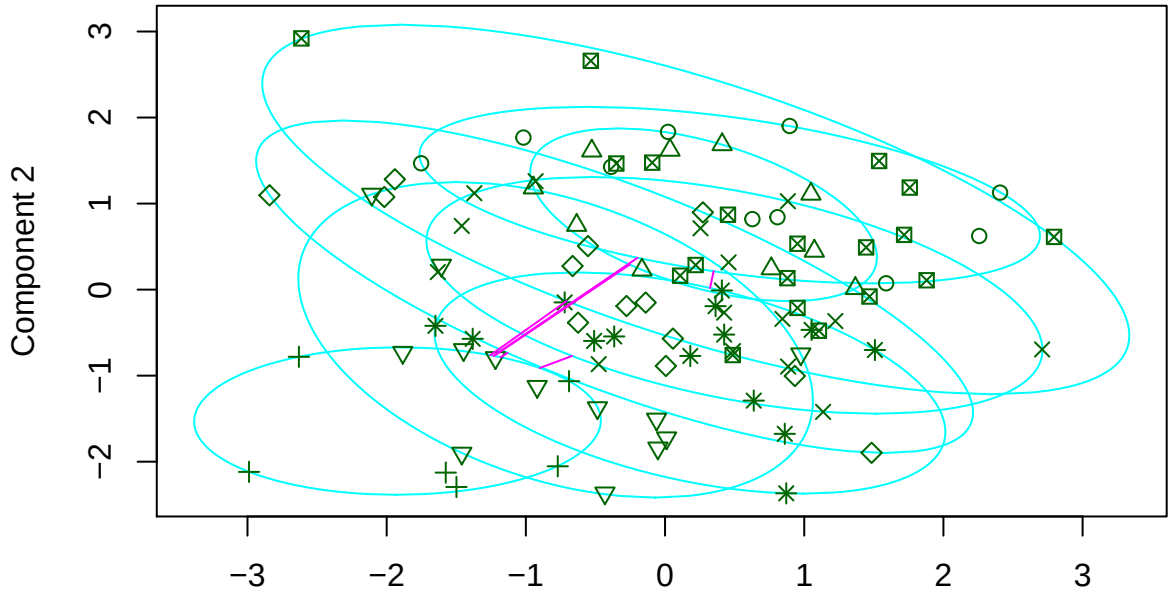
```
##       age              HR            MeanBP           SpO2
##  Min.   : 7836   Min.   : 39.2   Min.   : 37.3   Min.   : 89.0
##  1st Qu.:18581   1st Qu.: 80.0   1st Qu.: 68.9   1st Qu.: 96.0
##  Median :24284   Median : 90.5   Median : 79.0   Median : 98.1
##  Mean   :23343   Mean   : 90.1   Mean   : 80.3   Mean   : 97.5
##  3rd Qu.:28210   3rd Qu.:100.8   3rd Qu.: 88.0   3rd Qu.: 99.8
##  Max.   :33384   Max.   :144.0   Max.   :131.0   Max.   :100.0
##  Arterial_lactate   Creatinine         SOFA
##  Min.   : 0.60    Min.   : 0.30   Min.   : 0.00
##  1st Qu.: 1.10    1st Qu.: 0.70   1st Qu.: 5.00
##  Median : 1.50    Median : 0.95   Median : 7.00
##  Mean   : 1.91    Mean   : 1.79   Mean   : 6.97
##  3rd Qu.: 2.12    3rd Qu.: 1.88   3rd Qu.: 9.00
##  Max.   :10.89    Max.   :11.30   Max.   :15.00
```

```
lots_of_clusters(d,max_k)
```

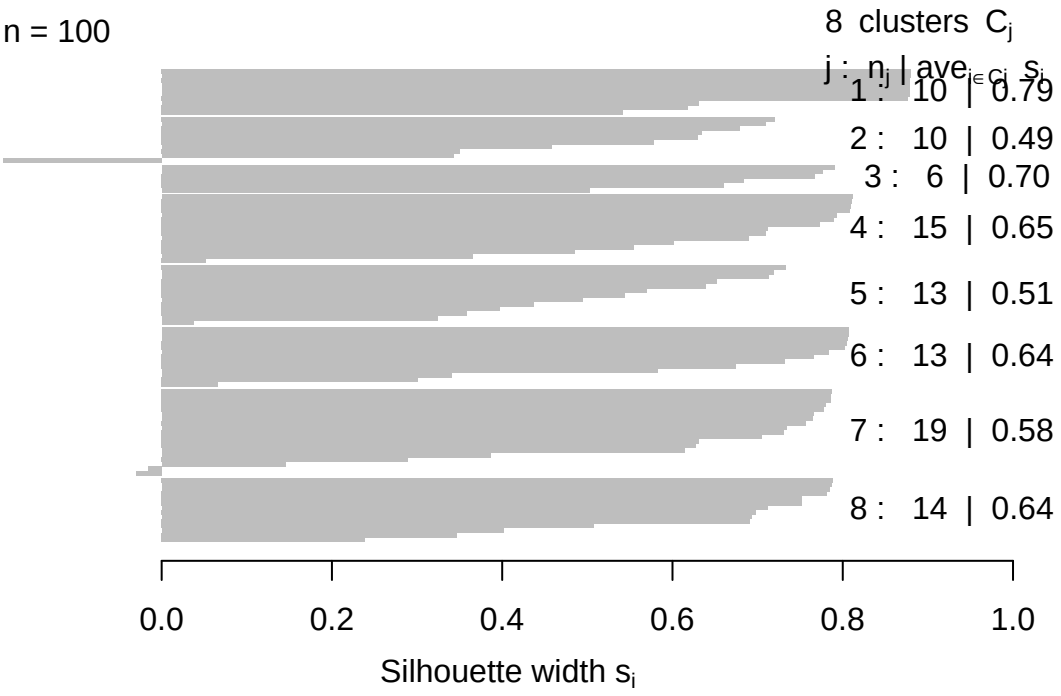



```
## number of clusters estimated by optimum average silhouette width: 8
```

clusplot(pam(x = d, k = pamk.best$nc))

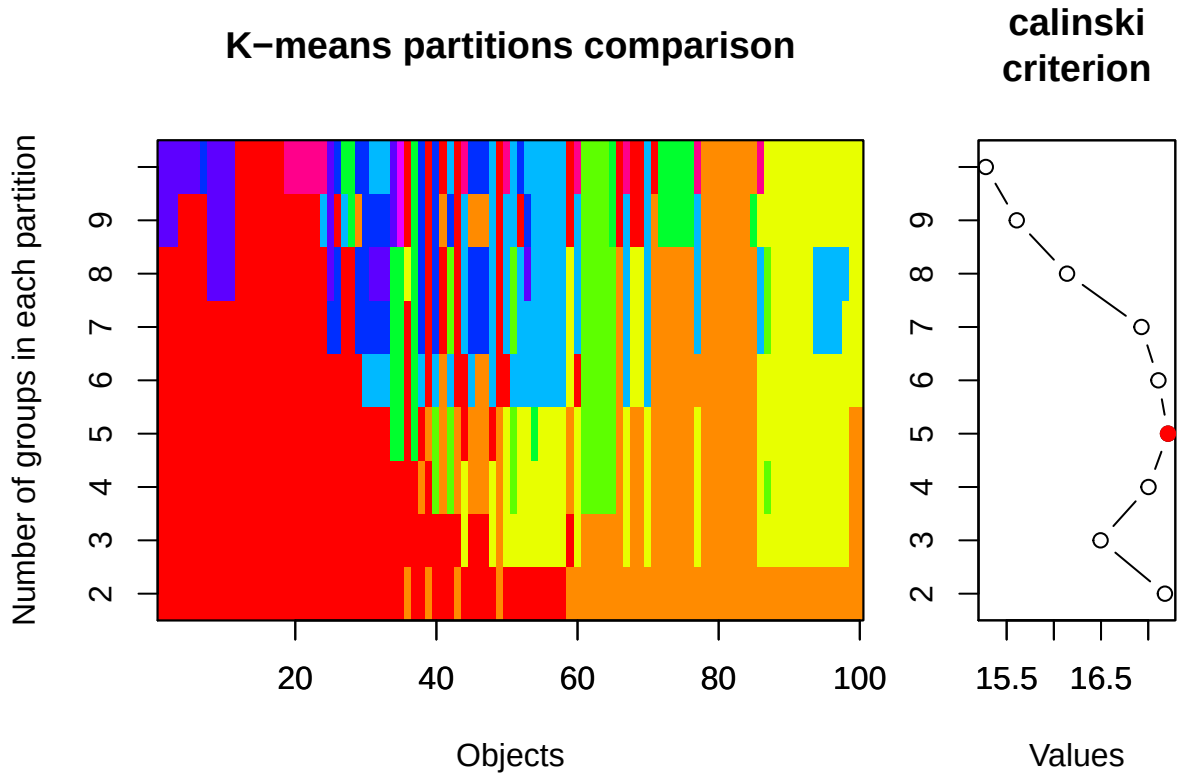


These two components explain 41.2 % of the point variability.

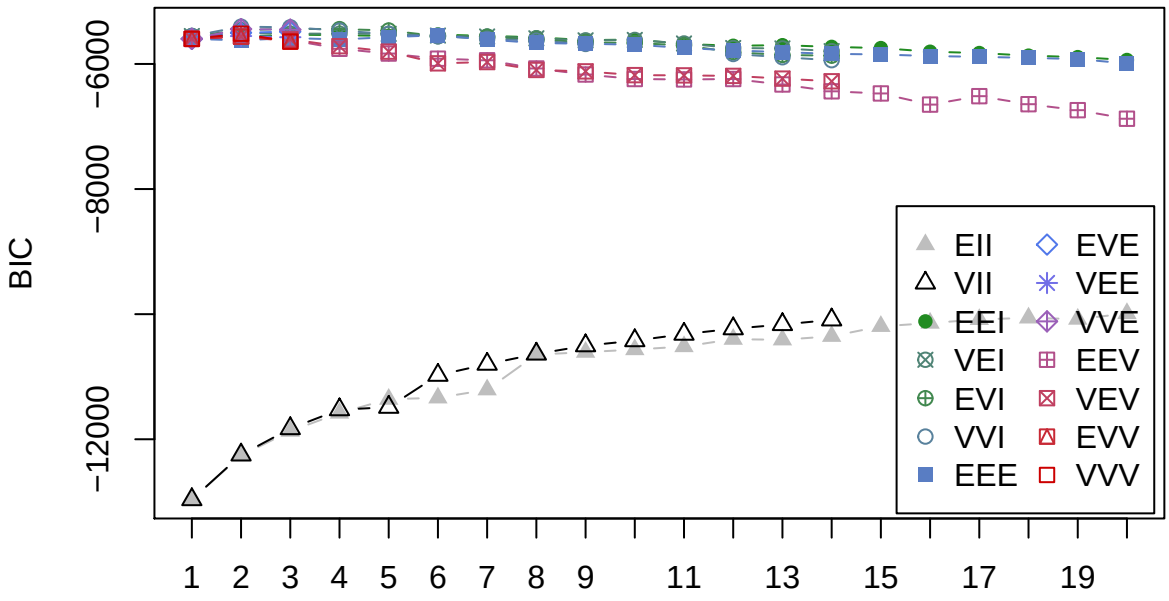


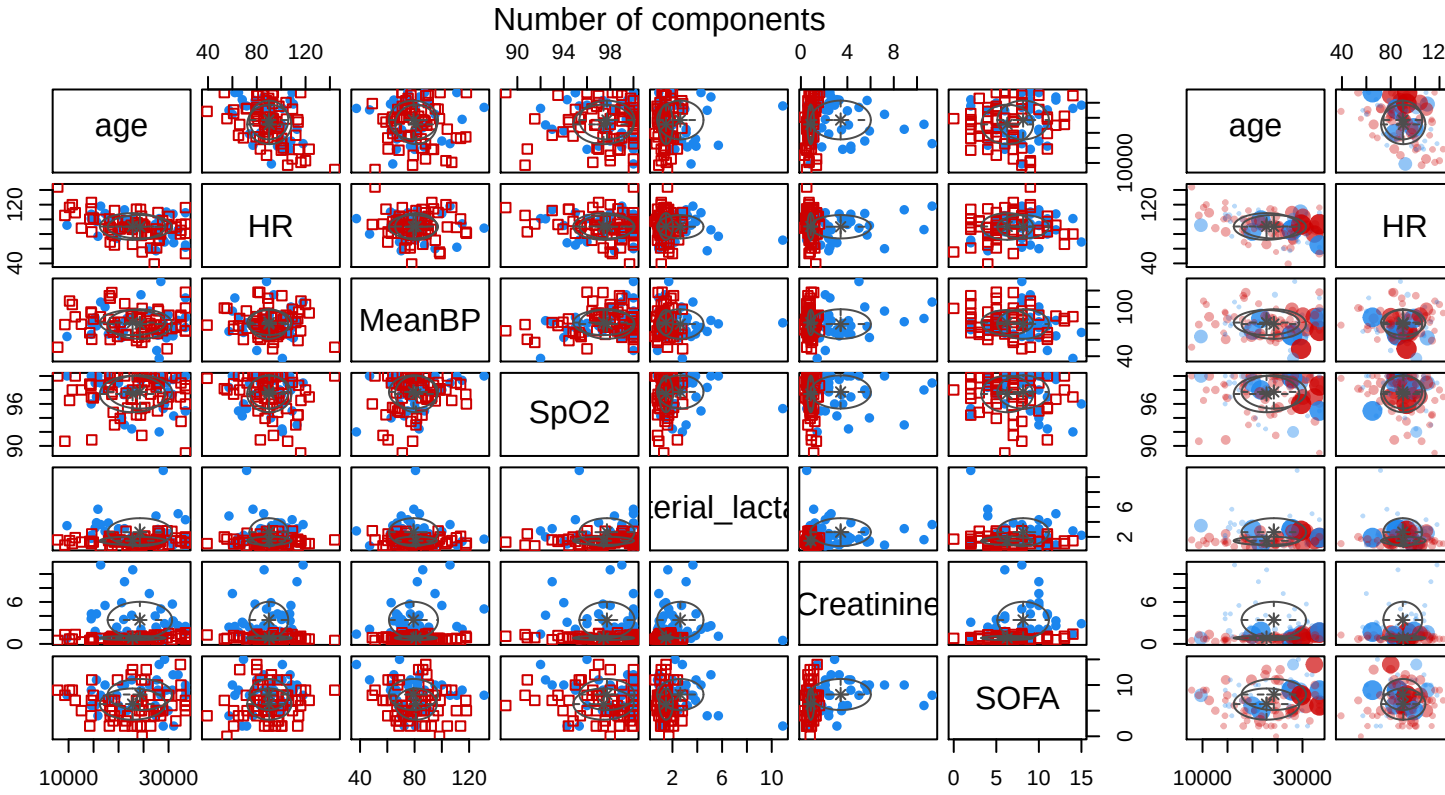
Average silhouette width : 0.62

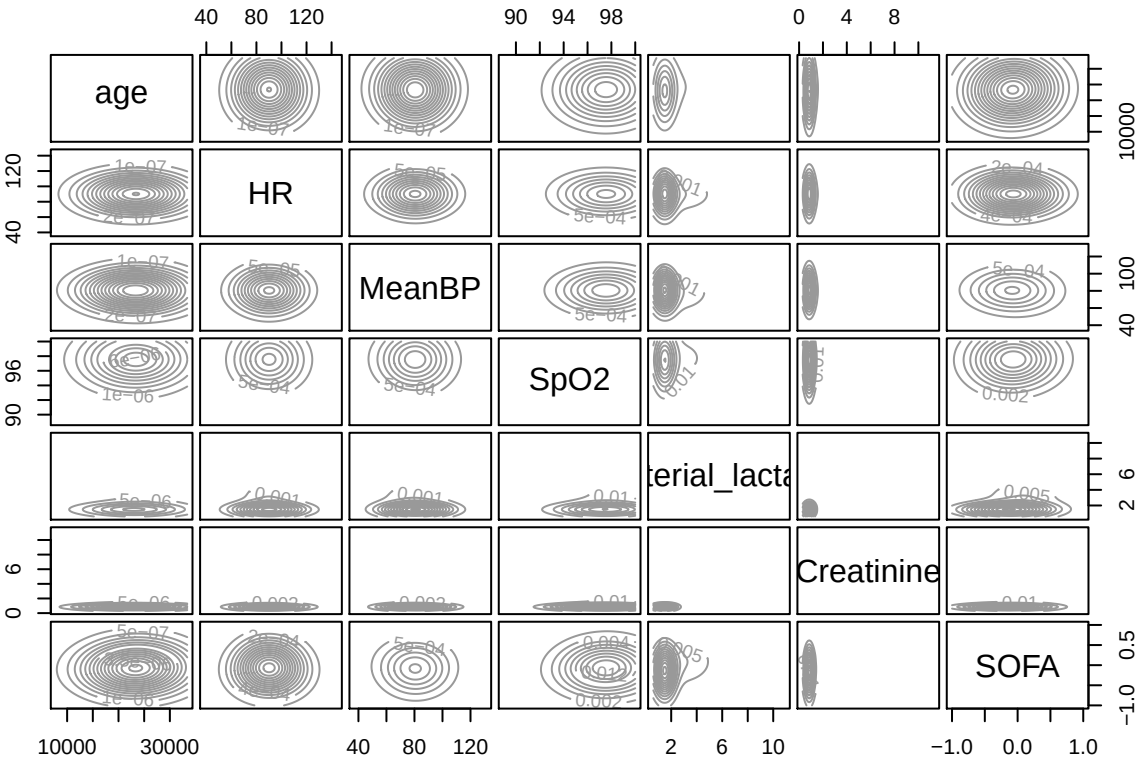
```
## silhouette-optimal number of clusters: 8
```



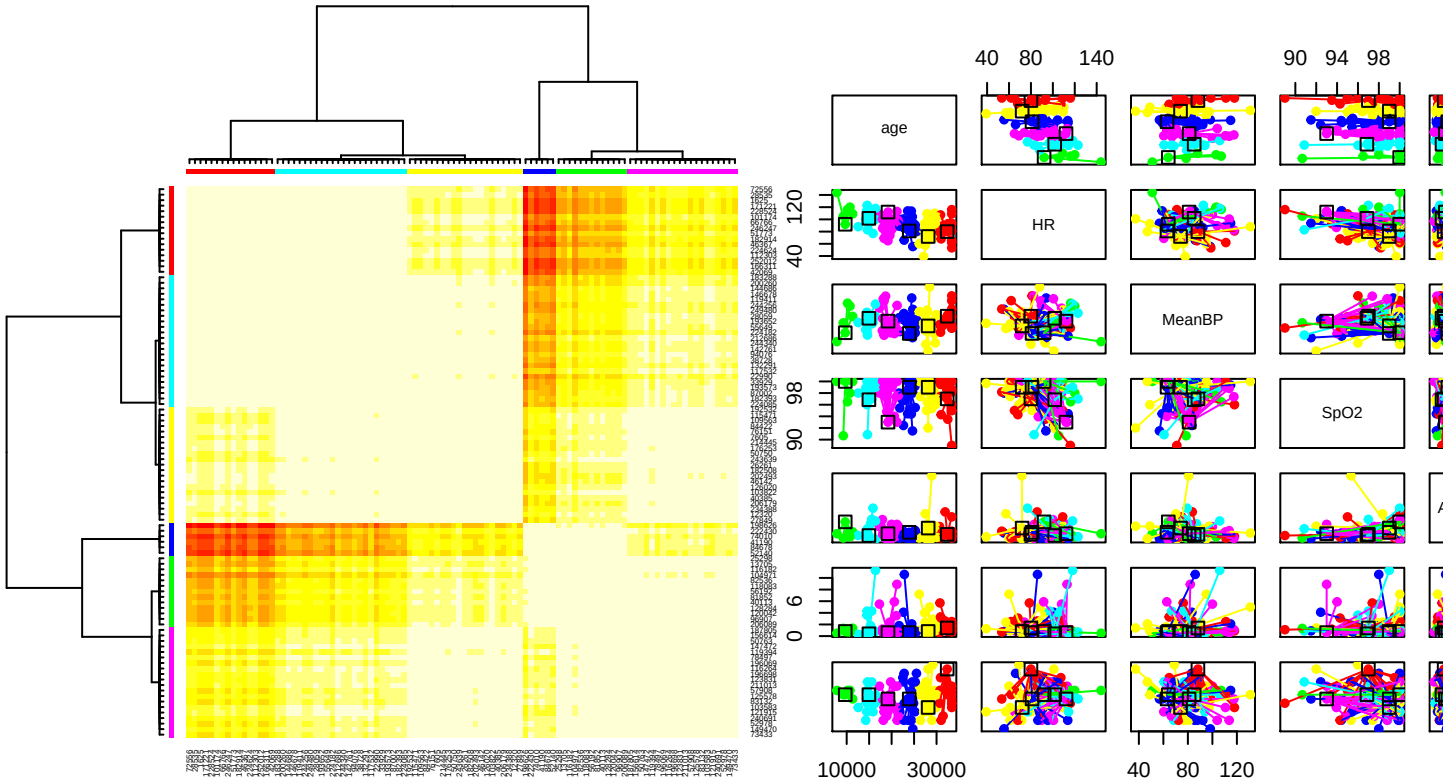

```
## Calinski criterion optimal number of clusters: 5
## model-based optimal number of clusters: 2
```

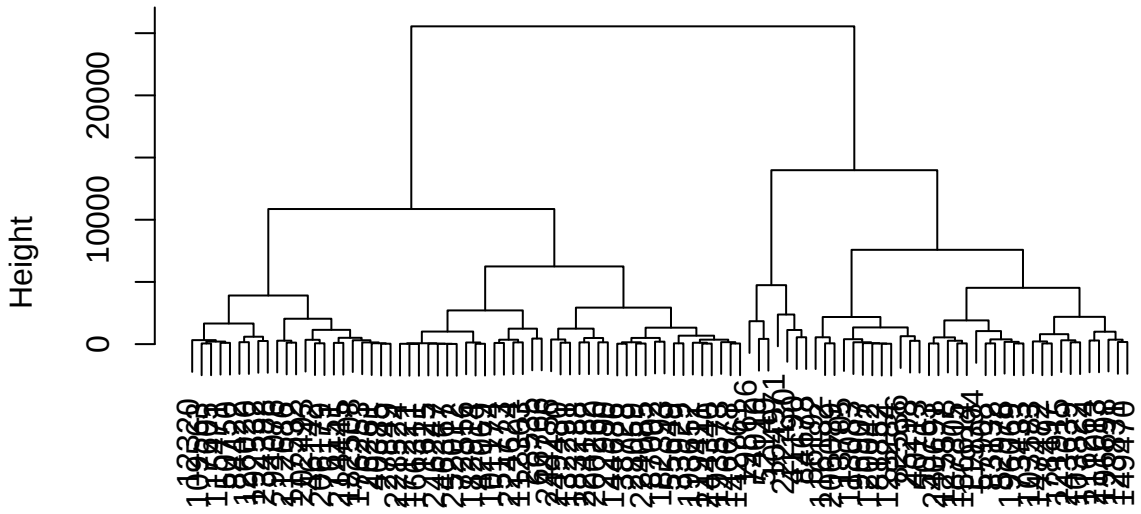

BIC
−6000
−8000
−12000
1 2 3 4 5 6 7 8 9 11 13 15 17 19
EII
VII
EEI
VEI
EVI
VVI
EEE
EVE
VEE
VVE
EEV
VEV
EVV
VVV


Number of components
age
HR
MeanBP
SpO2
Creatinine
SOFA




```
## affinity propogation optimal number of clusters: 6
```

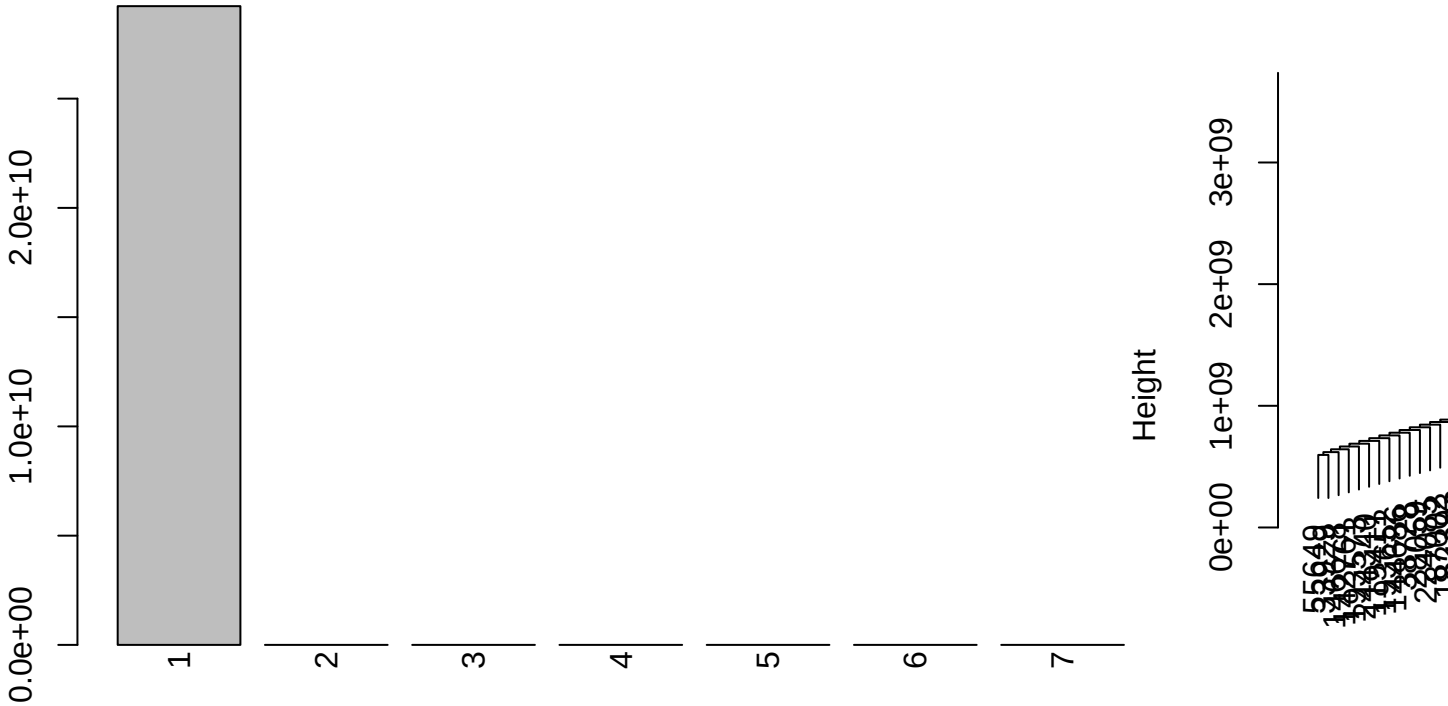


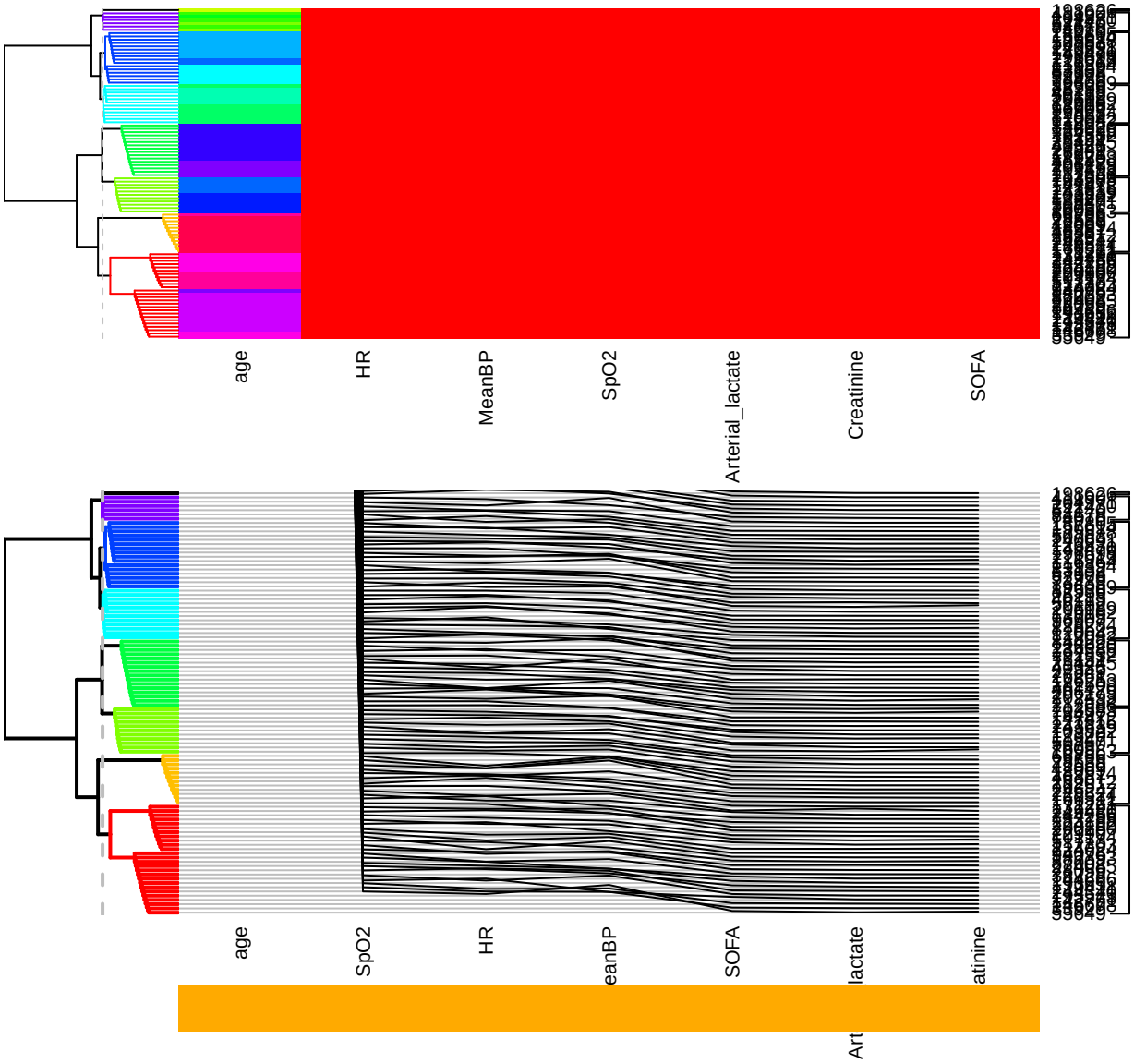



```
## 2/100
## 3/100
## 4/100
## 5/100
## 6/100
## 7/100
## 8/100
## 9/100
## 10/100
## 11/100
## 12/100
## 13/100
## 14/100
## 15/100
## 16/100
## 17/100
## 18/100
## 19/100
## 20/100
## 21/100
## 22/100
## 23/100
## 24/100
## 25/100
## 26/100
## 27/100
## 28/100
## 29/100
```

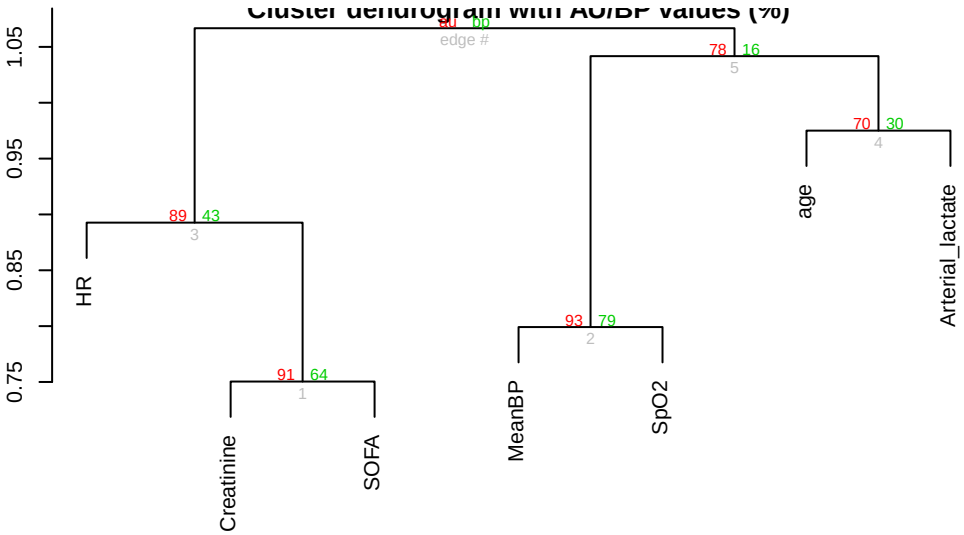
```
## 30/100
## 31/100
## 32/100
## 33/100
## 34/100
## 35/100
## 36/100
## 37/100
## 38/100
## 39/100
## 40/100
## 41/100
## 42/100
## 43/100
## 44/100
## 45/100
## 46/100
## 47/100
## 48/100
## 49/100
## 50/100
## 51/100
## 52/100
## 53/100
## 54/100
## 55/100
## 56/100
## 57/100
## 58/100
## 59/100
## 60/100
## 61/100
## 62/100
## 63/100
## 64/100
## 65/100
## 66/100
## 67/100
## 68/100
## 69/100
## 70/100
## 71/100
## 72/100
## 73/100
## 74/100
## 75/100
## 76/100
## 77/100
## 78/100
## 79/100
## 80/100
## 81/100
## 82/100
## 83/100
```

```
## 84/100
## 85/100
## 86/100
## 87/100
## 88/100
## 89/100
## 90/100
## 91/100
## 92/100
## 93/100
## 94/100
## 95/100
## 96/100
## 97/100
## 98/100
## 99/100
## 100/100
```






```
## Bootstrap (r = 0.5)... Done.
## Bootstrap (r = 0.6)... Done.
## Bootstrap (r = 0.7)... Done.
## Bootstrap (r = 0.8)... Done.
## Bootstrap (r = 0.9)... Done.
## Bootstrap (r = 1.0)... Done.
## Bootstrap (r = 1.1)... Done.
## Bootstrap (r = 1.2)... Done.
## Bootstrap (r = 1.3)... Done.
## Bootstrap (r = 1.4)... Done.
```


Cluster dendrogram with AU/BP values (%)
edge #
1.05
0.95
0.85
0.75
HR
Creatinine
SOFA
MeanBP
SpO2
age
Arterial_lactate