

SISG Bayes: Exercise 6

Ken Rice

2025-06-05

Clustering

A non-genetic example

We have data on 145 diabetics, that we will use for clustering

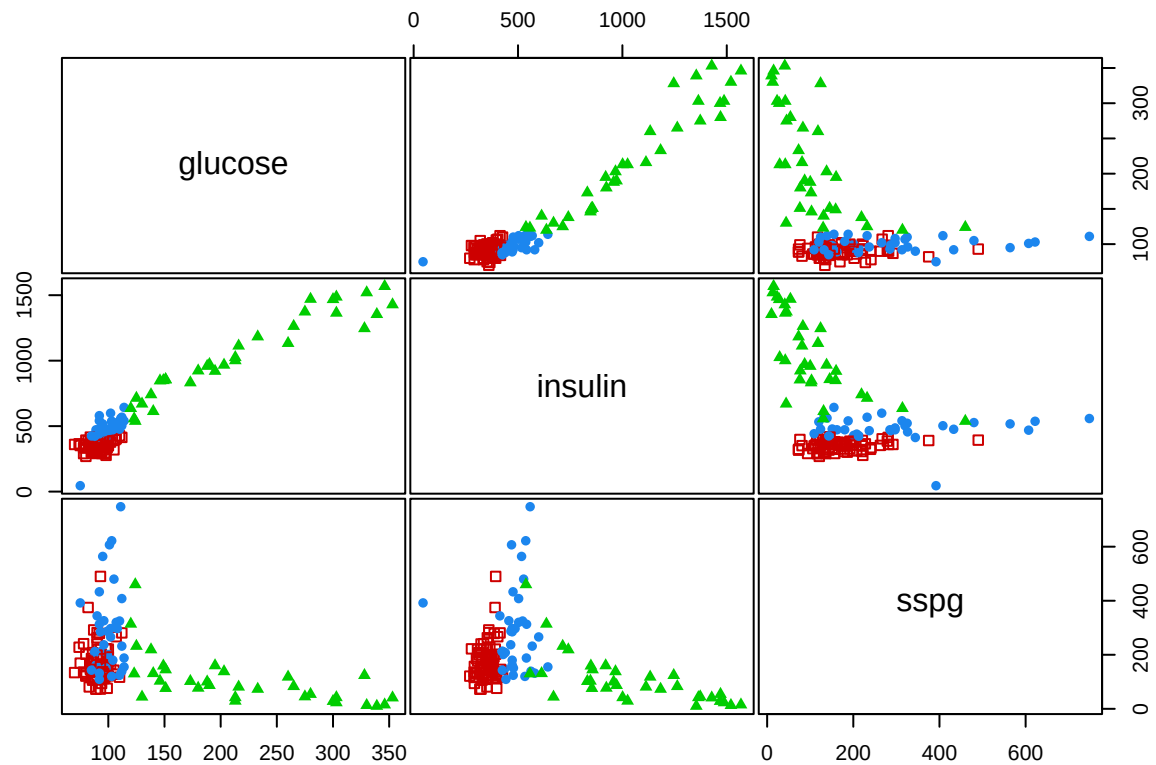
- glucose: after 3 hour tolerance test
- insulin: after 3 hour tolerance test
- sspg: steady state glucose

To indicate the success of the clustering, note that *in this example* we also not the class of their diabetes – which can be “Chemical”, “Normal”, or “Overt”.

```
#install.packages("mclust")
library("mclust")
data(diabetes)
summary(diabetes)
##      class      glucose      insulin      sspg
## Chemical:36  Min.   : 70   Min.   : 45.0   Min.   : 10.0
## Normal  :76  1st Qu.: 90   1st Qu.: 352.0  1st Qu.:118.0
## Overt   :33  Median : 97   Median : 403.0  Median :156.0
##                Mean   :122   Mean   : 540.8   Mean   :186.1
##                3rd Qu.:112   3rd Qu.: 558.0   3rd Qu.:221.0
##                Max.   :353   Max.   :1568.0   Max.   :748.0

#simple pairs plot in which the colors show the 3 classes:
## glucose: after 3 hour tolerance test
## insulin: after 3 hour tolerance test
## sspg: steady state glucose

clPairs(diabetes[,c("glucose","insulin","sspg")], diabetes$class)
```



Clustering is done by fitting a mixture of multivariate Normals to the data, in this case using the EM algorithm. We'll use BIC (Bayesian Information Criterion) to select the number of clusters – though this measure is only (very) approximately Bayesian;

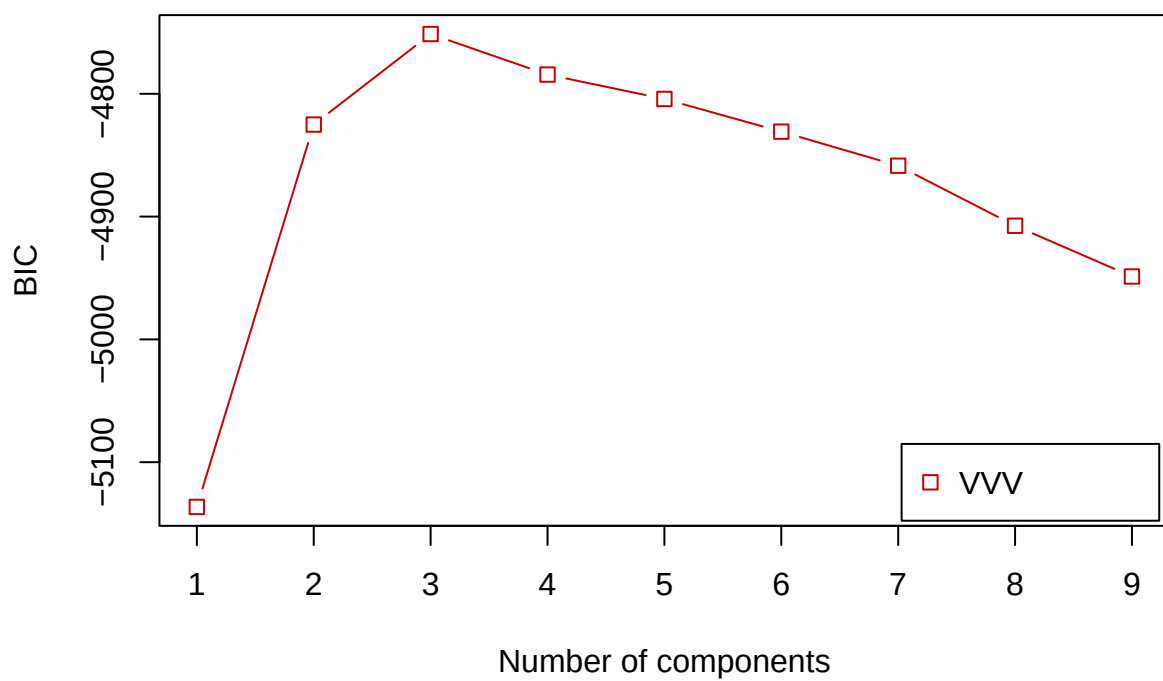
A first version, where we let each cluster's Gaussian distribution have its own mean and covariance structure:

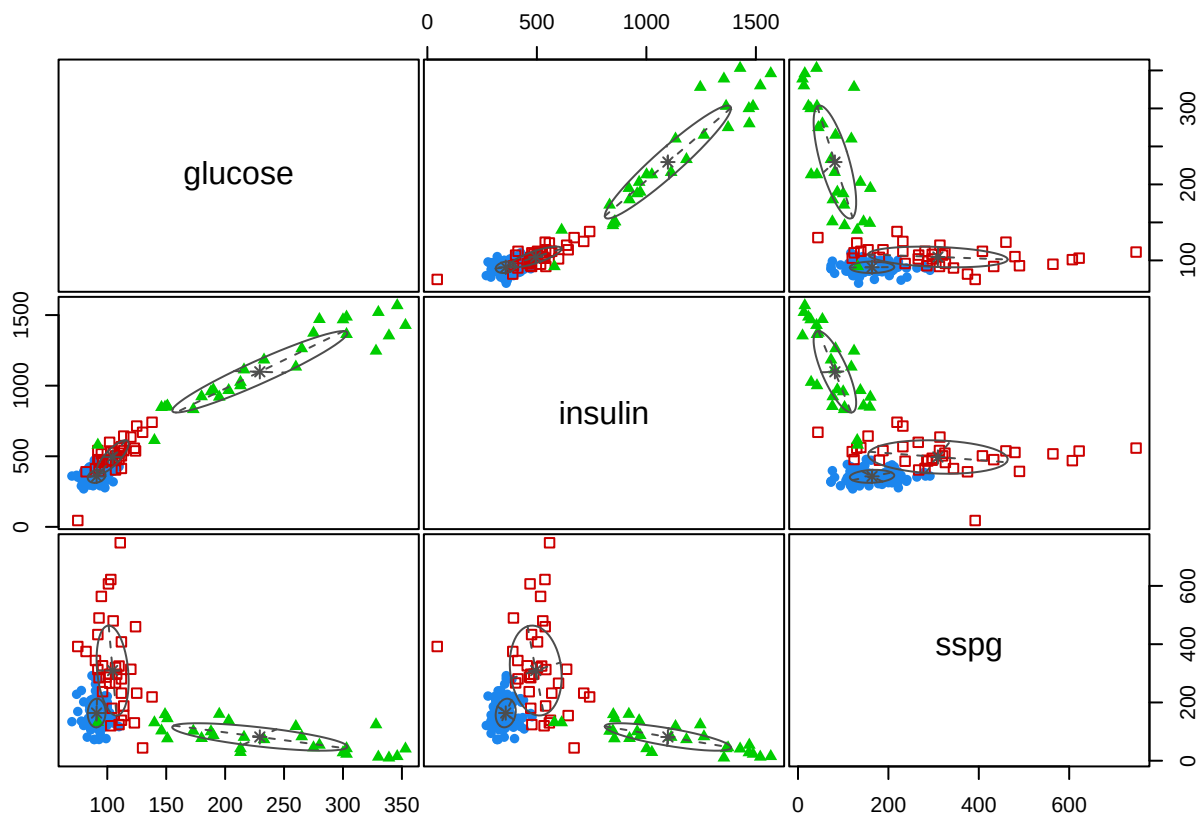
```
clust1 <- Mclust(data=diabetes[,c("glucose","insulin","sspg")], modelNames="VVV")
summary(clust1)
## -----
## Gaussian finite mixture model fitted by EM algorithm
## -----
##
## Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 3
## components:
##
## log-likelihood  n df      BIC      ICL
##      -2303.496 145 29 -4751.316 -4770.169
##
## Clustering table:
##  1  2  3
## 81 36 28
```

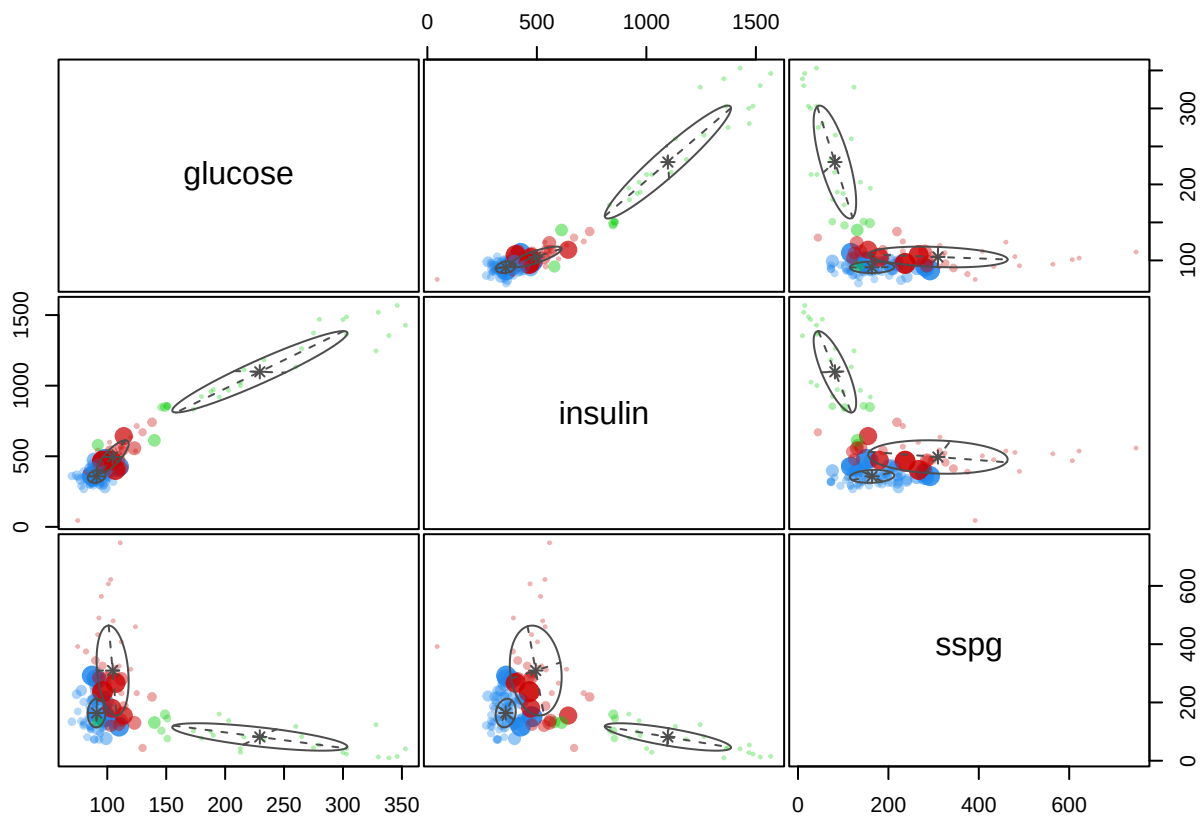
```
clust1$BIC
## Bayesian Information Criterion (BIC):
##      VVV
##  1 -5136.446
```

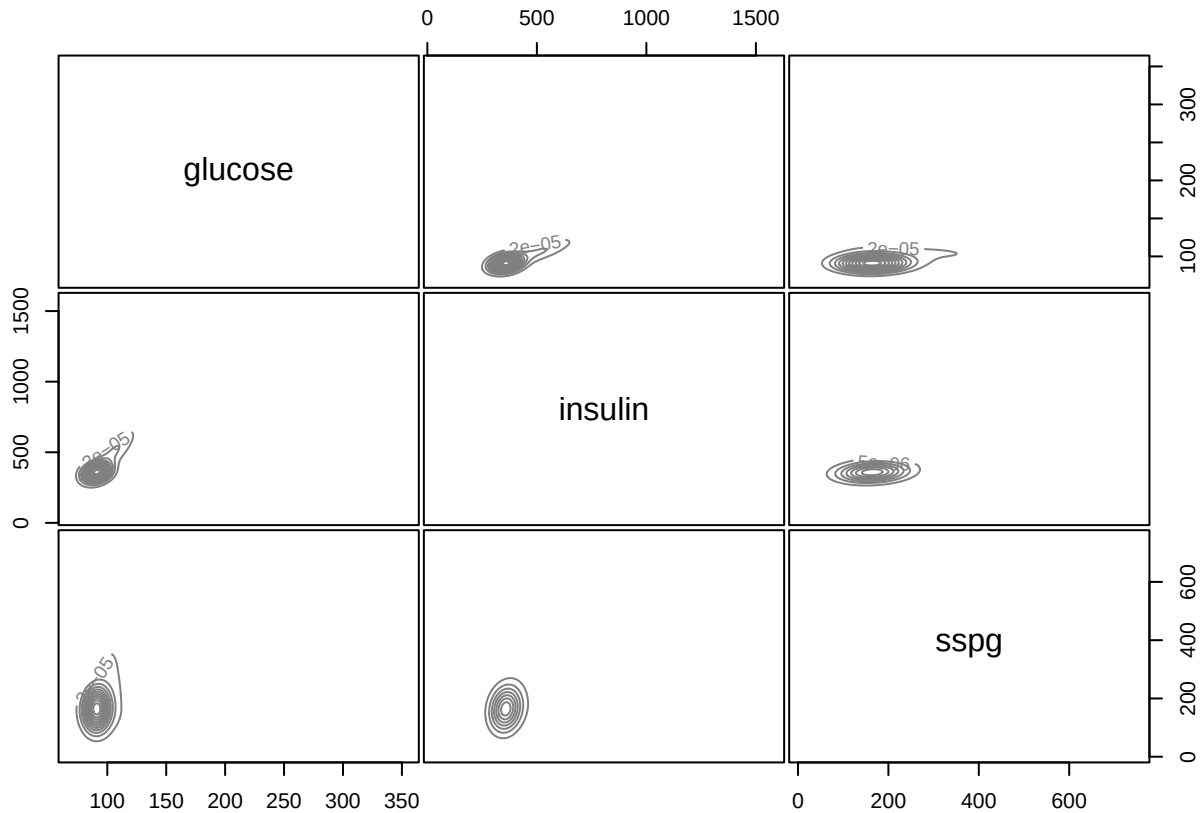
```
## 2 -4825.046
## 3 -4751.316
## 4 -4784.322
## 5 -4804.197
## 6 -4830.825
## 7 -4858.575
## 8 -4907.467
## 9 -4948.764
##
## Top 3 models based on the BIC criterion:
##      VVV,3      VVV,4      VVV,5
## -4751.316 -4784.322 -4804.197
```

```
plot(clust1)
```









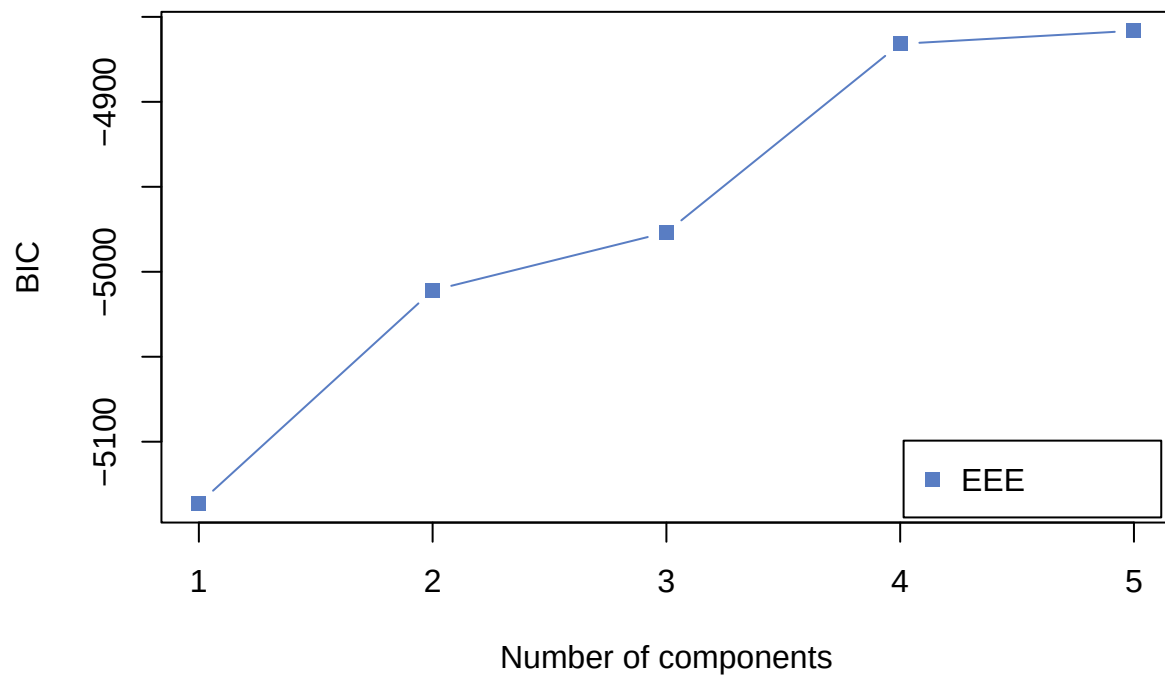
Compare this to a more restricted Gaussian mixture distribution, where the clusters are all the same shape and volume

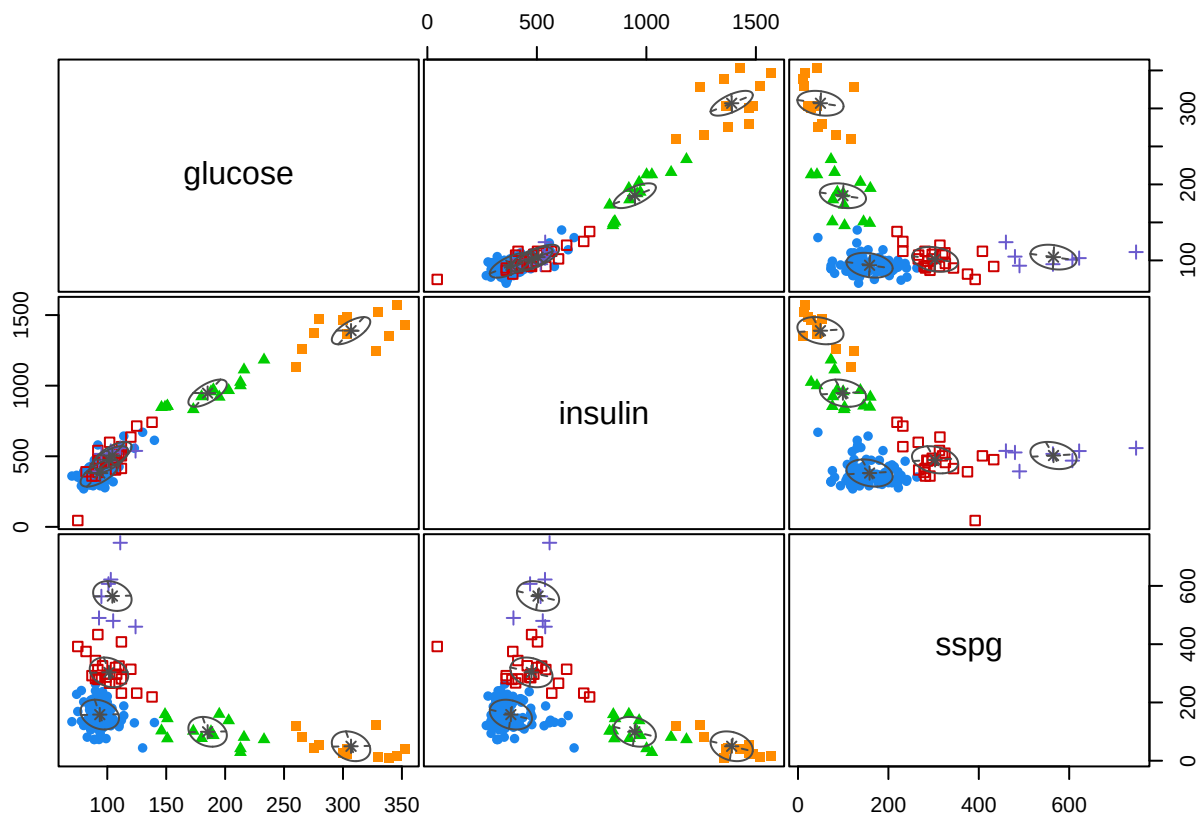
```
clust2 <- Mclust(data=diabetes[,c("glucose","insulin","sspg")], G=1:5, modelNames="EEE")
summary(clust2)
## -----
## Gaussian finite mixture model fitted by EM algorithm
## -----
##
## Mclust EEE (ellipsoidal, equal volume, shape and orientation) model with 5
## components:
##
## log-likelihood  n df      BIC      ICL
##      -2366.877 145 25 -4858.171 -4867.322
##
## Clustering table:
##  1  2  3  4  5
## 89 23 14  7 12

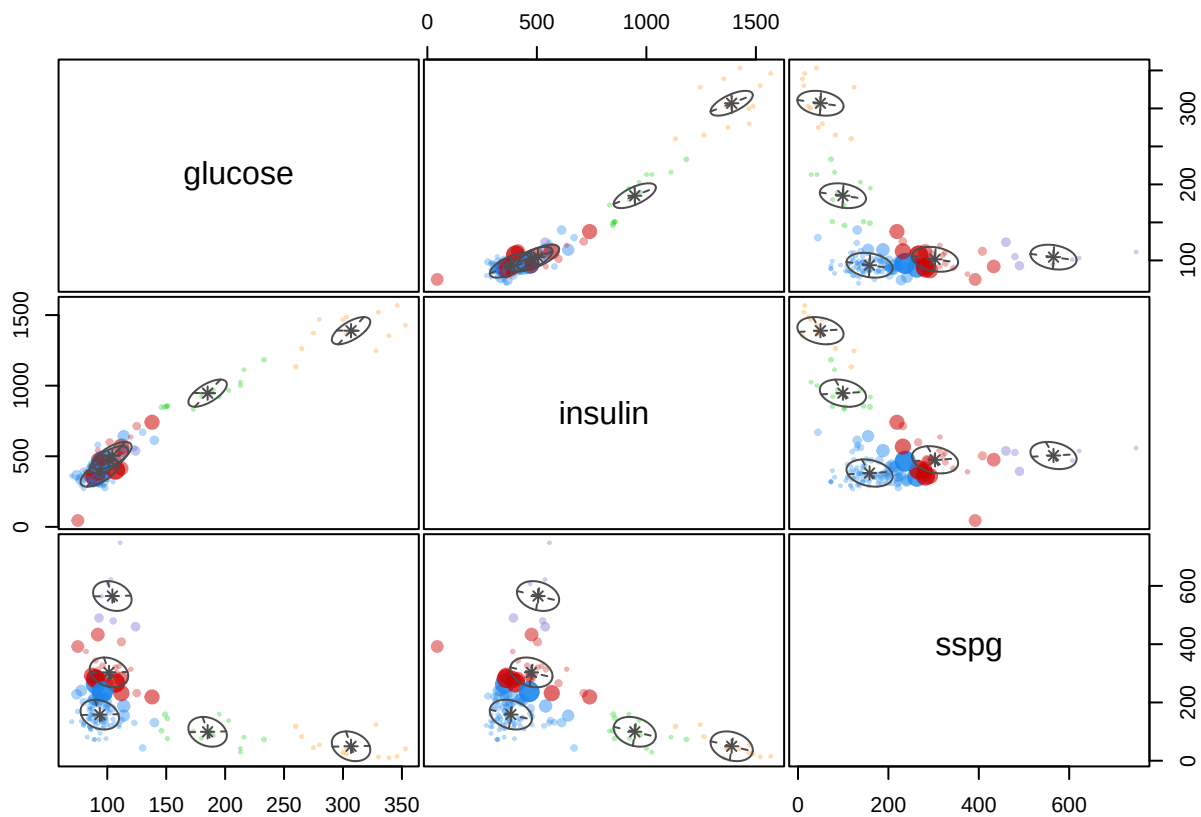
clust2$BIC
## Bayesian Information Criterion (BIC):
##      EEE
## 1 -5136.446
## 2 -5010.986
## 3 -4976.839
```

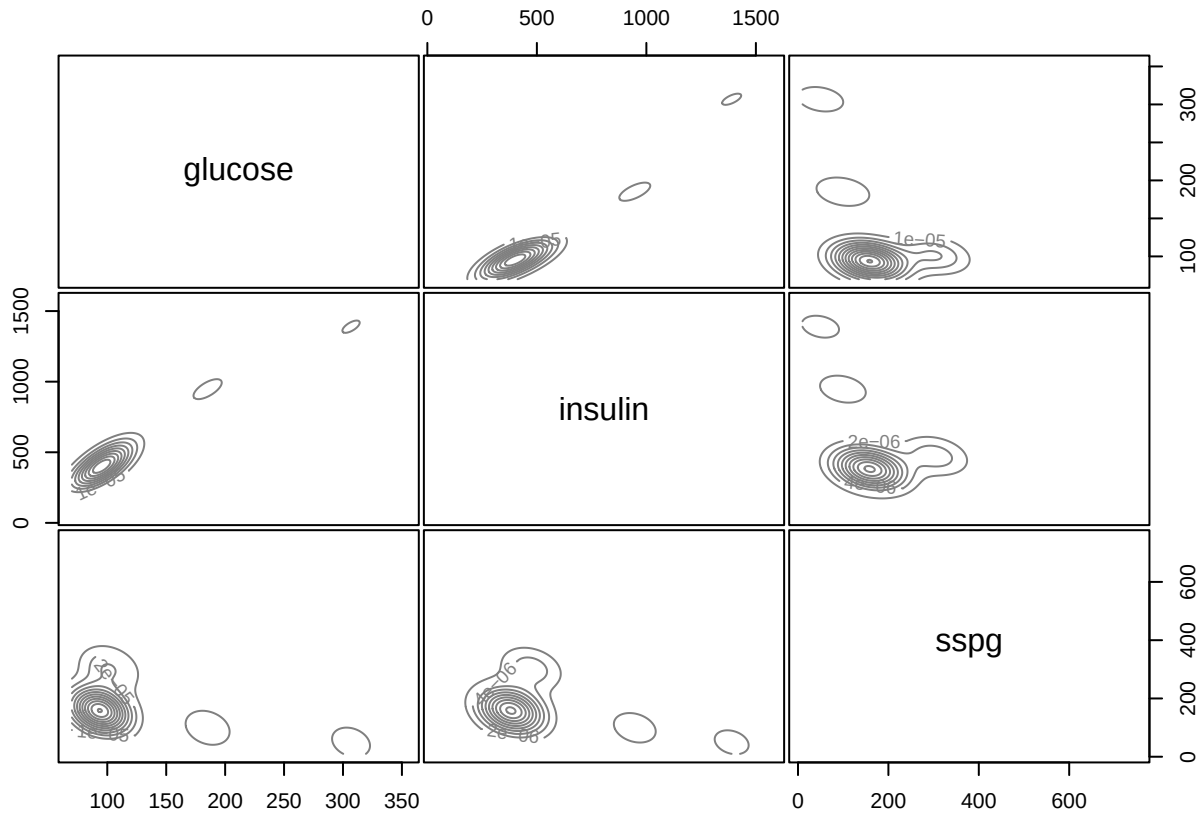
```
## 4 -4865.876
## 5 -4858.171
##
## Top 3 models based on the BIC criterion:
##   EEE,5   EEE,4   EEE,3
## -4858.171 -4865.876 -4976.839
```

```
plot(clust2)
```









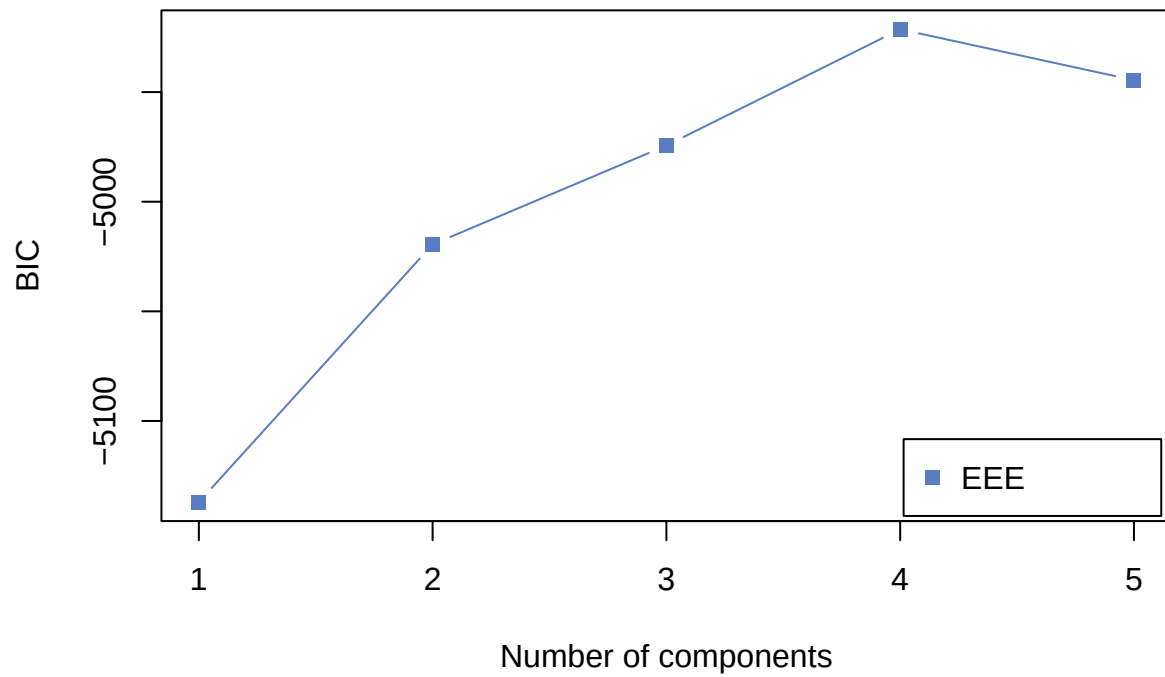
And bringing in some more Bayesian ideas, we can penalize the mixture likelihood, encouraging its components to have particular means and expressing the degree of shrinkage we want the prior (on the cluster-specific variance) to induce:

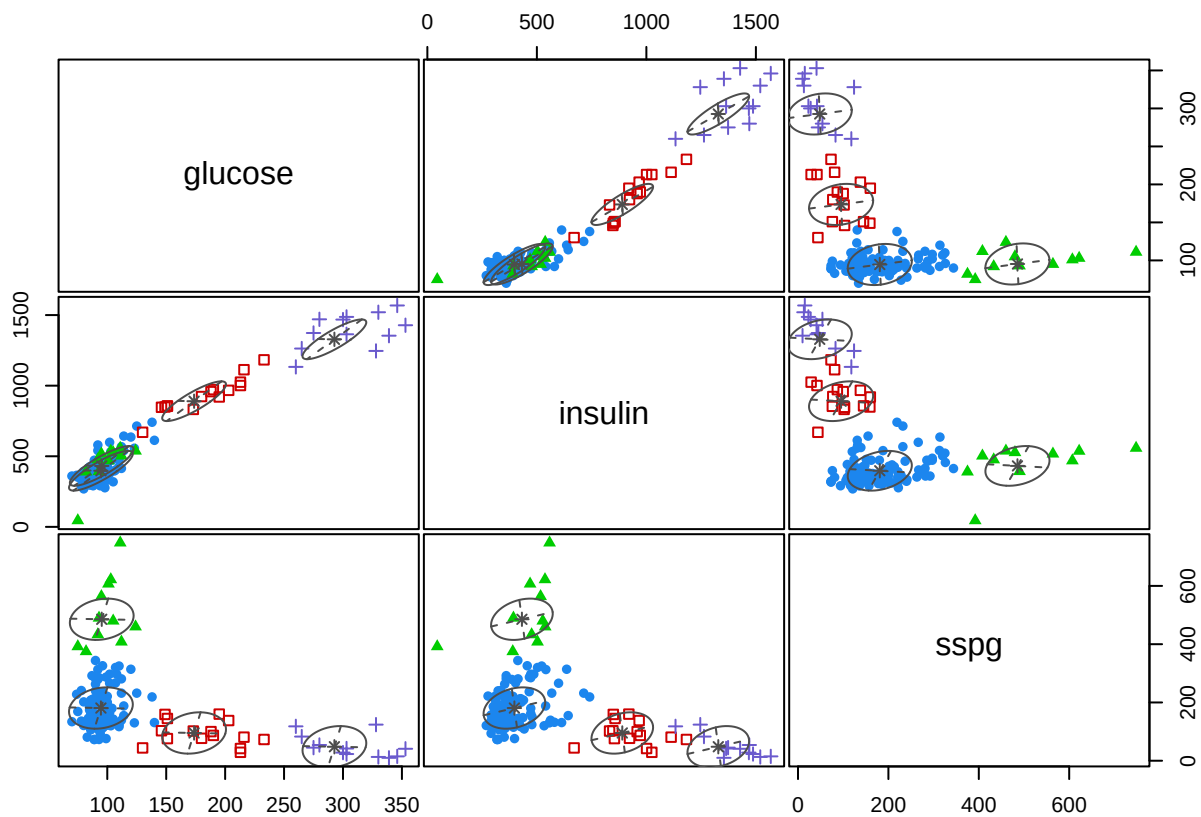
```
clust3 <- Mclust(data=diabetes[,c("glucose","insulin","sspg")], G=1:5, modelNames="EEE",
  prior=priorControl(shrinkage=0.5, mean=c(0,0,0) ))
summary(clust3)
## -----
## Gaussian finite mixture model fitted by EM algorithm
## -----
##
## Mclust EEE (ellipsoidal, equal volume, shape and orientation) model with 4
## components:
##
## Prior: defaultPrior(shrinkage = 0.5, mean = c(0, 0, 0))
##
## log-likelihood  n df      BIC      ICL
##      -2408.417 145 21 -4921.345 -4925.624
##
## Clustering table:
##   1  2  3  4
## 107 15 11 12

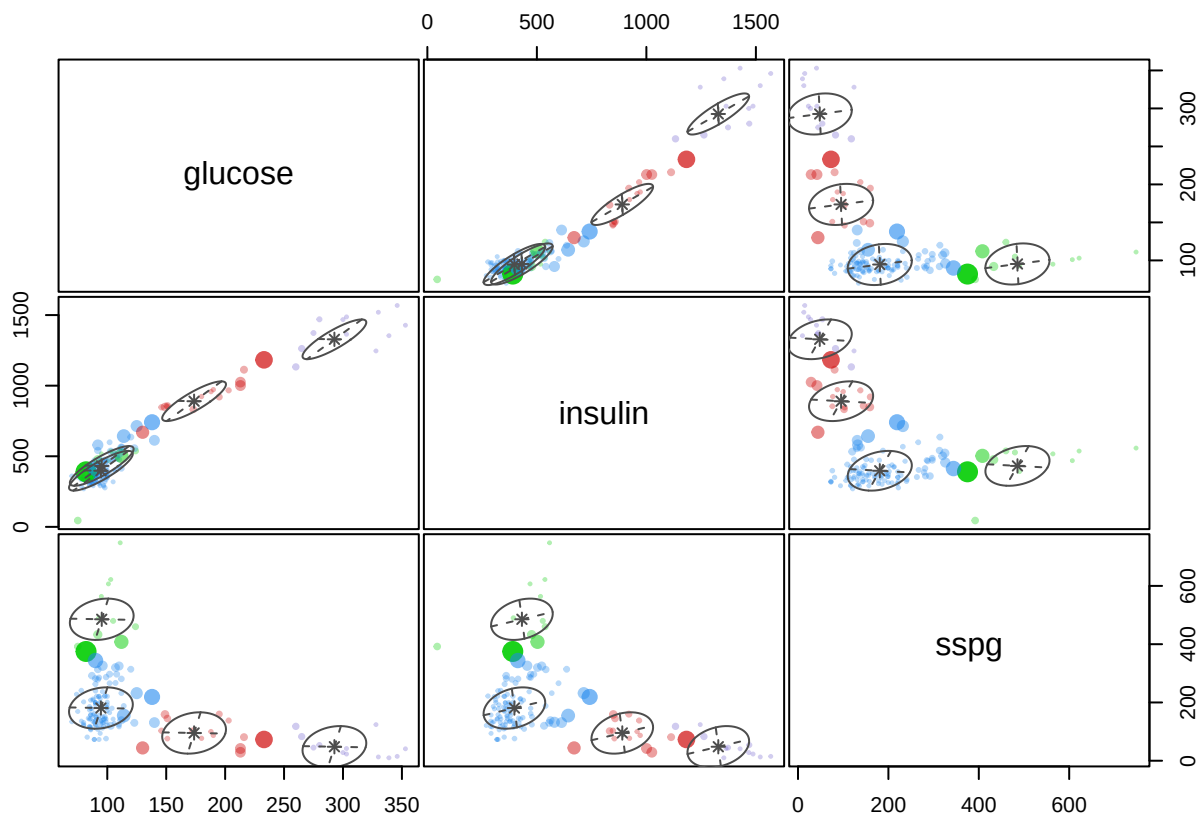
clust3$BIC
## Bayesian Information Criterion (BIC):
```

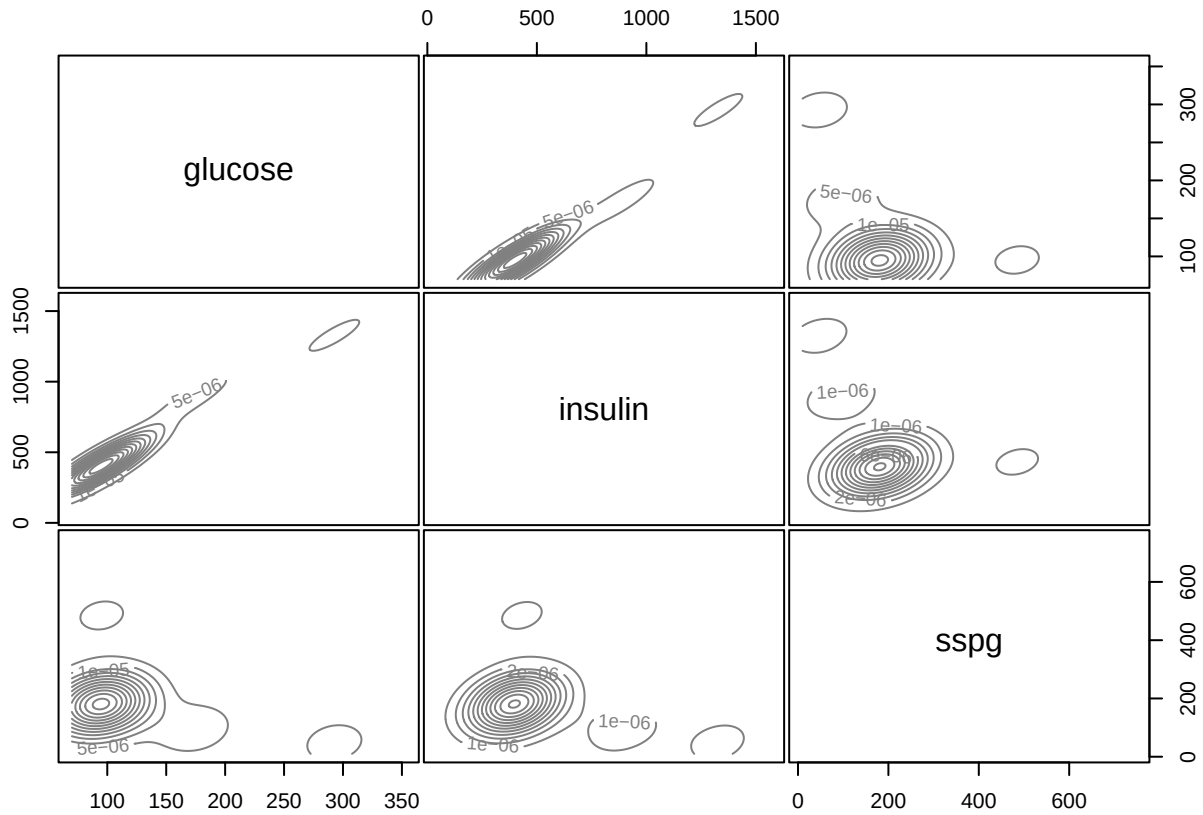
```
##      EEE
## 1 -5137.035
## 2 -5019.491
## 3 -4974.382
## 4 -4921.345
## 5 -4944.569
##
## Top 3 models based on the BIC criterion:
##      EEE,4      EEE,5      EEE,3
## -4921.345 -4944.569 -4974.382
```

```
plot(clust3)
```









Exercises

1. Specifying the “prior” had little impact on the results for the more restricted Gaussian mixture model. What happens if we use the same “prior” on the less restricted mixture model, where components have distinct covariances etc?