# Finite Mixture Models

## Example data: Banknote

STAT 32950-24620

Spring 2023 (5/4)



## Finite Mixture Models

A random vector $X$ with probability density (or mass) function $f(x), x \in \mathbb{R}^p$ is a $k$-mixture distribution if

$$f(x) = p_1 f_1(x) + \cdots + p_k f_k(x)$$

where

$$p_1 + \cdots + p_k = 1,$$

$k$ — number of mixtures

$p_i$ — membership probability



## Finite Mixture identifiability

In

$$f(x) = p_1 f_1(x) + \cdots + p_k f_k(x)$$

$f_i(x), x \in \mathbb{R}^p$ are probability density (or mass) functions, $\int_{\mathbb{R}^p} f_i(x)dx = 1$.

Identifiability:

$$f_i \neq f_j \quad \textit{when} \quad i \neq j, \qquad p_i > 0, \qquad i, j = 1, \cdots, k.$$



## Finite mixture model Example

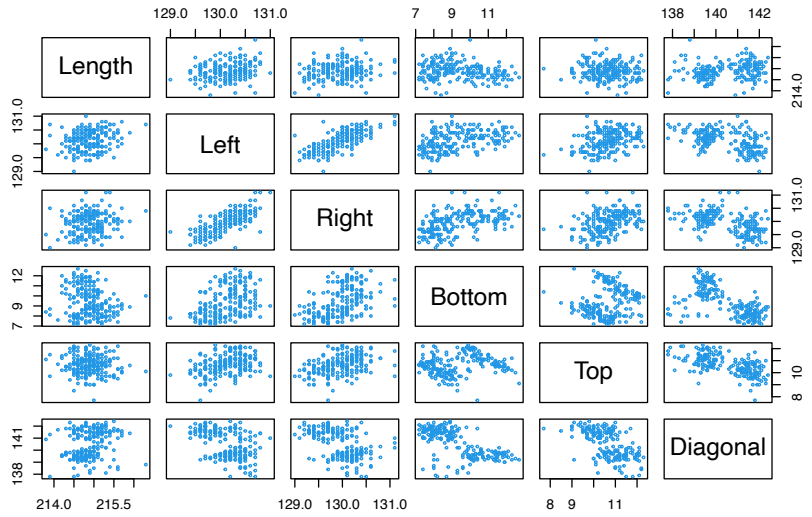
Example: Banknote data

```
# {mclust}:
library(mclust)
#Normal Mixture for Cluster, Classify, density est.
data(banknote)
str(banknote)
```

```
## 'data.frame':    200 obs. of  7 variables:
##  $ Status  : Factor w/ 2 levels "counterfeit",..: 2 2 2
##  $ Length  : num  215 215 215 215 215 ...
##  $ Left    : num  131 130 130 130 130 ...
##  $ Right   : num  131 130 130 130 130 ...
##  $ Bottom  : num  9 8.1 8.7 7.5 10.4 9 7.9 7.2 8.2 9.2 .
##  $ Top     : num  9.7 9.5 9.6 10.4 7.7 10.1 9.6 10.7 11
##  $ Diagonal: num  141 142 142 142 142 ...
```

## Visual check of data

```
pairs(banknote[2:7],cex=.4,col=4)
```






## Mixture model example setup

- Using the measurements only (variable 2 to 7)
- Treat the data as "unsupervised"
- Cluster anlaysis using finite Gaussian mixture model
- EM algorithm for parameter estimation
- Number of clusters $k = 2, \cdots, 9$ is selected by BIC

$$BIC = -2\ln(L) + k^* \ln(n)$$

where $L$ is the maximum likelihood of the model given the data,

$k^*$ is the number of parameters in the model.
(not to be confused with the number of mixtures $k$)



## EM algorithm estimation and model fit

```
mbank=Mclust(banknote[,2:7]) #BIC choose k from 2 to 9
summary(mbank)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VVE (ellipsoidal, equal orientation) model with 3
##
##  log-likelihood   n df   BIC   ICL
##          -663.4 200 53 -1608 -1608
##
## Clustering table:
##  1  2  3
## 18 98 84
```



## Number of clusters selcted

The chosen number of clusters is

$$k = 3$$

BIC is used to selection $k$.

Under various covariance structure assumptions

In terms of

- Shape (e.g. ellipsoidal)
- Volume (equal or unequal among clusters)
- Orientation (equal or unequal)

## BIC of selected mixture model

For the selected model of $k = 3$ mixtures,

Number of observations $n = 200$

Number of parameters $k^* = 53$

Log-likelihood $log(L) = -663.4$

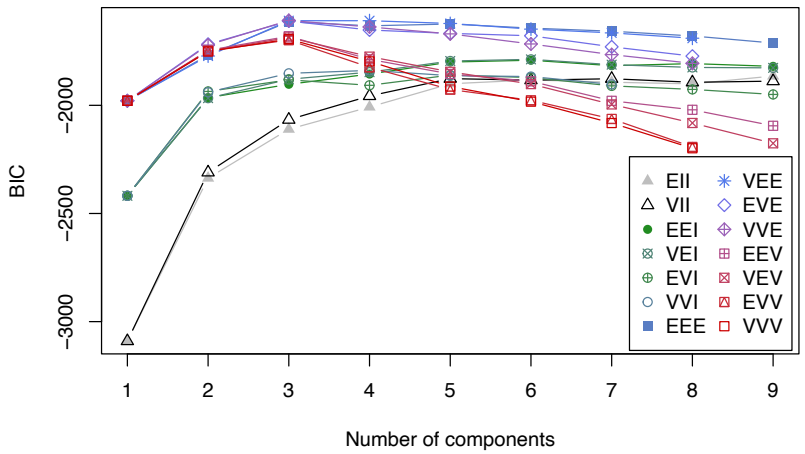
$$\Rightarrow \quad BIC = -2ln(L) + k * ln(n) = 1607.6 \approx 1608$$

Note: in `mclust` -BIC is used (but called BIC)

```
plot(mbank,what=c("BIC"))
```



## BIC plot for model selection






## Best candidates of model assumptions by BIC

```
summary(mclustBIC(banknote[,2:7]))
```

```
## Best BIC values:
##          VVE,3     VEE,4    VEE,3
## BIC      -1608 -1608.768 -1608.79
## BIC diff     0    -1.194    -1.22
```

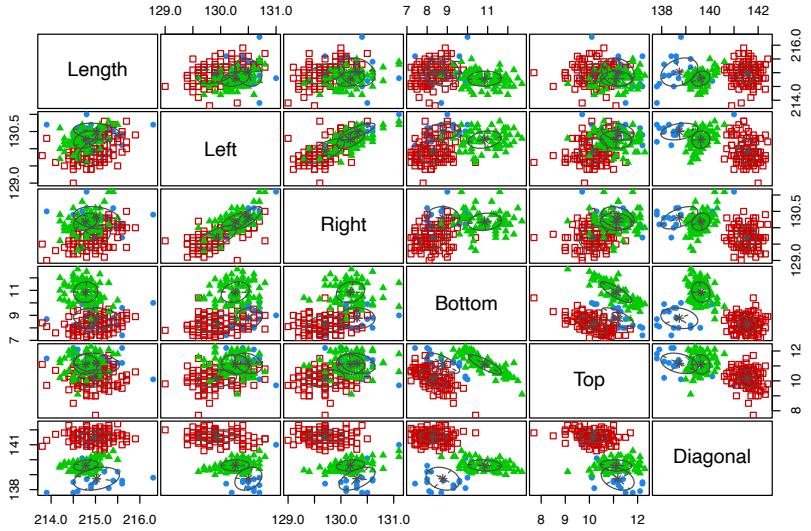
Model types:

- "EEE" = ellipsoidal, equal volume, shape, and orientation
- "EVE" = ellipsoidal, equal volume and orientation
- "VEE" = ellipsoidal, equal shape and orientation
- "VVE" = ellipsoidal, equal orientation

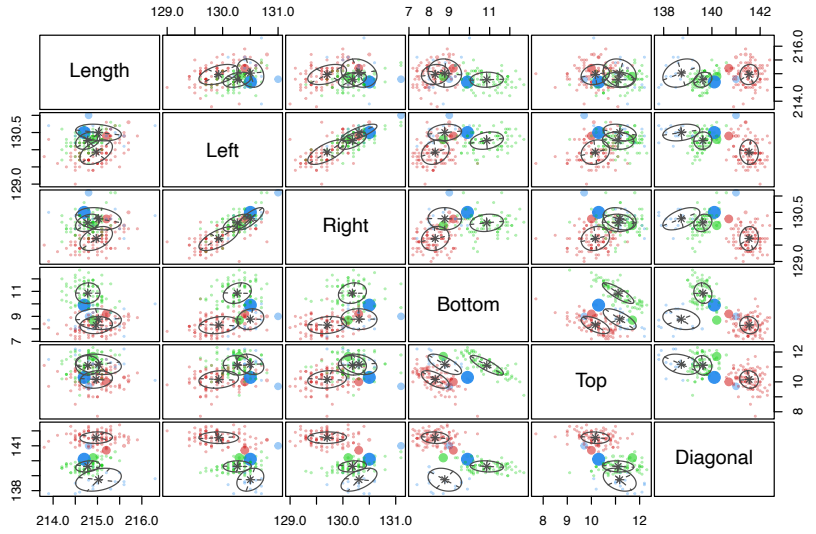
```
#??mclustModelNames # for model name definitions
```



```
plot(mbank,what=c("classification"))  # => k=3 selected
```

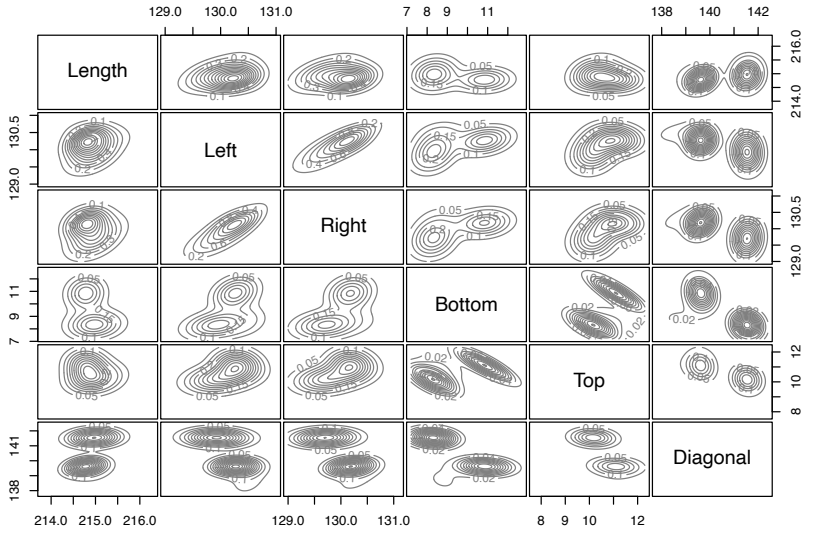


```
plot(mbank,what=c("uncertainty"))
```



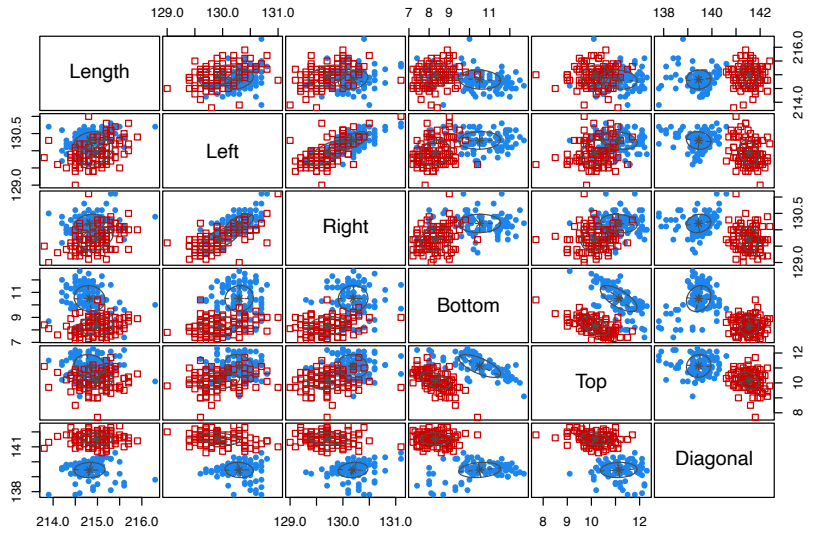



```
plot(mbank,what=c("density"))  # 3 clusters
```



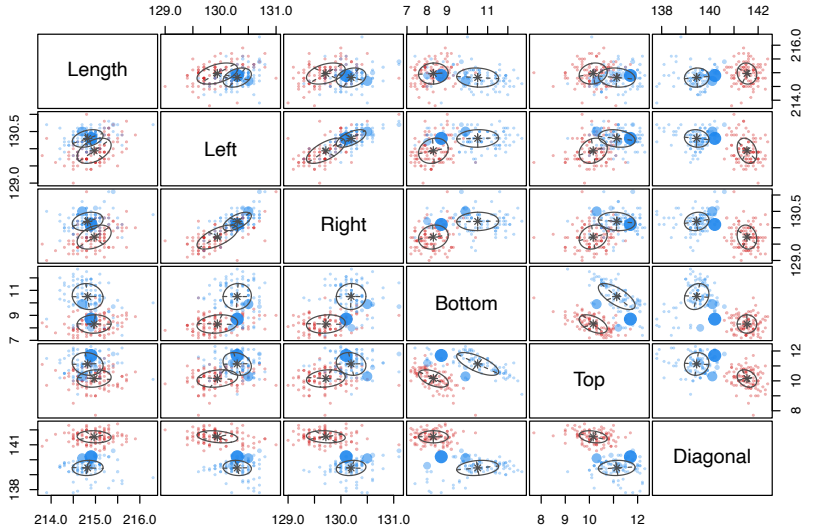



```
bank2=Mclust(banknote[,2:7],G=2)  # force k=2
plot(bank2, what = c("classification"))
```

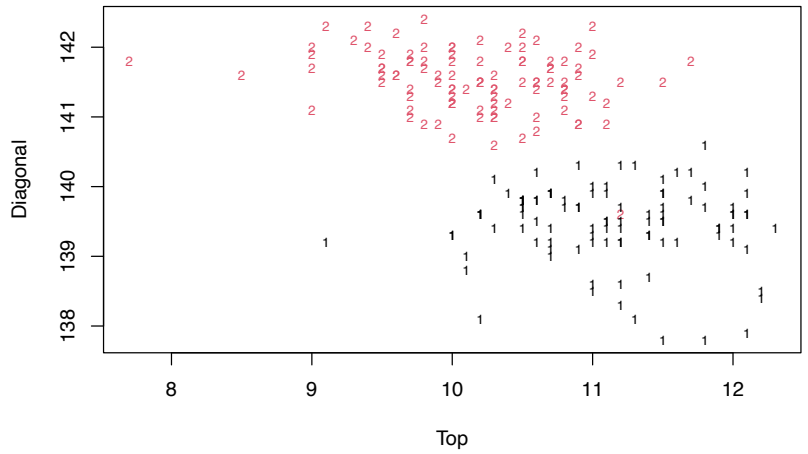





```
plot(bank2, what = c("uncertainty"))
```



```
par(mfrow=c(1,1))
plot(banknote[,6:7],type="n")
text(banknote[,6:7],label=as.numeric(banknote[,1]),
     col=as.numeric(banknote[,1]),cex=.7,lwd=2)
```






## Fix number of mixtures $k = 2$

```
summary(Mclust(banknote[,2:7], G=2))
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EVE (ellipsoidal, equal volume and orientation) m
##
##  log-likelihood   n df   BIC   ICL
##          -755.4 200 39 -1717 -1717
##
## Clustering table:
##   1   2
## 101  99
```



## Model assumptions for $k = 2$ mixtures

```
summary(mclustBIC(banknote[,2:7],G=2))
```

```
## Best BIC values:
##          EVE,2     VVE,2    EEV,2
## BIC      -1717 -1721.311 -1745.84
## BIC diff     0    -3.867   -28.39
```

- "EVE" = ellipsoidal, equal volume and orientation
- "VEE" = ellipsoidal, equal shape and orientation
- "EEV" = ellipsoidal, equal volume and equal shape



## Estimated mixture proportions (for $k = 2, 3$)

For $k = 2$

```
Mclust(banknote[,2:7], G=2)$parameters$pro
```

```
## [1] 0.5049 0.4951
```

For $k = 3$

```
Mclust(banknote[,2:7])$parameters$pro
```

```
## [1] 0.08988 0.49005 0.42006
```

```
#summary(Mclust(banknote[,2:7]), parameters=T)
```