ETC3250/5250: Introduction to Machine Learning

Model-based clustering

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 11a

Overview

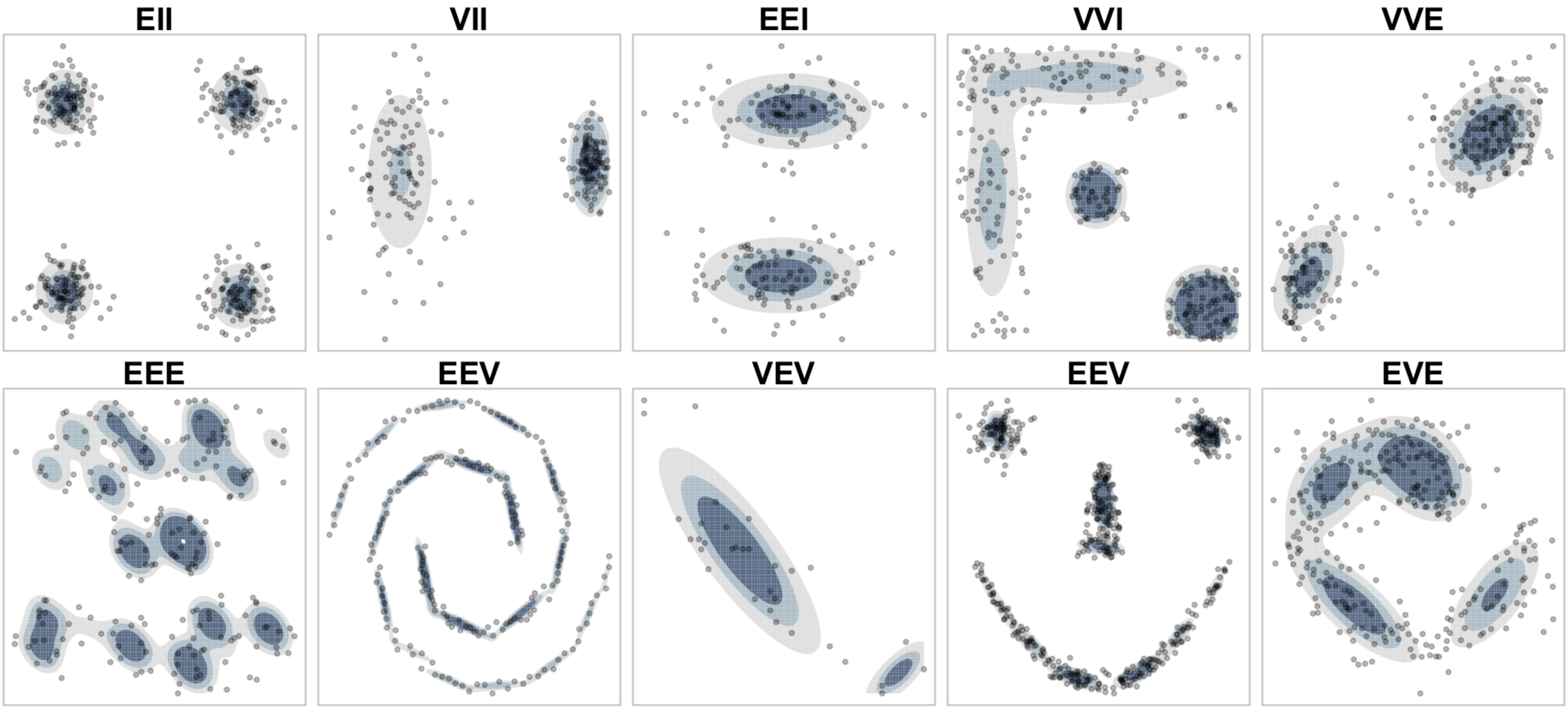
Model-based clustering makes an assumption about the distribution of the data, primarily

- Assumes the data is a sample from a Gaussian mixture model
- Requires the assumption that clusters have an elliptical shape
- The shape is determined by the variance-covariance of the clusters
- A variety of models is available by using different constraints on the variance-covariance

Model is

$$f(x_i) = \sum_{k=1}^{G} \pi_k f_k(x_i; \mu_k, \Sigma_k)$$

where f_k is usually a multivariate normal distribution. The parameters are estimated by maximum likelihood, and choice between models is made using BIC.



Source: Boehmke (2020) Hands-on machine learning

Variance-covariance specification

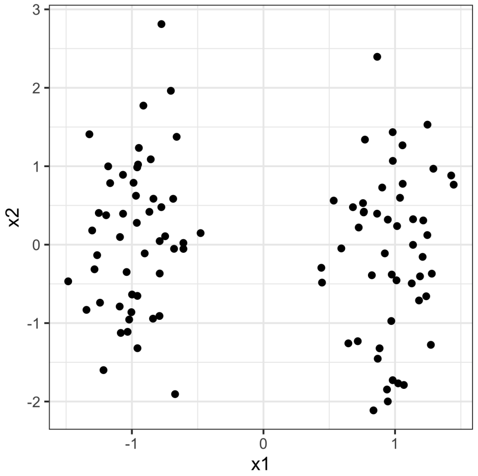
Constraints applied on cluster variance-covariance:

1. volume: each cluster has approximately the same size
2. shape: each cluster has approximately the same variance so that the distribution is spherical
3. orientation: each cluster is forced to be axis-aligned

Variance-covariance contraints

Model	Family	Volume	Shape	Orientation	Identifier
1	Spherical	Equal	Equal	NA	EII
2	Spherical	Variable	Equal	NA	VII
3	Diagonal	Equal	Equal	Axes	EEI
6	Diagonal	Variable	Variable	Axes	VVI
7	General	Equal	Equal	Equal	EEE
8	General	Equal	Variable	Equal	EVE
10	General	Variable	Variable	Equal	VVE
11	General	Equal	Equal	Variable	EEV
12	General	Variable	Equal	Variable	VEV
14	General	Variable	Variable	Variable	VVV

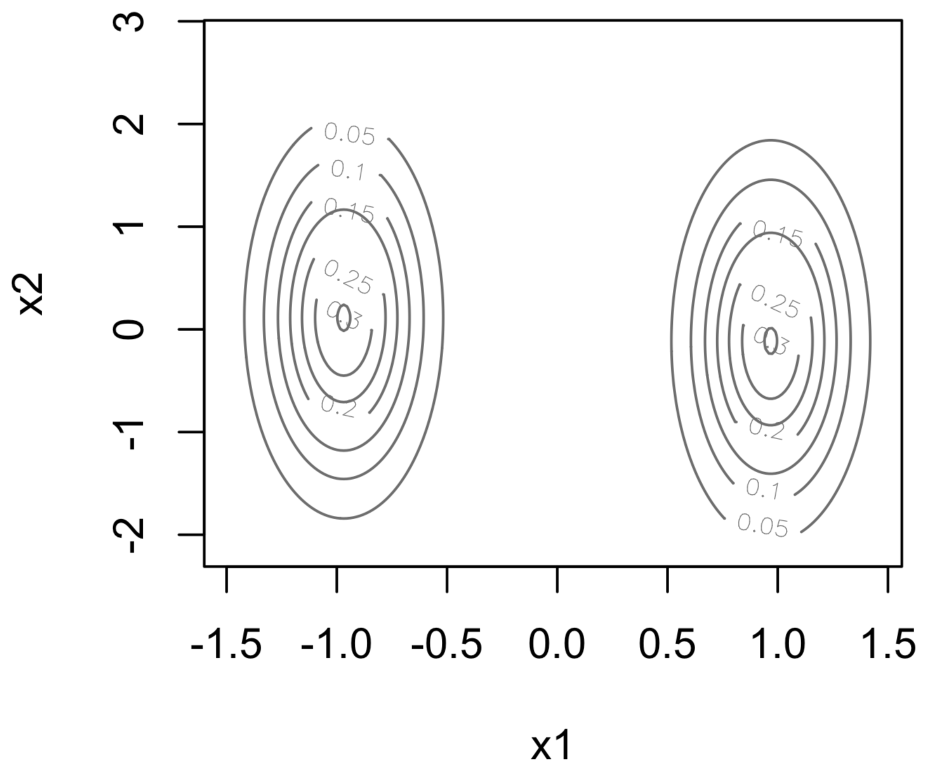
Example: nuisance variable



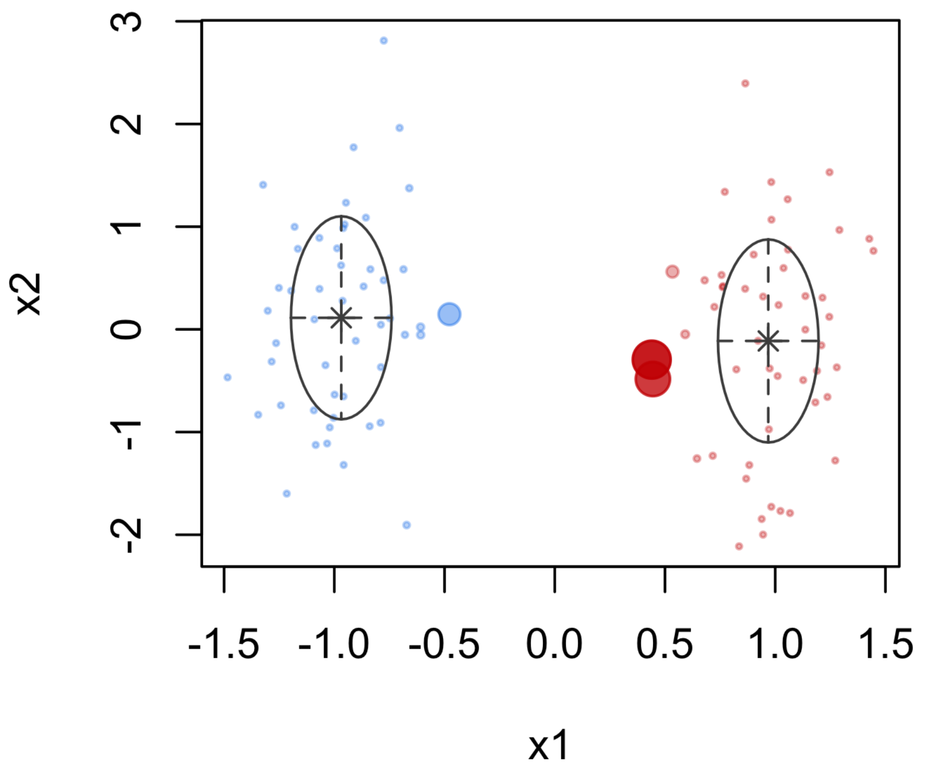

```
df_mc <- Mclust(df, G = 2)
summary(df_mc)
```

```
## -----------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## -----------------------------------------------
##
## Mclust EEI (diagonal, equal volume and shape) model 
##
##  log-likelihood   n df       BIC       ICL
##       -204.1509 100  7 -440.538 -440.538
##
## Clustering table:
##  1  2
## 50 50
```

```r
plot(df_mc, what = "density")
```




```r
plot(df_mc, what = "uncertainty")
```



Parameter estimates

Cluster means

```
options(digits=2)
df_mc$parameters$mean
```

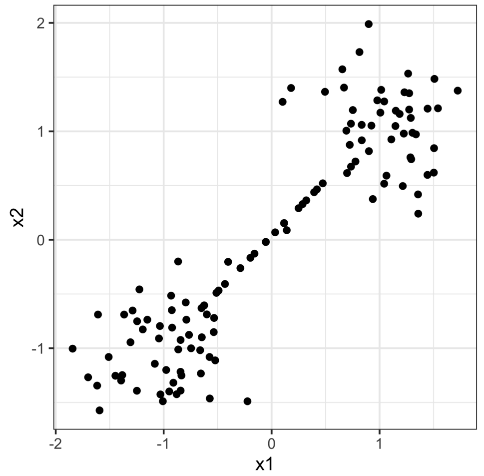
```
##      [,1]  [,2]
## x1 -0.97  0.97
## x2  0.11 -0.11
```

Cluster variances

```
df_mc$parameters$variance$sigma
```

```
## , , 1
##
##        x1   x2
## x1 0.052 0.00
## x2 0.000 0.98
##
## , , 2
##
##        x1   x2
## x1 0.052 0.00
## x2 0.000 0.98
```

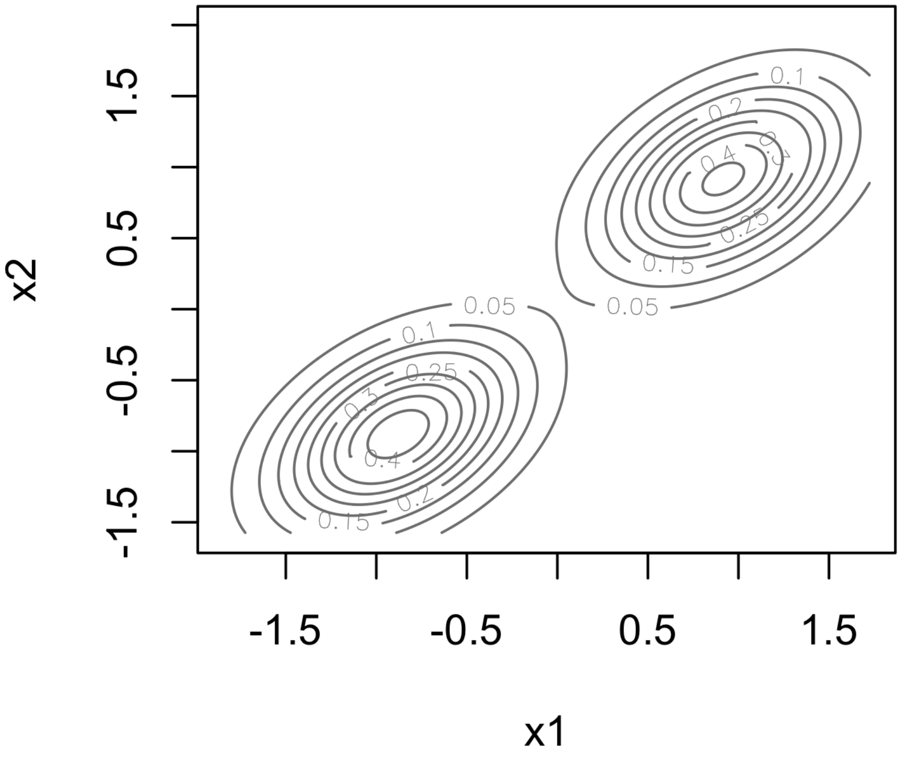
Example: nuisance observations



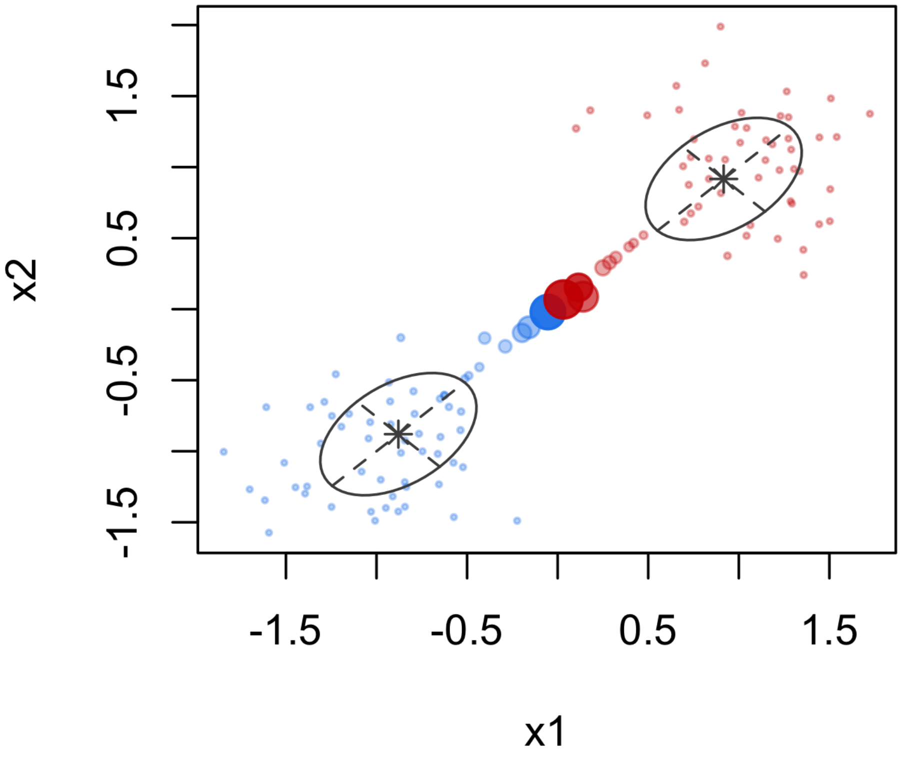

```r
df_mc <- Mclust(df, G = 2)
summary(df_mc)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEE (ellipsoidal, equal volume, shape and ori
## components:
##
##  log-likelihood   n df  BIC  ICL
##            -205 120  8 -447 -452
##
## Clustering table:
##  1  2
## 61 59
```

```r
plot(df_mc, what = "density")
```




```r
plot(df_mc, what = "uncertainty")
```



Parameter estimates

Cluster means

```
df_mc$parameters$mean
```

```
##     [,1] [,2]
## x1 -0.88 0.92
## x2 -0.88 0.92
```

Cluster variances

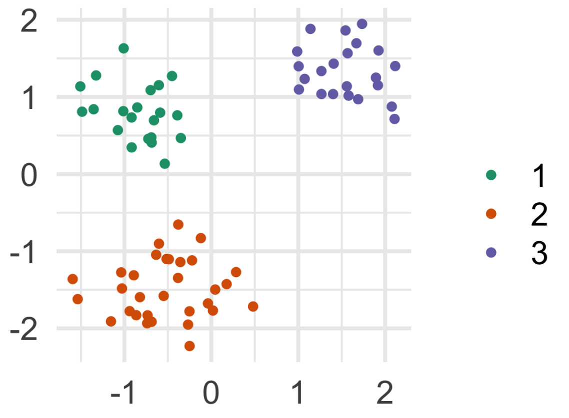
```
df_mc$parameters$variance$sigma
```

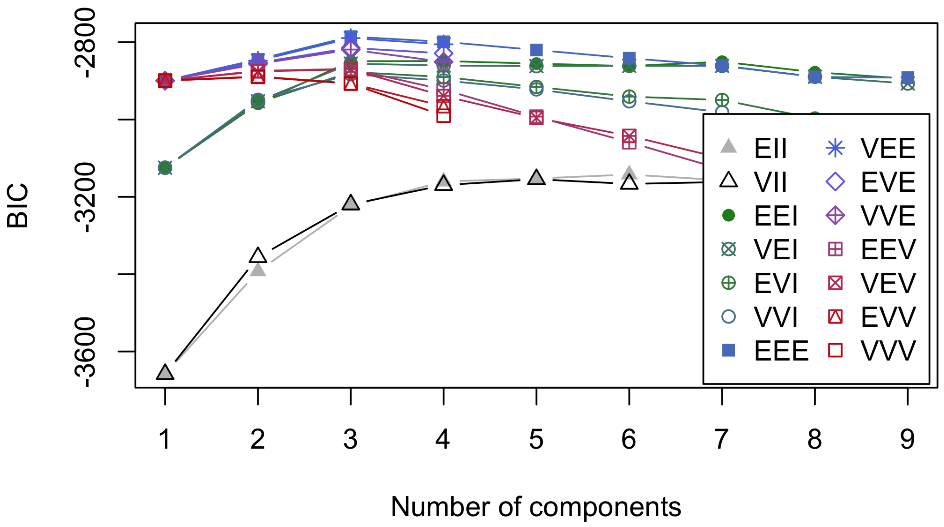
```
## , , 1
##
##        x1    x2
## x1 0.186 0.081
## x2 0.081 0.185
##
## , , 2
##
##        x1    x2
## x1 0.186 0.081
## x2 0.081 0.185
```

Example: flea with nuisance variables and observations

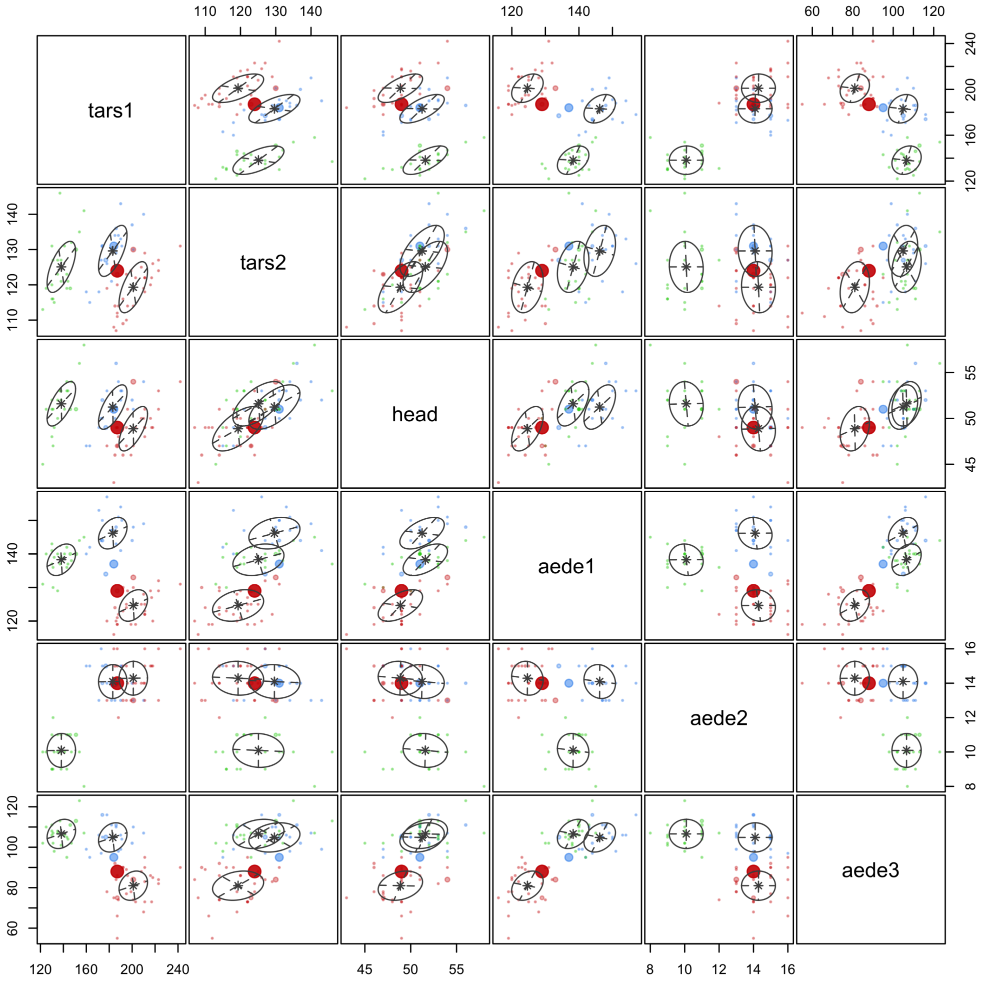
```
set.seed(6)
data(flea)
flea_mc <- Mclust(flea[,2:7])
summary(flea_mc)

## -----------------------------------------
## Gaussian finite mixture model fitted by EM al
## -----------------------------------------
##
## Mclust EEE (ellipsoidal, equal volume, shape a
## components:
##
##  log-likelihood  n df   BIC   ICL
##           -1305 74 41 -2786 -2786
##
## Clustering table:
##  1  2  3
## 21 31 22
```

Let model-based decide number of clusters, and variance-covariance parametrization.





- The spherical models are clearly inferior.
- Across many models, three clusters is where ther is a peak in BIC
- A few parametrisations are nearly equally as good, pick the simplest.



Parameter estimates

Cluster means

```
flea_mc$parameters$mean

##        [,1] [,2] [,3]
## tars1   183  201  138
## tars2   130  119  125
## head     51   49   52
## aede1   146  125  138
## aede2    14   14   10
## aede3   105   81  107
```

Cluster variances

```
flea_mc$parameters$variance$sigma[

##        tars1 tars2  head aede1
## tars1 154.70 56.36 19.99 21.83
## tars2  56.36 52.48 10.78  9.44
## head   19.99 10.78  5.87  6.22
## aede1  21.83  9.44  6.22 22.09
## aede2   0.15 -0.46 -0.22 -0.54
## aede3  20.13 11.51  4.61 11.22
```

Summary

- Model-based clustering provides a nice automated clustering, if the data has neatly separated clusters, even in the presence of nuisance variables.
- Non-elliptical clusters could be modeled by combining multiple ellipses.
- It is affected by nuisance observations, and has a parameter `noise` to attempt to filter these.
- It may not function so well if the data hasn't got separated clusters.
- k-means and Wards linkage hierarchical would yield similar results to constraining the variance-covariance model to EEI (or VII, EEE).
- Having a functional model for the clusters is useful.

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 11a