MONASH University

ETC3250/5250: Introduction to Machine Learning

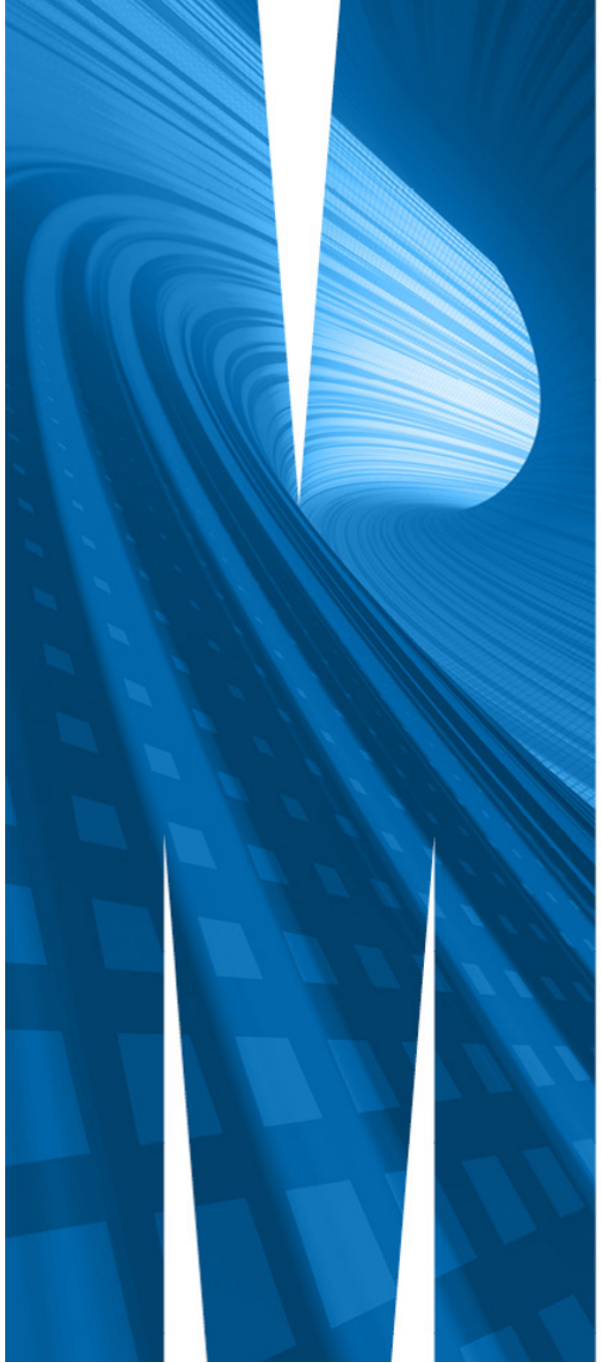
Model-based clustering

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 11b



Overview

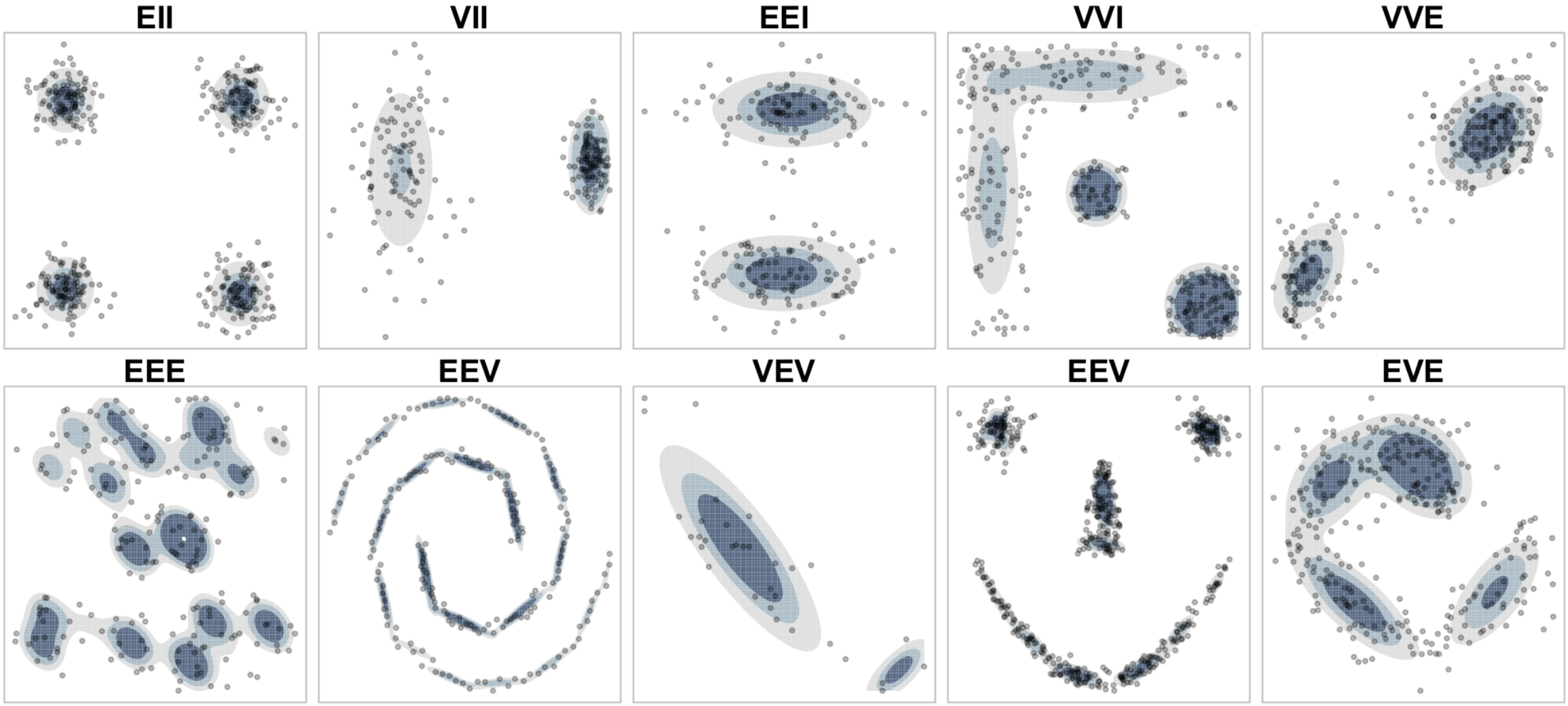
Model-based clustering makes an assumption about the distribution of the data, primarily

- 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

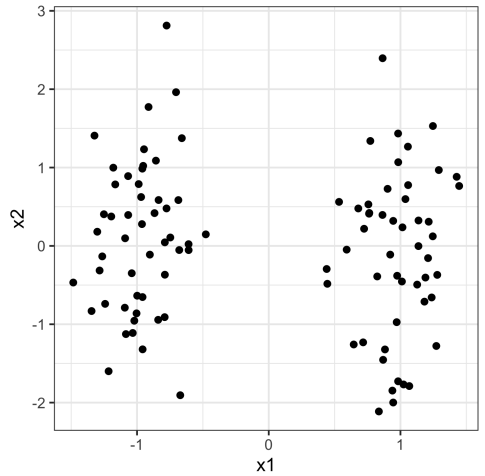
Constraints applied on cluster variance-covariance:

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

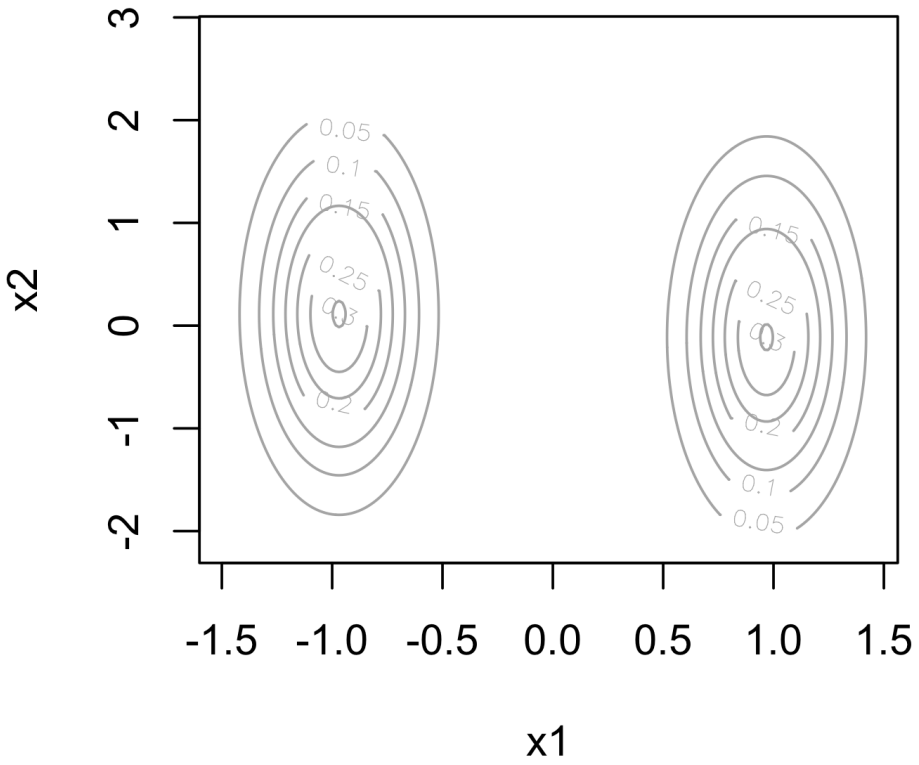
Model	Family	Volume	Shape	Orientation	Identifier
1	Spherical	Equal	Equal	NA	EII
2	Spherical	Variable	Equal	NA	VII
3	Diagonal	Equal	Equal	Axes	EEI
6	Diagonal	Variable	Variable	Axes	VVI
7	General	Equal	Equal	Equal	EEE
8	General	Equal	Variable	Equal	EVE
10	General	Variable	Variable	Equal	VVE
11	General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

Example: nuisance variable

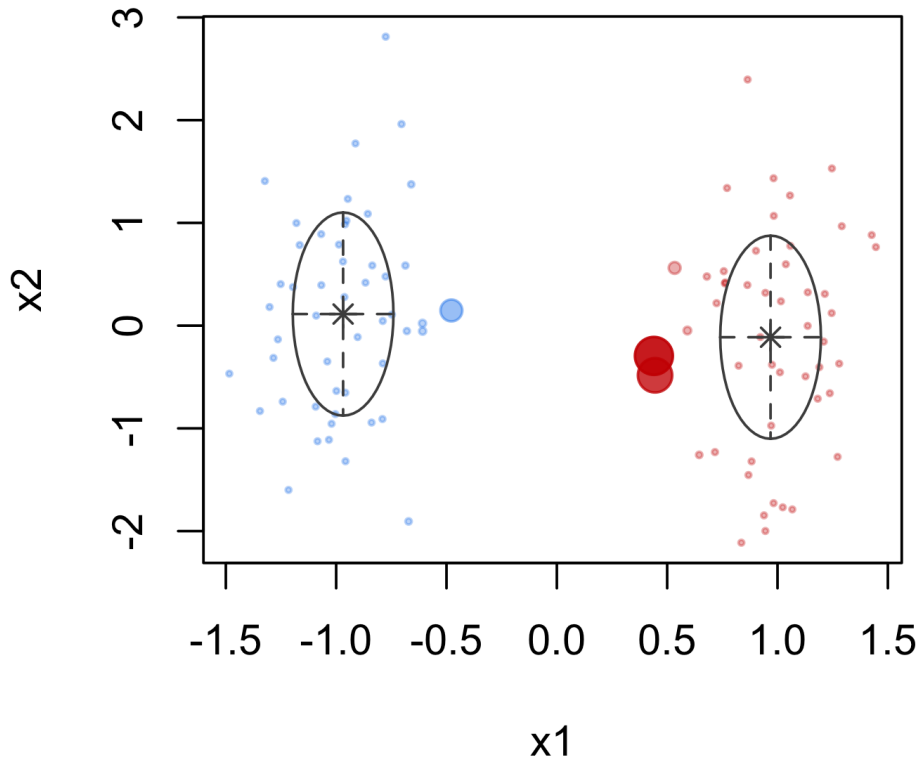



```
df_mc <- Mclust(df, G = 2)
summary(df_mc)

## ----------------------------------------------------
## 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
```

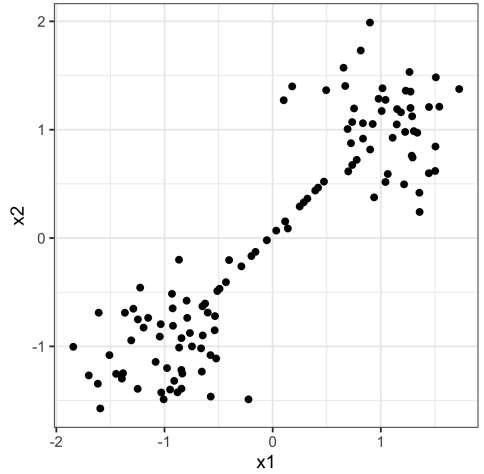
```
plot(df_mc, what = "density")
```




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

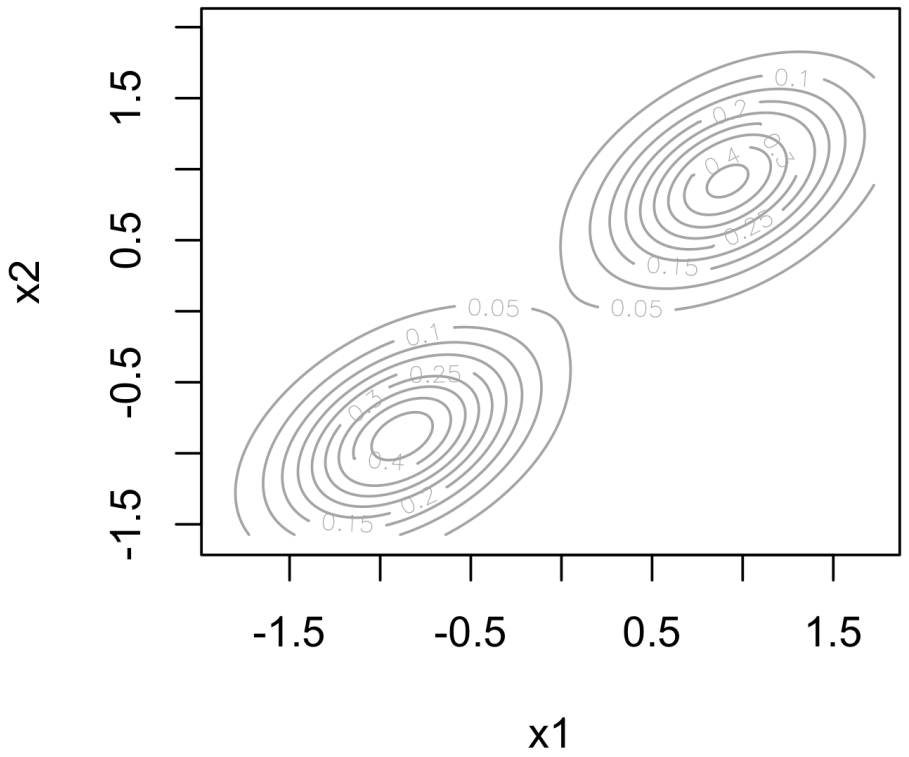


Example: nuisance observations

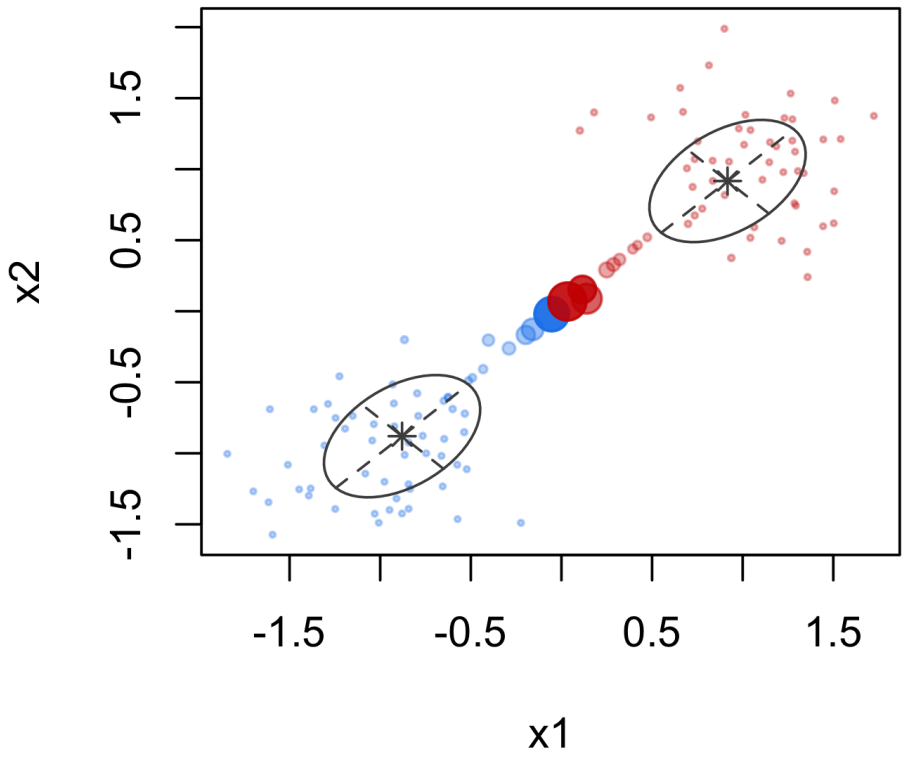



```
df_mc <- Mclust(df, G = 2)
summary(df_mc)

## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEE (ellipsoidal, equal volume, shape and ori
## components:
##
##  log-likelihood   n df       BIC       ICL
##       -204.5104 120  8 -447.3208 -451.7038
##
## Clustering table:
##  1  2
## 61 59
```

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



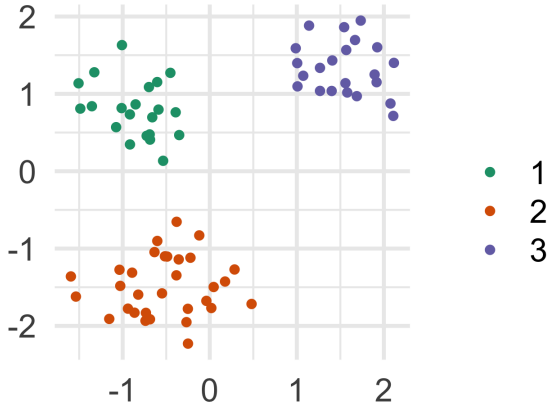

```
plot(df_mc, what = "uncertainty")
```

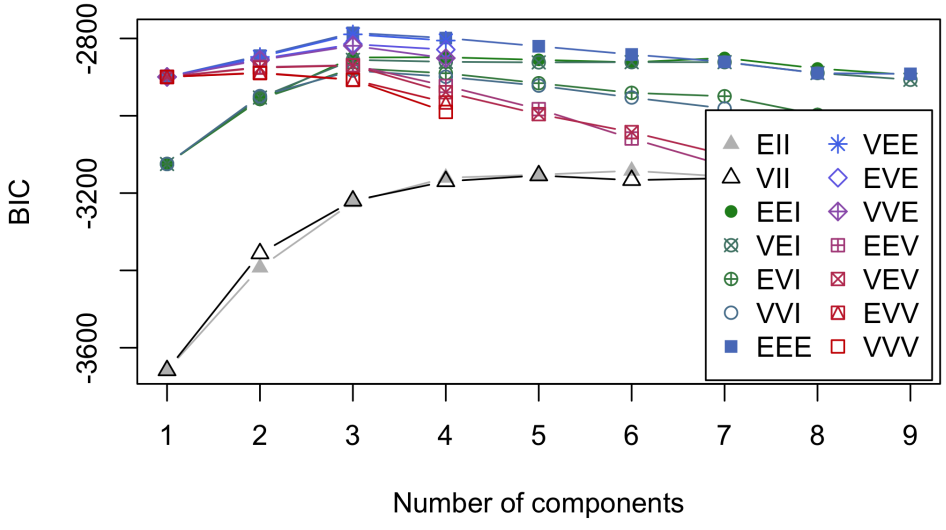


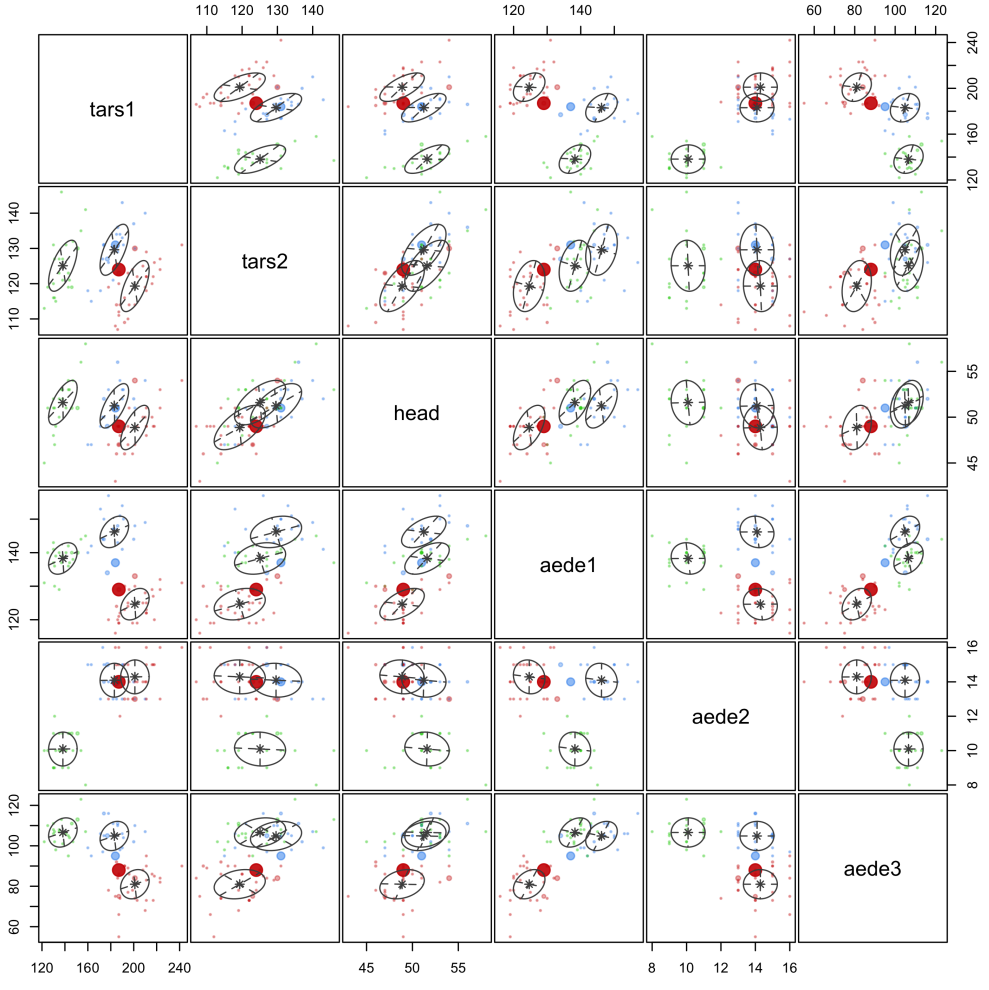
```
set.seed(6)
data(flea)
flea_mc <- Mclust(flea[,2:7])
summary(flea_mc)

## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM alg
## ----------------------------------------------------
##
## Mclust EEE (ellipsoidal, equal volume, shape a
## components:
##
##  log-likelihood  n df       BIC       ICL
##       -1304.552 74 41 -2785.572 -2785.574
##
## Clustering table:
##  1  2  3
## 21 31 22
```

Example: flea with nuisance variables and observations

original units



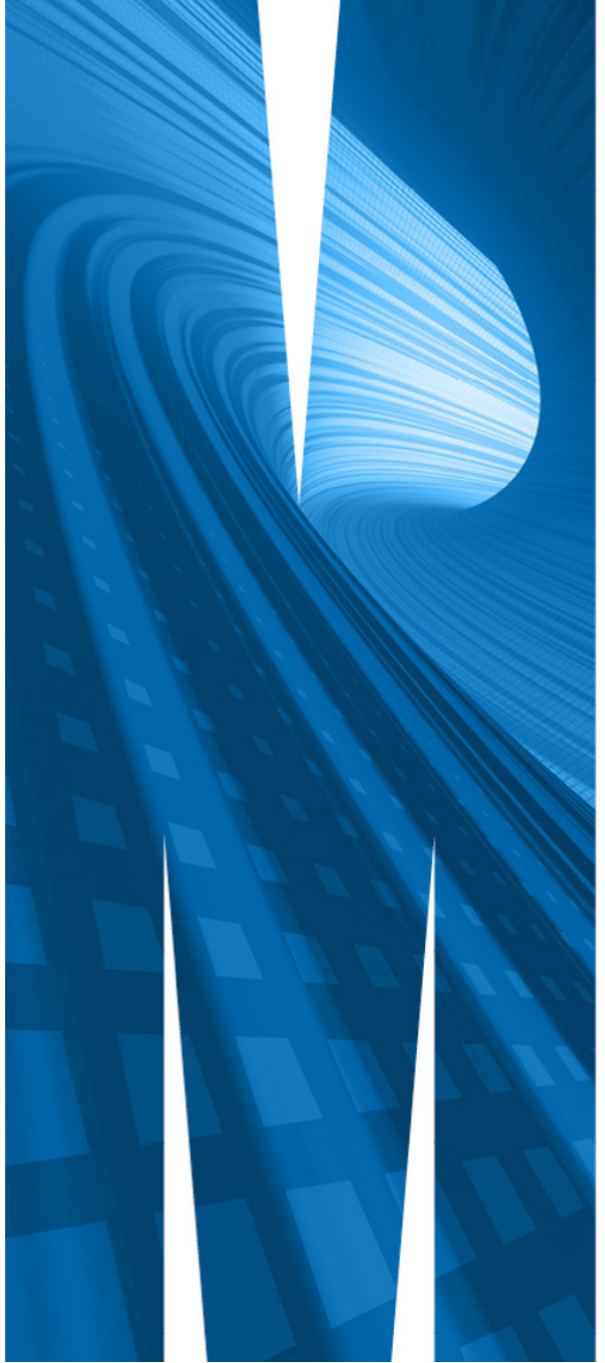

BIC
-2800
-3200
-3600
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
VEE
EVE
VVE
EEV
VEV
EVV
VVV


tars1
tars2
head
aede1
aede2
aede3

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.

MONASH University



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 11b