


## An Introduction to R: Statistics, Graphics and Data Mining


**Daniele Soria**  
 d.soria@cs.nott.ac.uk  
 IMA tutorial, 2 March 2010



## Outline

- n Why R?
- n Download and install
- n Data manipulation
- n Basic statistics
- n Graphics
- n Clustering

Daniele Soria    An Introduction to R: Statistics, Graphics and Data Mining    2




## Why R?

- n It allows users to add additional functionality by defining new functions.
- n Its graphical facilities produce publication-quality graphs which can include mathematical symbols.
- n And most importantly...

IT'S FREE!! J


Daniele Soria    An Introduction to R: Statistics, Graphics and Data Mining    3



## Download and install

- n <http://www.r-project.org>
- n On the left select **CRAN**
- n Scroll down to find UK and select the **University of Bristol** link
- n Select the OS you are using
- n For Windows select **base**
- n **Download R 2.10.1 for Windows** and install it


Daniele Soria    An Introduction to R: Statistics, Graphics and Data Mining    4



## Let's get started

- n In R-2.10.1 folder, several packages (libraries) are already present
- n Command line interface
- n Follow suggestion and type **demo()**
- n Type **demo(graphics)**
- n Script files may be written in R
  - n I use an external editor, called Crimson Editor

Daniele Soria    An Introduction to R: Statistics, Graphics and Data Mining    5



## Basic stuff

- n Create
  - n Variable
  - n Vector
  - n Sequence
  - n Replicate a value
- n Length, max, min, mean, summary
- n Dealing with missing values
- n Vectors and matrices

Daniele Soria    An Introduction to R: Statistics, Graphics and Data Mining    6

## Load data

- n UCI Machine Learning Repository
  - n Download IRIS data
- n `read.table(options)`
- n Dataset information
  - n Names of variables
  - n Size
  - n Tables
  - n `ifelse` command

## Graphics

- n Pie charts
- n Histograms
  - n Add mean and median vertical lines
  - n Add legend
- n Boxplots
- n Matrix of scatter plots
- n 3D plots

## Data Mining

- n Clustering
  - n K-means
  - n PAM
  - n Fuzzy c-means
- n Plots of clustering results
- n Principal component analysis

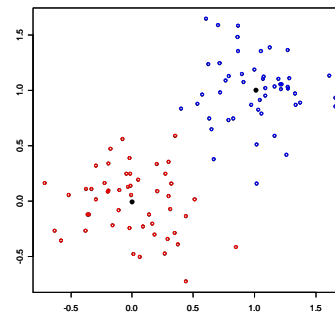
## KM method

- n **Minimization** of the objective function

$$J(V) = \sum_{j=1}^k \sum_{i=1}^{c_j} \|x_i - v_j\|^2$$

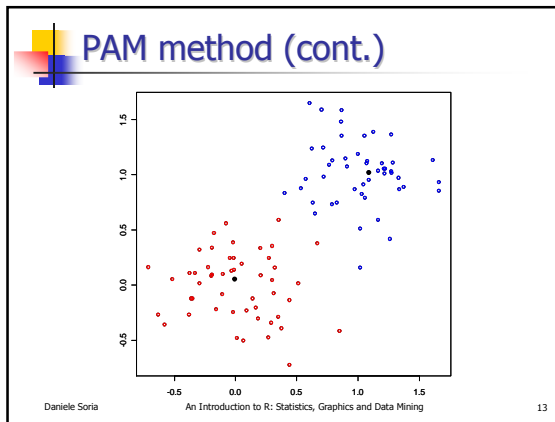
- n  $\|x_i - v_j\|$ : Euclidean distance between  $x_i$  and  $v_j$
- n  $c_j$ : data points in cluster  $j$
- n  $v_j$  can be calculated as  $v_j = \frac{1}{c_j} \sum_{i=1}^{c_j} x_i$ ,  $j = 1, \dots, k$

## KM method (cont.)



## PAM method

- n Based on the search for  $k$  representative objects (**medoids**) among the observations
- n Minimum **sum of dissimilarities** among the observations to their closest medoid
- n  $k$  clusters are constructed by assigning each observation to the nearest medoid



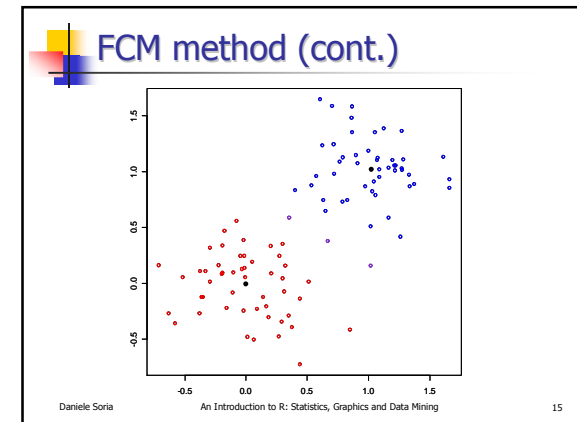
### FCM method

- Minimization of the objective function

$$J(U, V) = \sum_{i=1}^n \sum_{j=1}^c (\mu_{ij})^m \|x_i - v_j\|^2$$

- $X = \{x_1, x_2, \dots, x_n\}$ :  $n$  data points
- $V = \{v_1, v_2, \dots, v_c\}$ :  $c$  cluster centres
- $U = (\mu_{ij})_{n \times c}$ : fuzzy partition matrix
  - $\mu_{ij}$ : membership degree of  $x_i$  to  $v_j$
- $m$ : fuzziness index (often  $m=2$ )

Daniele Soria  
An Introduction to R: Statistics, Graphics and Data Mining 14



### Additional resources

- R-Seek Search Engine
  - <http://www.rseek.org>
- J.H. Maindonald and W.J. Braun. *Data Analysis and Graphics Using R – An Example-Based Approach*. Cambridge University Press, 2003.
- R Reference Card
  - <http://cran.r-project.org/doc/contrib/Short-refcard.pdf>
- Use Google to look for specific commands and functions

Daniele Soria  
An Introduction to R: Statistics, Graphics and Data Mining 16

### Thank you!

Questions?

Contact: [d.soria@cs.nott.ac.uk](mailto:d.soria@cs.nott.ac.uk)

Daniele Soria  
An Introduction to R: Statistics, Graphics and Data Mining 17