MIXMOD
A software for model-based classification for quantitative and qualitative data

C.Biernacki[1] G.Celeux[2] G.Govaert[3] F.Langrognet[4]

[1]UMR824 - Univ. Lille & CNRS - Lille (France)

[2]INRIA Orsay (France)

[3]UMR6599 - UTC & CNRS - Compiègne (France)

[4]UMR6623- Univ. Franche-Comté & CNRS - Besançon (France)

June 2008

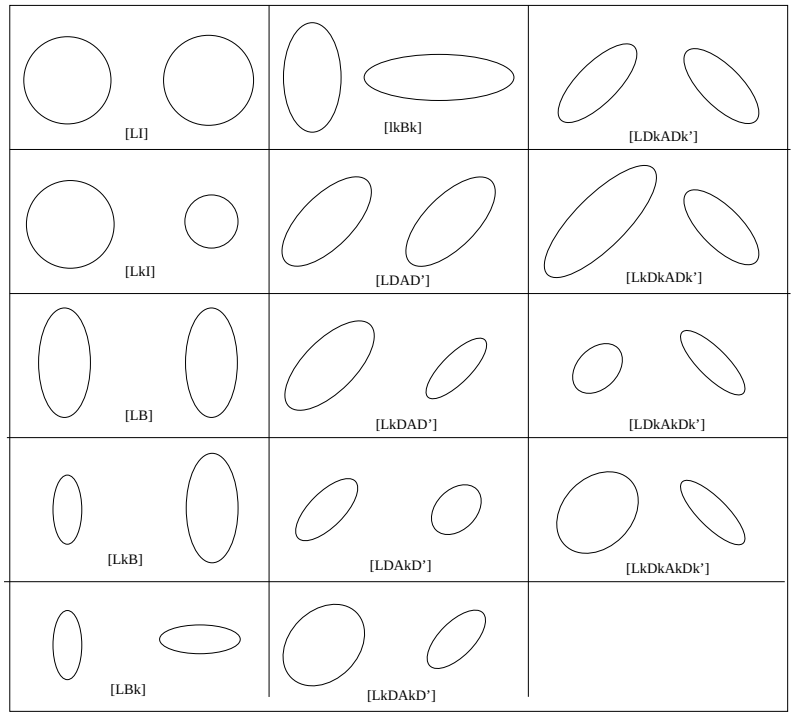
MIXMOD
A software for model-based classification for quantitative and qualitative data

MIXMOD

Parsimonious models

Gaussian models

14 Gaussian models
based on eigenvalue decomposition of variance matrices
(see Celeux and Govaert 1995)



High Dimensional Data

8 specific models for High Dimensional Data
(see Bouveyron 2007)

Mutinomial models

5 Multinomial models
based on a parameterization of Bernoulli distributions
(see Celeux and Govaert 1991)

Other functionalities

Algorithms

To maximize Likelihood or Completed Likelihood

- EM (Expectation Maximisation)
- SEM (Stochatitic EM)
- CEM (Classification EM)

Criteria

- BIC (Bayesian Information Criterion)
- ICL (Integrated Completed Likelihood)
- NEC (Normalized Entropy Criterion)
- CV (Cross Validation)

Initializations and Strategies

- 6 initializations

  Ex : 'random', 'short runs of EM'
- Chained algorithms

  Ex : 100 iterations of **SEM** and 50 iterations of **EM**

Others

- Partial labeling of individuals
- Weighted individuals

MIXMOD

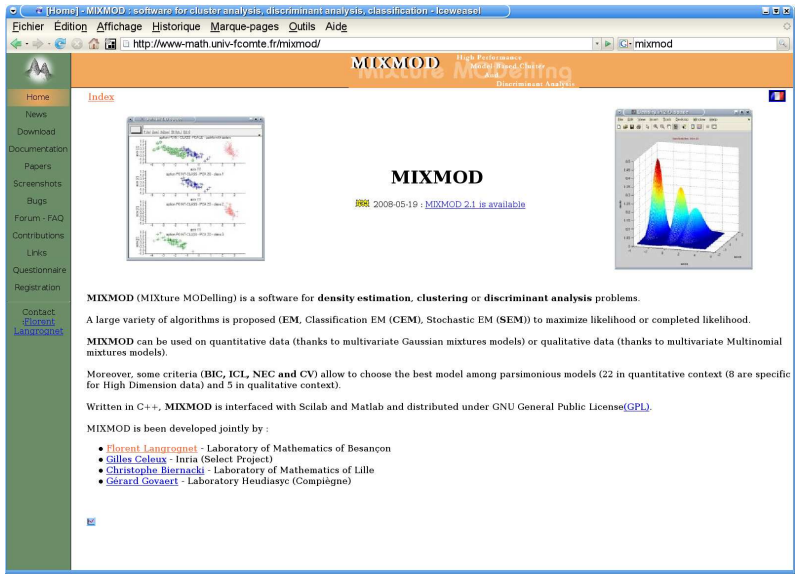
Software overview

Historical background

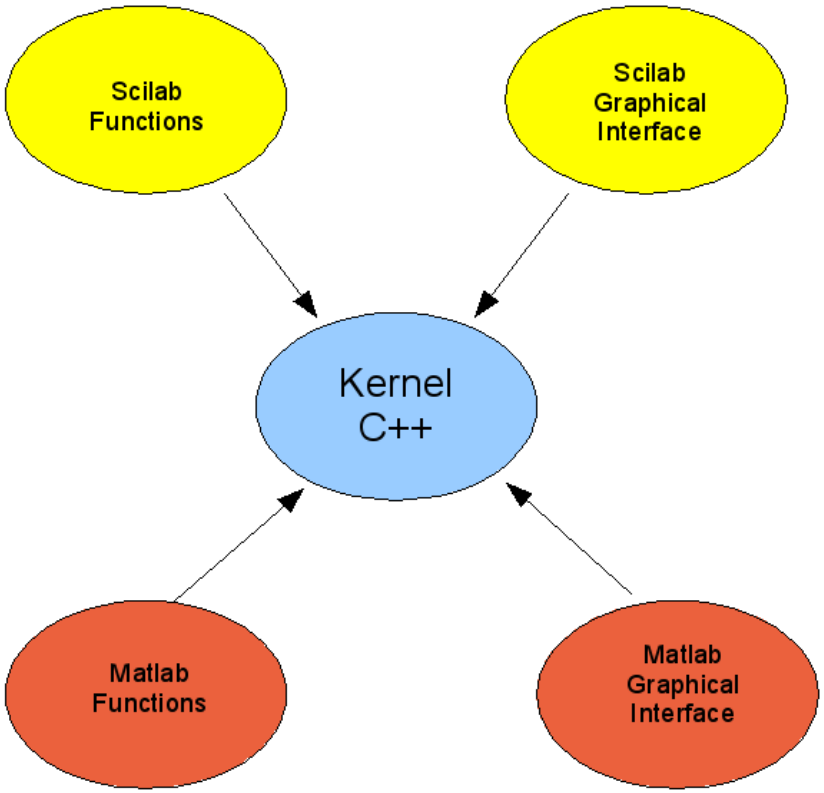
- Sep. 2000 : Mixmod 1.0
- Feb. 2007 : Mixmod 2.0 (with qualitative data processing)
- May 2008 : Mixmod 2.1 (with models for High Dimension data)

Distribution

- http ://www-math.univ-fcomte.fr/mixmod/
  800 visits and 300 downloads per month
- GNU GPL License
- Windows and Linux packages (source and binary)
- Documentations (statistical, userguide, quickStart...)

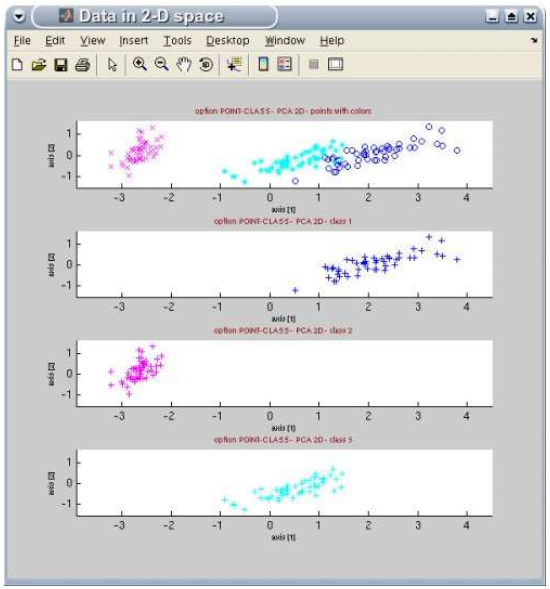


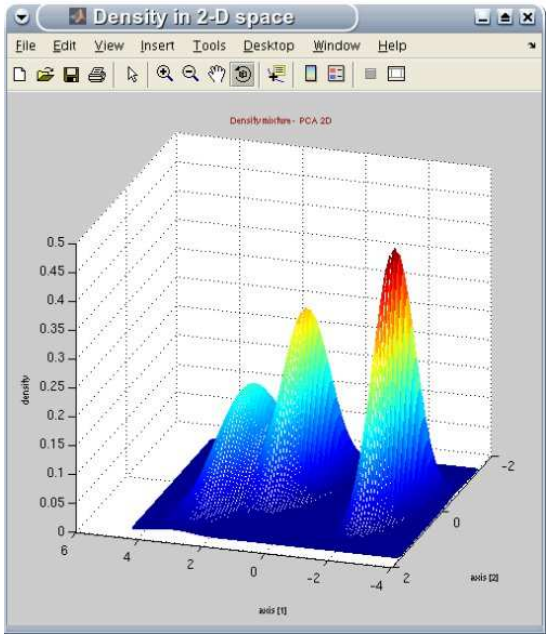
Software Architecture

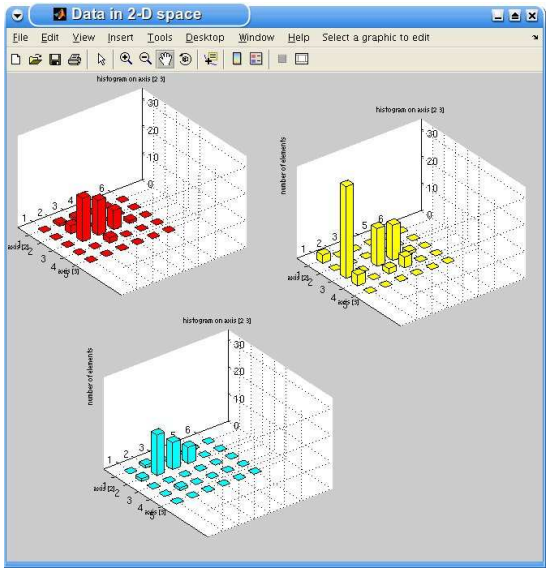


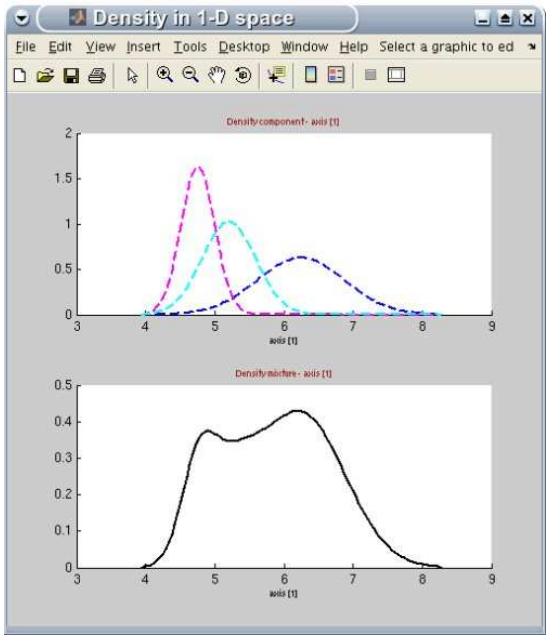
- Mixmod Kernel : C++
- Two levels of use :
  - graphical interfaces (Scilab, Matlab)
  - functions for Scilab and Matlab
- Mixmod library

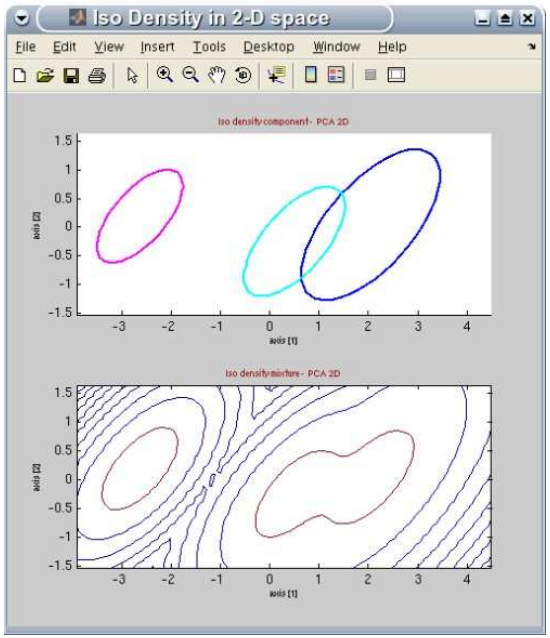
Screenshots

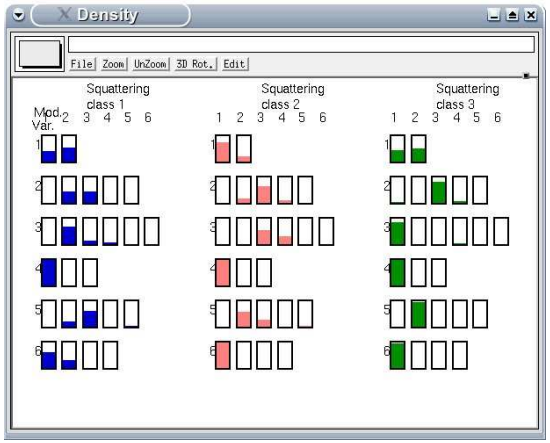












MIXMOD

Supervised classification with qualitative data

Puffins

Data

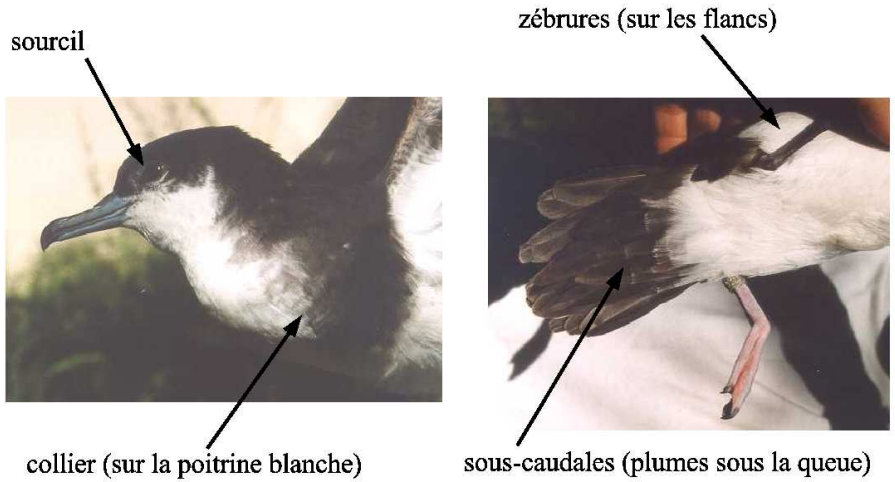
- Number of individuals : $n = 69$
- Number of species (or clusters) : $K = 2$
- Number of variables : $d = 6$
- Individual i : $(x_i, z_i) = ((x_i^j)_{j=1,\ldots,d}, z_i)$

n^o	z_i	x_i^1	x_i^2	x_i^3	x_i^4	x_i^5	x_i^6
1	1	1	2	2	1	2	2
2	1	2	1	3	1	3	1
⋮	⋮			⋮			
68	2	1	4	1	1	2	1
69	2	1	3	1	1	2	1

Observations

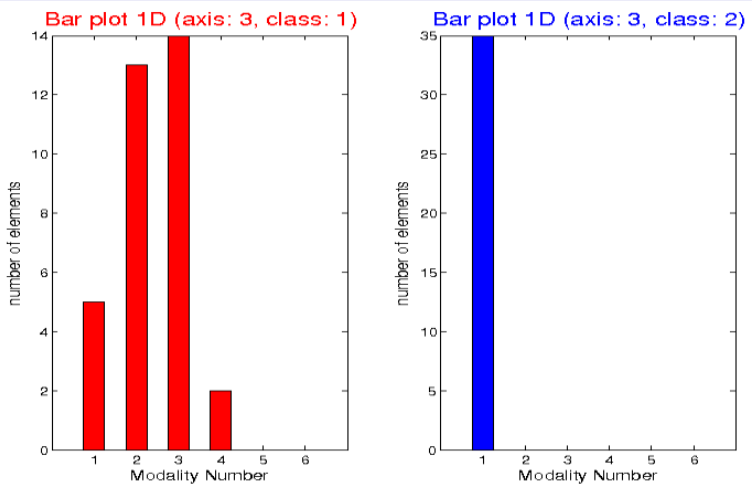
Variables description

variable	number of response levels	values
sex	2	2 values (male, female)
eyebrows	5	5 values (none -> heavy)
collar	6	6 values (none -> unless)
strips	3	3 values (none, . . .)
subcaudal	5	5 values (black, white, . . .)
liseret	4	4 values (. . .)

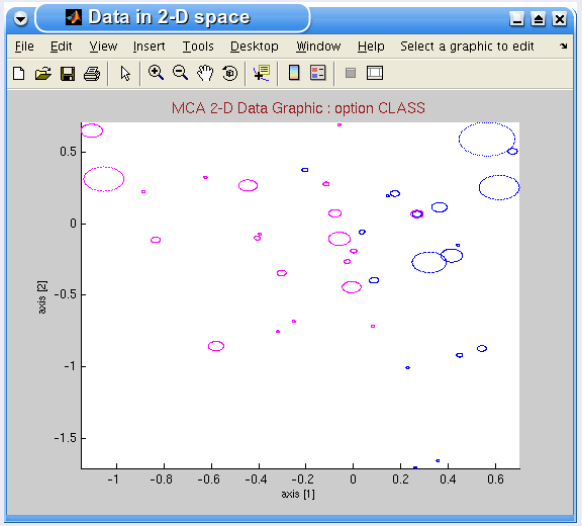


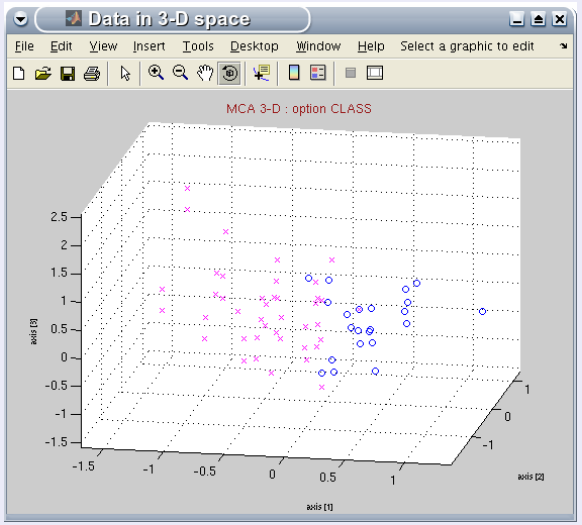
Data visualisation

barplots



scatterplots





Discriminant Analysis with MIXMOD (step 1)

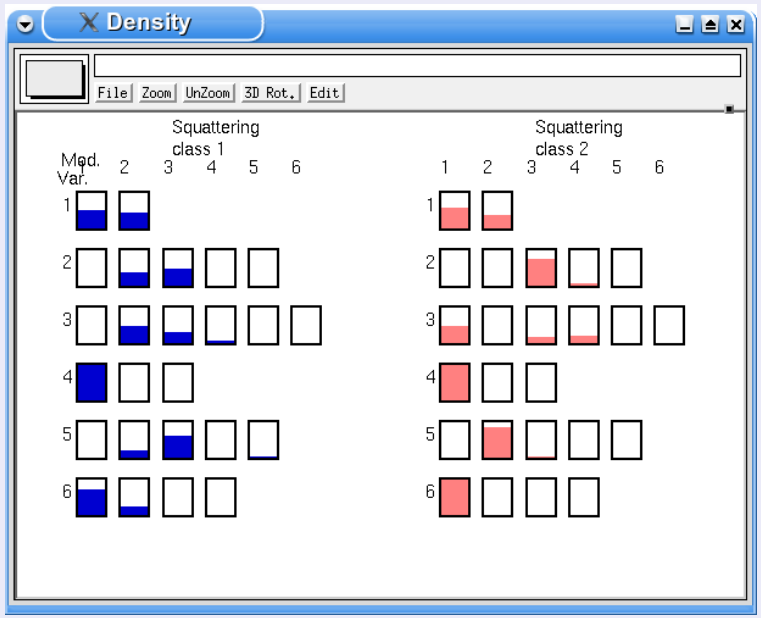
Classification rule

Estimated parameters

$$\hat{p}_1 = 0.49, \quad \hat{p}_2 = 0.51$$

$$\hat{\alpha}_1 = \begin{bmatrix} \hat{\alpha}_1^1 = & 0.53 & 0.47 & & & & \\ \hat{\alpha}_1^2 = & 0.00 & 0.44 & 0.56 & 0.00 & 0.00 & \\ \hat{\alpha}_1^3 = & 0.00 & 0.62 & 0.30 & 0.08 & 0.00 & 0.00 \\ \hat{\alpha}_1^4 = & 1.00 & 0.00 & 0.00 & & & \\ \hat{\alpha}_1^5 = & 0.00 & 0.20 & 0.72 & 0.00 & 0.08 & \\ \hat{\alpha}_1^6 = & 0.81 & 0.19 & 0.00 & 0.00 & & \end{bmatrix}$$

$$\hat{\alpha}_2 = \begin{bmatrix} \hat{\alpha}_2^1 = & 0.59 & 0.41 & & & & \\ \hat{\alpha}_2^2 = & 0.00 & 0.00 & 0.89 & 0.11 & 0.00 & \\ \hat{\alpha}_2^3 = & 0.69 & 0.00 & 0.15 & 0.16 & 0.00 & 0.00 \\ \hat{\alpha}_2^4 = & 1.00 & 0.00 & 0.00 & & & \\ \hat{\alpha}_2^5 = & 0.00 & 0.97 & 0.03 & 0.00 & 0.00 & \\ \hat{\alpha}_2^6 = & 1.0 & 0.00 & 0.00 & 0.00 & & \end{bmatrix}$$



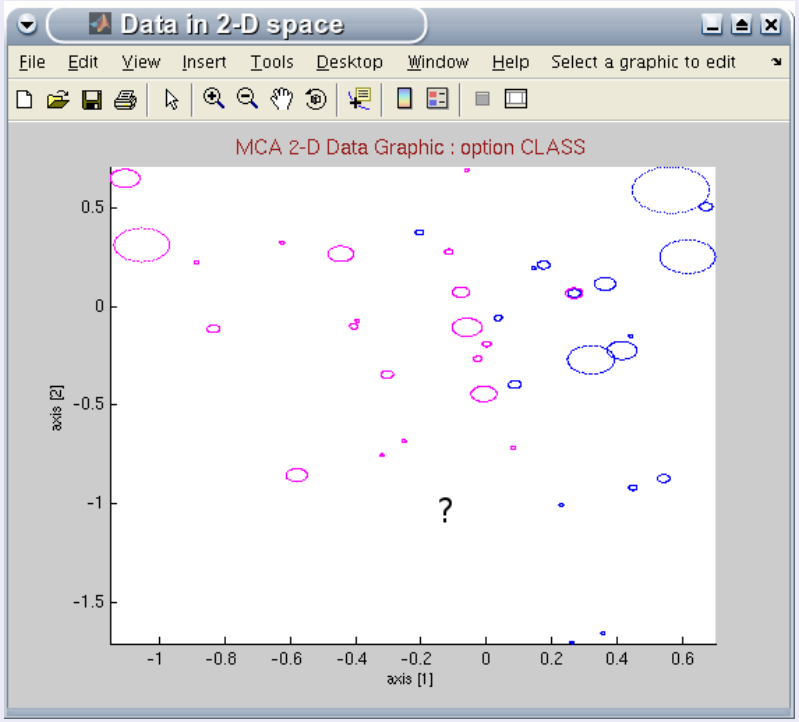
Quality of the classification rule

- Cross Validation rate (leave-one-out) : 94.8%
- Reclassification by MAP method : 96.7%

Discriminant Analysis with MIXMOD (step 2)

Classification of a new observation

n^o	z_i	x_i^1	x_i^2	x_i^3	x_i^4	x_i^5	x_i^6
70	?	1	3	1	1	5	2



Results

$z_{70} = 2$

$p(z_{70} = 1) = 0.06$ and $p(z_{70} = 2) = 0.94$

MIXMOD

Perspectives

New functionalities

- Semi-supervised classification
- Heterogeneous data (quantitative and qualitative data)
- ...

Software evolutions

- New graphical user interface (without Scilab and Matlab)
- Double license distribution (GPL and ...)
- ...

Thank you for your attention

Web : http ://www-math.univ-fcomte.fr/mixmod/

Authors :

- christophe.biernacki@math.univ-lille1.fr
- gilles.celeux@math.u-psud.fr
- gerard.govaert@hds.utc.fr
- florent.langrognet@univ-fcomte.fr

ANNEX

Finite mixture models
Latent Class Model

Finite mixture models

- $\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_n})$: **n** independent vectors in $\mathbf{R^d}$.
  Each $\mathbf{x}_i$ arises from a probability distribution with density : $f(x) = \sum_{k=1}^{K} p_k f_k(x)$
  - **K** : number of groups
  - $\mathbf{p_k}$: mixing proportions
  - $\mathbf{f_k}(.)$: densities of components
    - Gaussian densities for quantitative data
    - Multinomial densities for qualitative data
- $\mathbf{z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$: a sample of indicator vectors or labels

- Unsupervised classification : **z** is unknown
- Supervised classification : **z** is known
  Estimate $\mathbf{z_{n+1}}$, the label of a new observation $\mathbf{x_{n+1}}$

Qualitative data : Latent Class Model (LCM)

Data

$\boldsymbol{x}_i = (x_i^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$

- $\mathbf{m_j}$: the number of response levels of the variable j
- $\begin{cases} x_i^{jh} = 1 & \text{if } i \text{ has level } h \text{ for variable } j \\ x_i^{jh} = 0 & \text{otherwise.} \end{cases}$

Example : $X_i \begin{cases} x_i^1 = 10 \\ x_i^2 = 0010 \\ x_i^3 = 010 \end{cases}$ $(d = 3, m_1 = 2, m_2 = 4, m_2 = 3)$

Multivariate mutinomial distribution

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_k p_k m_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) = \sum_k p_k \prod_{j,h} (\alpha_k^{jh})^{x_i^{jh}}$$

$\boldsymbol{\theta} = (p_1, \ldots, p_K, \alpha_1^{11}, \ldots, \alpha_K^{dm_d})$: the parameter of the latent class model

- α_k^{jh} : probability that variable j has level h in cluster k
- p_k : mixing proportions

LMC : Reparameterization

$\forall k, j : (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j}) \longrightarrow (a_k^{j1}, \ldots, a_k^{jm_j}, \varepsilon_k^{j1}, \ldots, \varepsilon_k^{jm_j})$

- $a_k^{jh} = \begin{cases} 1 & \text{if } h = \arg\max_h \alpha_k^{jh} \\ 0 & \text{otherwise} \end{cases}$ *center*
- $\varepsilon_k^{jh} = \begin{cases} 1 - \alpha_k^{jh} & \text{if } a_k^{jh} = 1 \\ \alpha_k^{jh} & \text{if } a_k^{jh} = 0. \end{cases}$ *scattering*

Example : $(0.2, 0.7, 0.1) \longrightarrow (0, 1, 0, \quad 0.2, 0.3, 0.1)$.

Five latent class models

Different constraints to the scattering parameters ε_k^{jh}.

model	the scattering is depending on. . .
$[\varepsilon_k^{jh}]$	clusters, variables and levels
$[\varepsilon_k^{j}]$	clusters and variables
$[\varepsilon_k]$	clusters
$[\varepsilon^{j}]$	variables
$[\varepsilon]$	none