# Rmixmod
# Le package R de MIXMOD®


Rencontres R 2012 - Bordeaux

Florent Langrognet

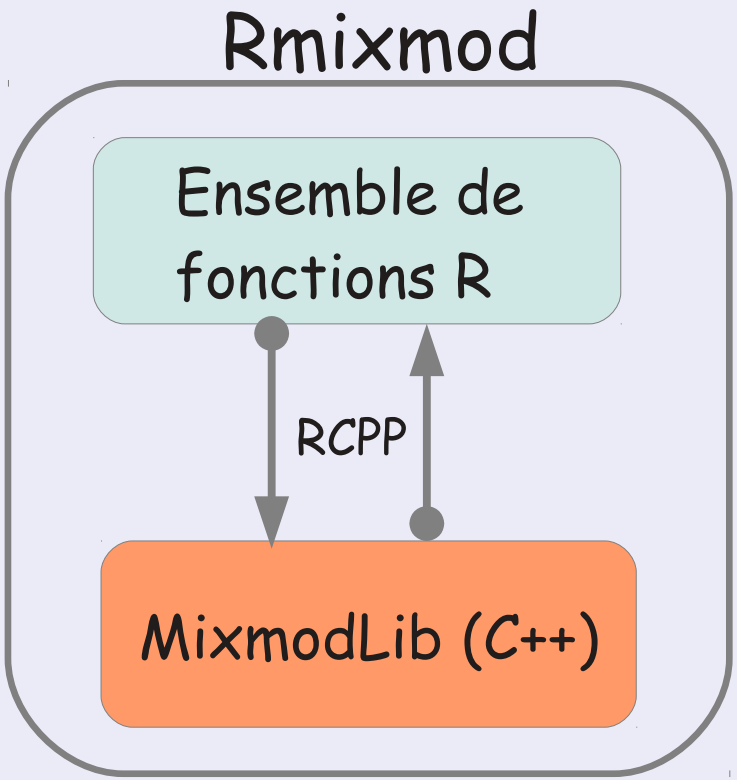
Laboratoire de Mathématiques de Besançon

# Rmixmod

# Rmixmod

# Rmixmod : le package R de Mixmod

## Architecture




## Avantages

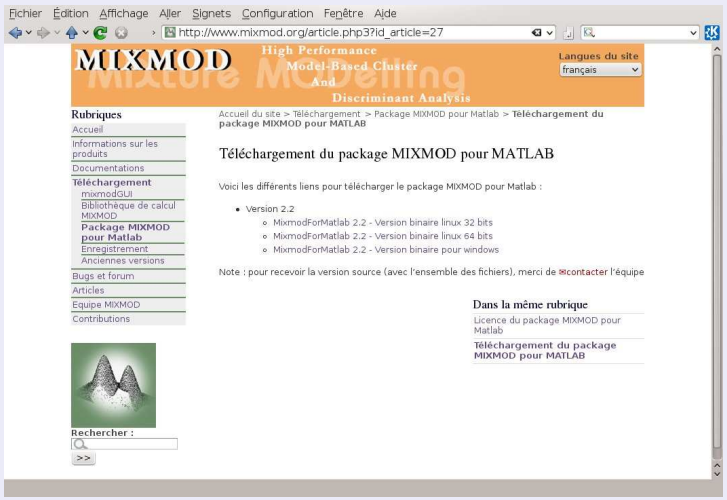
- Atouts de R
  - Environnement R reconnu, efficace et apprécié
  - Interface avec d'autres packages
  - Outils de visualisation
- Atouts de mixmodLib (C++)
  - Développée depuis 2001
  - Largement diffusée, et utilisée
  - Eprouvée, robuste, rapide

# Rmixmod

# MIXMOD

## Fiche d'identité

- Projet débuté en 2001
- Compétences complémentaires en informatique et statisitiques
- Diffusion : www.mixmod.org
- Licence : GNU GPL
- Rencontres Mixmod
- Forum d'utilisateurs, accompagnement, demandes d'évolutions, . . .

# Rmixmod

# Fonctionnalités (1)

## Problématiques traitées

- Classification non supervisée
- Classification supervisée (analyse discriminante)
- Estimation de densité

## Cadre de travail - Type de données traitées

Modèles de mélanges

- Gaussiens (données quantitatives)
- Multinomiaux (données qualitatives)
- Modèles spécifiques pour les données en grande dimension

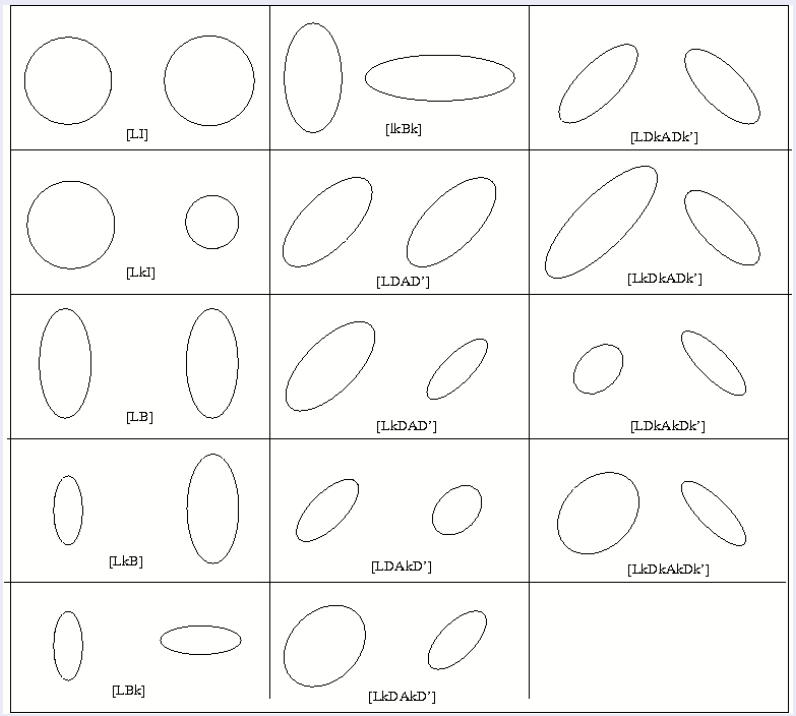
# Fonctionnalités (2)

## Modèles et métriques

### Données quantitatives

14 modèles gaussiens

basés sur la décomposition en valeur sigulière de la matrice de variance




### Données quantitatives en grande dimension

8 modèles spécifiques pour la grande dimension

### Données qualitatives

5 modèles multinomiaux

basés sur une reparamétrisation de la distribution de Bernoulli

# Fonctionnalités (3)

## Algorithmes

Maximisation de la vraisemblance (ou vraisemblance complétée)

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

## Critères

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

## Initialisations et Stratégies

- 6 initialisations

  Ex : 'random', 'short runs of EM',...
- Algorithmes chaînés

  Ex : 100 iterations de **SEM** puis 50 iterations de **EM**

## Et aussi...

- Connaissance partielle des labels des individus (semi-supervisé)
- Individus pondérés

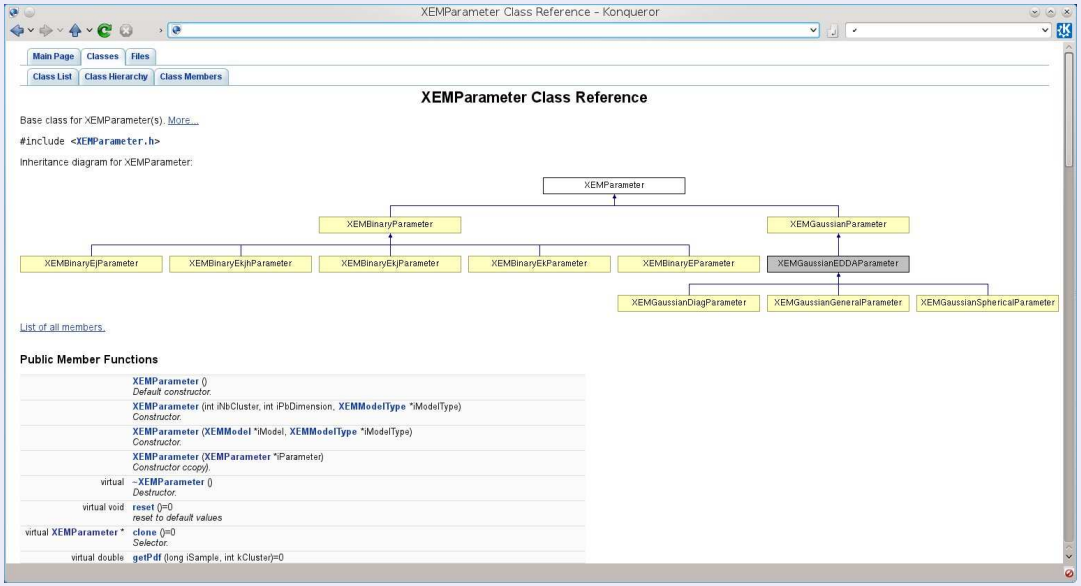
# Rmixmod

# L'ensemble logiciel MIXMOD (1)

## mixmodLib

Bibliothèque de calcul

- Rapide, robuste, éprouvée
- Ensemble de classes C++
- Env. 500 téléchargements par an depuis 2001

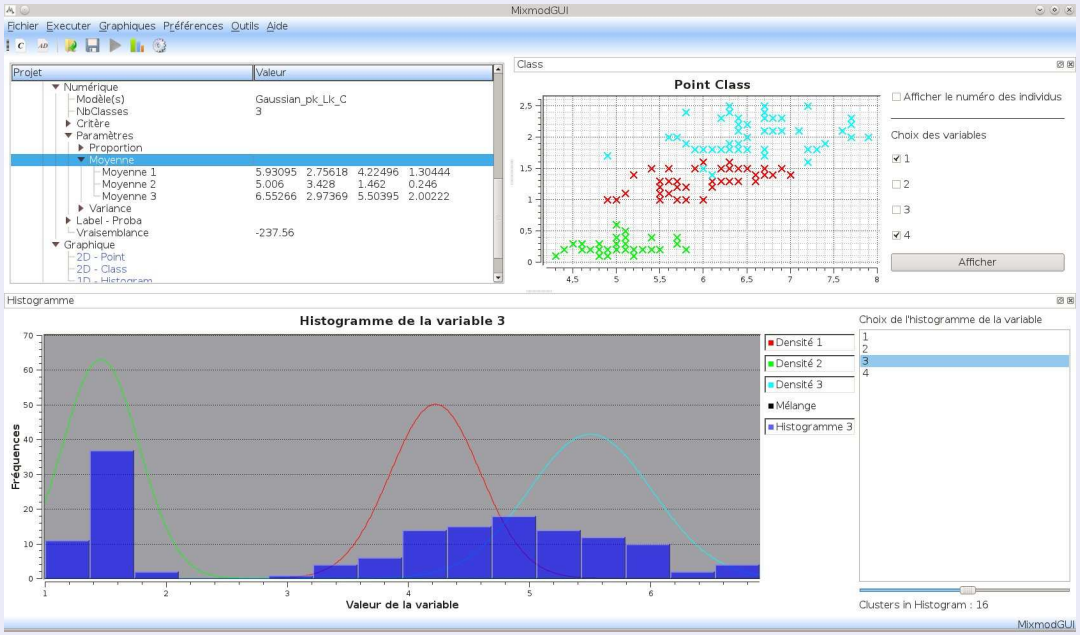


# L'ensemble logiciel MIXMOD (2)

## mixmodGUI

### Interface graphique

- Conviviale
- Entrées/Sorties XML
- Utilisation des bibilothèques QT et Qwt
- Disponible depuis 2011 env. 500 téléchargements sur un an

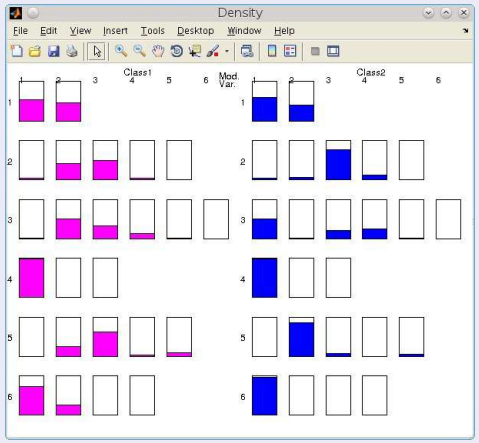


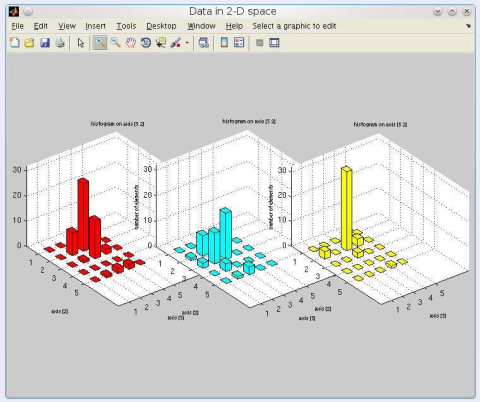
# L'ensemble logiciel MIXMOD (3)

## mixmodForMatlab

Package pour Matlab

- Interface de mixmodLib pour Matlab
- Ensemble de fonctions Matlab :
  - Classification supervisée et non supervisée
  - Outils de visualisation
- Disponible depuis 2003
  env. 300 téléchargements par an



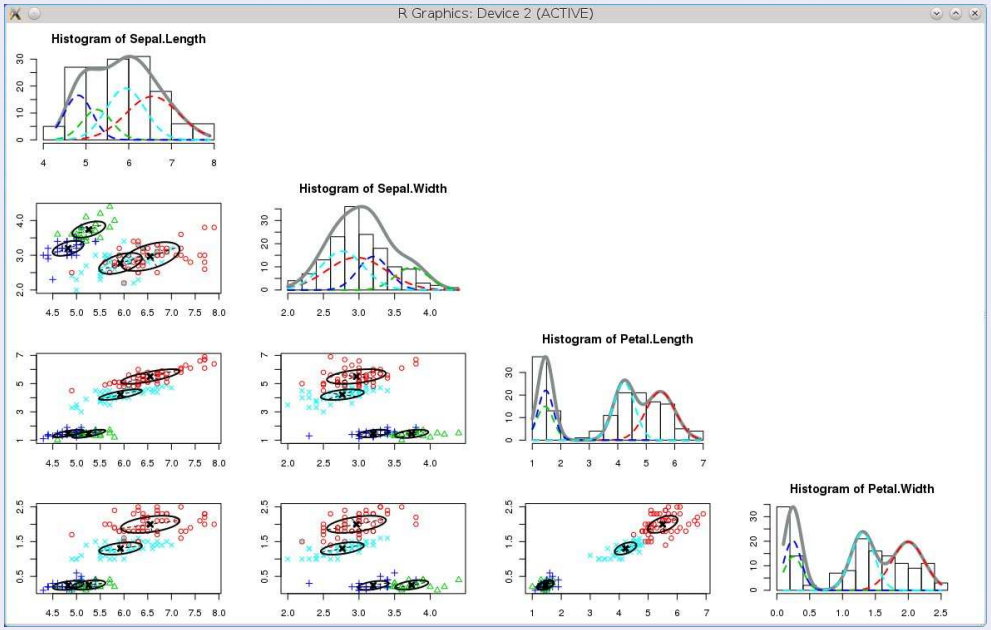



# L'ensemble logiciel MIXMOD (4)

## Rmixmod

Package pour R

- Interface de mixmodLib pour R
- Ensemble de fonctions R :
  - Classification supervisée et non supervisée
  - Outils de visualisation
- Disponible depuis 2012

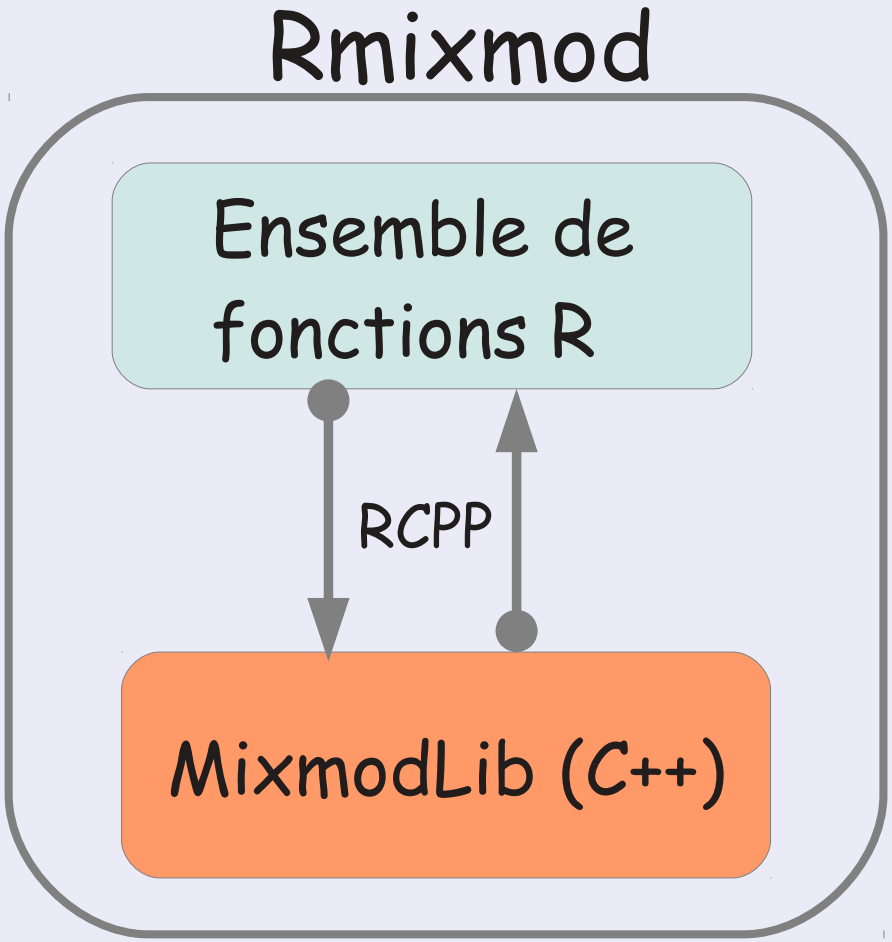


# Rmixmod

# Rmixmod

# Rmixmod

## Architecture



# Classes - Fonctions

## Classes Rmixmod

**Classes (S4)**

Mixmod
MixmodCluster [<-Mixmod]
MixmodLearn [<-Mixmod]
MixmodPredict
MixmodResults
MixmodDAResults [<-MixmodResults]
Model
MultinomialModel [<-Model]
GaussianModel [<-Model]
Parameter
GaussianParameter [<-Parameter]
MultinomialParameter [<-Parameter]
Strategy

## Fonctions Rmixmod

**Fonctions**

mixmodCluster
mixmodLearn
mixmodPredict

mixmodStrategy
mixmodGaussianModel
mixmodMultinomialModel
sortByCriterion
nbFactorFromData

summary
print
hist
histCluster
plot
PlotCluster
barplot
barplotCluster

# Rmixmod

# Classification non supervisée

## mixmodCluster

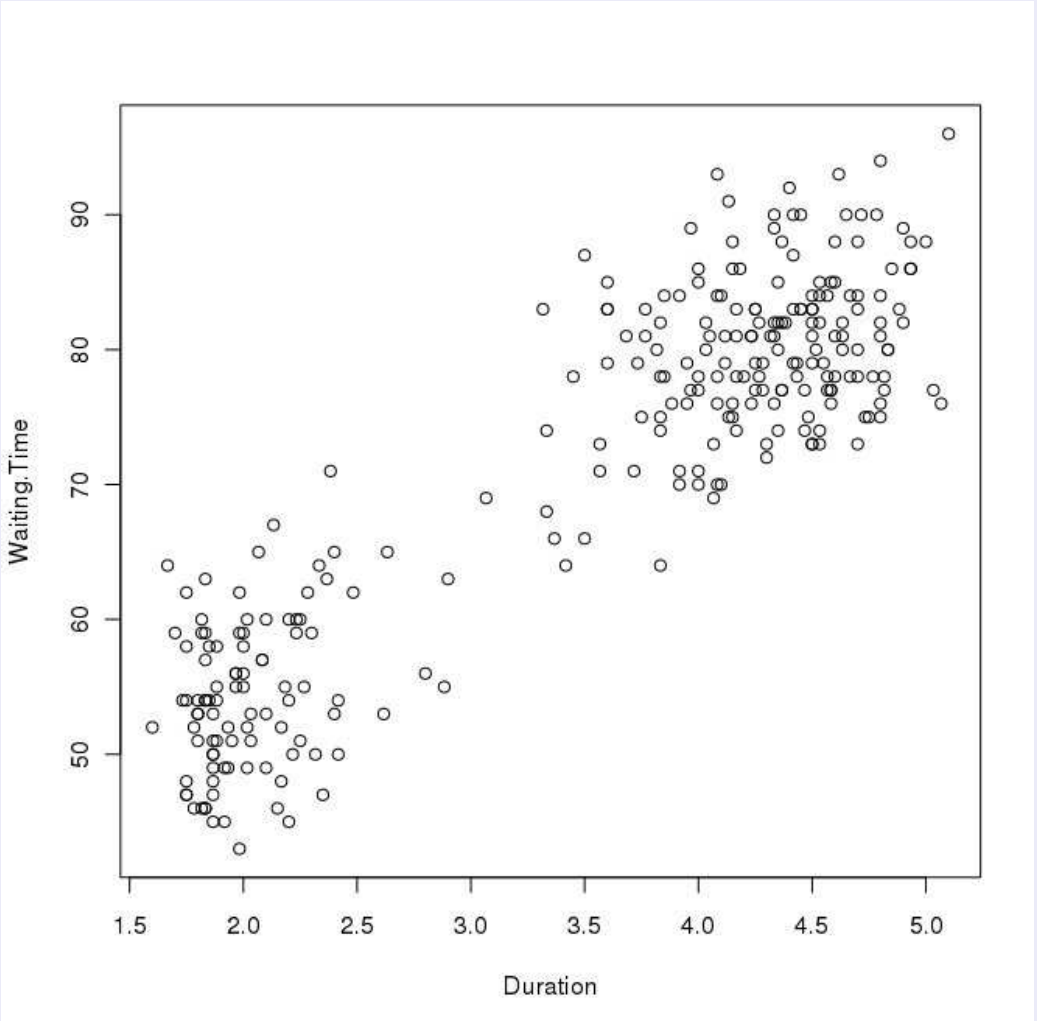
*mixmodCluster(data, nbCluster, dataType=NULL, models=NULL, strategy=mixmodStrategy(), criterion="BIC", weight=NULL, knownLabels=NULL)*

- Entrées
  - **data** (matrice ou *data frame*)
  - **nbCluster** (vecteur d'entiers)
  - dataType (chaîne de caractères)
  - models : liste des modèles gaussiens ou multinomiaux (objet [Model])
  - strategy : stratégie Mixmod (initialisation + algorithme) (objet [Strategy])
  - criterion : liste de critères de sélection (chaîne de caractères)
  - weight : vecteur de poids
  - knownLabels : vecteur des labels connus
- Sortie : Un objet [MixmodCluster]
  Contenant notamment 2 objets [MixmodResults] :
  - bestResult : le meilleur modède
  - results : liste des modèles triés (selon le $1^{er}$ crtière)

# Classification non supervisée

## Illustration

Geyser (données quantitatives)



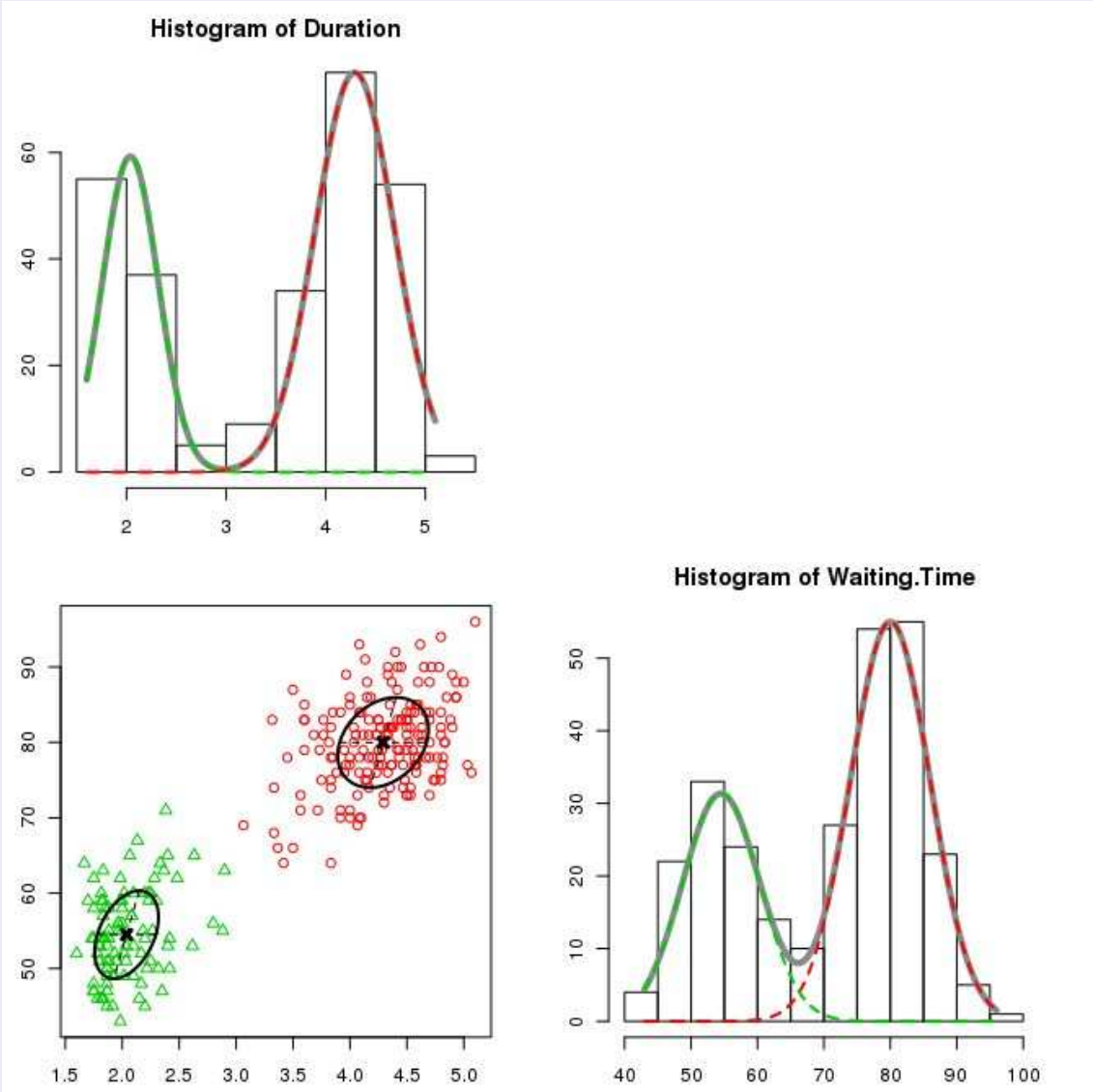
# Illustration 1

## Commandes Rmixmod

```
> data(geyser)
> out<-mixmodCluster(geyser, nbCluster=2)
> summary(out)
*************************************************************
* Number of samples     =  272
* Problem dimension     =  2
*************************************************************
*       Number of cluster =  2
*              Model Type =  Gaussian_pk_Lk_C
*               Criterion =  BIC(2322.9719)
*              Parameters =  list by cluster
*                  Cluster  1 :
                         Proportion =  0.6429
                              Means =  4.2922 79.9964
                          Variances = |     0.1453     0.8301 |
                                      |     0.8301    40.9022 |
*                  Cluster  2 :
                         Proportion =  0.3571
                              Means =  2.0397 54.5171
                          Variances = |     0.0984     0.5618 |
                                      |     0.5618    27.6831 |
*           Log-likelihood =  -1136.2599
*************************************************************
```

# Illustration 1

## Plot



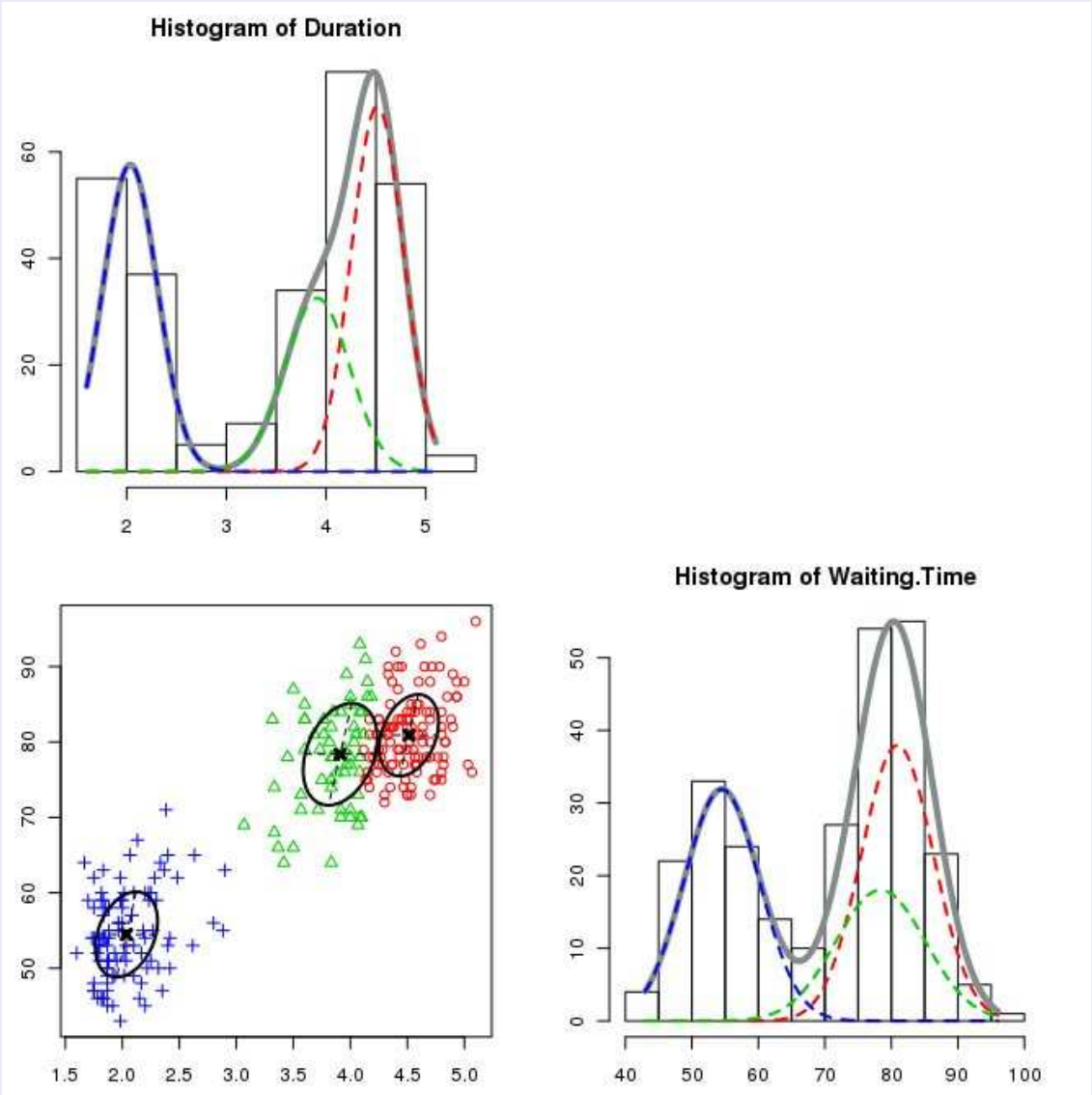
# Plusieurs modèles et plusieurs critères

## Commandes Rmixmod

```
> out<-mixmodCluster(geyser, nbCluster=2:3, criterion=c("BIC","ICL"),
+ models=mixmodGaussianModel())
> summary(out)
*************************************************************
* Number of samples     =  272
* Problem dimension     =  2
*************************************************************
*       Number of cluster =  3
*              Model Type =  Gaussian_p_L_C
*               Criterion =  BIC(2312.6006) ICL(2377.6923)
*              Parameters =  list by cluster
*                  Cluster  1 :
                         Proportion =  0.3333
                              Means =  3.9765 78.7195
                          Variances = |     0.0798     0.5341 |
                                      |     0.5341    34.2108 |
*                  Cluster  2 :
                         Proportion =  0.3333
                              Means =  4.5545 81.0528
                          Variances = |     0.0798     0.5341 |
                                      |     0.5341    34.2108 |
*                  Cluster  3 :
                         Proportion =  0.3333
                              Means =  2.0390 54.5083
                          Variances = |     0.0798     0.5341 |
                                      |     0.5341    34.2108 |
*          Log-likelihood =  -1131.0742
*************************************************************
```

# Illustration 1 bis

## plot (3 classes)



# Illustration 1 bis

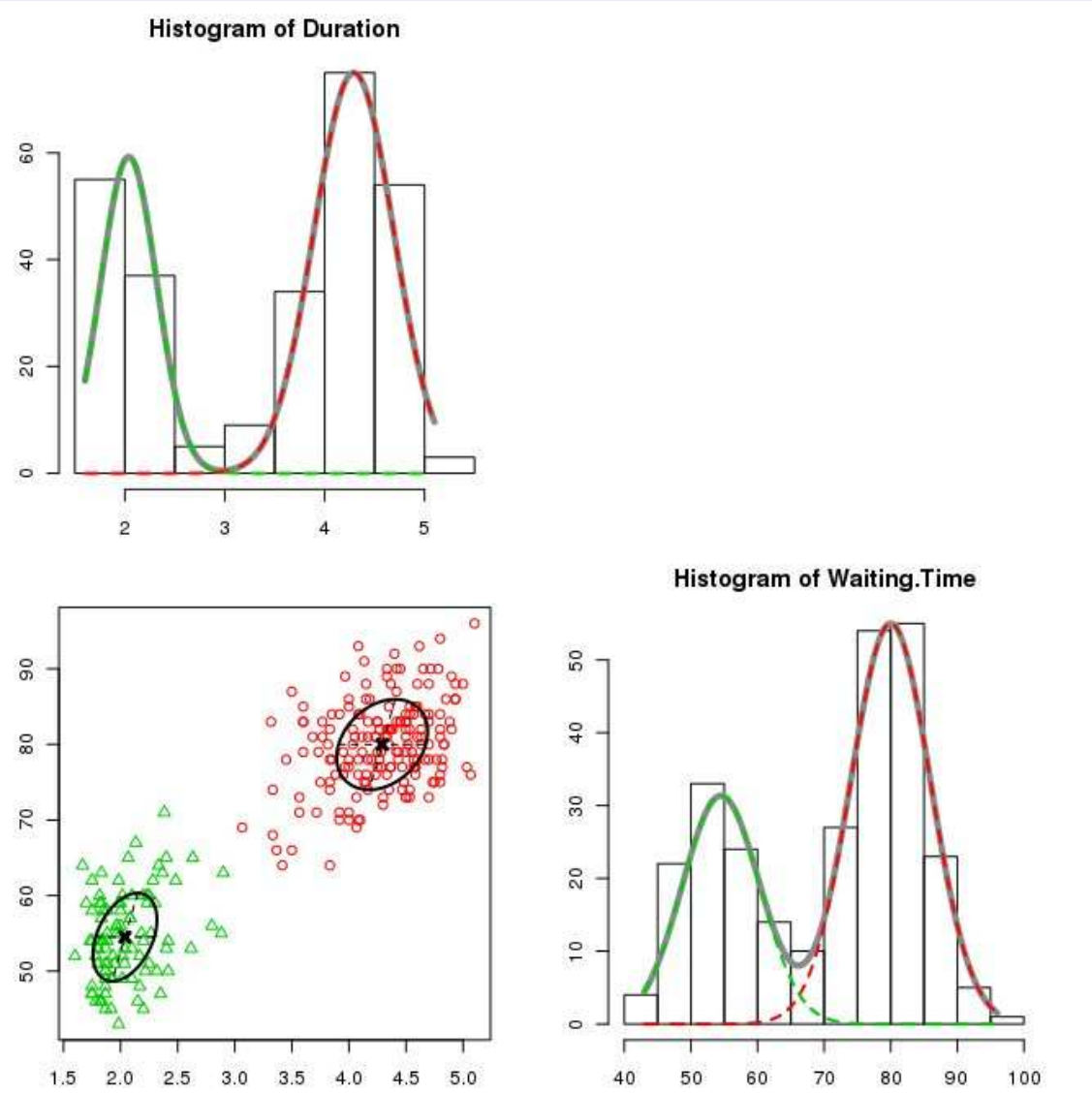
## Tri selon ICL

```
> out_icl<-sortByCriterion(out, "ICL")
> summary(out_icl)
*************************************************************
* Number of samples     =  272
* Problem dimension     =  2
*************************************************************
*       Number of cluster =  2
*              Model Type =  Gaussian_pk_Lk_D_Ak_D
*               Criterion =  BIC(2320.2833) ICL(2320.5794)
*              Parameters =  list by cluster
*                  Cluster  1 :
                         Proportion =  0.6432
                              Means =  4.2915 79.9893
                          Variances = |     0.1588     0.6810 |
                                      |     0.6810    35.7667 |
*                  Cluster  2 :
                         Proportion =  0.3568
                              Means =  2.0387 54.5041
                          Variances = |     0.0783     0.6467 |
                                      |     0.6467    33.8930 |
*           Log-likelihood =  -1132.1126
*************************************************************
```

# Illustration 1 bis

## plot (2 classes)

2 classes



# Rmixmod

# Classification supervisée : 1$^{re}$ étape

## mixmodLearn

*mixmodLearn(data, knownLabels, dataType=NULL, models=NULL, criterion="CV", nbCVBlocks=10, weight=NULL)*

- Entrées
  - **data** : échantillon d'apprentissage (matrice ou *data frame*)
  - **knownLabels** : étiquettes de l'échantillon d'apprentissage (vecteur d'entiers)
  - dataType (chaîne de caractères)
  - models : liste des modèles gaussiens ou multinomiaux (objet [Model])
  - criterion : liste de critères de sélection (chaîne de caractères
  - nbCVBlocks : nombre de blocs pour la Validation Croisée
  - weight : vecteur de poids
- Sortie : Un objet [MixmodLearn]
  Contenant notamment 2 objets [MixmodResults] :
  - bestResult : le meilleur modède
  - results : liste des modèles triés (selon le 1$^{er}$ crtière)

# Classification supervisée : $2^e$ étape

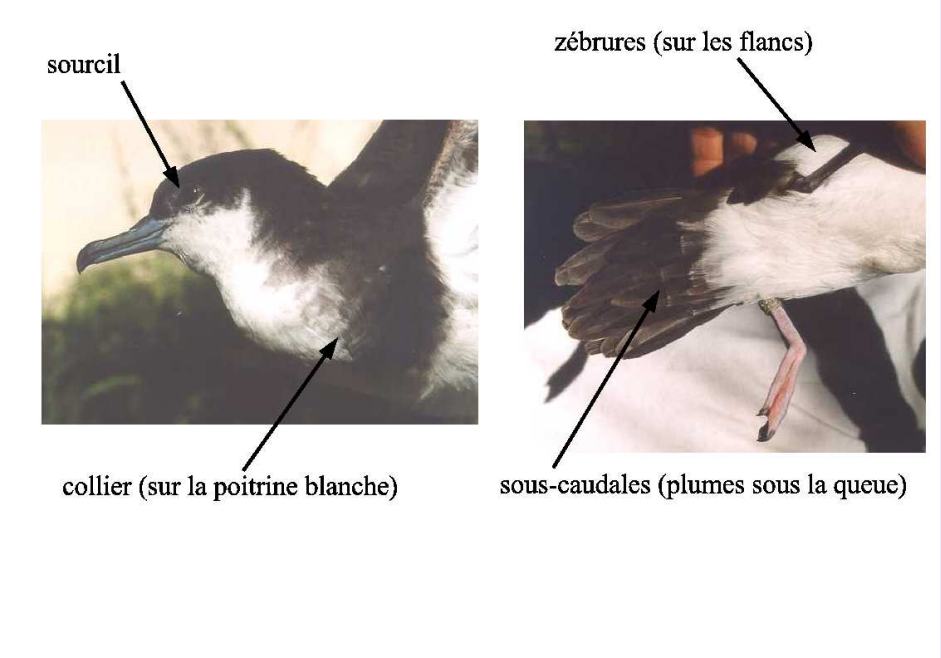
## mixmodPredict

*mixmodPredict(data, classificationRule)*

- Entrées
  - **data** : individus à classer (matrice ou *data frame*)
  - **classificationRule** : règle de classement issue de l'étape mixmodLearn (objet [MixmodResults])
- Sortie : Un objet [MixmodPredict] avec :
  - labels : labels obtenus par la règle de classement
  - proba : probabilité d'appartenance à chaque classe

# Classification supervisée

## Illustration (données qualitatives) : puffins



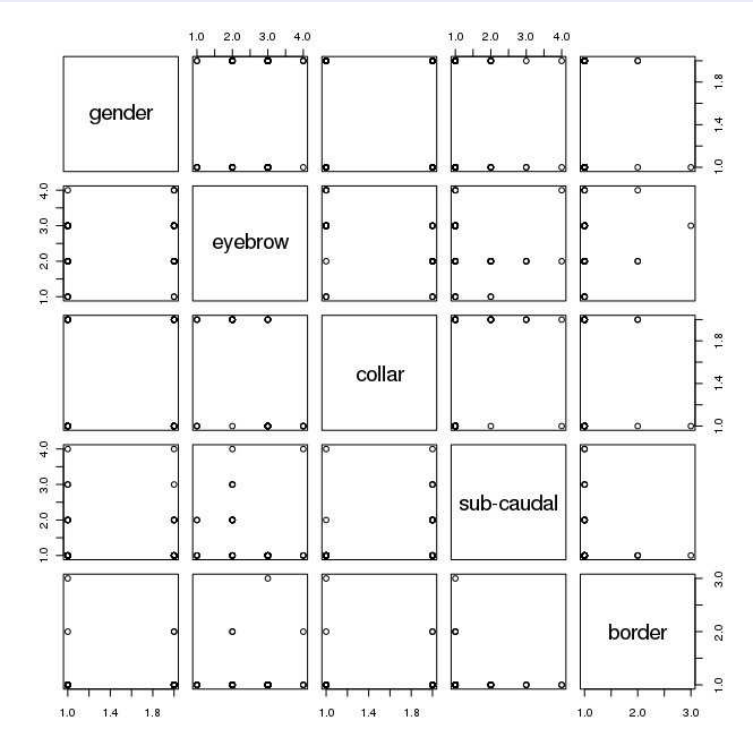

| variable | nombre de niveaux de réponse | valeurs |
|---|---|---|
| sexe | 2 | mâle, femelle |
| sourcils | 5 | absent -> très prononcé |
| collier | 6 | absent -> continu |
| sous-caudales | 5 | blanc, noir, noir&blanc, noir&BLANC, NOIR&blanc |
| liseret | 4 | absent, . . . , beaucoup |

# Observations

## Données

- Nombre d'individus : $n = 69$
- Nombre d'espèces (classes) : $K = 2$
- Nombre de variables : $d = 5$
- Individu $i$ : $(x_i, z_i) = ((x_i^j)_{j=1,\ldots,d}, z_i)$

| n$^o$ | $z_i$ | $x_i^1$ | $x_i^2$ | $x_i^3$ | $x_i^4$ | $x_i^5$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 2 | 1 | 2 | 1 | 3 | 3 | 1 |
| ⋮ | ⋮ | | | ⋮ | | |
| 68 | 2 | 1 | 4 | 1 | 2 | 1 |
| 69 | 2 | 1 | 3 | 1 | 2 | 1 |



# Illustration 2 - Apprentissage

## Apprentissage (1)

```
> data(birds)
> learn<-mixmodLearn(birds, birdsLabel)
> summary(learn)
*************************************************************
* Number of samples     =  69
* Problem dimension     =  5
*************************************************************
*       Number of cluster =  2
*              Model Type =  Binary_pk_Ekjh
*               Criterion =  CV(0.9855)
*              Parameters =  list by cluster
*                  Cluster  1 :
                        Proportion =  0.6667
                            Center =  1.0000 3.0000 1.0000 1.0000 1.0000
                           Scatter = |     0.4787     0.4787 |
                                     |     0.0691     0.0266     0.1862     0.0904 |
                                     |     0.1660     0.1532     0.0043     0.0043     0.0043 |
                                     |     0.0383     0.0043     0.0043     0.0255     0.0043 |
                                     |     0.0780     0.0496     0.0284 |
*                 Cluster  2 :
                        Proportion =  0.3333
                            Center =  2.0000 2.0000 2.0000 2.0000 1.0000
                           Scatter = |     0.3958     0.3958 |
                                     |     0.1354     0.1562     0.0104     0.0104 |
                                     |     0.0500     0.0750     0.0083     0.0083     0.0083 |
                                     |     0.3417     0.5333     0.1333     0.0500     0.0083 |
                                     |     0.0694     0.0556     0.0139 |
*          Log-likelihood =  -203.5308
*************************************************************
```

# Illustration 2 - Apprentissage

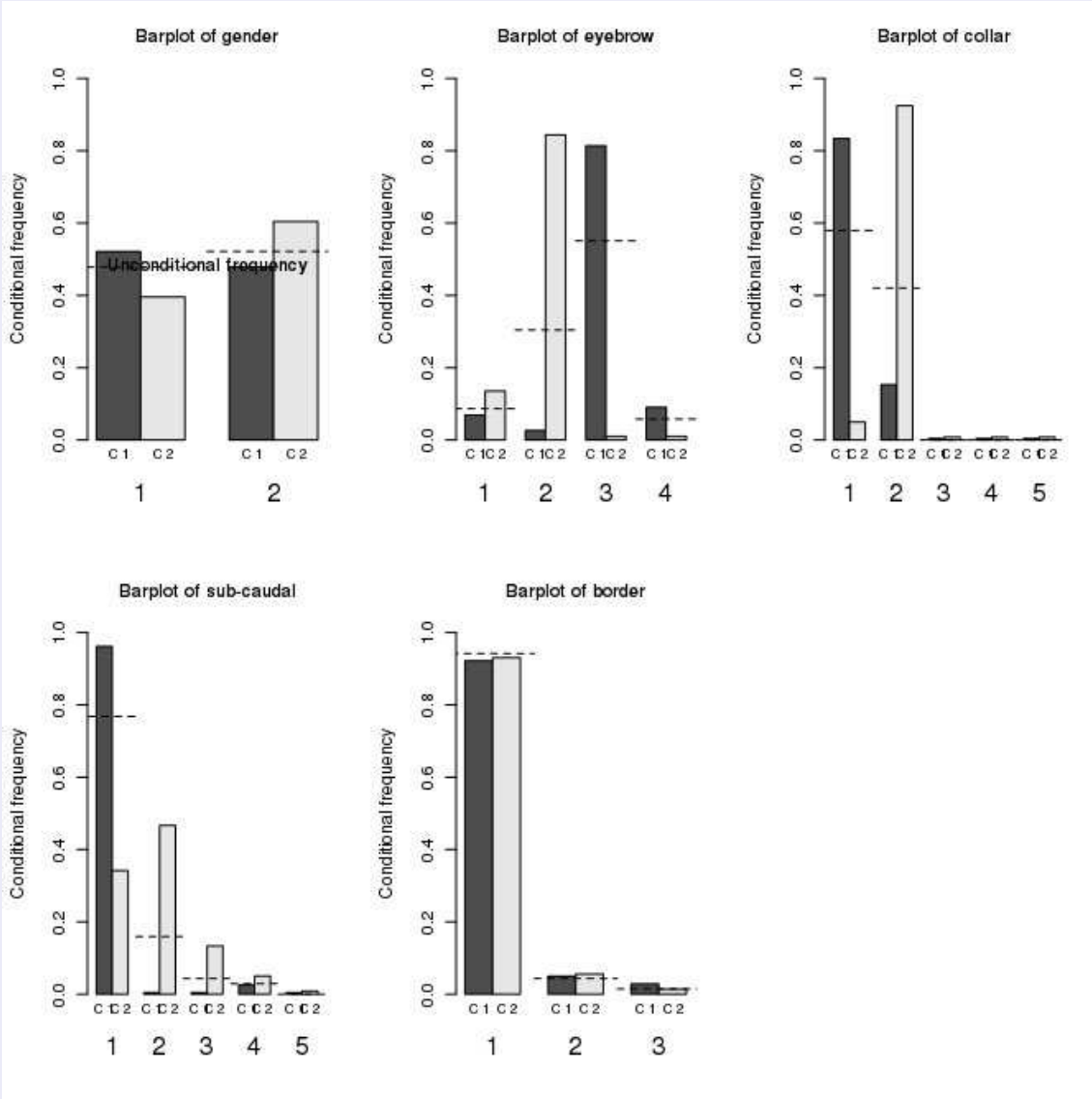
## Apprentissage (2)

```
> learn@bestResult
* nbCluster   =  2
* model name  =  Binary_pk_Ekjh
* criterion   =  CV(0.9855)
* likelihood  =  -203.5308
****************************************
* number of modalities =  2 4 5 5 3
*** Cluster 1
* proportion =  0.6667
* center     =  1.0000 3.0000 1.0000 1.0000 1.0000
* scatter    = |     0.4787     0.4787 |
               |     0.0691     0.0266     0.1862     0.0904 |
               |     0.1660     0.1532     0.0043     0.0043     0.0043 |
               |     0.0383     0.0043     0.0043     0.0255     0.0043 |
               |     0.0780     0.0496     0.0284 |
*** Cluster 2
* proportion =  0.3333
* center     =  2.0000 2.0000 2.0000 2.0000 1.0000
* scatter    = |     0.3958     0.3958 |
               |     0.1354     0.1562     0.0104     0.0104 |
               |     0.0500     0.0750     0.0083     0.0083     0.0083 |
               |     0.3417     0.5333     0.1333     0.0500     0.0083 |
               |     0.0694     0.0556     0.0139 |
****************************************
* Classification with CV:
           | Cluster 1 | Cluster 2 |
---------- ---------- ----------
 Cluster 1 |        45 |         0 |
 Cluster 2 |         1 |        23 |
---------- ---------- ----------
* Error rate with CV =  1.45 %

* Classification with MAP:
           | Cluster 1 | Cluster 2 |
---------- ---------- ----------
 Cluster 1 |        46 |         0 |
 Cluster 2 |         0 |        23 |
---------- ---------- ----------
* Error rate with MAP =  0.00 %
****************************************
```

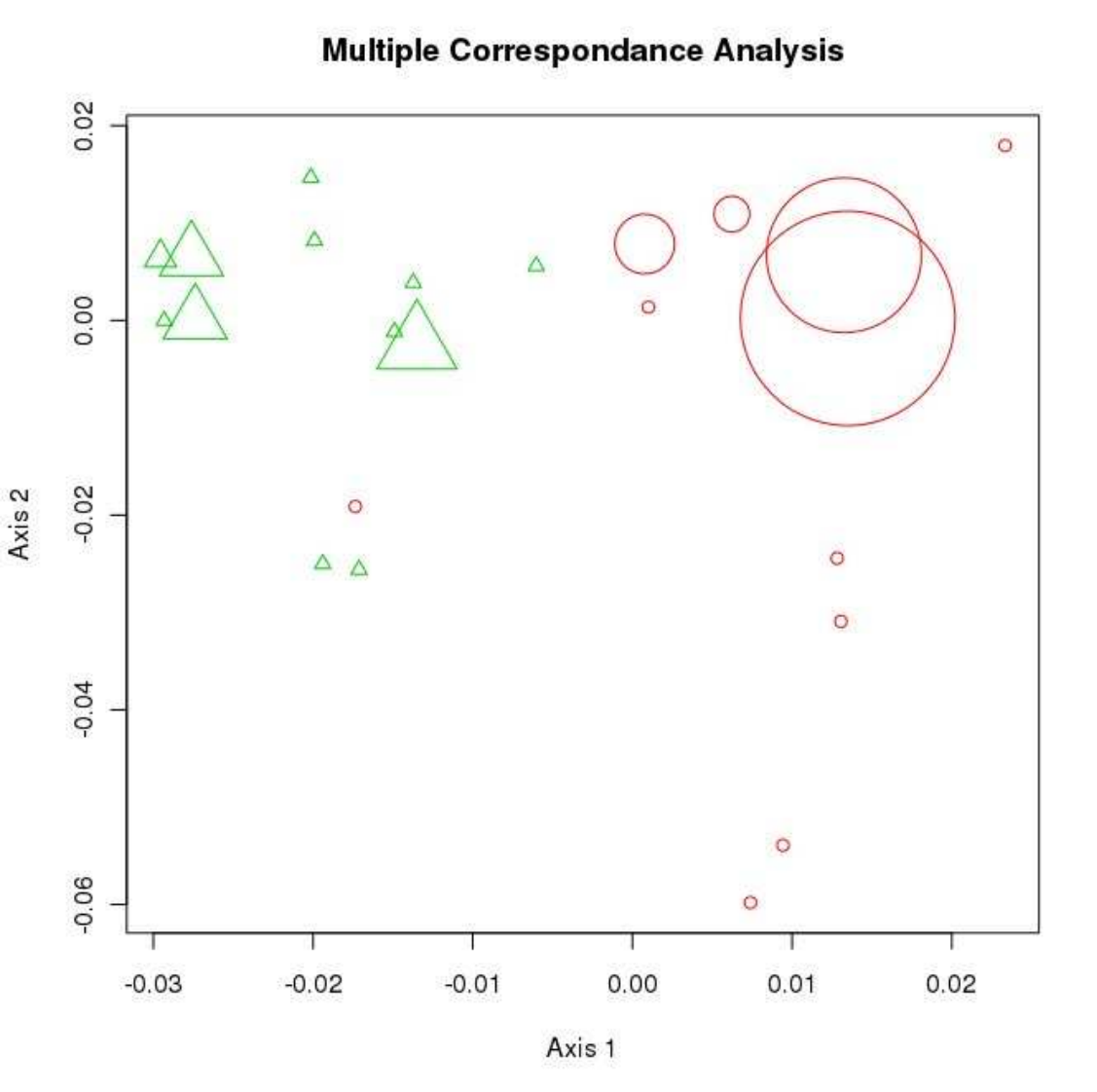
# Illustration 2 - Apprentissage

## Visualisation (1)



# Illustration 2 - Apprentissage

## Visualisation (2)



# Illustration 2 - Classement

## Classement

```
> other_birds
   gender         eyebrow collar sub-caudal border
1    male poor pronounced dotted      white    few
2  female            none dotted      black   none
3  female      pronounced   none      white   none
4    male      pronounced dotted      white   none
5    male      pronounced dotted      white   none
6    male      pronounced dotted      white   none
7  female      pronounced dotted      white   none
8  female poor pronounced dotted      white   none
9  female poor pronounced dotted      white   none
10 female poor pronounced dotted      white   none
> predict<-mixmodPredict(other_birds, classificationRule=learn["bestResult"])
> summary(predict)
*************************************************************
* partition     =  2 2 1 1 1 1 1 2 2 2
* probabilities = |   0.0334   0.9666 |
                  |   0.0012   0.9988 |
                  |   0.9998   0.0002 |
                  |   0.9896   0.0104 |
                  |   0.9896   0.0104 |
                  |   0.9896   0.0104 |
                  |   0.9828   0.0172 |
                  |   0.0226   0.9774 |
                  |   0.0226   0.9774 |
                  |   0.0226   0.9774 |
*************************************************************
```

# Rmixmod

# Perspectives

## Perspectives

- Intégrer toutes les fonctionnalités actuelles de mixmodLib.
  Reste à faire :
  - Initialisations *PARAM (/USER)* et *LABEL*
  - Modèles Haute Dimension pour la classification supervisée
- . . . et les prochaines fonctionnalités
- Activer une communauté autour de Rmixmod
- Inciter les contributions

Les idées, les contributions sont les bienvenues

# Rmixmod

## Ressources

- Site web : http ://www.mixmod.org
- Forum de discussion : https ://gforge.inria.fr/forum/forum.php ?forum_id=10462
- contact :
  - contact@mixmod.org
  - florent.langrognet@univ-fcomte.fr

FIN

Merci de votre attention