Mixmod

Un logiciel de classification supervisée et non supervisée pour données quantitatives et qualitatives

Décembre 2008

PLAN

PLAN

Fiche d'identité

Partenariat

- 4 auteurs et des compétences complémentaires en statistiques et informatique
  - C. Biernacki (Labo. Paul Painlevé - Université Lille 1/CNRS)
  - G. Celeux (Projet Select - INRIA Saclay)
  - G. Govaert (Labo. Heudiasyc - UTC/CNRS)
  - F. Langrognet (Labo. de math. de Besançon - Université Franche-Comté / CNRS)
- 5 instituts de recherche

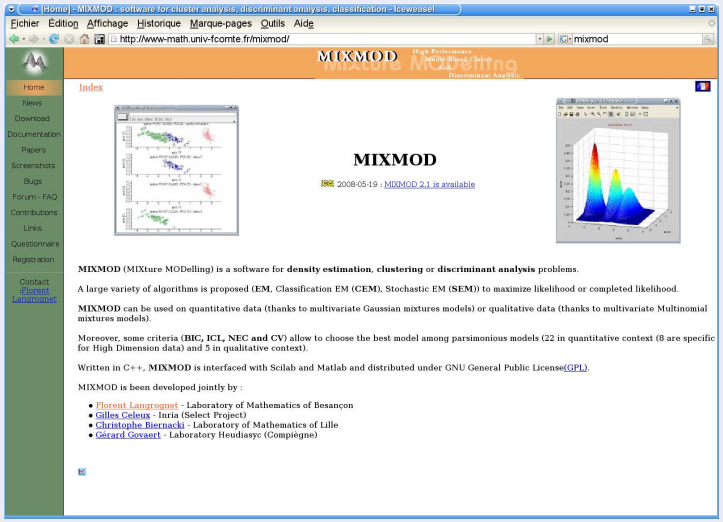
Historique

Janvier 2001	: Début du projet Mixmod
Fin 2001	: Mixmod 1.0
Février 2007	: Mixmod 2.0 (traitement des données qualitatives)
Mai 2008	: Mixmod 2.1 (traitement des données en grande dimension)

Fiche d'identité

Distribution : *www-math.univ-fcomte.fr/mixmod*

- Rubriques (eng/fr)
  - Téléchargement
  - Documentations
  - Bugs
  - FAQ/forum (google group)
  - Questionnaire
  - ...



- Packages pour Linux et Windows (source et binaire)
- 700 visites et 250 téléchargements par mois

Licence

- GNU GPL
- Autre licence si besoin ...

Fiche d'identité

Problématiques traitées

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

Cadre de travail - Type de données traitées

Modèles de mélanges

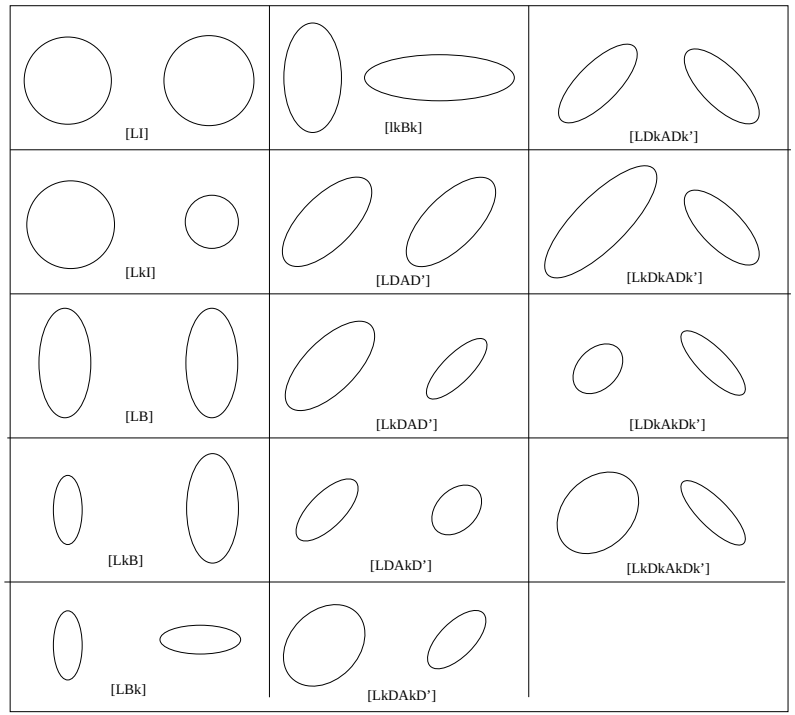
- Gaussiens (données quantitatives)
- Multinomiaux (données qualitatives)
- Modèles spécifiques pour les données en grande dimension

Modèles parcimonieux

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

Principales fonctionnalités

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

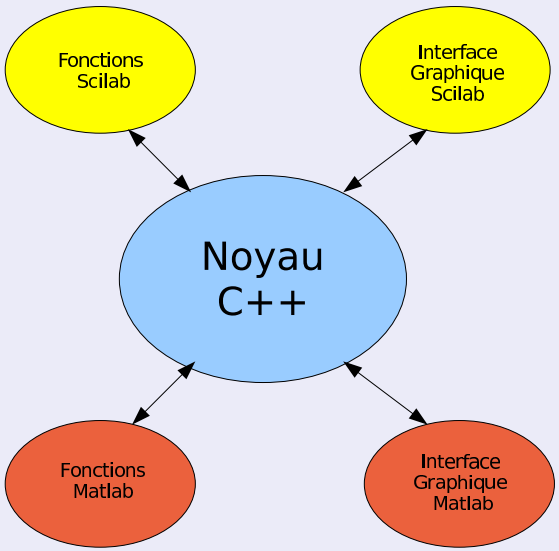
  Ex : 100 iterations de **SEM** puis 50 iterations de **EM**

Et aussi...

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

Architecture, utilisations et utilisateurs

Architecture

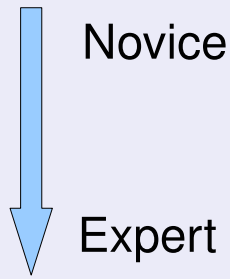


Domaines

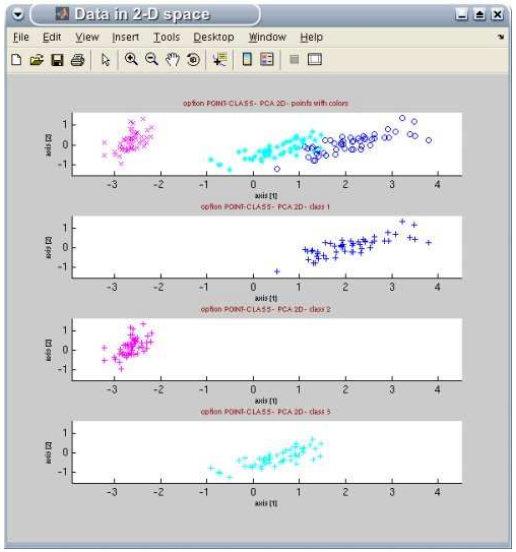
- Finance
- Biologie
- Reconnaissance de formes
- Santé
- ...

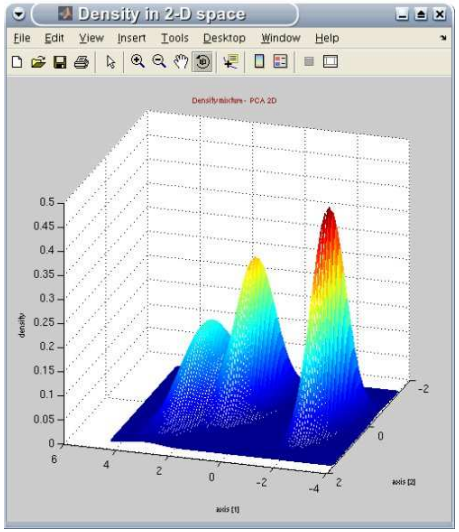
Utilisations, utilisateurs

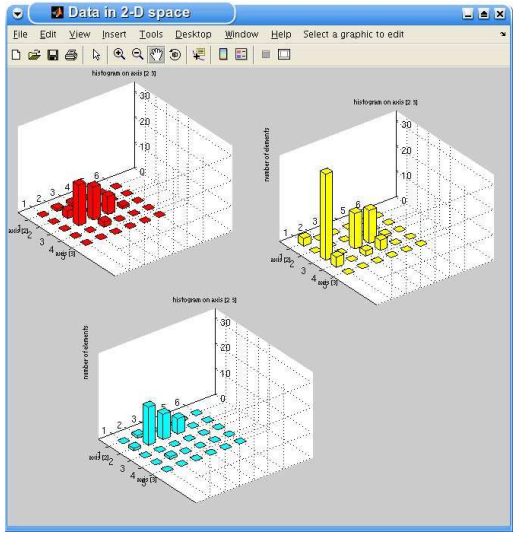
- Interfaces graphiques pour Scilab et Matlab
- Fonctions pour Scilab et Matlab
- Bibliothèque de calcul (C++)

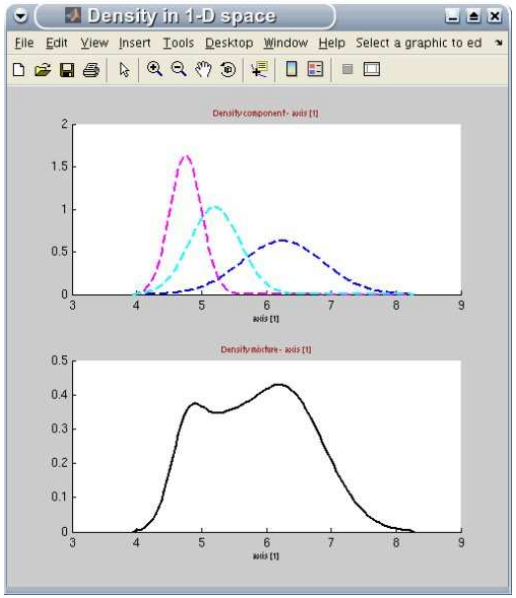


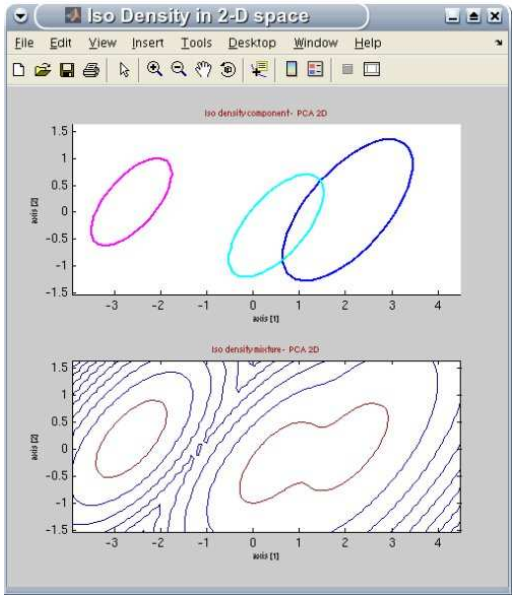
Quelques captures d'écran...

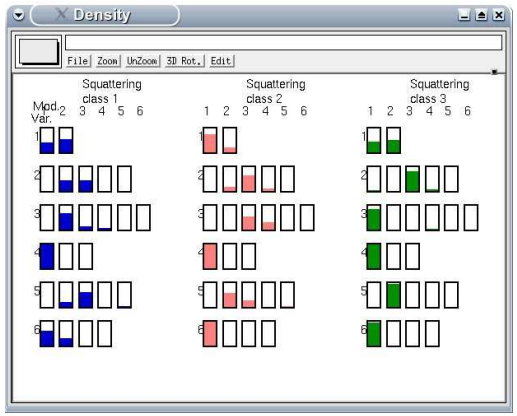












PLAN

Exemple 1
Classification non supervisée sur données quantitatives

Hypothèses

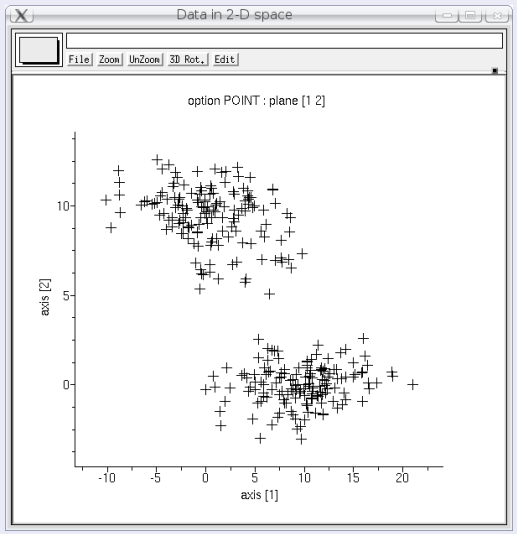
- nombre de classes : 3
- modèle gaussien : proportions libres, variances de même orientation et volumes libres (Modèle $p_k \lambda_k C$)

Objectifs

- classer les individus
- caractériser les classes

Données

300 individus en dimension 2



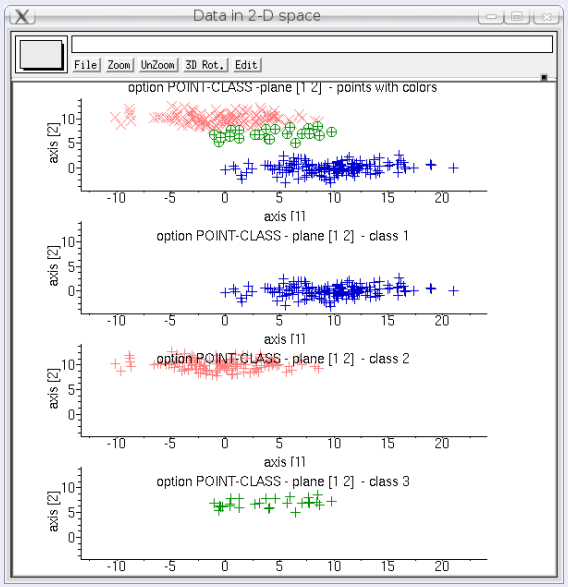
Résultats

Résultats

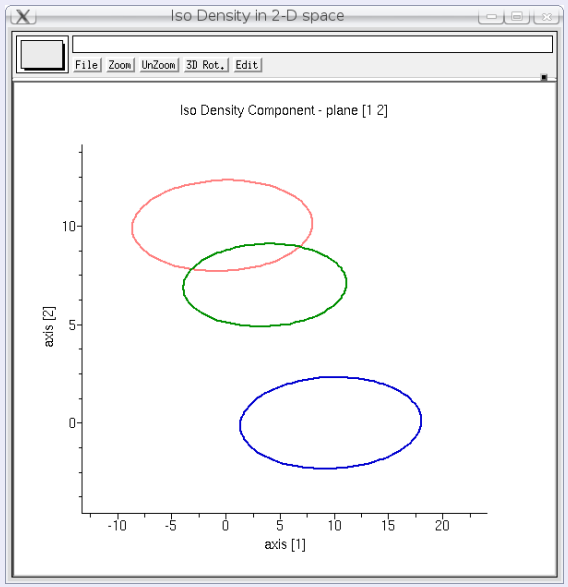
data = read('DATA/mesDonnes.dat',300,2) ;

out = mixmod(data, 3) ;

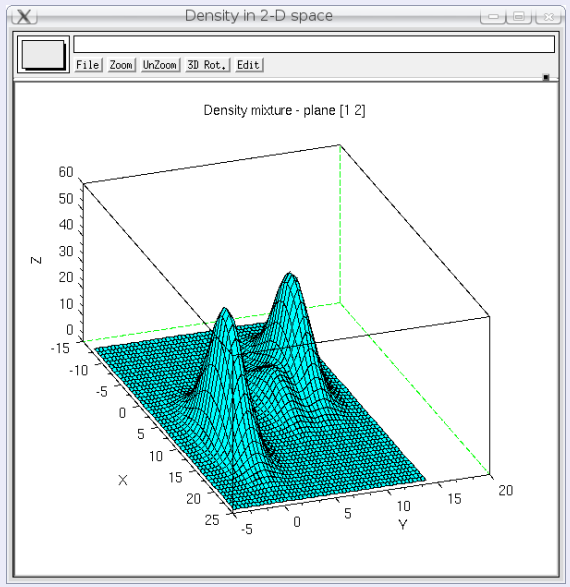
mixmodView(out, '[1 2]', poinClass')



mixmodView(out, '[1 2]', 'isoDensityComponent')



mixmodView(out, '[1 2]', 'densityMixture')



$p_1 = 0.5$
$p_2 = 0.4$
$p_3 = 0.1$

Comment obtient-on ces résultats ?

ou ...les dessous de l'algorithme EM

Maximisation de la vraisemblance par l'algorithme EM

- EM est très dépendant de l'initialisation
- EM couverge vers un maximum local

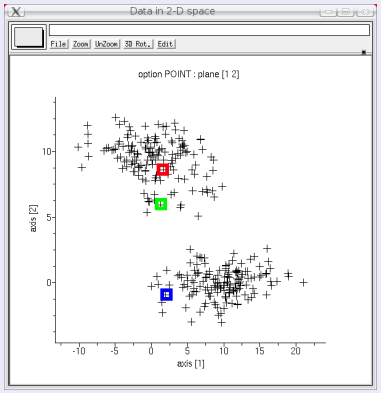
Outils disponibles dans Mixmod

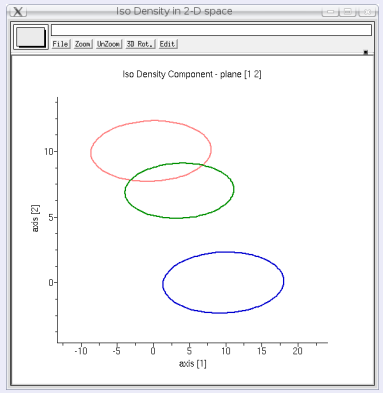
- Pour éviter de converger vers des maxima locaux
- Pour accélérer la vitesse de converge (vers le maximum global !)

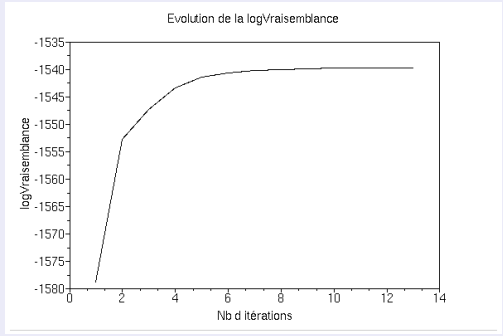
Algorithme EM et initialisation

Vitesse de convergence

- Initialisation n^{o}1

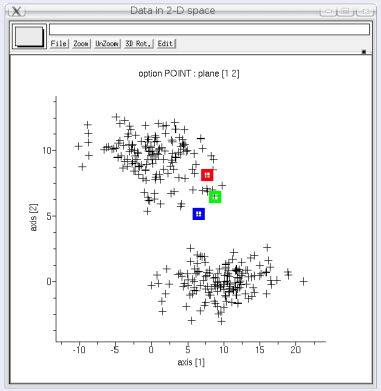


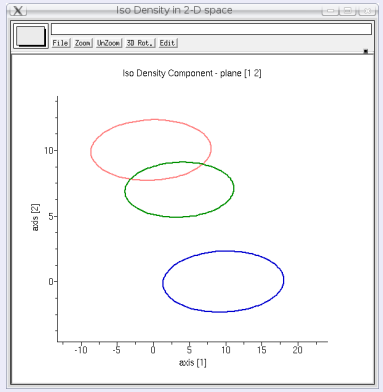


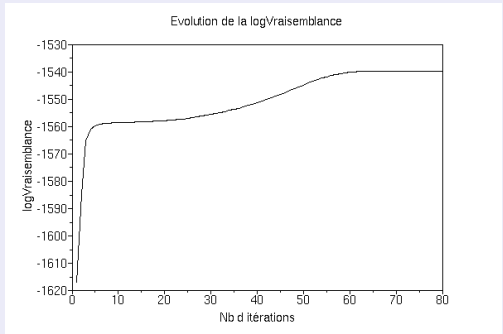


$LL = -1539$ en 10 itérations

- Initialisation n^{o}2





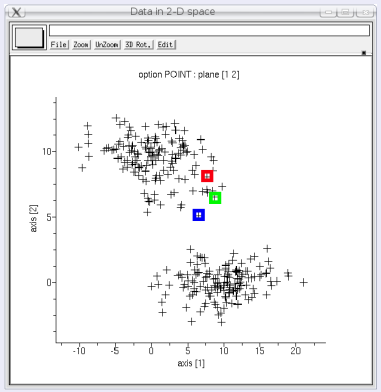


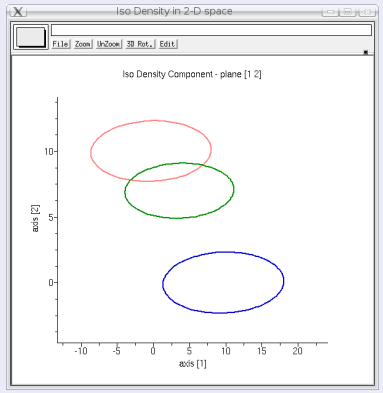
$LL = -1539$ en 80 itérations

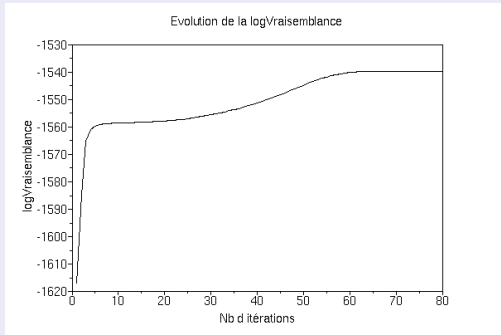
Algorithme EM et initialisation

Convergence vers des maxima différents

- Initialisation n^{o}2

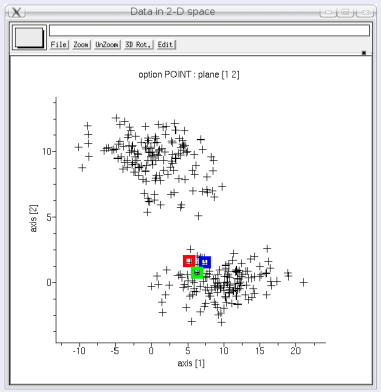


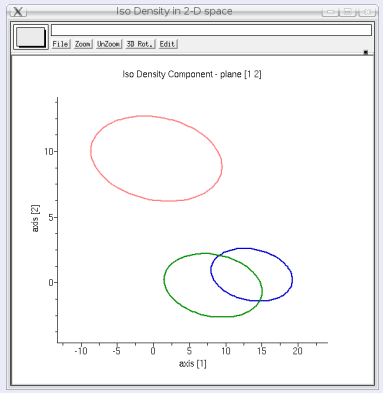


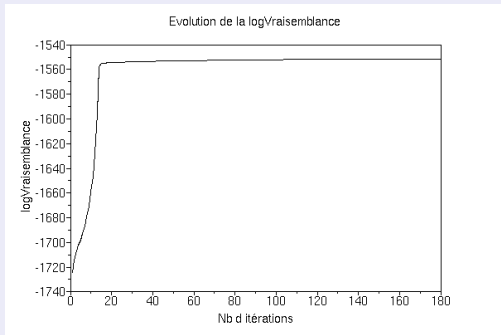


$LL = -1539$

- Initialisation n^{o}3







$LL = -1551$

Des outils pour éviter de converger vers un maximum local (1)

6 initialisations

- Information a prioiri
  - Une estimation des paramètres du modèle (proportions, moyennes, dispersions) (*USER*)
  - Une connaissance partielle de certains labels (*USER_PARTITION*)
- Pas d'information a priori
  - *Random* : meilleure configuration (Max LL) de n tirages au hasard d'individus pour intialiser les centres (n=5)
  - *Small_EM* : meilleure configuration de n tirages au hasard suivis de m itérations de EM (m<=10 et nb total de EM durant l'initialisation=50)
  - *SEM* : meilleure configuration parmi les n étapes de SEM après tirage au hasard (n=500)
  - *CEM* : meilleure configuration de n tirages au hasard suivis de m itérations de CEM (n=10, m=50)

Des outils pour éviter de converger vers un maximum local (2)

Algorithmes dans Mixmod

- 3 algorithmes : EM, SEM, CEM
- 3 possibilités d'arrêt de l'algorithme
  - après un nombre donné d'itérations
  - à la stationnarité de la vraisemblance
  - après un nombre donné d'itérations ou à la stationnarité de la vraisemblance
- Possibilité de chaîner des algorithmes dans Mixmod

Exemple de stratégie pour éviter les maxima locaux

- Initialisation : Small_EM
- 100 itérations de SEM
- itérations de EM jusqu'à stationnarité de la vraisemblance

Exemple 2
Classification supervisée sur données qualitatives

Puffins

Données

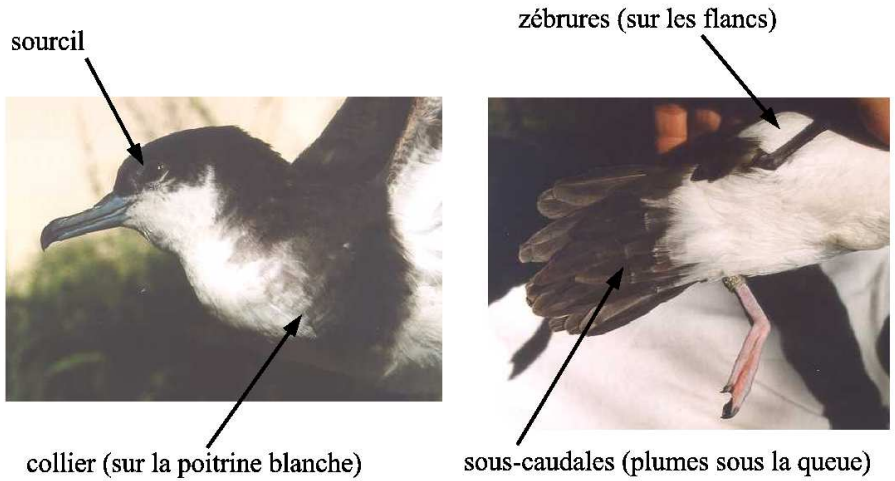
- Nombre d'individus : $n = 69$
- Nombre d'espèces (classes) : $K = 2$
- Nombre de variables : $d = 6$
- Individu i : $(x_i, z_i) = ((x_i^j)_{j=1,\ldots,d}, z_i)$

n^o	z_i	x_i^1	x_i^2	x_i^3	x_i^4	x_i^5	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

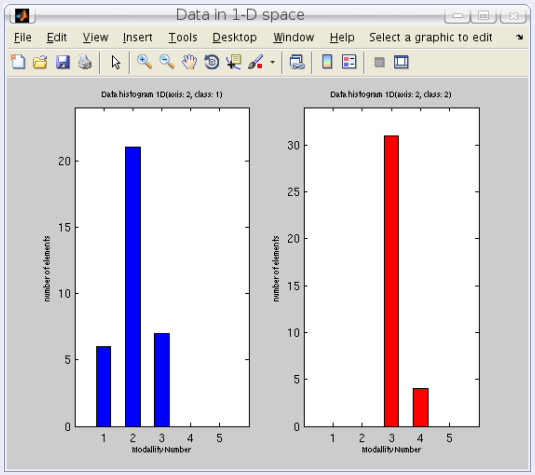
Description des variables

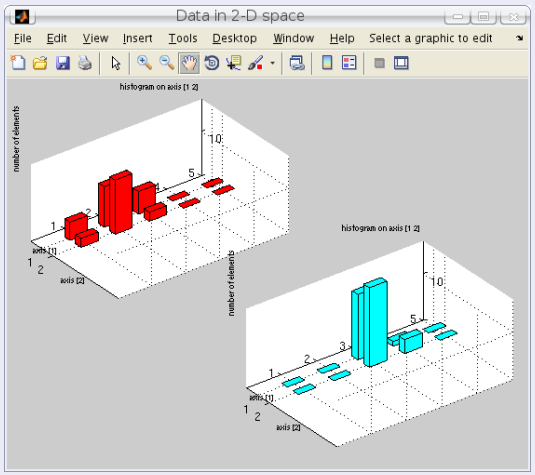
variable	nombre de niveaux de réponse	valeurs
sexe	2	mâle, femelle
sourcils	5	absent -> très prononcé
collier	6	absent -> continu
zébrures	3	absent, peu , présence forte
sous-caudales	5	blanc, noir, noir&blanc, noir&BLANC, NOIR&blanc
liseret	4	absent, . . . , beaucoup



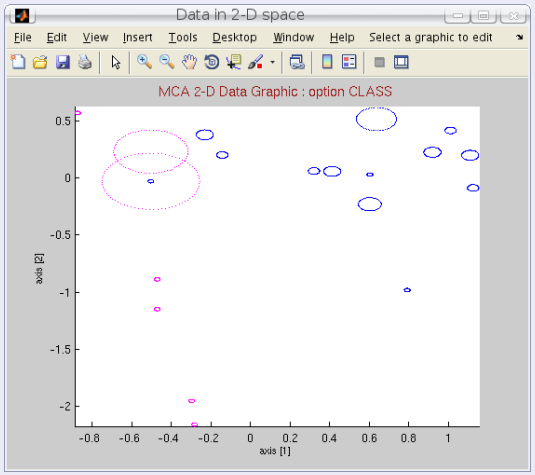
Visualisation des données avec Mixmod

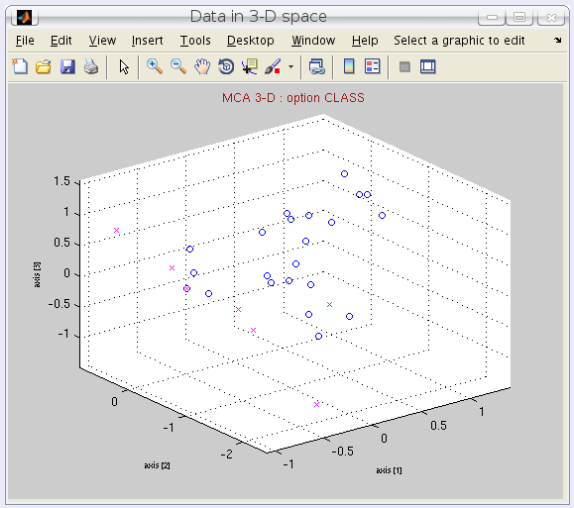
barplots





scatterplots





Classification supervisée - étape 1

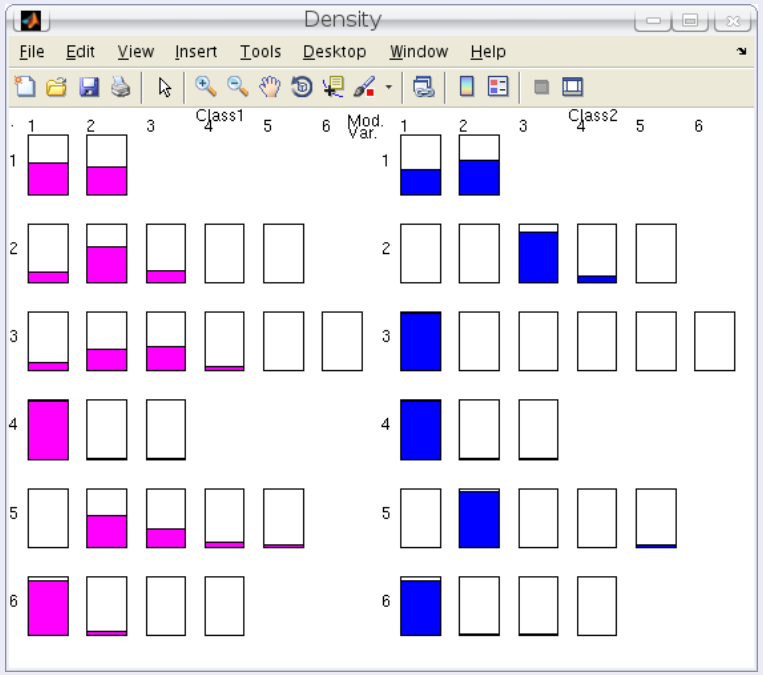
Règle de classement

Paramètres estimés

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

$$\hat{\alpha}_1 = \begin{bmatrix} \hat{\alpha}_1^1 = & 0.53 & 0.47 & & & & \\ \hat{\alpha}_1^2 = & 0.18 & 0.61 & 0.20 & 0.00 & 0.00 & \\ \hat{\alpha}_1^3 = & 0.15 & 0.37 & 0.41 & 0.06 & 0.00 & 0.00 \\ \hat{\alpha}_1^4 = & 0.98 & 0.01 & 0.01 & & & \\ \hat{\alpha}_1^5 = & 0.00 & 0.55 & 0.32 & 0.10 & 0.04 & \\ \hat{\alpha}_1^6 = & 0.94 & 0.06 & 0.00 & 0.00 & & \end{bmatrix}$$

$$\hat{\alpha}_2 = \begin{bmatrix} \hat{\alpha}_2^1 = & 0.43 & 0.57 & & & & \\ \hat{\alpha}_2^2 = & 0.00 & 0.00 & 0.87 & 0.13 & 0.00 & \\ \hat{\alpha}_2^3 = & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ \hat{\alpha}_2^4 = & 0.98 & 0.01 & 0.01 & & & \\ \hat{\alpha}_2^5 = & 0.00 & 0.97 & 0.00 & 0.00 & 0.03 & \\ \hat{\alpha}_2^6 = & 0.94 & 0.03 & 0.03 & 0.00 & & \end{bmatrix}$$



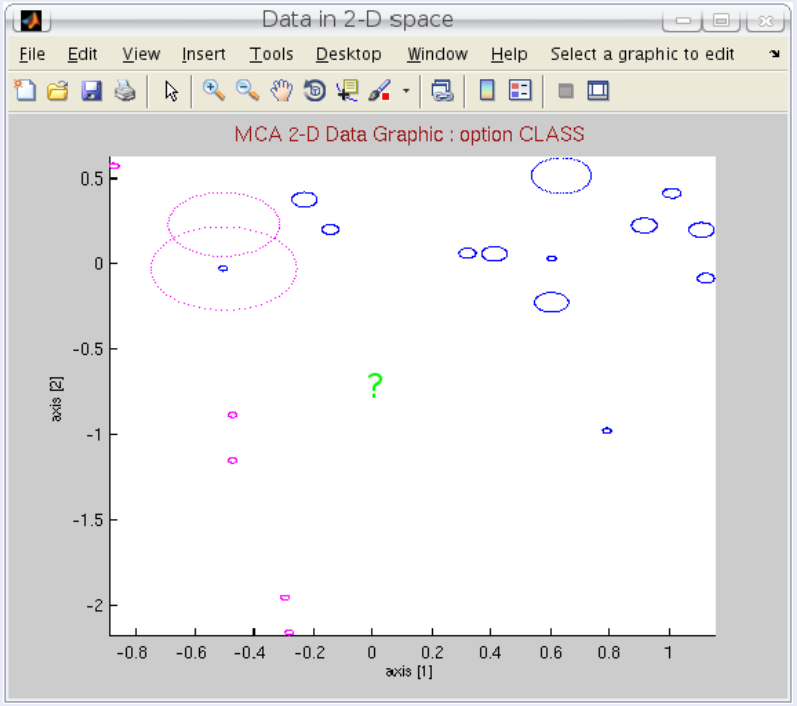
Mesure de la qualité de la règle de classement

- Taux de reclassement par Validation Croisée : 97.0% (67 sur 69)
- Taux de reclassement par MAP : 98.5% (68 sur 69)

Classification supervisée - étape 2

Classement d'un nouvel individu

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



Resultat - Application de la règle de classement

$z_{70} = 2$

$P(z_{70} = 1) = 0.06$ - $P(z_{70} = 2) = 0.94$

Modèle de mélanges parcimonieux

- Objectif : proposer des modèles 'plus simples' (avec moins de paramètres à estimer)
- Moyen : imposer des contraintes raisonnables sur p_k et α_k
- Garde-fou : critère de choix de modèles

Reparamétrisation de α_k

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

- $a_k^{jh} = \begin{cases} 1 & \text{si } h = \arg\max_h \alpha_k^{jh} \\ 0 & \text{sinon} \end{cases}$ *centre*
- $\varepsilon_k^{jh} = \begin{cases} 1 - \alpha_k^{jh} & \text{si } a_k^{jh} = 1 \\ \alpha_k^{jh} & \text{si } a_k^{jh} = 0. \end{cases}$ *dispersion*

Exemple : $\alpha_k^{jh} = (0.2\ 0.7\ 0.1) \longrightarrow \begin{cases} a_k^j = (0\ 1\ 0) \\ \varepsilon_k^j = (0.2\ 0.3\ 0.1) \end{cases}$

Modèle de classes latentes

10 modèles

- Contraintes imposées au paramètre de dispersion ε_k^{jh} : 5 modèles

modèle	la dispersion peut dépendre de ...	dispersions identiques pour ...
$[\varepsilon]$	rien	toutes les classes, variables et tous les niv. de réponse
$[\varepsilon_k]$	classes	toutes les variables et tous les niveaux de réponse
$[\varepsilon^j]$	variables	toutes les classes et tous les niveaux de réponse
$[\varepsilon_k^j]$	classes et variables	tous les niveaux de réponse
$[\varepsilon_k^{jh}]$	classes, variables et niveaux de réponse	dispersions totalement libres

- Contraintes sur les proportions (p_k) (libres ou non) : 10 modèles

Nombre de paramètres à estimer (dl)

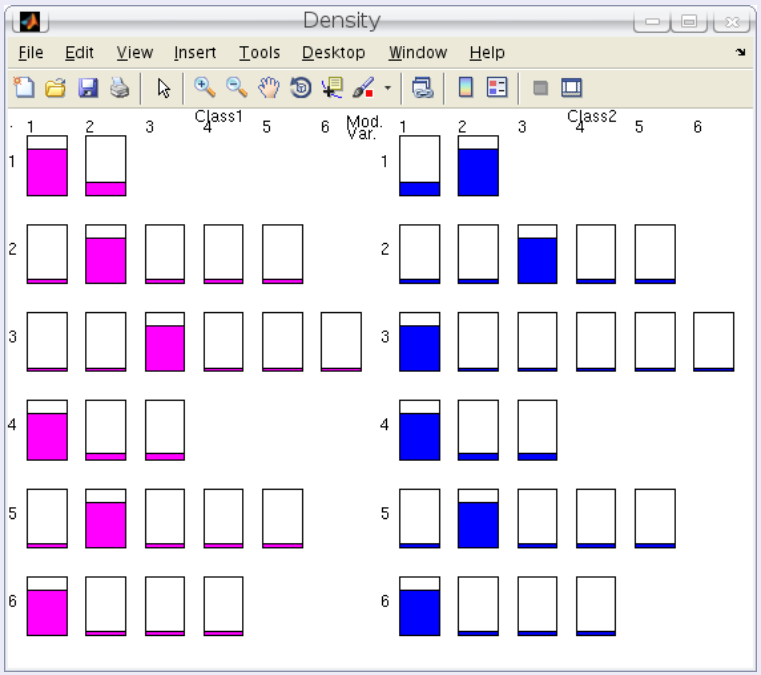
modèle	dl	Exemple (oiseaux)
$[p\varepsilon]$	1	1
$[p\varepsilon_k]$	K	2
$[p\varepsilon^j]$	d	6
$[p\varepsilon_k^j]$	Kd	12
$[p\varepsilon_k^{jh}]$	$K\sum_{j=1}^{d}(m_j-1)$	38

Note : ajouter $K-1$ pour les modèles à proportions libres.

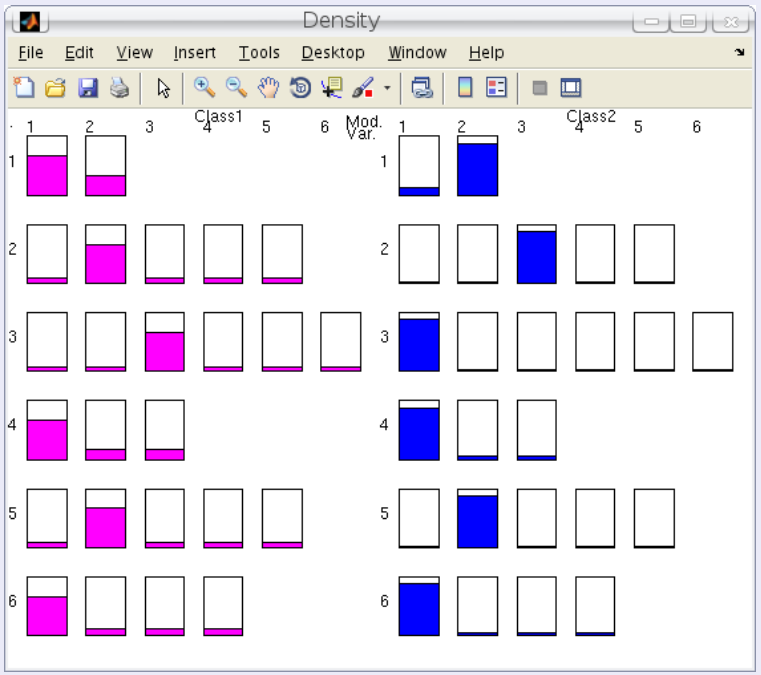
Modèle de classes latentes

Paramètres estimés

Modèle $[p\varepsilon]$



Modèle $[p\varepsilon_k]$



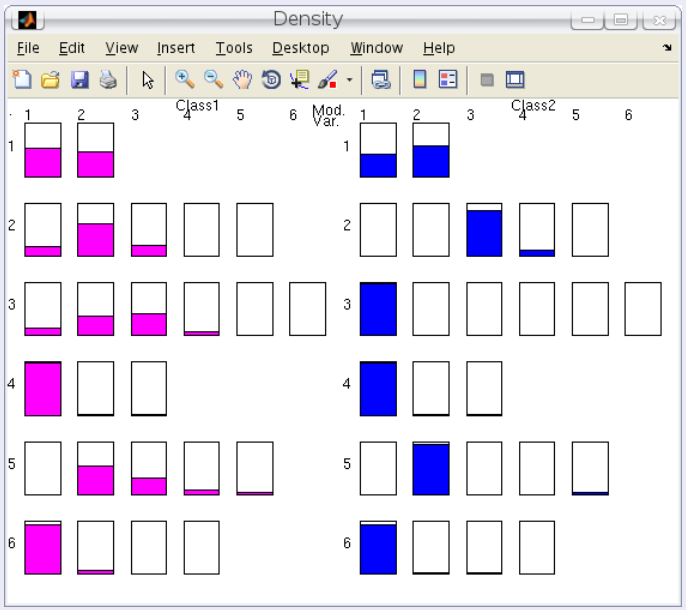
Choix de modèles

Critères

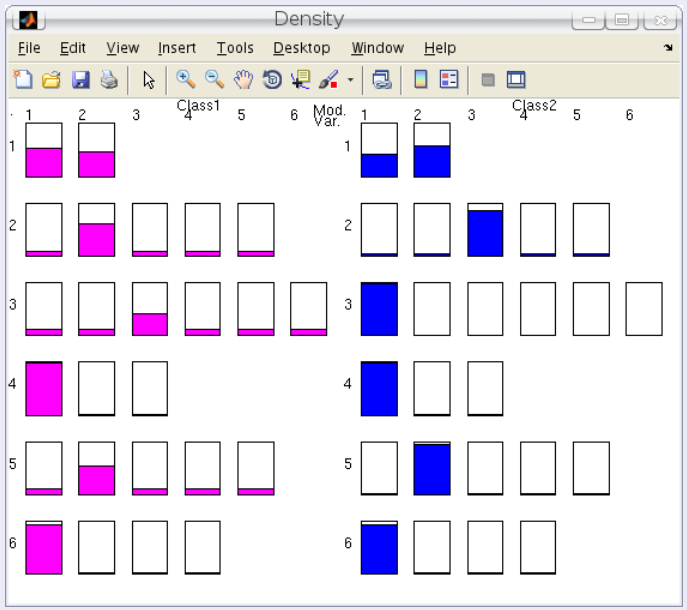
- CV (Validation Croisée) : taux de 'bon' reclassement à maximiser
- BIC à minimiser ($BIC = -2.LL + dl.log(n)$)

	$[p\varepsilon]$	$[p\varepsilon_k]$	$[p\varepsilon^j]$	$[p\varepsilon_k^j]$	$[p\varepsilon_k^{jh}]$	$[p_k\varepsilon]$	$[p_k\varepsilon_k]$	$[p_k\varepsilon^j]$	$[p_k\varepsilon_k^j]$	$[p_k\varepsilon_k^{jh}]$
LL	-344	-329	-311	-281	-239	-344	-329	-311	-281	-239
CV	87.0%	88.5%	84.1%	92.8%	97.0%	87.0%	85.5%	85.5%	92.8%	97.0%
BIC	693	667	649	613	640	697	671	653	617	645

Choix par CV : $[p\varepsilon_k^{jh}]$



Choix par BIC : $[p\varepsilon_k^{j}]$



PLAN

Perspectives (1)

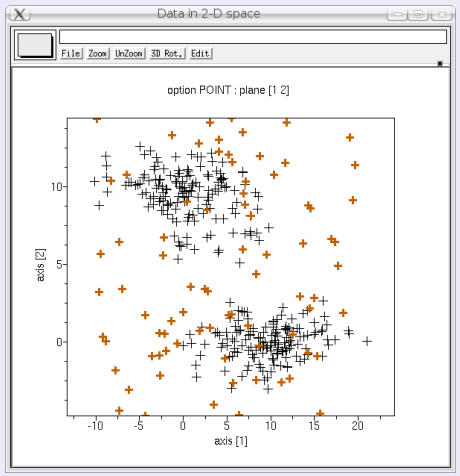
Nouvelles fonctionnalités

- Classification semi-supervisée
- Traitement des données mixtes (quantitatives et qualitatives)

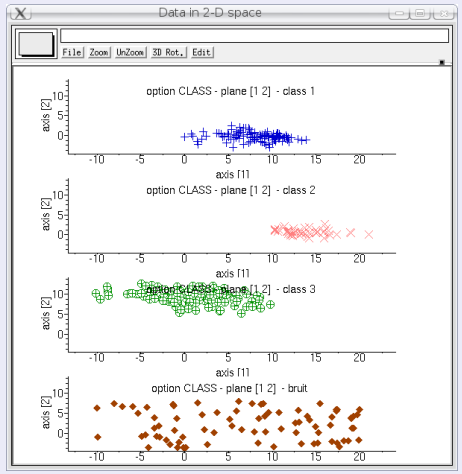
individu	hauteur	poids	sexe
1	172.5	66.3	1
2	167.1	54.1	2
.	.	.	.
.	.	.	.
.	.	.	.

- Traitement des données bruitées

Données avec bruit



Classification avec présence d'une classe de bruit



- . . .

Perspectives (2)

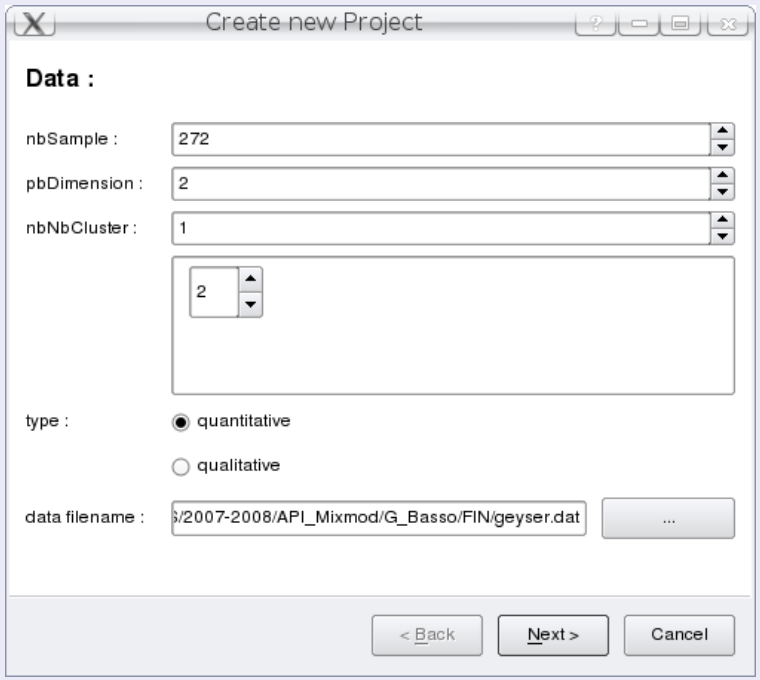
Evolutions informatiques

- Travail de fonds sur l'amélioration des performances
- Evolution de la licence Mixmod
  Vers un modèle à double licence (GPL + autre licence) ?
- Utilisation de Mixmod
  - Simplifier l'utilsation des fonctions Mixmod pour Scilab et Matlab
  - Fonctions Mixmod pour R
  - Création de communautés Mixmod/Scilab, Mixmod/Matlab et Mixmod/R
  - Développement d'une interface graphique Mixmod (logiciel *standalone*)

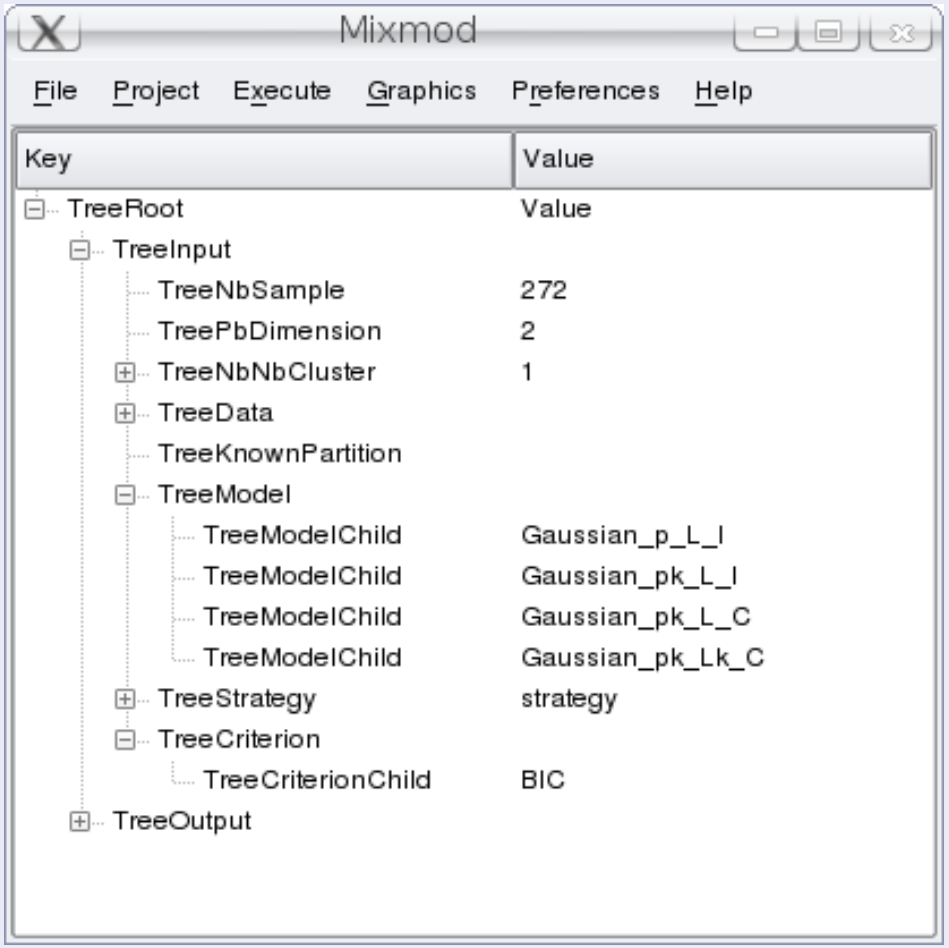
Perspectives (3)

Interface graphique pour Mixmod

Nouveau projet



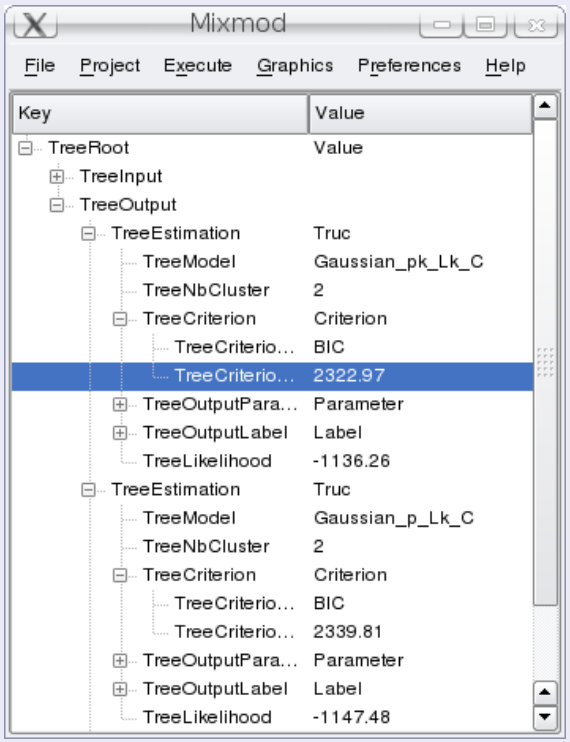
Input



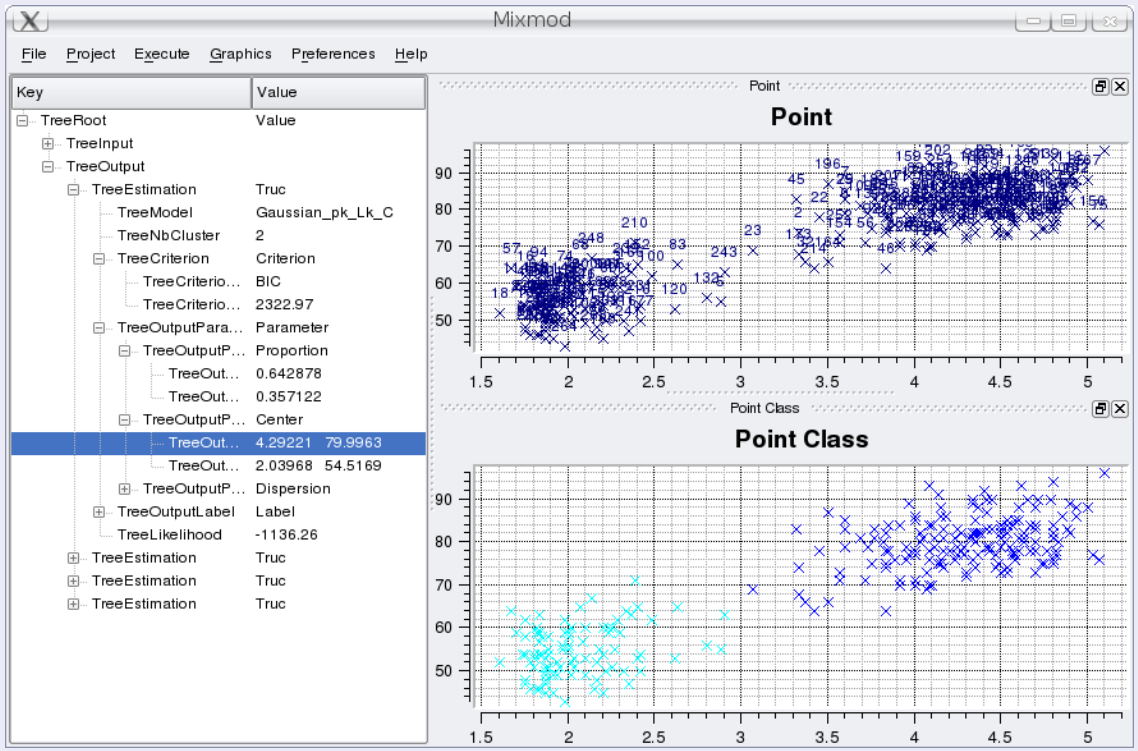
Perspectives (4)

Interface graphique pour Mixmod

Outputs



Graphiques



FIN

Merci de votre attention

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