Mixmod

Logiciel de classification supervisée et non supervisée

Vu du côté utilisateurs et développeurs

3^{e} rencontre Mixmod
2 décembre 2010

PLAN

1. Comment est développé Mixmod ?
2. Principales fonctionnalités
3. Comment utiliser Mixmod ?
   - La situation actuelle (avant Mixmod 3.0)
   - Les nouveautés et les perspectives
   - Roadmap

PLAN

L'équipe Mixmod

Un Comité de Pilotage

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)

L'équipe de développement

- F. Langrognet (à 60% depuis 10 ans)
- Ingénieurs contractuels (INRIA)
  6 années ingénieurs depuis 2002
- Stagiaires

Besoins - Demandes

- Intégrer de nouvelles fonctionnalités :
  - modèles de grande dimension pour la classification non supervisée
  - données mixtes
  - données bruitées
  - ...
- Répondre aux demandes des utilisateurs :
  - installation
  - utilisation
- Maintenance
- Amélioration des performances (temps de calcul)
- Amélioration de la robustesse (traitement amélioré des underflow et overflow, ...)
- Adapter les utilisations de Mixmod :
  - interface graphique
  - fonctions pour R
  - adapter les fonctions pour Scilab, Matlab

Ressources vs Besoins

Besoins et demandes croissants

- 250 téléchargements pas mois
- Hétérogénéité croissante des utilisateurs
- Hétérogénéité des plateformes informatiques (OS, architecture, ...) et des logiciels tiers (Scilab, Matlab)
- Taille du logiciel (70 000 lignes de code)

Ressources

- Ressources :
  Au 1/11/2010 : fin de contrat du dernier CDD INRIA
  Depuis : un seul ingénieur/developpeur (à 60% sur Mixmod)
- Nombre de contributions faible

PLAN

Fonctionnalités

Problématiques traitées

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

Cadre de travail - Type de données traitées

Modèles de mélanges

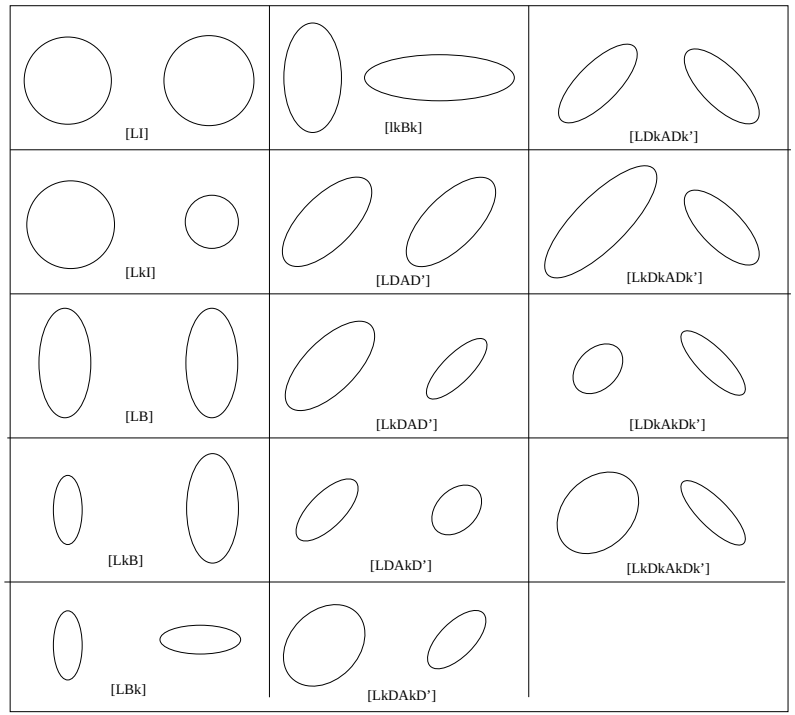
- Gaussiens (données quantitatives)
- Multinomiaux (données qualitatives)
- Modèles spécifiques pour les données en grande dimension

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

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

Exemple
Classification non supervisée sur données quantitatives

Hypothèses

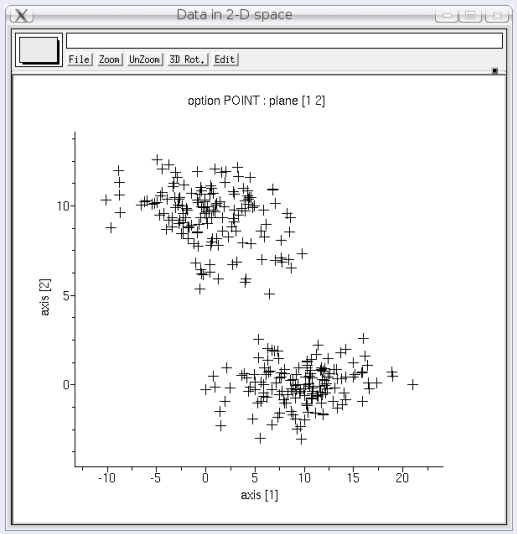
- nombre de classes : 3
- modèle gaussien : proportions libres, variances de même orientation et volumes libres (Modèle $p_k \lambda_k C$)

Objectifs

- classer les individus
- caractériser les classes

Données

300 individus en dimension 2

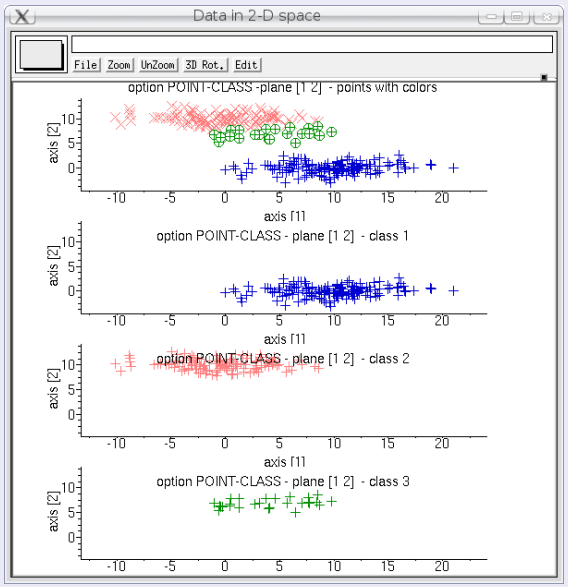


Résultats

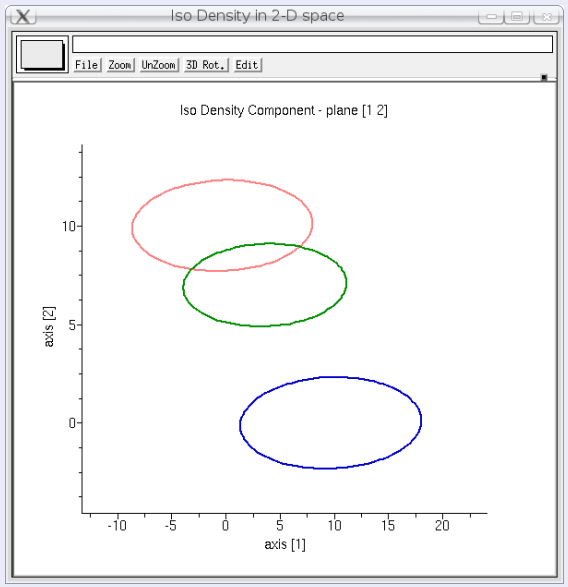
data = read('DATA/mesDonnes.dat',300,2) ;

out = mixmod(data, 3) ;

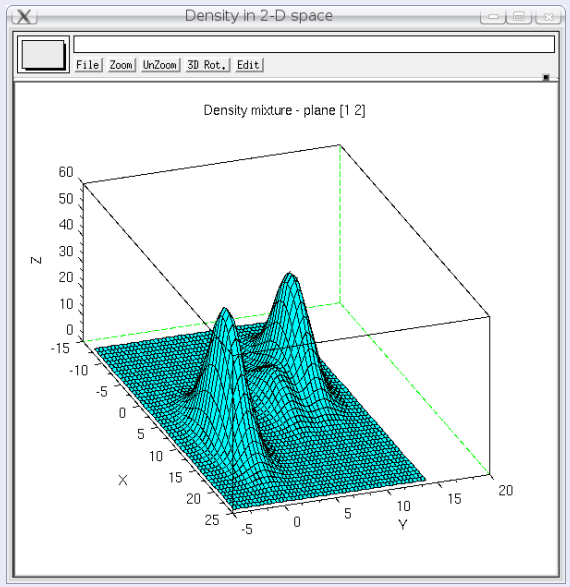
mixmodView(out, '[1 2]', poinClass')



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

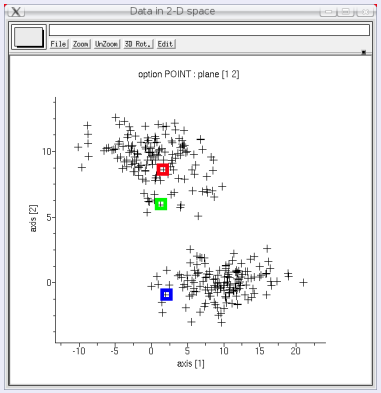
Outils disponibles dans Mixmod

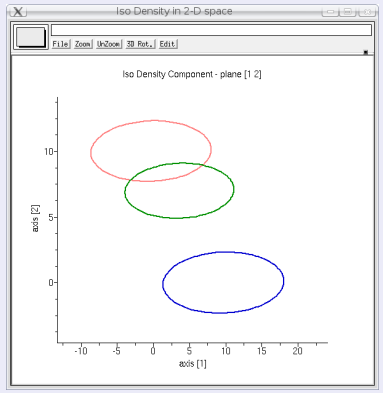
- Pour éviter de converger vers des maxima locaux
- Pour accélérer la vitesse de converge (vers le maximum global !)

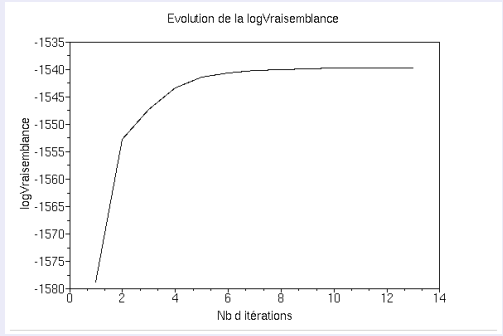
Algorithme EM et initialisation

Vitesse de convergence

- Initialisation n^{o}1

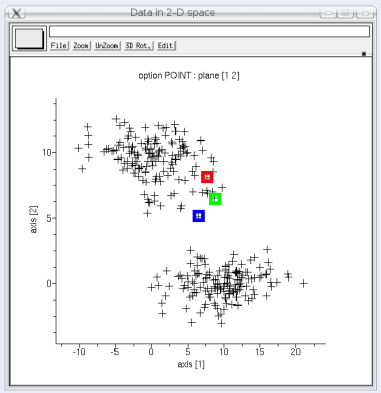


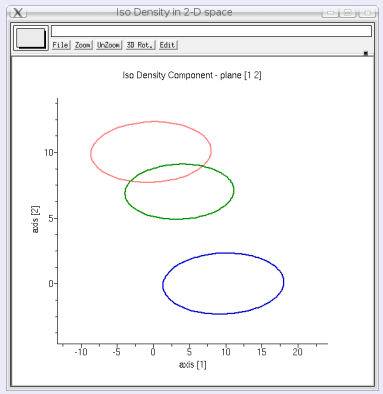


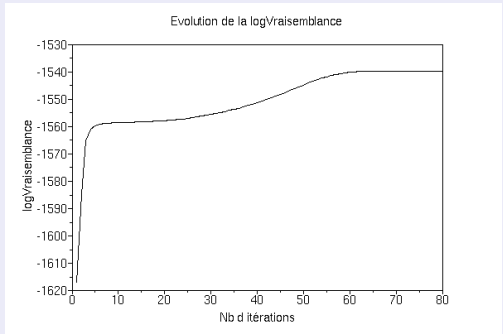


$LL = -1539$ en 10 itérations

- Initialisation n^{o}2





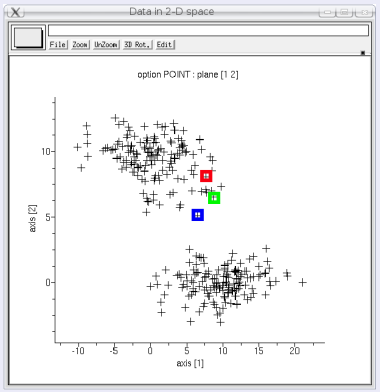


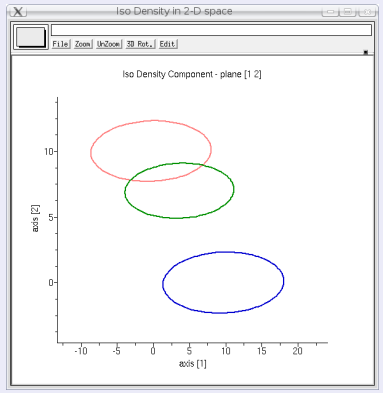
$LL = -1539$ en 80 itérations

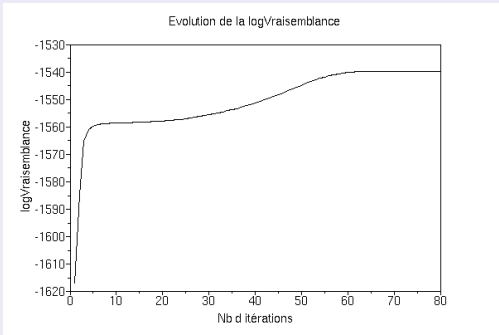
Algorithme EM et initialisation

Convergence vers des maxima différents

- Initialisation n^{o}2

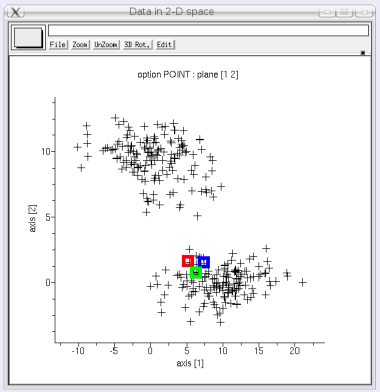


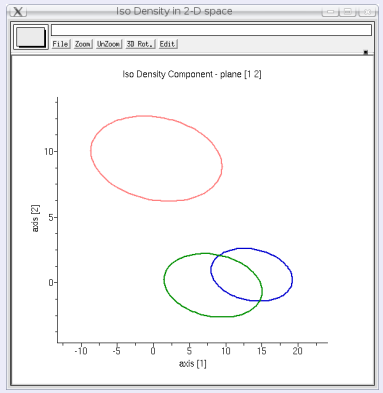


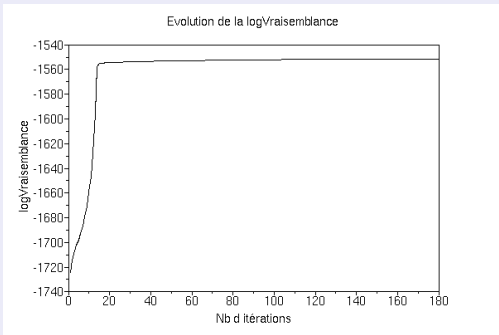


$LL = -1539$

- 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 // PARAMETER*)
  - Une connaissance partielle de certains labels *(USER_PARTITION // 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

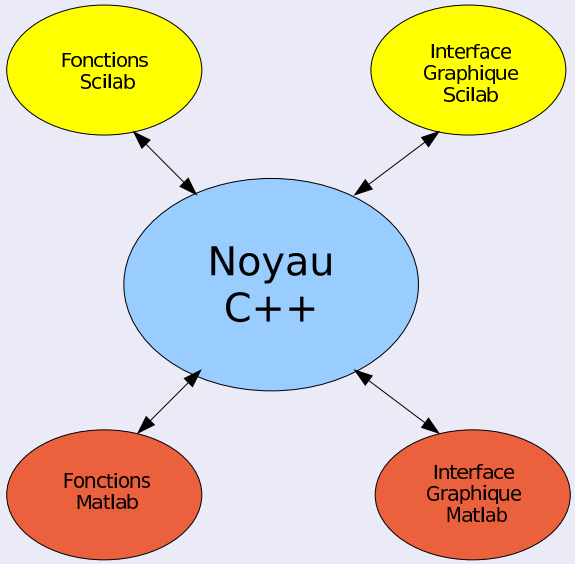
- Initialisation : Small_EM
- 100 itérations de SEM
- itérations de EM jusqu'à stationnarité de la vraisemblance

PLAN

PLAN

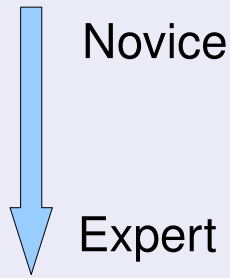
Architecture, utilisations et utilisateurs

Architecture



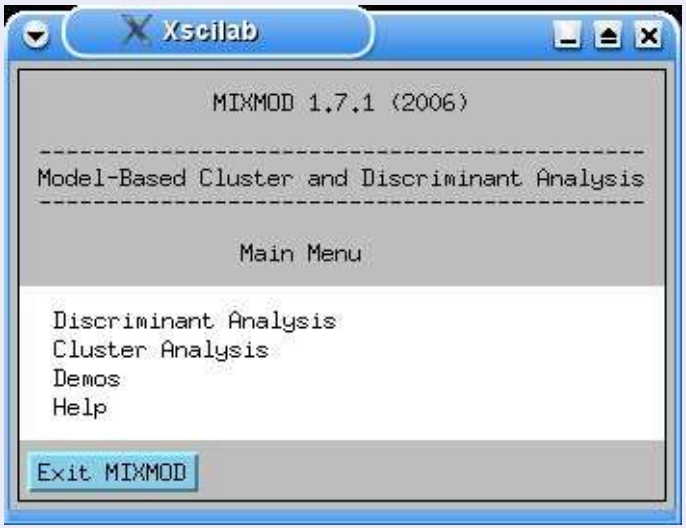
Utilisations

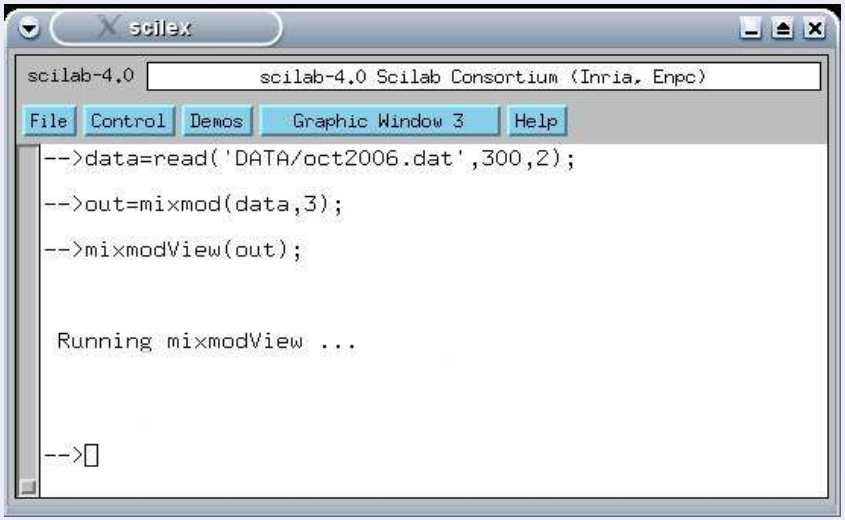
- Interfaces graphiques pour Scilab et Matlab
- Fonctions pour Scilab et Matlab
- Mode expert et Bibliothèque de calcul (C++)

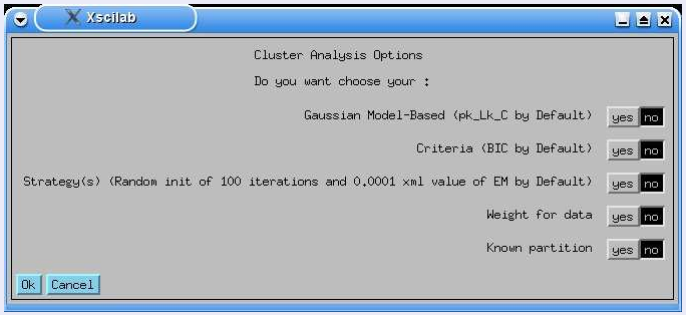


Outils pour Scilab et Matlab

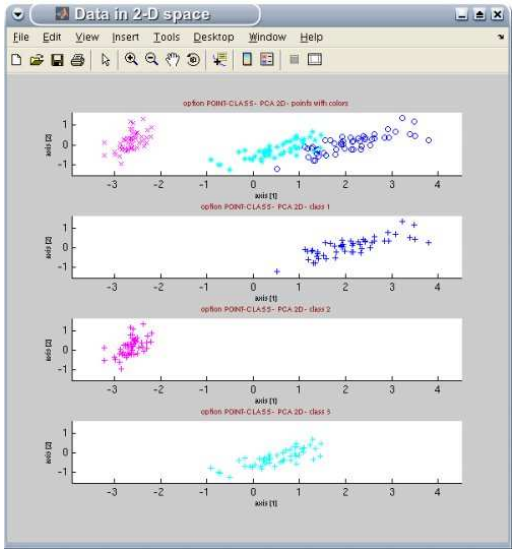
Interface graphique et Fonctions

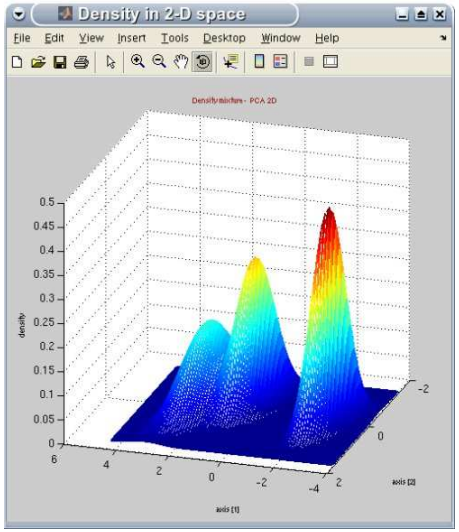


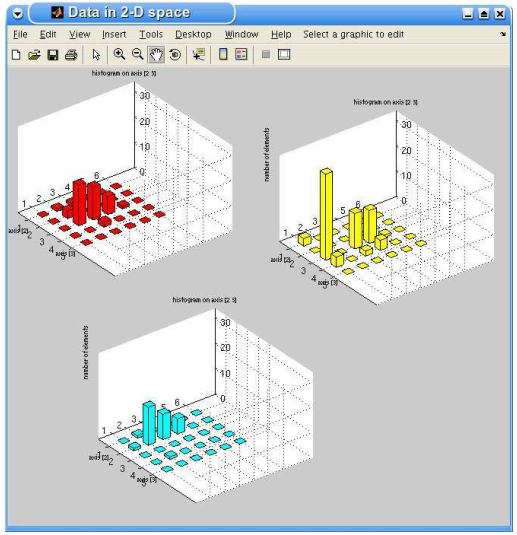


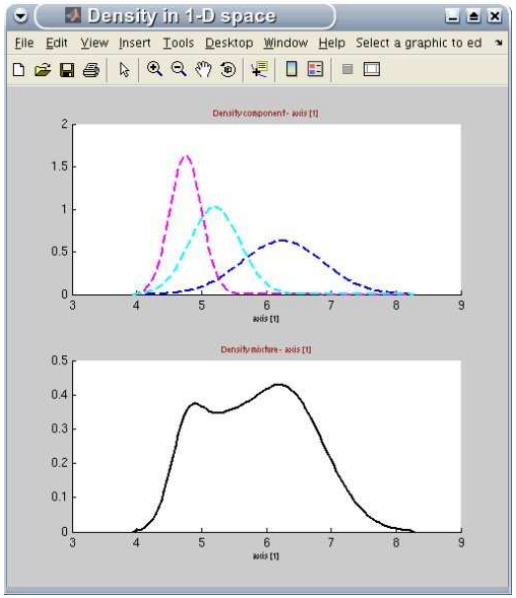


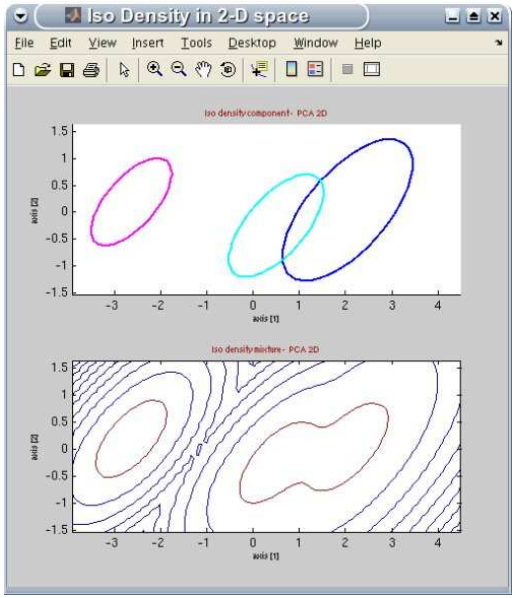
Outils de visualisation pour Scilab/Matlab

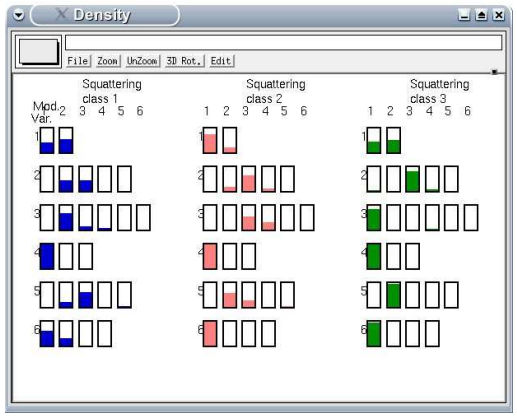






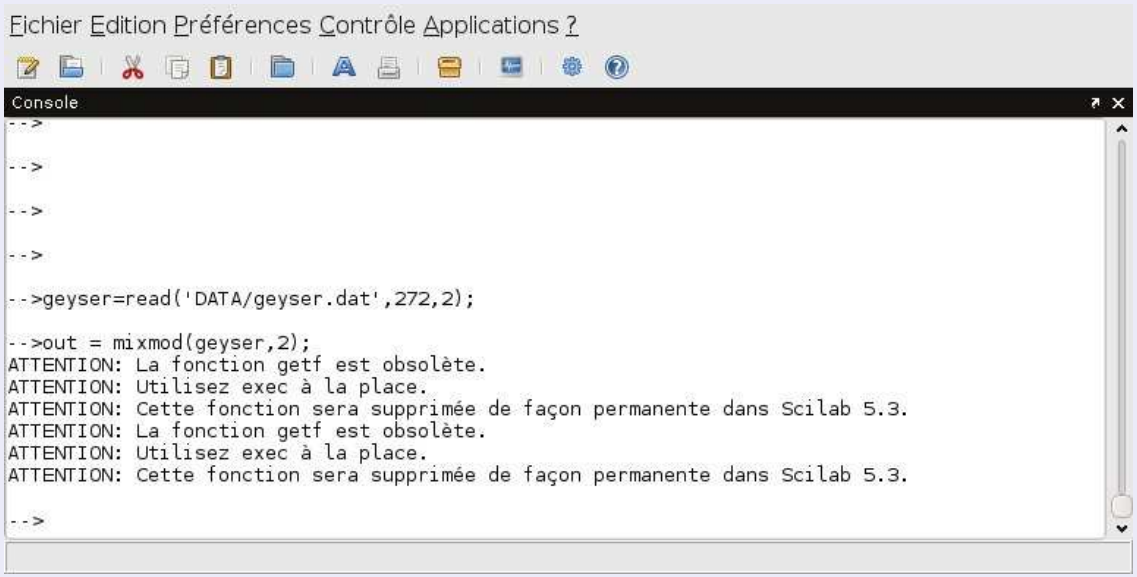






Limites et faiblesses

- Convivialité de l'interface graphique (Scilab/Matlab)
- Quelques difficultés à utiliser les fonctions pour Scilab/Matlab
- Absence de fonctions pour R
- Difficulté à suivre les évolutions de Scilab/Matlab



PLAN

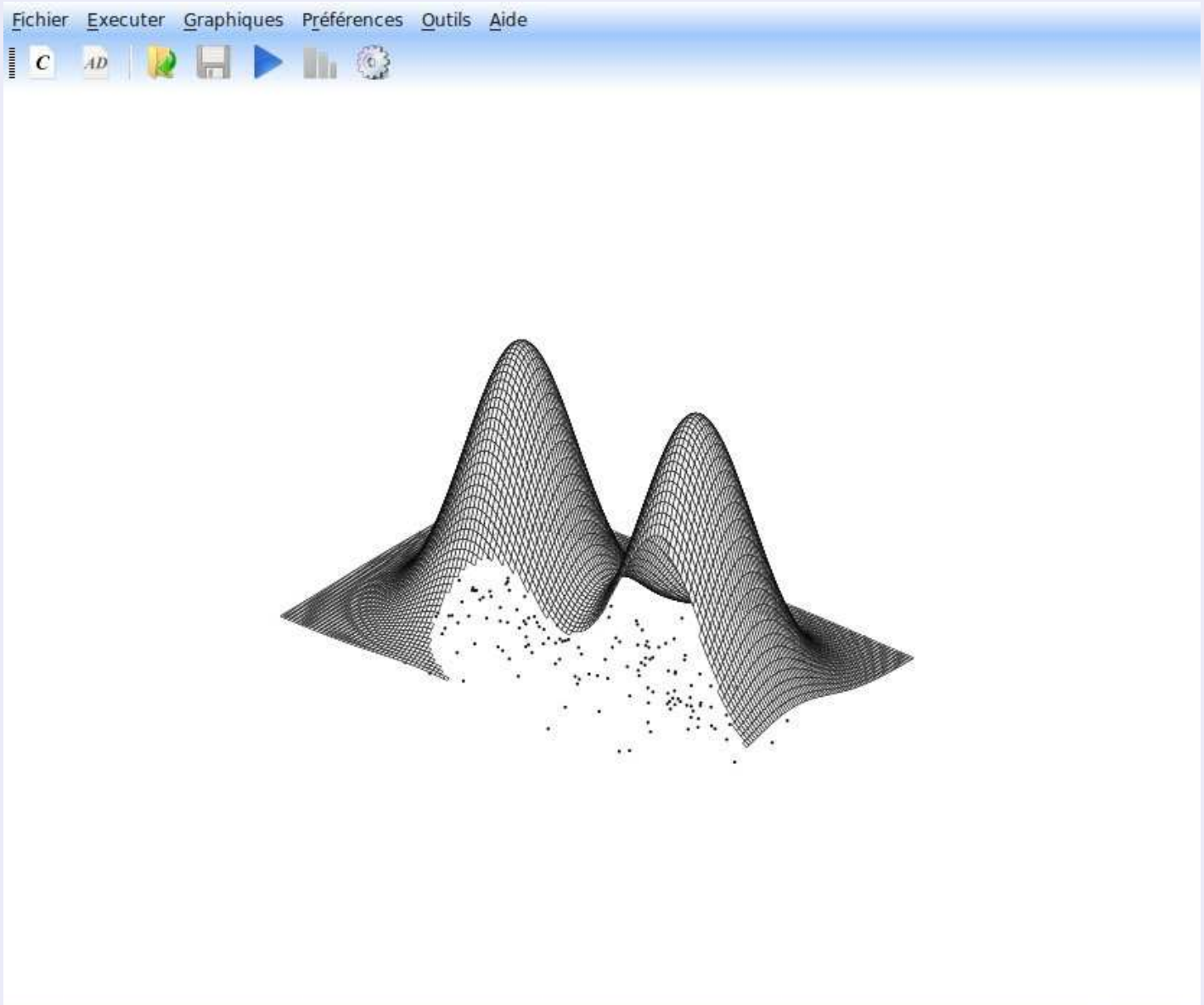
Nouveautés et perspectives

Pour les logiciels tiers (Scilab, Matlab, R)

- Des fonctions ré-écrites pour Scilab et Matlab et des fonctions pour R
  - Une fonction par fonctionnalité (mixmod_cluster, mixmod_DA)
  - Une utilisation simplifiée
- Les fonctions graphiques (visualisation) ne seront plus maintenues par l'équipe Mixmod
  Création de communautés :
  - Mixmod/Scilab
  - Mixmod/Matlab
  - Mixmod/R
- Plus d'interface graphique pour Scilab/Matlab

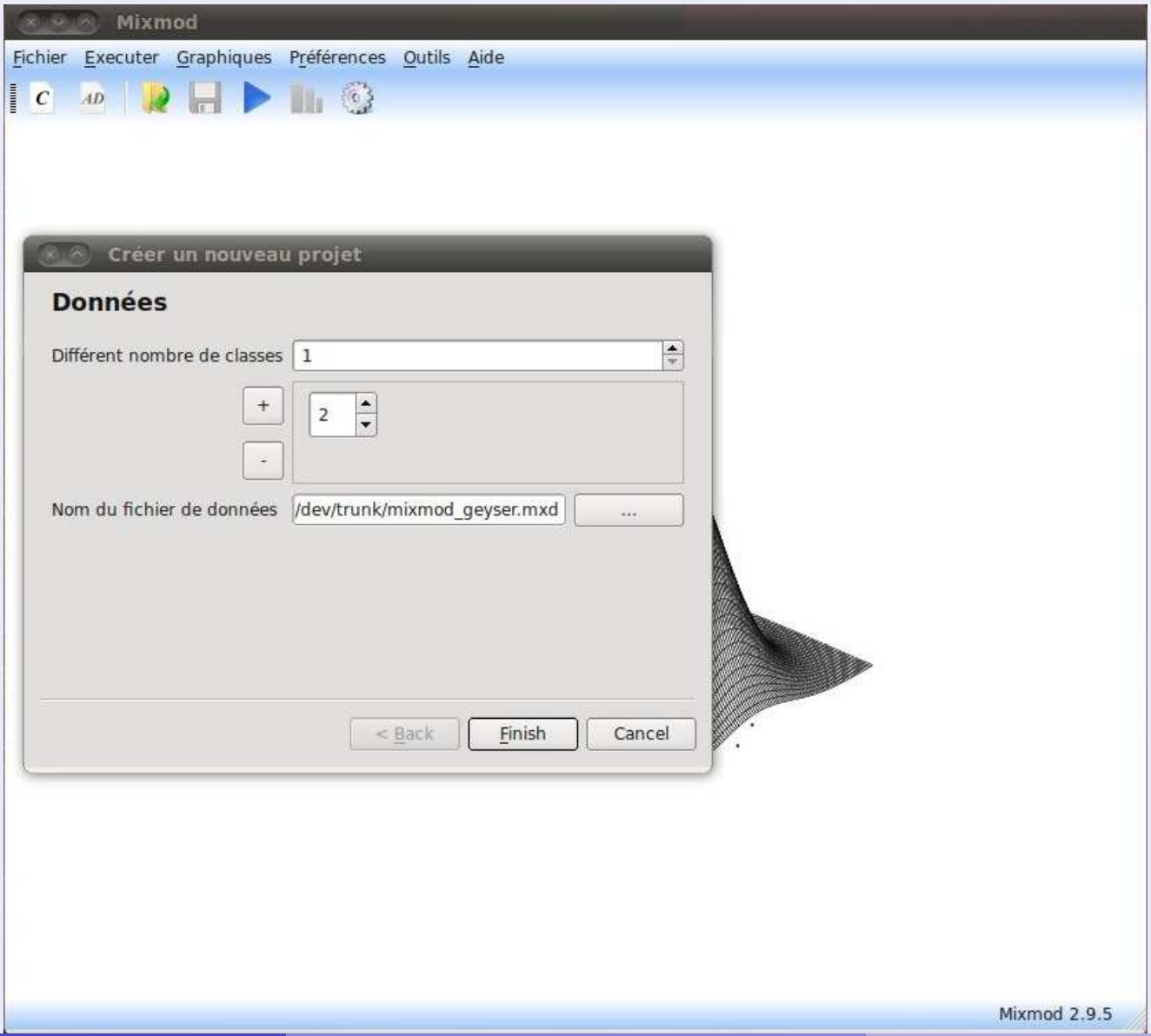
Nouveautés et perspectives

Une interface graphique : mixmodGUI



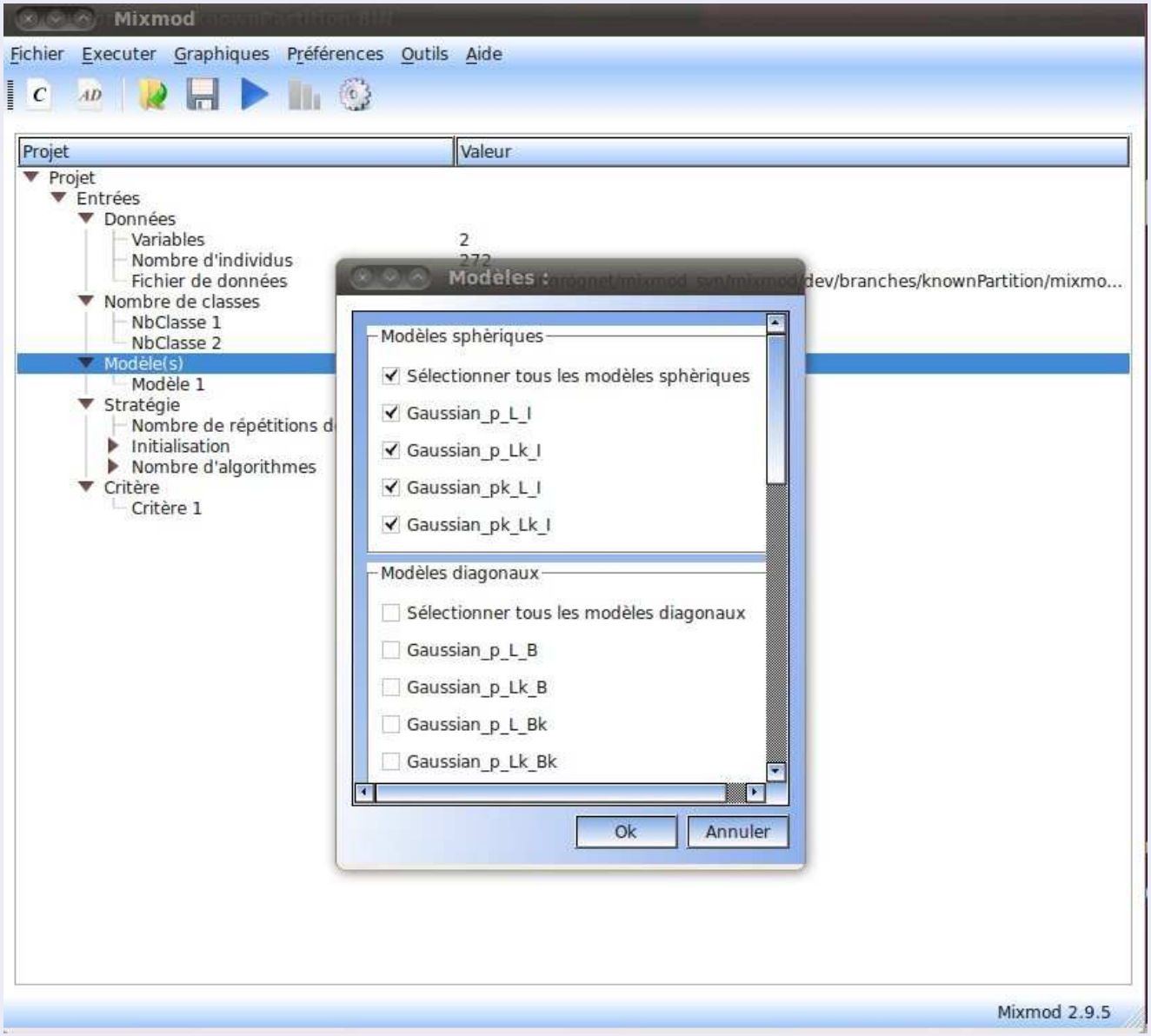
mixmodGUI

Création d'un nouveau Projet (classification)



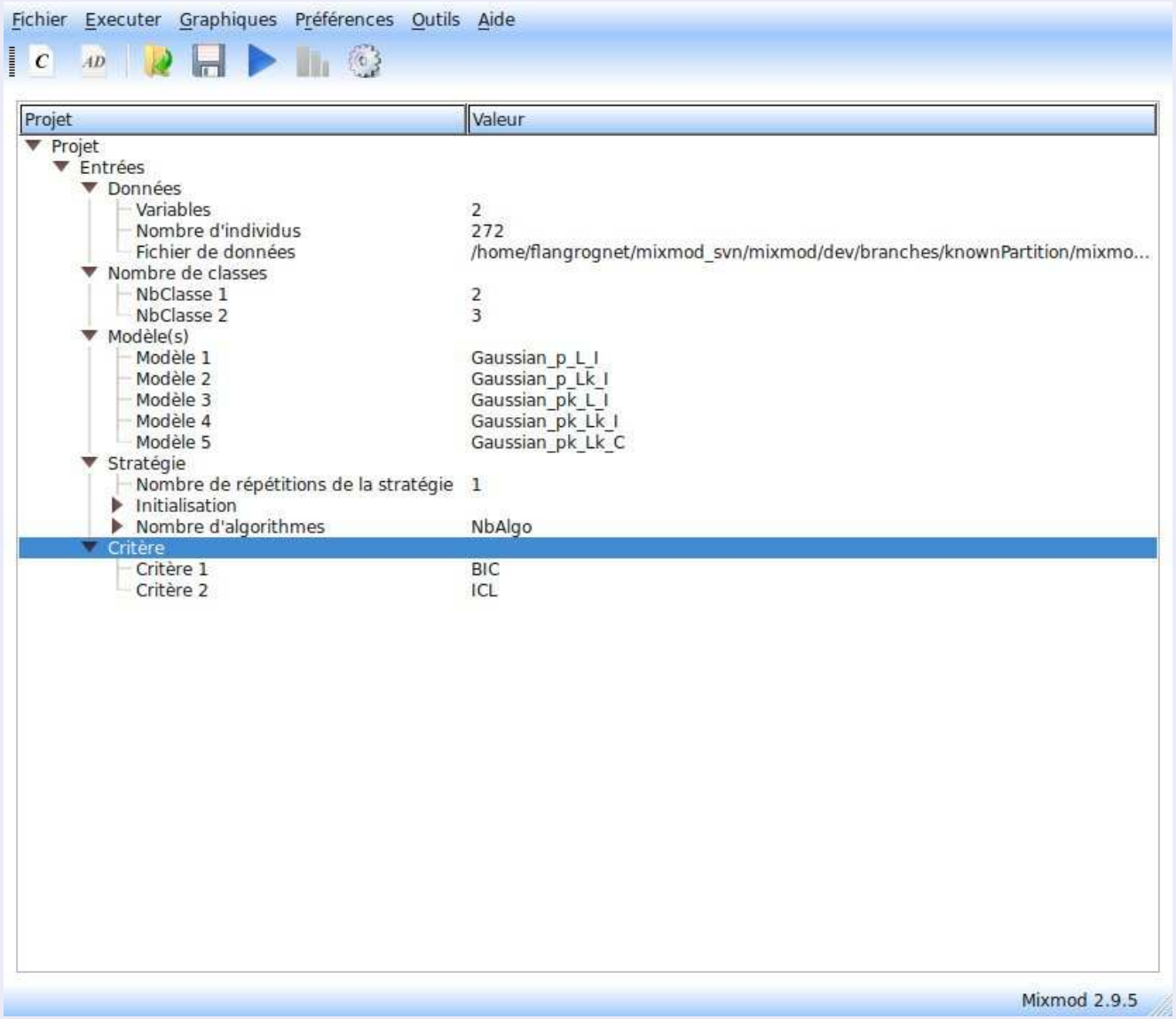
mixmodGUI

Choix de modèles

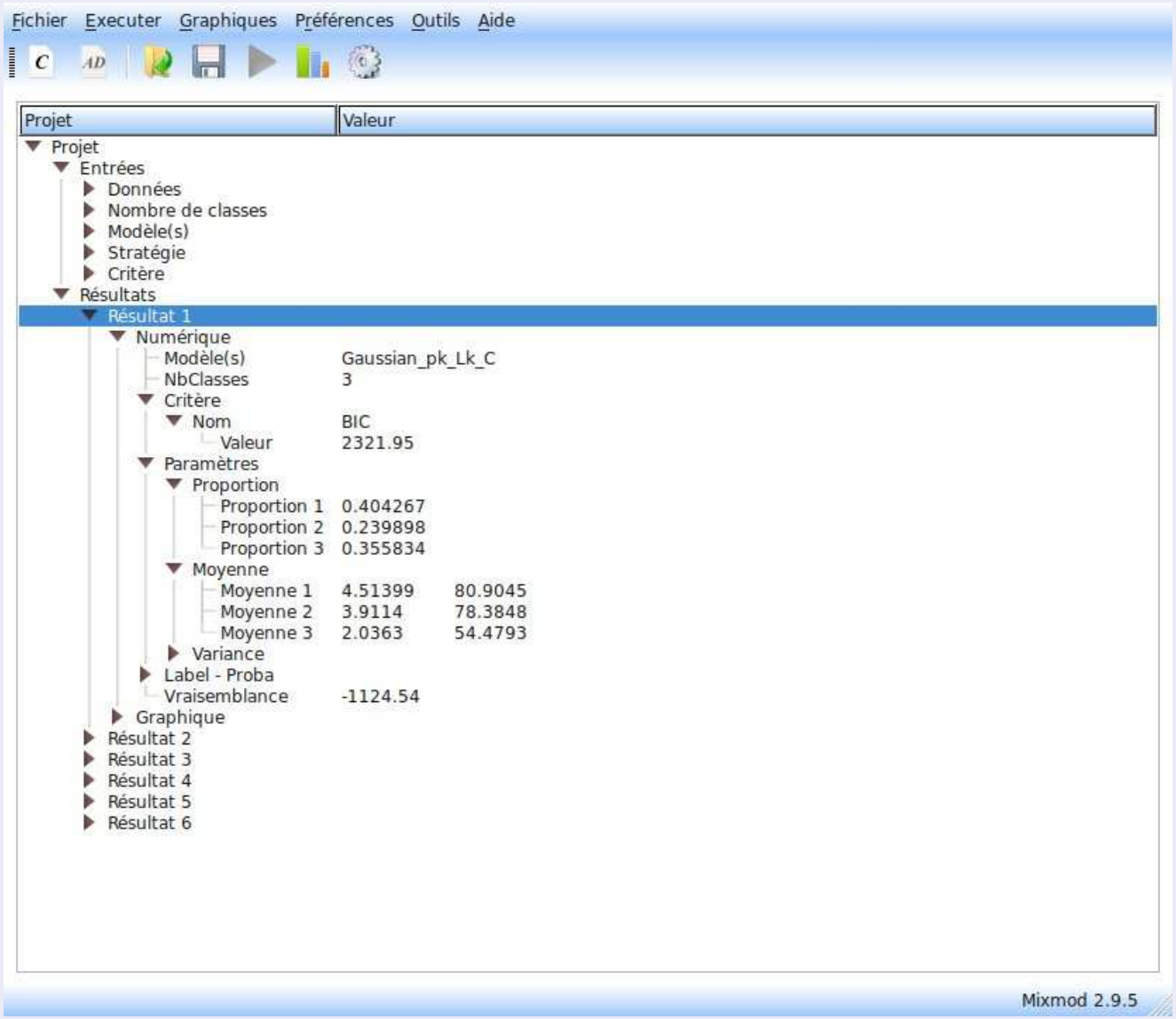


mixmodGUI

Vue des inputs

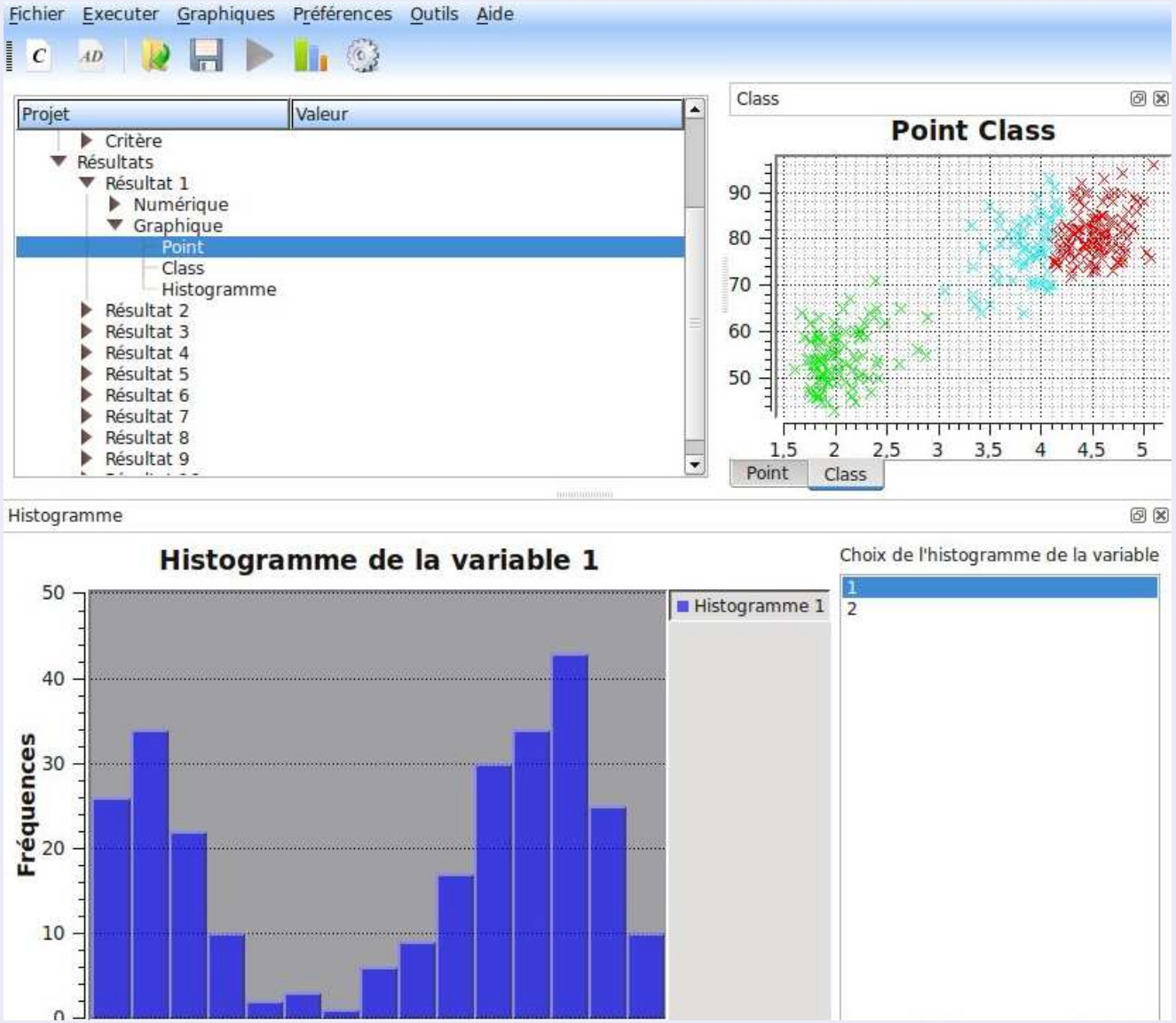


Résultats numériques



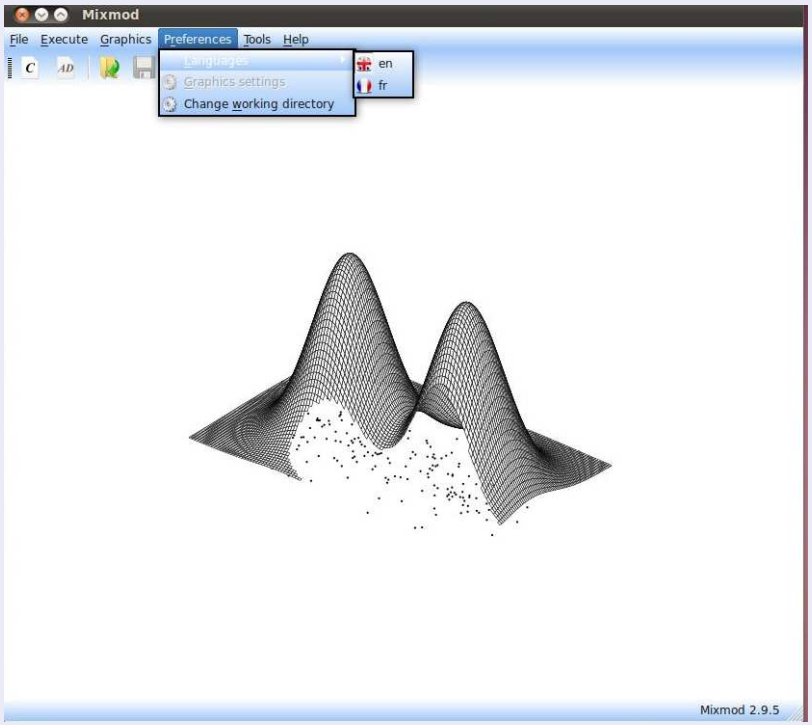
mixmodGUI

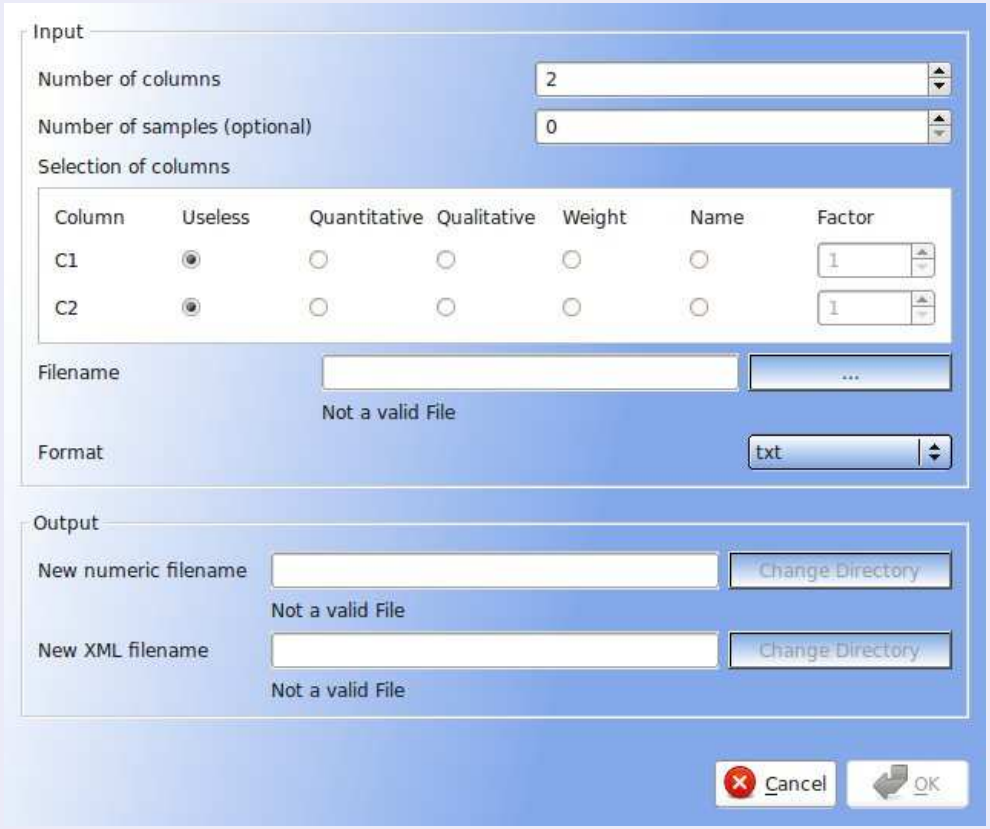
Graphiques



Réglages des préférences

Choix de la langue

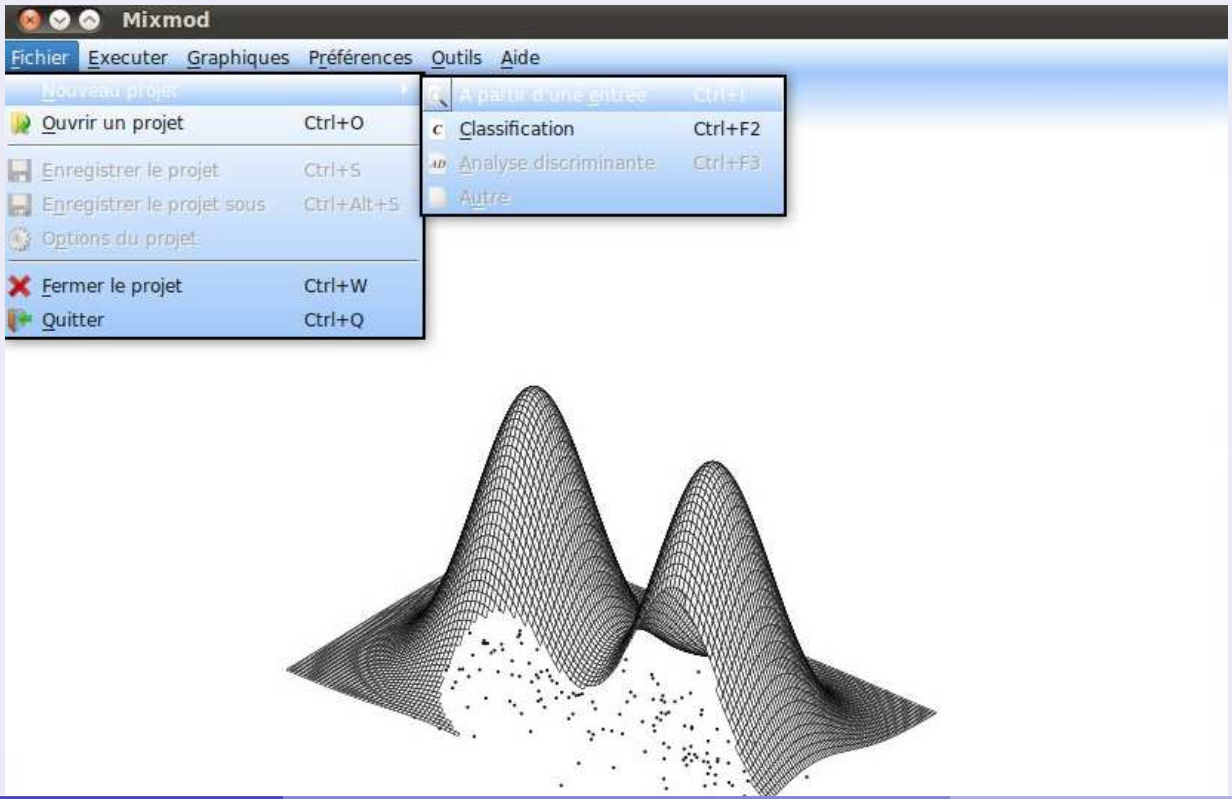




projets dans mixmodGUI

- Sauvegarde d'un projet (input/output)
- Création d'un nouveau projet à partir d'inputs enregistrées

Seuls les projets de type 'classification' (non supervisée) sont disponibles dans mixmodGUI 2.9.5



Nouveaux formats de fichiers : XML

Séparation entre les données et de leur description

Avantages :

- Souplesse
- Evolutivité
- Vérification syntaxique, sémantique
- Echanges facilités avec d'autres applications

Notes :

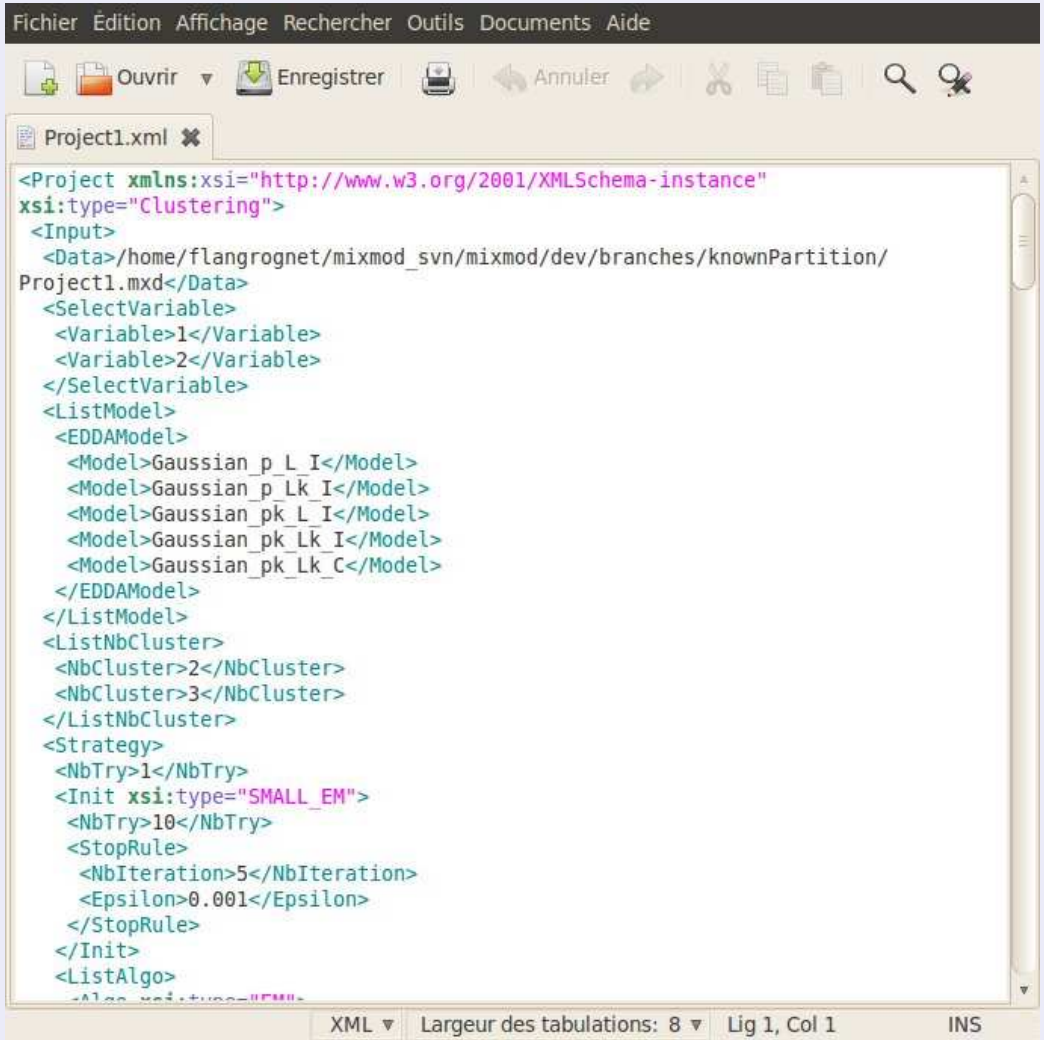
- L'usage de tels formats est transparent pour l'utilisateur de mixmodGUI
- L'utilisateur de Mixmod en 'mode expert' (ligne de commande) devra utiliser ces nouveaux formats. Un 'convertisseur' sera proposé.

Fichiers Mixmod au format XML

Quels fichiers ?

Fichier .mixmod

Description des inputs et outputs




```
<Project xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:type="Clustering">
 <Input>
  <Data>/home/flangrognet/mixmod_svn/mixmod/dev/branches/knownPartition/
Project1.mxd</Data>
  <SelectVariable>
   <Variable>1</Variable>
   <Variable>2</Variable>
  </SelectVariable>
  <ListModel>
   <EDDAModel>
    <Model>Gaussian_p_L_I</Model>
    <Model>Gaussian_p_Lk_I</Model>
    <Model>Gaussian_pk_L_I</Model>
    <Model>Gaussian_pk_Lk_I</Model>
    <Model>Gaussian_pk_Lk_C</Model>
   </EDDAModel>
  </ListModel>
  <ListNbCluster>
   <NbCluster>2</NbCluster>
   <NbCluster>3</NbCluster>
  </ListNbCluster>
  <Strategy>
   <NbTry>1</NbTry>
   <Init xsi:type="SMALL_EM">
    <NbTry>10</NbTry>
    <StopRule>
     <NbIteration>5</NbIteration>
     <Epsilon>0.001</Epsilon>
    </StopRule>
   </Init>
   <ListAlgo>
```

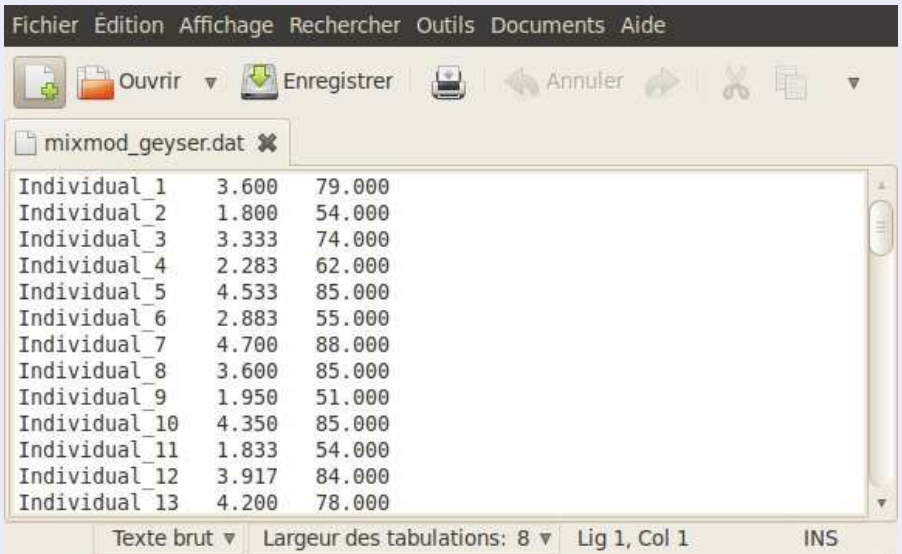
Fichiers Mixmod au format XML

Quels fichiers ?

Les données

Fichier .mxd
Description des données

Fichier .txt
Les Données

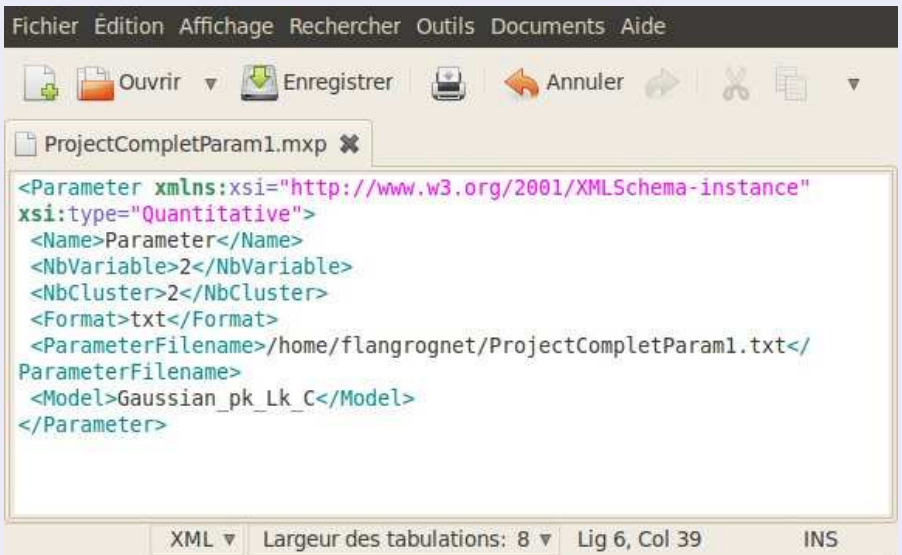


Fichiers Mixmod au format XML

Quels fichiers ?

Les paramètres

Fichier .mxp
Description des paramètres



Fichier .txt
Les paramètres

A partir de 2011 : plusieurs produits

Un produit par besoin

- mixmodGUI
- Packages (fonctions et noyau de calcul) pour
  - Scilab
  - Matlab
  - R
- Bibliothèque mixmod (C++)

Et aussi ...

- Communautés de développeurs/utilisateurs pour
  - Scilab
  - Matlab
  - R
- Encourager les contributions

PLAN

Roadmap : MixmodGUI (1)

Fonctionnalités statistiques

		Mixmod 2.9.5
Supervised classification		N
Unsupervised classification		
Algorithms	EM	Y
	CEM	Y
	SEM	Y
Models	28 gaussian models	Y
	10 binary models	Y
Initialisation	RANDOM	Y
	SMALL_EM	Y
	CEM	Y
	SEM	Y
	PARAMETER	N
	PARTITION	N
Criteria	BIC	Y
	ICL	Y
	NEC	Y
Others	Known Partition	In Progress
	Weight	N
	Variable selection	N
	Individual Selection	N

Roadmap : MixmodGUI (2)

GUI

		Mixmod 2.9.5
Open/Close/Save	New project GUI	Y
	New project from Input (XML)	Y
	Open project (XML)	Y
	Close project	Y
	Save project (XML)	Y
XML format (input, output, data)		Y
Input tree modification	Model	Y
	Strategy :	
	Number of strategy repeats	Y
	Initialisation (repeats, type, stopRule)	In Progress
	Algorithms	Y
	Criterion	Y
	Partition	In Progress
Run Computation		Y
Output	Output tree sort (by criterion)	Y
	Delete output	Y
	Numeric values	Y
Others	Data converter (to XML Mixmod format)	Y
	Set working directory	Y
	Error treatment	In Progress
	Multi langages	Y

Roadmap : MixmodGUI (3)

Visualisation

		Mixmod 2.9.5
Quantitative data		In Progress
1D	Histogram	Y
	Density component	Y
	Density mixture	Y
2D	Point	Y
	Class	Y
Qualitative data		In Progress
1D	Histogram	Y

Roadmap

- de mixmodGUI 2.9.5 à mixmodGUI 3.0 : convergence vers une version stable
  - Phase de tests
  - Amélioration de l'ergonomie (retour des utilisateurs)
  - Amélioration des graphiques
- mixmodGUI 3.0 : Classification non supervisée
- mixmodGUI 3.1 : Classification non supervisée et classification supervisée

FIN

Merci de votre attention

http ://www.mixmod.org/