

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

- 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

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. Langrognnet** (Labo. de math. de Besançon - Université Franche-Comté / CNRS)

L'équipe de développement

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

- 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

- 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

Fonctionnalités

Problématiques traitées

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

Cadre de travail - Type de données traitées

Modèles de mélanges

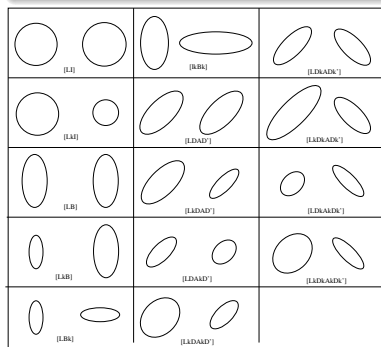
- Gaussiens (données quantitatives)
- Multinomiaux (données qualitatives)
- Modèles spécifiques pour les données en grande dimension

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** (Stochastic 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

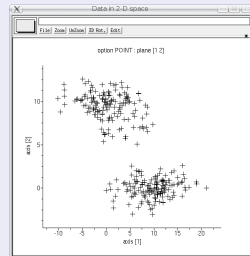
- **nombre de classes** : 3
- **modèle gaussien** : proportions libres, variances de même orientation et volumes libres (Modèle $p_k \lambda_k C$)

Objectifs

- **classer** les individus
- **caractériser** les classes

Données

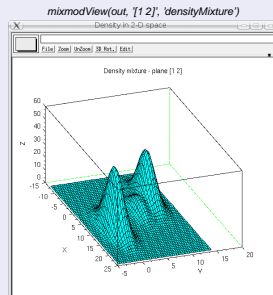
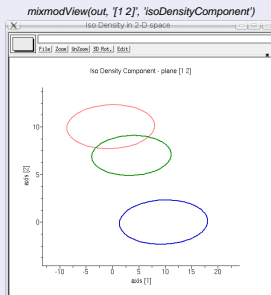
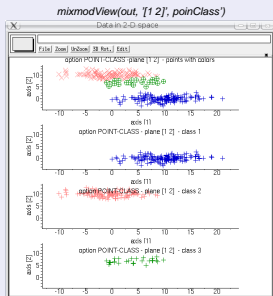
300 individus en dimension 2



Résultats

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

```
out = mixmod(data, 3);
```



$$p_1 = 0.5$$

$$p_2 = 0.4$$

$$p_3 = 0.1$$

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 converge vers un **maximum local**

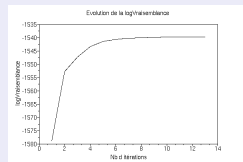
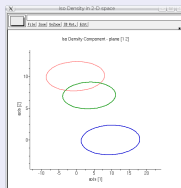
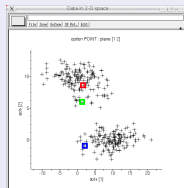
Outils disponibles dans Mixmod

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

Algorithme EM et initialisation

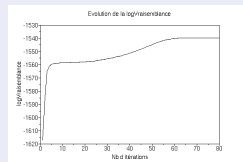
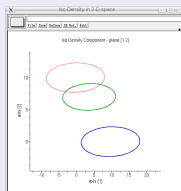
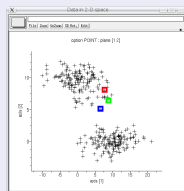
Vitesse de convergence

● Initialisation n°1



LL = -1539 en 10 itérations

● Initialisation n°2

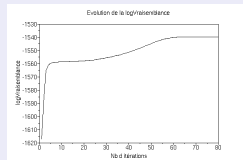
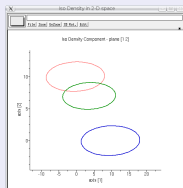
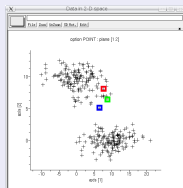


LL = -1539 en 80 itérations

Algorithme EM et initialisation

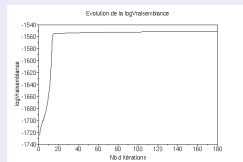
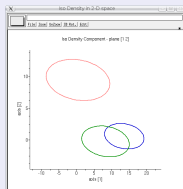
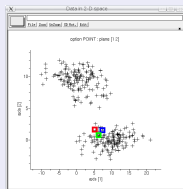
Convergence vers des maxima différents

● Initialisation n°2



LL = -1539

● Initialisation n°3



LL = -1551

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

6 initialisations

● Information a priori

- ▶ 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 initialiser 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

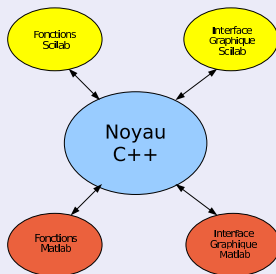
- 1 Comment est développé Mixmod ?
- 2 Principales fonctionnalités
- 3 Comment utiliser Mixmod ?
 - La situation actuelle (avant Mixmod 3.0)
 - Les nouveautés et les perspectives
 - Roadmap

PLAN

- 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

Architecture, utilisations et utilisateurs

Architecture

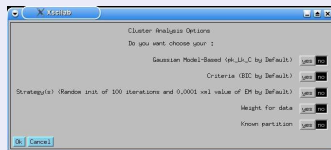
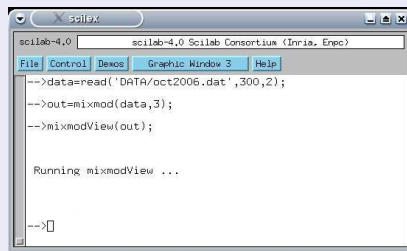
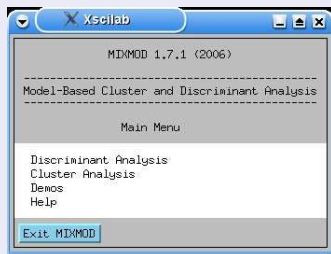


Utilisations

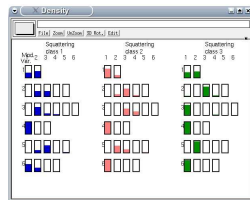
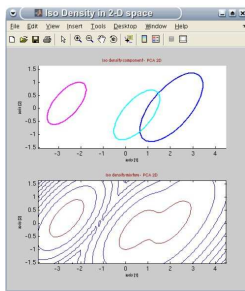
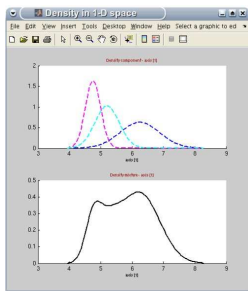
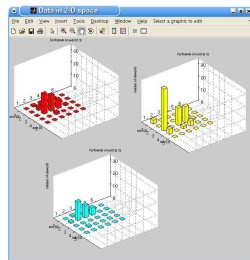
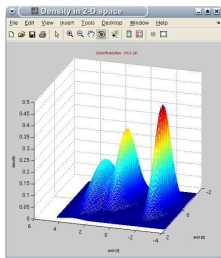
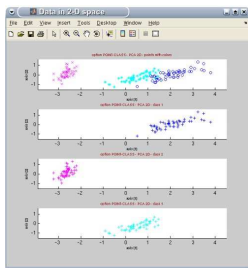
- Interfaces graphiques pour **Scilab** et **Matlab**
- Fonctions pour **Scilab** et **Matlab**
- **Mode expert** et **Bibliothèque** de calcul (C++)



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



```

Echier Edition Préférences Contrôle Applications ?
---
-->
-->
-->
-->
-->geyser=read('DATA/geyser.dat',272,2);
-->out = mixmod(geyser,2);
ATTENTION: La fonction getf est obsolète.
ATTENTION: Utilisez exec à la place.
ATTENTION: Cette fonction sera supprimée de façon permanente dans Scilab 5.3.
ATTENTION: La fonction getf est obsolète.
ATTENTION: Utilisez exec à la place.
ATTENTION: Cette fonction sera supprimée de façon permanente dans Scilab 5.3.
-->

```

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

Nouveautés et perspectives

Pour les logiciels tiers (Scilab, Matlab, R)

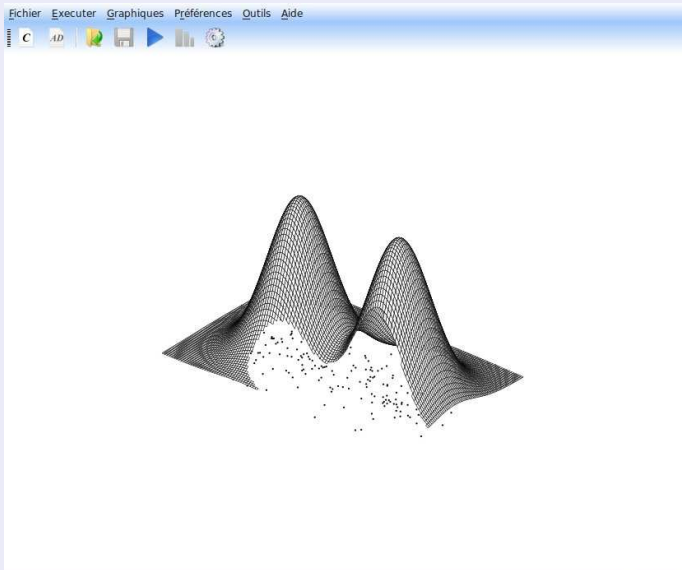
- Des **fonctions ré-écrites** pour Scilab et Matlab et des **fonctions pour R**
 - ▶ Une fonction par fonctionnalité (mixmod_cluster, mixmod_DA)
 - ▶ Une utilisation simplifiée
- Les **fonctions graphiques** (visualisation) ne seront **plus maintenues** par l'équipe Mixmod

Création de communautés :

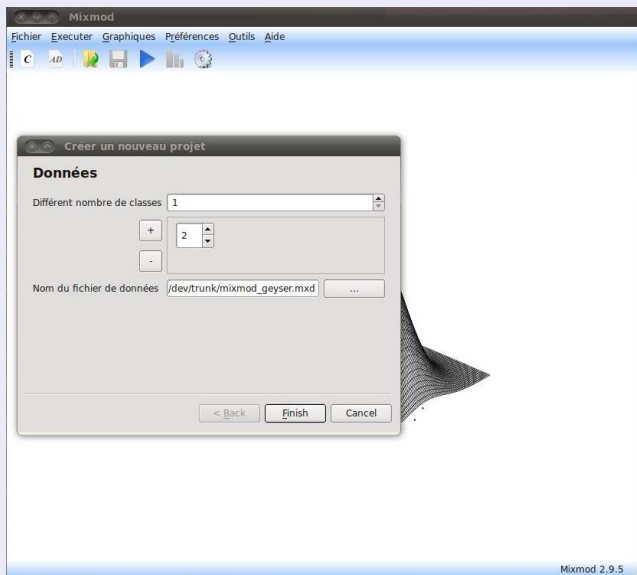
 - ▶ Mixmod/Scilab
 - ▶ Mixmod/Matlab
 - ▶ Mixmod/R
- **Plus d'interface graphique** pour Scilab/Matlab

Nouveautés et perspectives

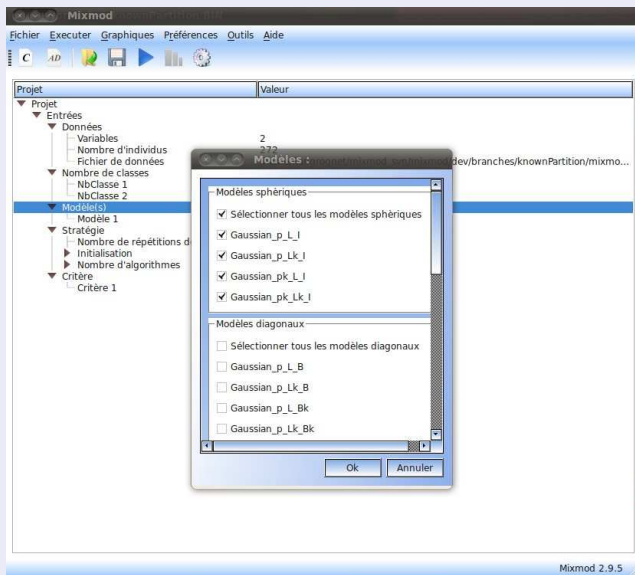
Une interface graphique : mixmodGUI



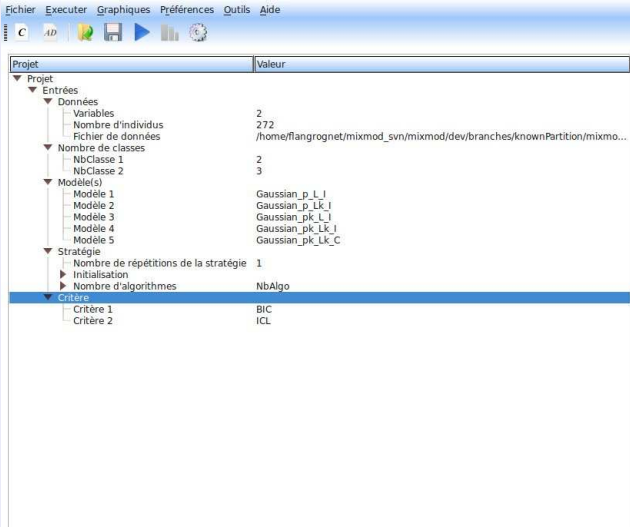
Création d'un nouveau Projet (classification)



Choix de modèles



Vue des inputs



The screenshot shows the Mixmod GUI interface with the 'Vue des inputs' window open. The window title is 'Fichier Executer Graphiques Préférences Outils Aide'. The main content area is a table with two columns: 'Projet' and 'Valeur'. The table is organized into a tree structure under the 'Projet' column. The 'Critère' section is highlighted in blue.

Projet	Valeur
▼ Projet	
▼ Entrées	
▼ Données	
Variables	2
Nombre d'individus	272
Fichier de données	/home/flangrognnet/mixmod_svn/mixmod/dev/branches/knownPartition/mixmo...
▼ Nombre de classes	
NbClasse 1	2
NbClasse 2	3
▼ Modèle(s)	
Modèle 1	Gaussian_p_L_I
Modèle 2	Gaussian_p_Lk_I
Modèle 3	Gaussian_pk_L_I
Modèle 4	Gaussian_pk_Lk_I
Modèle 5	Gaussian_pk_Lk_C
▼ Stratégie	
Nombre de répétitions de la stratégie	1
Initialisation	
Nombre d'algorithmes	NbAlgo
▼ Critère	
Critère 1	BIC
Critère 2	ICL

Mixmod 2.9.5

Résultats numériques

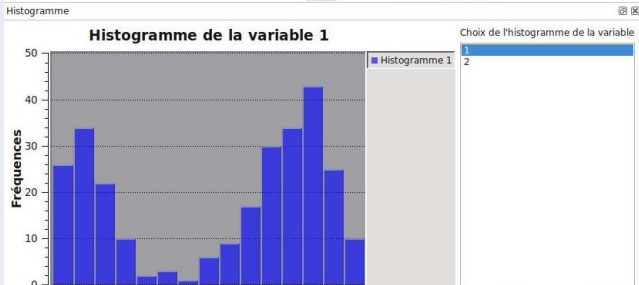
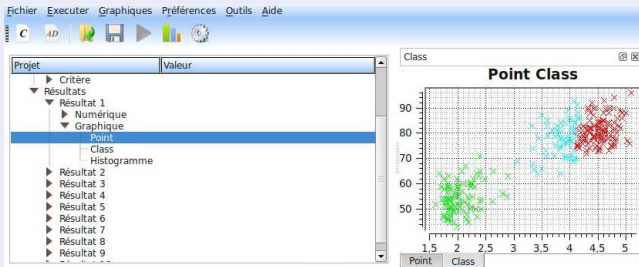
Fichier Exécuter Graphiques Préférences Outils Aide

Projet Valeur

- Projet
 - Entrées
 - Données
 - Nombre de classes
 - Modèle(s)
 - Stratégie
 - Critère
 - Résultats
 - Résultat 1
 - Numérique
 - Modèle(s) Gaussian_pk_Lk_C
 - NbClasses 3
 - Critère
 - Nom
 - Valeur BIC 2321.95
 - Paramètres
 - Proportion
 - Proportion 1 0.404267
 - Proportion 2 0.239898
 - Proportion 3 0.355834
 - Moyenne
 - Moyenne 1 4.51399 80.9045
 - Moyenne 2 3.9114 78.3848
 - Moyenne 3 2.0363 54.4793
 - Variance
 - Label - Proba
 - Vraisemblance -1124.54
 - Graphique
 - Résultat 2
 - Résultat 3
 - Résultat 4
 - Résultat 5
 - Résultat 6

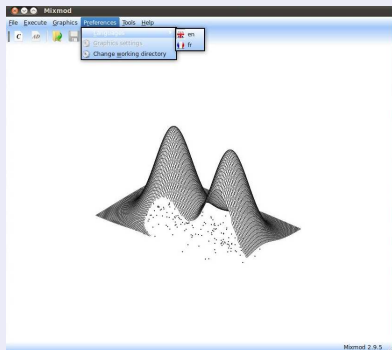
Mixmod 2.9.5

Graphiques



Réglages des préférences

Choix de la langue



Input

Number of columns

Number of samples (optional)

Selection of columns

Column	Useless	Quantitative	Qualitative	Weight	Name	Factor
C1	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text" value="1"/>
C2	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text" value="1"/>

Filename
Not a valid File

Format

Output

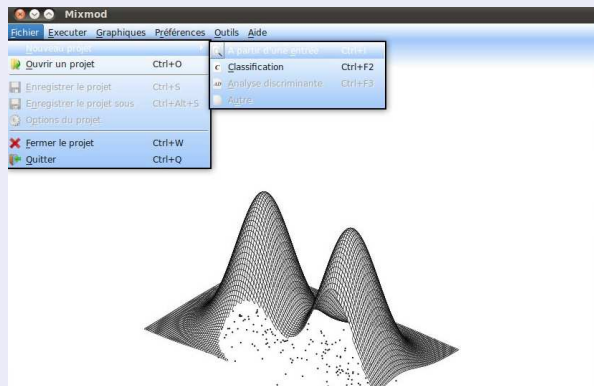
New numeric filename
Not a valid File

New XML filename
Not a valid File

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 :

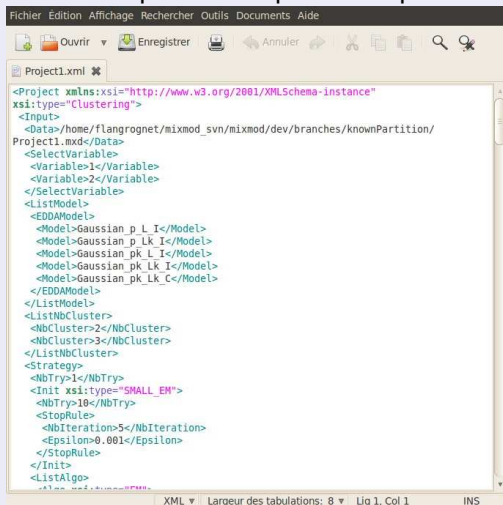
- L'usage de tels formats est transparent pour l'utilisateur de mixmodGUI
- L'utilisateur de Mixmod en 'mode expert' (ligne de commande) devra utiliser ces nouveaux formats. Un 'convertisseur' sera proposé.

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/flangrognnet/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>
      <Algo>refinitive</Algo>
    </ListAlgo>
  </Input>


```

Fichiers Mixmod au format XML

Quels fichiers ?

Les données

Fichier .mxd Description des données



The screenshot shows the Mixmod interface with the file 'mixmod_geyser.mxd' open. The XML content is as follows:

```
<Data xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">  
<NbSample>272</NbSample>  
<Format>txt</Format>  
<DataFilename>/home/flangrognnet/mixmod_geyser.dat</DataFilename>  
<ListColumn>  
<Column Num="1" xsi:type="Individual"/>  
<Column Num="2" xsi:type="Quantitative"/>  
<Column Num="3" xsi:type="Quantitative"/>  
</ListColumn>  
</Data>
```

The status bar at the bottom indicates 'XML', 'Largeur des tabulations: 8', 'Lig 4, Col 34', and 'INS'.

Fichier .txt Les Données



The screenshot shows the Mixmod interface with the file 'mixmod_geyser.dat' open. The raw text data is displayed as follows:

Individual 1	3.600	79.000
Individual 2	1.800	54.000
Individual 3	3.333	74.000
Individual 4	2.283	62.000
Individual 5	4.533	85.000
Individual 6	2.883	55.000
Individual 7	4.700	88.000
Individual 8	3.600	85.000
Individual 9	1.950	51.000
Individual 10	4.350	85.000
Individual 11	1.833	54.000
Individual 12	3.917	84.000
Individual 13	4.200	78.000

The status bar at the bottom indicates 'Texte brut', 'Largeur des tabulations: 8', 'Lig 1, Col 1', and 'INS'.

Fichiers Mixmod au format XML

Quels fichiers ?

Les paramètres

Fichier .mxp

Description des paramètres



```
<Parameter xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:type="Quantitative">
  <Name>Parameter</Name>
  <NbVariable>2</NbVariable>
  <NbCluster>2</NbCluster>
  <Format>txt</Format>
  <ParameterFilename>/home/flangrognnet/ProjectCompletParam1.txt</
ParameterFilename>
  <Model>Gaussian_pk_Lk_C</Model>
</Parameter>
```

Fichier .txt

Les paramètres



```
0.642875
4.29222 79.9964
0.145342 0.830098
0.830098 40.9021

0.357125
2.03969 54.5171
0.0983696 0.561822
0.561822 27.6831
```

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

- 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

Roadmap : MixmodGUI (1)

Fonctionnalités statistiques

Mixmod 2.9.5

		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
	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

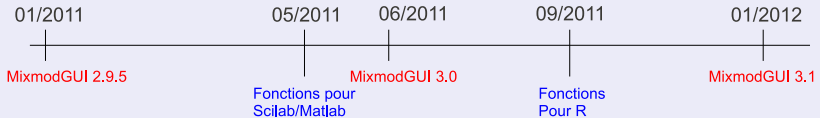
Open/Close/Save	New project GUI	Y
	New project from Input (XML)	Y
	Open project (XML)	Y
	Close project	Y
	Save project (XML)	Y
XML format (input, output, data)		Y
Input tree modification	Model	Y
	Strategy :	
	Number of strategy repeats	Y
	Initialisation (repeats, type, stopRule)	In Progress
	Algorithms	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 languages	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/>