Le logiciel MIXMOD d'analyse de mélange pour la classification et l'analyse discriminante

Christophe Biernacki[1], Gilles Celeux[2], Anwuli Echenim[23], Gérard Govaert[4], Florent Langrognet[3]

[1] UMR 8524, CNRS & Université de Lille 1, 59655 Villeneuve d'Ascq, France
[2] NRIA Futurs, 91405 Orsay, France
[3] UMR 6623, CNRS & Université de Franche-Comté, 25030 Besançon, France
[4] UMR 6599, CNRS & UTC, 60205 Compiègne, France

E-mails : christophe.biernacki@math.univ-lille1.fr, gilles.celeux@inria.fr, anwuli.echenim@univ-fcomte.fr, gerard.govaert@utc.fr, florent.langrognet@univ-fcomte.fr

Résumé Le logiciel MIXMOD est dévolu à l'analyse de mélanges de lois de probabilité sur des données multidimensionnelles dans un but d'estimation de densité, de classification ou d'analyse discriminante. Il propose un choix important d'algorithmes pour estimer les paramètres d'un mélange (EM, Classification EM, *Stochastic* EM). Il est possible de combiner ces algorithmes de multiples façons pour obtenir un maximum local pertinent de la vraisemblance ou de la vraisemblance complétée d'un modèle. Pour des variables quantitatives, MIXMOD utilise des mélanges de lois normales multidimensionnelles. Il propose ainsi quatorze modèles gaussiens différents selon des hypothèses faites sur les éléments spectraux des matrices de variance des composants. Pour des variables qualitatives, MIXMOD utilise des mélanges de lois multinomiales multidimensionnelles sous une hypothèse d'indépendance conditionnelle des variables sachant le composant du mélange. Grâce à une reparamétrisation des probabilités multinomiales, il propose cinq modélisations différentes. Par ailleurs, différents critères d'information sont proposés pour choisir un modèle parcimonieux et permettent notamment de choisir un nombre de composants pertinents. L'emploi de l'un ou l'autre de ces critères dépend de l'objectif poursuivi (estimation de densité, classification supervisée ou non). Écrit en C++, MIXMOD possède des interfaces avec SCILAB et MATLAB. Le logiciel, sa documentation statistique et son guide d'utilisation sont disponibles à l'adresse suivante :
`http ://www-math.univ-fcomte.fr/mixmod/index.php`

Mots-clés : modèles gaussiens, modèles multinomiaux, algorithmes de type EM, sélection de modèles.

1 Introduction

Par leur flexibilité, les mélanges finis de distributions de probabilité sont devenus un outil populaire pour modéliser une grande variété de phénomènes aléatoires. En particulier, ils constituent un outil de choix pour l'estimation de densité, la classification et l'analyse discriminante. Les modèles de mélange sont alors utilisés dans un nombre croissant de disciplines comme l'astronomie, la biologie, la génétique, l'économie, les sciences de l'ingénieur et le marketing. On a vu ainsi se développer plusieurs logiciels dédiés à ce

modèle. MIXMOD est l'un de ces logiciels et il a été essentiellement conçu pour traiter des problèmes de classification supervisée ou non. Cet article vise à décrire les caractéristiques statistiques de ce logiciel d'analyse de données multidimensionnelles.

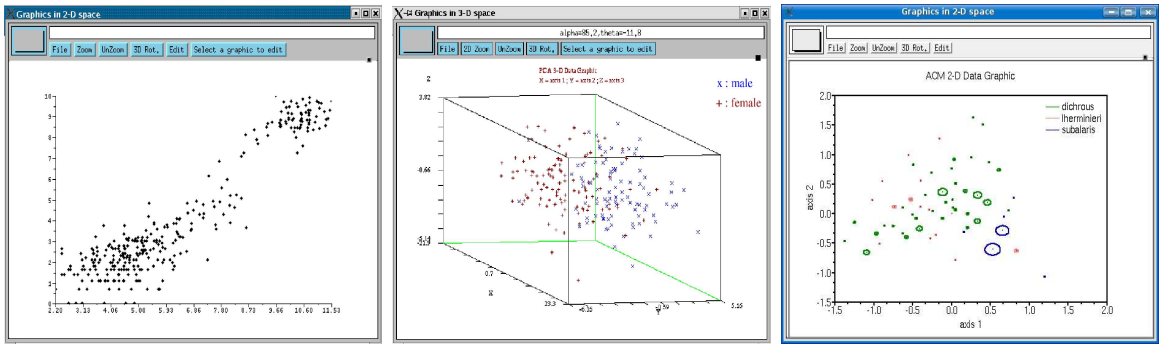
MIXMOD est un logiciel libre, disponible sous licence GPL pour les systèmes d'exploitation Linux, Unix et Windows. Le noyau du logiciel, écrit en C++, est disponible en tant que bibliothèque et exécutable mais il est également accessible via des interfaces avec les logiciels MATLAB et SCILAB. Il a été développé conjointement par l'Inria, le laboratoire de mathématiques de l'université de Besançon, le laboratoire Heudiasyc de l'UTC Compiègne et le laboratoire de mathématiques de l'université Lille 1.

Dans sa version actuelle, MIXMOD propose des modèles de mélanges gaussiens multidimensionnels ainsi que des mélanges multinomiaux multidimensionnels (modèle dit des classes latentes). Les principales caractéristiques de cette version sont les suivantes :

- trois niveaux d'utilisation du débutant à l'expert ;
- quatorze modèles de mélanges gaussiens tirés de paramétrisations différentes des matrices de variance des composants ;
- cinq modèles de mélanges multinomiaux tirés de paramétrisations différentes des probabilités multinomiales ;
- l'estimation des paramètres du mélange par l'algorithme EM ou des variantes de cet algorithme, permettant différentes stratégies d'initialisation ;
- la mise à disposition de nombreux critères pour sélectionner un modèle fiable en fonction de l'objectif d'estimation de densité, de classification ou de classement ;
- de nombreux graphiques 1D, 2D et 3D dont des densités, des isodensités, des descriptions d'un classifieur dans un espace factoriel.

Cet article ne vise pas à remplacer le guide d'utilisation ni la documentation statistique de MIXMOD que l'on peut trouver sur le site web. Il a pour but de fournir une vue synthétique des caractéristiques de MIXMOD en associant une présentation statistique courte d'exemples d'utilisation.

Le premier jeu de données introduit à cette fin concerne la classification non supervisée. La figure 1 (a) montre la logpopulation *versus* la logdensité (par habitants/km^2) pour 312 villes de trois départements français (Biernacki et al., 2000), la *Seine-Saint-Denis* et les *Hauts de Seine*, dans l'agglomération parisienne et le département rural de *Haute-Corse*.



(a) (b) (c)

FIG. 1 – Données sélectionnées : (a) Les départements français pour la classification, (b) les oiseaux *borealis* pour l'analyse discriminante dans la cadre continu et (c) les oiseaux *puffins* pour l'analyse discriminante dans le cadre qualitatif.

Le deuxième jeu de données a pour but d'illustrer les caractéristiques de MIXMOD dans un contexte d'analyse discriminante. Il concerne 204 oiseaux de la sous-espèce *borealis* de la famille *Procellaridae* (Pétrel) dont cinq mesures morphologiques sont disponibles (Biernacki et al., 2002) : culmen (longueur du cou), tarsus, taille des ailes et de la queue et largeur du cou. La figure 1 (b) représente les mâles (55%) et les femelles (45%) dans le premier espace 3D de l'analyse en composantes principales.

L'objectif du troisième jeu de données est aussi d'illustrer l'analyse discriminante mais cette fois sur des données qualitatives. Il s'agit de 132 puffins issus de trois sous-espèces différentes (*dichrous*, *lherminieri* et *subalaris*) sur lesquels six mesures morphologiques ont été relevées : sexe, sourcil, collier, zébrures, sous-caudales et liseret. La figure 1(c) représente les individus de chaque sous-espèce dans le premier plan de l'analyse des correspondances multiples.

2 Caractéristiques techniques de MIXMOD

Le développement du logiciel a commencé en 2001 et la dernière version de MIXMOD (MIXMOD 2.0) est composé de 50 classes C++ (25000 lignes), et 20000 lignes de code SCILAB et MATLAB. Sur le site web dédié à MIXMOD

(`http ://www-math.univ-fcomte.fr/mixmod/index.php`), on accède à l'ensemble des ressources : téléchargement, documentations (userguide, statistique et logicielle), forum de discussion, nouvelles, ...

2.1 Modes opératoires de MIXMOD

Le logiciel MIXMOD peut être utilisé de trois façons.

- MIXMOD par l'intermédiaire d'une interface graphique.
  La fonction `mixmodGraph`, disponible dans SCILAB et dans MATLAB, est le moyen le plus simple d'accéder aux principales fonctionnalités de MIXMOD (même si certaines d'entre elles ne sont pas disponibles via cette interface).
- MIXMOD en tant que fonction SCILAB ou MATLAB.
  Grâce à une syntaxe simple (deux entrées seulement sont obligatoires), cette fonction permet de résoudre l'ensemble des problématiques de classification supervisée ou non. Associée à d'autres fonctions écrites pour SCILAB et MATLAB, comme la fonction `mixmodView` permettant de visualiser les résultats, elle représente un outil de choix alliant performance et convivialité.
- MIXMOD en tant que bibliothèque de calcul ou exécutable.
  Les fonctionnalités de MIXMOD sont disponibles par l'intermédiaire d'une bibliothèque C++ que l'utilisateur peut intégrer à tout programme. De plus, celui-ci peut également lancer MIXMOD en ligne de commande (sous Linux, Unix ou Windows) après avoir défini un fichier d'entrée. Les fichiers résultats pourront être alors sauvegardés ou réutilisés par une autre application.

Dans cet article, les exemples sont présentés avec la fonction `mixmod` dans l'environnement SCILAB.

2.2 Représentation des données dans MIXMOD

MIXMOD accepte trois structures de données complémentaires selon les données disponibles :

- représentation standard : chaque individu est représenté par une ligne et chaque variable par une colonne ;
- partition : chaque ligne est le vecteur indicateur d'appartenance d'un individu aux différentes classes. La coordonnée j est `1` si l'individu appartient à la classe j, et `0` sinon. Une ligne de "0" indique un individu de classe inconnue ;
- poids : chaque ligne donne le poids d'un individu.

2.3 Performances de MIXMOD (vitesse d'exécution)

Au-delà des fonctionnalités, l'objectif de MIXMOD est d'être un outil de choix pour les gros jeux de données. À ce titre, un effort particulier et continu est engagé pour atteindre les meilleures performances possibles. Le choix du langage de programmation du noyau MIXMOD (C++) a évidemment été fait en ce sens. Notons à titre d'illustration que MIXMOD 2.0 est approximativement 10 fois plus rapide que la première version.

3 Les modèles statistiques

Nous décrivons maintenant les modèles de mélange disponibles dans MIXMOD. Nous présentons d'abord les modèles de mélange gaussien pour le traitement de données quantitatives, puis le modèle des classes latentes pour le traitement des données qualitatives.

3.1 Quatorze modèles de mélange gaussien

3.1.1 Paramétrisation spectrale des matrices de variance

Dans MIXMOD, les observations continues $\in \mathbb{R}^d$, issues de d variables quantitatives sont supposées provenir d'un mélange de densité :

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \varphi(\boldsymbol{x};\boldsymbol{\mu}_k,\Sigma_k) \tag{1}$$

où $p_k \geq 0$ pour $k = 1,\ldots,K$ avec $\sum_{k=1}^{K} p_k = 1$ et représentent les proportions du mélange, $\varphi(\boldsymbol{x};\boldsymbol{\mu},\Sigma)$ est la densité d'une distribution gaussienne multivariée de moyenne $\boldsymbol{\mu}$ et de matrice de variance Σ, et $\boldsymbol{\theta} = (p_1,\ldots,p_K,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K,\Sigma_1,\ldots,\Sigma_K)$ représente le vecteur des paramètres à estimer.

Pour ce modèle, la densité du k^{e} composant est la densité gaussienne

$$\varphi(\boldsymbol{x};\boldsymbol{\mu}_k,\Sigma_k) = (2\pi)^{-d/2}|\Sigma_k|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)'\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}. \tag{2}$$

Cette densité gaussienne modélise une classe ellipsoïdale de centre $\boldsymbol{\mu}_k$ et dont les caractéristiques géométriques peuvent être associées à la décomposition spectrale de la matrice de variance Σ_k.

Suivant Banfield and Raftery (1993) et Celeux and Govaert (1995), chaque matrice de variance des composants du mélange peut s'écrire :

$$\Sigma_k = \lambda_k D_k A_k D_k' \tag{3}$$

avec $\lambda_k = |\Sigma_k|^{1/d}$, D_k étant la matrice des vecteurs propres de Σ_k et A_k étant une matrice diagonale, telle que $|A_k| = 1$, dont la diagonale est constituée des valeurs propres normalisées de Σ_k rangées en ordre décroissant. Le paramètre λ_k caractérise le *volume* de la k^{e} classe, D_k son *orientation* et A_k sa *forme*. En permettant à ces paramètres de varier ou non entre les classes, on obtient des modèles plus ou moins parcimonieux, faciles à interpréter et utiles pour appréhender des situations de classification pertinentes que l'on soit dans un cadre supervisé ou non. Ainsi supposer que les paramètres λ_k, D_k et A_k dépendent ou non des classes conduit à huit modèles généraux. Par exemple, des volumes différents, des formes et des orientations égales s'obtiennent en supposant que $A_k = A$ (A inconnu) et $D_k = D$ (D inconnu) pour chaque composant du mélange. Ce modèle est noté $[\lambda_k DAD']$. Avec cette convention, $[\lambda D_k A_k D_k']$ indique un modèle dont les composants ont des volumes égaux, des formes et des orientations différentes. D'autres familles d'intérêt se restreignent à des matrices de variance Σ_k diagonales. Par la paramétrisation (3), cela signifie que les matrices d'orientation D_k sont des matrices de permutation. Dans cet article, ces matrices de variance diagonales sont par convention notées $\Sigma_k = \lambda_k B_k$, B_k étant une matrice diagonale avec $|B_k| = 1$. Cette paramétrisation particulière conduit à quatre modèles : $[\lambda B]$, $[\lambda_k B]$, $[\lambda B_k]$ et $[\lambda_k B_k]$. La dernière famille des modèles suppose des formes sphériques, c'est-à-dire telle que $A_k = I$, I désignant la matrice identité. Dans ce cas, deux modèles parcimonieux peuvent être considérés : $[\lambda I]$ et $[\lambda_k I]$. Au total, quatorze modèles sont ainsi obtenus.

Remarquons que, dans la suite, les modèles $[\lambda DAD']$ et $[\lambda D_k A_k D_k']$ pourront aussi être désignés sous une forme plus compacte $[\lambda C]$ et $[\lambda C_k]$ respectivement. De même, les modèles $[\lambda_k C]$ et $[\lambda_k C_k]$ désigneront les modèles $[\lambda_k DAD']$ et $[\lambda_k D_k A_k D_k']$ respectivement.

3.1.2 Contraintes sur les proportions

En dehors des caractéristiques géométriques, les proportions du mélange p_k constituent des paramètres importants. Deux hypothèses les concernant sont considérées dans MIXMOD : soit, elles sont supposées égales, soit elles dépendent des composants.

Combinant ces hypothèses avec celles qui ont produit quatorze modèles, on obtient vingt-huit modèles notés $[p\lambda I]$, $[p_k \lambda I]$, $[p\lambda_k DAD']$, etc., avec la même convention. Tous ces modèles, schématisés dans le tableau 1, sont disponibles dans MIXMOD dans les contextes non supervisés et supervisés.

3.1.3 Liens avec des critères classiques de classification

Les différents modèles présentés n'ont pas seulement une interprétation géométrique simple. Ils jettent aussi une lumière nouvelle sur des critères classiques de classification. Par exemple, le critère de l'algorithme des centres mobiles de Ward (1963) peut facilement se déduire du modèle de mélange gaussien le plus simple $[p\lambda I]$.

Le modèle $[p\lambda DAD']$ correspond au critère suggéré par Friedman and Rubin (1967), et les modèles $[p\lambda_k DAD']$, $[p\lambda_k DA_k D']$ et $[p\lambda_k D_k A_k D_k']$ correspondent à d'autres critères bien connus de classification (voir par exemple Scott and Symons (1971); Diday and Govaert (1974); Maronna and Jacovkis (1974); Schroeder (1976)). En analyse discriminante,

Modèle	Famille	Prop.	Volume	Forme	Orient.
$[p\lambda DAD']$	Général	Égal	Égal	Égal	Égal
$[p\lambda_k DAD']$			Variable	Égal	Égal
$[p\lambda DA_k D']$			Égal	Variable	Égal
$[p\lambda_k DA_k D']$			Variable	Variable	Égal
$[p\lambda D_k AD'_k]$			Égal	Égal	Variable
$[p\lambda_k D_k AD'_k]$			Variable	Égal	Variable
$[p\lambda D_k A_k D'_k]$			Égal	Variable	Variable
$[p\lambda_k D_k A_k D'_k]$			Variable	Variable	Variable
$[p\lambda B]$	Diagonal	Égal	Égal	Égal	Axes
$[p\lambda_k B]$			Variable	Égal	Axes
$[p\lambda B_k]$			Égal	Variable	Axes
$[p\lambda_k B_k]$			Variable	Variable	Axes
$[p\lambda I]$	Sphérique	Égal	Égal	Égal	NA
$[p\lambda_k I]$			Variable	Égal	NA
$[p_k\lambda DAD']$	Général	Variable	Égal	Égal	Égal
$[p_k\lambda_k DAD']$			Variable	Égal	Égal
$[p_k\lambda DA_k D']$			Égal	Variable	Égal
$[p_k\lambda_k DA_k D']$			Variable	Variable	Égal
$[p_k\lambda D_k AD'_k]$			Égal	Égal	Variable
$[p_k\lambda_k D_k AD'_k]$			Variable	Égal	Variable
$[p_k\lambda D_k A_k D'_k]$			Égal	Variable	Variable
$[p_k\lambda_k D_k A_k D'_k]$			Variable	Variable	Variable
$[p_k\lambda B]$	Diagonal	Variable	Égal	Égal	Axes
$[p_k\lambda_k B]$			Variable	Égal	Axes
$[p_k\lambda B_k]$			Égal	Variable	Axes
$[p_k\lambda_k B_k]$			Variable	Variable	Axes
$[p_k\lambda I]$	Sphérique	Variable	Égal	Égal	NA
$[p_k\lambda_k I]$			Variable	Égal	NA

TAB. 1 – Caractéristiques et identifiants des vingt-huit modèles de mélange gaussien disponibles dans MIXMOD.

les modèles $[p\lambda C]$ et $[p\lambda_k C_k]$ définissent respectivement l'analyse discriminante linéaire et l'analyse discriminante quadratique (voir par exemple McLachlan, 1992).

3.2 Cinq modèles de mélanges multinomiaux

De la même façon que le modèle gaussien est souvent retenu pour modéliser chaque composant du mélange lorsque les variables sont continues, le choix du modèle loglinéaire (Agresti, 1990; Bock, 1986) s'impose assez naturellement lorsque les variables sont qualitatives. Le modèle loglinéaire complet ou saturé pour lequel chaque classe suit une distribution multinomiale à 2^d valeurs n'a pas d'intérêt dans le cadre d'un mélange puisqu'il conduit à des modèles non identifiables. Il faut se restreindre à des modèles loglinéaires suffisamment contraints pour les rendre identifiables. L'exemple le plus simple et le plus répandu est le modèle d'indépendance qui suppose que conditionnellement à l'appartenance à une classe, les variables qualitatives sont indépendantes. Le modèle de mélange associé à cette distribution est appelé modèle des classes latentes (Lazarfield and Henry, 1968; Goodman, 1974). C'est le modèle retenu dans MIXMOD pour traiter les données qualitatives.

On supposera dans ce paragraphe que toutes les variables sont des variables qualitatives à m_j modalités et que les données sont constituées d'un échantillon $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ où $\boldsymbol{x}_i = (x_i^{jh}; j = 1, \ldots, p; h = 1, \ldots, m_j)$ avec

$$\begin{cases} x_i^{jh} = 1 & \text{si } i \text{ prend la modalité } h \text{ pour la variable } j \\ x_i^{jh} = 0 & \text{sinon.} \end{cases}$$

3.2.1 Le modèle des classes latentes

Si on note α_k^{jh} la probabilité que la variable $\mathbf{x}^j$ prenne la modalité h lorsque l'individu est dans la classe k, la probabilité du mélange s'écrit alors :

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

où le paramètre $\boldsymbol{\theta}$ est défini par les proportions $\mathbf{p} = (p_1, \ldots, p_g)$ et par les paramètres $\boldsymbol{\alpha}_k = (\alpha_k^{jh}; j = 1, \ldots, p; h = 1, \ldots, m_j)$, vérifiant $\alpha_k^{jh} \in]0,1[$ et $\sum_h \alpha_k^{jh} = 1 \quad \forall k, j$.

3.2.2 Modèles parcimonieux

Le nombre de paramètres nécessaires au modèle des classes latentes que l'on vient d'étudier, égal à $(K-1) + K * \sum_j (m_j - 1)$, est généralement beaucoup plus petit que le nombre de paramètres nécessaires au modèle loglinéaire complet, égal à $\prod_j m_j$. Par exemple, pour un nombre de classes K égal à 5, un nombre de variables qualitatives d égal à 10 et si le nombre de modalités m_j est égal à 4 pour toutes les variables, on obtient respectivement 154 et 10^6 paramètres. Ce nombre de paramètres peut toutefois se révéler encore beaucoup trop grand et des modèles plus parcimonieux sont alors nécessaires.

Pour ceci, on contraint tous les vecteurs $(\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})$ à prendre la forme

$$(\frac{\varepsilon_k^j}{m_j - 1}, \frac{\varepsilon_k^j}{m_j - 1}, .., \frac{\varepsilon_k^j}{m_j - 1}, 1 - \varepsilon_k^j, \frac{\varepsilon_k^j}{m_j - 1}, ..., \frac{\varepsilon_k^j}{m - 1})$$

avec $\varepsilon_k^j < \frac{m_j - 1}{m_j}$ (Celeux and Govaert, 1991). Les vecteurs des probabilités sont alors simplement caractérisés par une modalité majoritaire (le mode) à laquelle est associé le terme $1 - \varepsilon_k^j$ et un terme de dispersion ε_k^j. Le modèle alors obtenu est noté $[\varepsilon_k^j]$.

Comme pour le modèle de mélange gaussien, il est possible d'imposer des contraintes supplémentaires ; on obtient alors les modèles suivants :

- modèle $[\varepsilon_k]$: la dispersion ne dépend pas de la variable ;
- modèle $[\varepsilon_j]$: la dispersion ne dépend pas de la classe ;
- modèle $[\varepsilon]$: la dispersion ne dépend ni de la variable, ni de la classe.

Enfin, par analogie, le modèle des classes latentes initial, qui n'impose aucune contrainte sur les probabilités associées à chaque modalité sera noté $[\varepsilon_k^{jh}]$.

Un bilan du nombre de paramètres associés à chacun de ces modèles disponibles dans MIXMOD est donné dans le tableau 2.

	Proportions égales	Proportions libres
$[\varepsilon_k^{jh}]$	$K \sum_{j=1}^d (m_j - 1)$	$(K-1) + K \sum_{j=1}^d (m_j - 1)$
$[\varepsilon_k^j]$	Kd	$(K-1) + Kd$
$[\varepsilon_k]$	K	$(K-1) + K$
$[\varepsilon^j]$	d	$(K-1) + d$
$[\varepsilon]$	1	$(K-1) + 1$

TAB. 2 – Nombre de paramètres des dix modèles des classes latentes disponibles dans MIXMOD.

4 Modèle de mélange pour la classification

4.1 Le problème de classification

Le contenu de cette section est général mais, pour simplifier l'exposé, on se place dans le cadre continu. Les données considérées dans MIXMOD pour la classification sont dans cette section n vecteurs $\mathbf{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ de $\mathbb{R}^d$. Cela signifie que l'on se place dans ce paragraphe dans la situation de données décrites par des variables quantitatives. Le cas où les données sont qualitatives ne sera pas détaillé ici, mais il n'induit pas de difficulté particulière. Le but de la classification est d'estimer une partition inconnue $\mathbf{z}$ de $\mathbf{x}$ en K classes, $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ désignant n vecteurs indicateurs ou étiquettes $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$, $i = 1, \ldots, n$ avec $z_{ik} = 1$ si $\boldsymbol{x}_i$ appartient à la k^{e} classe et 0 sinon. La vision du problème de la classification par un modèle de mélange est d'associer chaque classe à un composant du mélange. En règle générale, toutes les étiquettes $\mathbf{z}_i$ sont inconnues. Cependant, un étiquetage partiel des données est possible, et MIXMOD permet de traiter les cas où l'ensemble de données $\mathbf{x}$ est divisé en deux sous-ensembles $\mathbf{x} = \{\mathbf{x}^\ell, \mathbf{x}^u\}$ où $\mathbf{x}^\ell = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ $(1 \leq m \leq n)$ sont des données aux étiquettes connues $\mathbf{z}^\ell = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$, tandis que $\mathbf{x}^u = \{\boldsymbol{x}_{m+1}, \ldots, \boldsymbol{x}_n\}$ sont d'étiquettes inconnues $\mathbf{z}^u = \{\mathbf{z}_{m+1}, \ldots, \mathbf{z}_n\}$. De plus, MIXMOD permet de spécifier un poids pour chaque unité statistique. Cette possibilité est utile notamment pour le traitement de données groupées ou des fréquences.

Dans le contexte du modèle de mélange de MIXMOD, les données complètes $(\boldsymbol{x}_i, \mathbf{z}_i)$ $(i = 1, \ldots, n)$ sont supposées provenir de la distribution de probabilité $\prod_{k=1}^{K} (p_k \varphi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k))^{z_{ik}}$. Dans ce contexte statistique, MIXMOD considère deux approches du maximum de vraisemblance (m. v.) : la première, dite approche de mélange, consiste à maximiser en $\boldsymbol{\theta}$ la densité des données observées, et la seconde, l'approche de classification, consiste à maximiser en $\boldsymbol{\theta}$ et $\mathbf{z}^u$ la densité des données complètes.

4.2 Estimation par l'approche de mélange

Cette approche consiste à maximiser en $\boldsymbol{\theta} = (p_1, \ldots, p_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1, \ldots, \Sigma_K)$ la logvraisemblance observée :

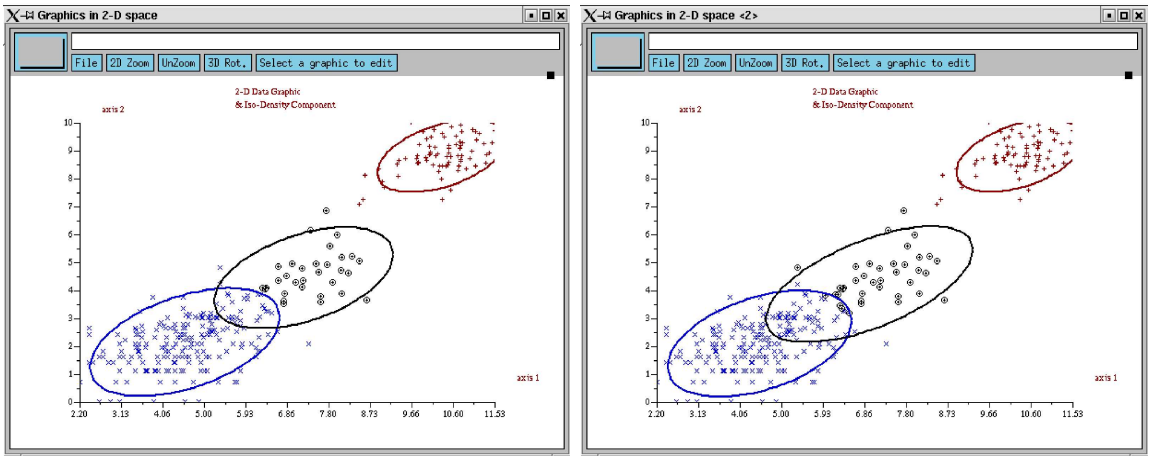
$$L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}^\ell) = \sum_{i=1}^{m} \sum_{k=1}^{K} z_{ik} \ln \left(p_k \varphi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k)\right) + \sum_{i=m+1}^{n} \ln \left(\sum_{k=1}^{K} p_k \varphi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k) \right). \qquad (4)$$

Une partition $\hat{\mathbf{z}}^u$ est déduite de l'estimateur m.v. $\hat{\boldsymbol{\theta}}$ par une procédure du *Maximum A Posteriori* (MAP) qui affecte chaque $\boldsymbol{x}_i$ de $\mathbf{x}^u$ au composant k de probabilité conditionnelle :

$$t_k(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}) = \frac{\hat{p}_k \varphi(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k)}{\sum_{k'=1}^{K} \hat{p}_{k'} \varphi(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_{k'}, \hat{\Sigma}_{k'})} \qquad (5)$$

la plus grande que $\boldsymbol{x}_i$ soit issue de ce composant. La maximisation de $L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}^\ell)$ peut être réalisée dans MIXMOD par l'algorithme EM de Dempster et al. (1977) ou par une version stochastique de EM, l'algorithme SEM (voir par exemple Celeux and Diebolt, 1985; McLachlan and Krishnan, 1997). Dans la section 7, on décrit trois façons différentes de combiner ces algorithmes. Bien sûr, l'estimateur $\hat{\boldsymbol{\theta}}$, et par conséquent $\hat{\mathbf{z}}^u$, dépend du modèle de mélange considéré et du nombre de classes.

Exemple 1 (Départements français) *La figure 2 (a) décrit la partition et les isodensités des composants estimés par l'algorithme EM pour le mélange gaussien* $[p_k \lambda_k D A_k D']$ *avec trois composants.*



(a) (b)

FIG. 2 – Partition estimée et isodensités des composants pour les départements français : (a) par EM et (b) par CEM.

4.3 Estimation par l'approche de classification

La seconde approche de MIXMOD est une approche de classification où les vecteurs indicateurs $\mathbf{z}^u$, de l'origine inconnue du composant du mélange, sont traités comme des paramètres. Cette approche vise à maximiser la logvraisemblance complétée :

$$CL(\boldsymbol{\theta}, \mathbf{z}^u; \mathbf{x}, \mathbf{z}\ell) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln \left(p_k \varphi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k) \right) \tag{6}$$

à la fois en $\boldsymbol{\theta}$ et en $\mathbf{z}^u$. Le critère CL peut être maximisé par une version de classification de l'algorithme EM, l'algorithme CEM (Celeux and Govaert, 1992) qui inclut une étape de classification (étape C) entre les étapes E et M. Dans la section 7, on décrit diverses stratégies pour calculer l'estimateur de $\boldsymbol{\theta}$ utilisant cet algorithme.

Exemple 2 (Départements français) *La figure 2 (b) décrit la partition et les isodensités estimées par l'algorithme CEM pour le mélange gaussien* $[p_k \lambda_k D A_k D']$ *avec trois composants. Cette solution est à comparer à celle obtenue par l'algorithme EM décrite figure 2 (a).*

5 Modèles génératifs d'analyse discriminante

Les données considérées par MIXMOD pour l'analyse discriminante constituent un échantillon d'apprentissage de n vecteurs $(\mathbf{x}, \mathbf{z}) = \{(\boldsymbol{x}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{z}_n)\}$, où $\boldsymbol{x}_i$ appartient à $\mathbb{R}^d$, et $\boldsymbol{z}_i$ est le vecteur indicateur de la classe de l'unité statistique i.

Le but est de construire à partir de cet ensemble d'apprentissage, un classifieur pour prédire la classe $\mathbf{z}_{n+1}$ d'une nouvelle observation décrite par le vecteur $\boldsymbol{x}_{n+1}$ de $\mathbb{R}^d$ et d'origine inconnue. Notons qu'il est également possible de pondérer les données dans le contexte de l'analyse discriminante sous MIXMOD. Les hypothèses statistiques sont les mêmes que celles utilisées pour la classification non supervisée.

Dans ce contexte supervisé, le paramètre $\boldsymbol{\theta}$ est estimé par la maximisation de la vraisemblance complétée (6). Comme $\mathbf{z}$ est parfaitement connu, l'obtention de l'estimation $\hat{\boldsymbol{\theta}}$ du m.v. se réduit à une étape de maximisation. Tout nouveau point $\boldsymbol{x}$ peut être affecté à l'une des K classes par la procédure MAP avec $\hat{\boldsymbol{\theta}}$.

En résumé, l'analyse discriminante est réalisée dans MIXMOD par les deux étapes suivantes :

- **étape M :** Calcul de l'estimateur m.v. $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ par la maximisation de la logvraisemblance complétée (6).
- **étape MAP :** Affectation de tout nouveau point $\boldsymbol{x}$ à l'une des K classes par la règle suivante :

$$k(\boldsymbol{x}) = \arg\max_k t_k(\boldsymbol{x}; \hat{\boldsymbol{\theta}}).$$

Exemple 3 (Oiseaux *borealis*) *La figure 3 décrit les frontières de classification, les isodensités des composants et les individus dans le premier plan principal de l'ACP pour le modèle de mélange gaussien le plus général* $[p_k \lambda_k D_k A_k D_k']$.

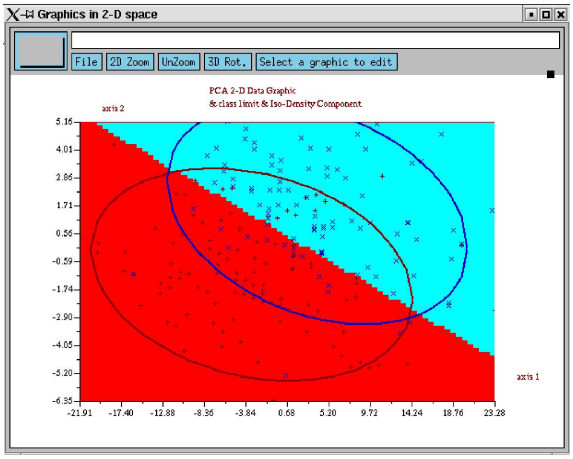
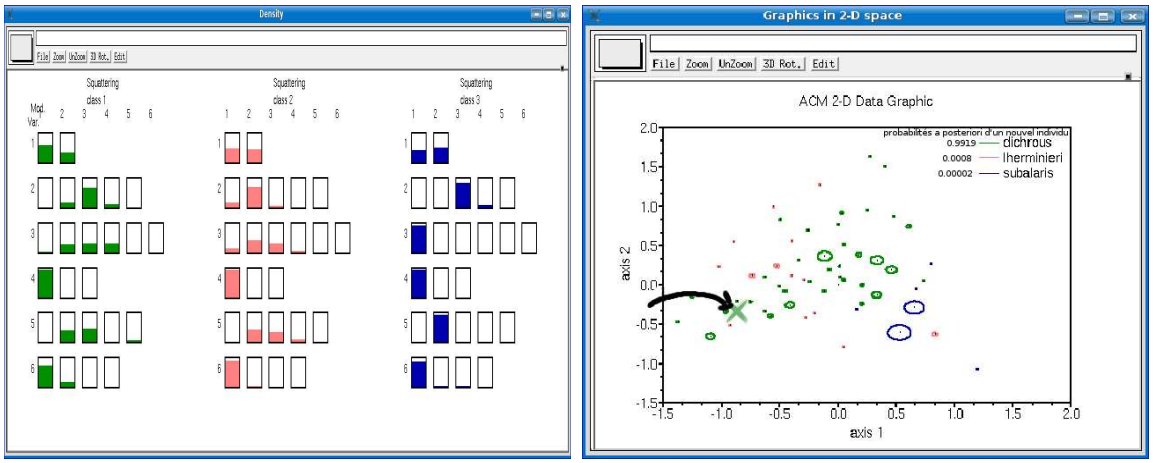


FIG. 3 – Limites des classes, isodensités des composants et individus pour les données des oiseaux avec le modèle $[p_k \lambda_k D_k A_k D_k']$ dans le premier plan principal de l'ACP.

Exemple 4 (Oiseaux *puffins*) *La figure 4 représente la dispersion des individus autour de chaque variable et de chaque modalité ainsi que la classification d'un individu supplémentaire dans le premier plan de l'ACM.*



(a) (b)

FIG. 4 – Illustration de l'analyse discriminante pour les données qualitatives : (a) dispersion et (b) nouvel individu dans le premier plan de l'ACM.

6 Les algorithmes de MIXMOD

6.1 L'algorithme EM

L'algorithme EM vise à maximiser la vraisemblance du mélange dans un contexte non supervisé. Partant d'une valeur initiale arbitraire $\boldsymbol{\theta}^0$, la q^{e} itération de l'algorithme EM consiste à effectuer les deux étapes E et M maintenant décrites :

- **Étape E :** Calcul des probabilités conditionnelles $t_{ik}^q = t_k(\boldsymbol{x}_i; \boldsymbol{\theta}^{q-1})$ que $\boldsymbol{x}_i$ appartienne à la k^{e} classe $(i = m+1, \ldots, n)$ en utilisant la valeur courante $\boldsymbol{\theta}^{q-1}$ du paramètre.
- **Étape M :** l'estimateur m.v. $\boldsymbol{\theta}^q$ de $\boldsymbol{\theta}$ est actualisé en utilisant les probabilités conditionnelles t_{ik}^q comme poids. Cette étape dépend bien sûr du modèle utilisé. Les formules détaillées pour les quatorze mélanges gaussiens disponibles dans MIXMOD sont données dans (Celeux and Govaert, 1995).

6.2 L'algorithme SEM

Dans la version stochastique SEM de l'algorithme EM, une étape S est incorporée entre les étapes E et M. Il s'agit d'une restauration aléatoire des labels inconnus des composants selon leur distribution conditionnelle courante. À l'étape M, l'estimateur du paramètre $\boldsymbol{\theta}$ est actualisé en maximisant la vraisemblance complétée associée à cette restauration des données manquantes. L'algorithme SEM n'assure pas une convergence ponctuelle. Il engendre une chaîne de Markov dont la distribution stationnaire est plus ou moins concentrée autour de l'estimateur du maximum de vraisemblance. Un estimateur naturel tiré de la suite $(\boldsymbol{\theta}^q)_{q=1,\ldots,Q}$ engendrée par SEM est sa moyenne $\sum_{q=r+1}^{Q} \boldsymbol{\theta}^q/(Q-r)$ (amputée des r premières valeurs de chauffe de l'algorithme). Un estimateur alternatif est obtenu par sélection de la valeur de la suite SEM fournissant la plus grande vraisemblance.

6.3 L'algorithme CEM

L'algorithme Classification EM (CEM) incorpore une étape de classification entre les étapes E et M de EM. Cette étape de classification consiste à affecter chaque point à l'un des K composants par la procédure MAP. Comme pour SEM, l'étape M consiste alors à actualiser l'estimateur du paramètre $\boldsymbol{\theta}$ en maximisant la vraisemblance complétée associée à la restauration des données manquantes.

CEM est un algorithme de type *K-means* et, contrairement à EM, il converge en un nombre fini d'itérations. L'algorithme CEM ne maximise pas la vraisemblance observée L (4), mais vise à maximiser la vraisemblance complétée CL (6) en $\boldsymbol{\theta}$ et en $\mathbf{z}^u$. En conséquence, CEM qui n'est pas destiné à maximiser la vraisemblance de $\boldsymbol{\theta}$, produit des estimateurs biaisés des paramètres. Ce biais est d'autant plus fort que les composants du mélange sont imbriqués et que les proportions sont déséquilibrées (McLachlan and Peel, 2000, Section 2.21).

6.4 Fonctions M et MAP

Ces deux fonctions sont surtout utiles pour l'analyse discriminante. La fonction M est dévolue à l'estimation du maximum de vraisemblance du paramètre $\boldsymbol{\theta}$ d'un mélange dont les labels $\mathbf{z}$ sont connus. Cette fonction de maximisation est simplement l'étape M utilisée par les algorithmes SEM et CEM. La fonction MAP a été décrite dans la section 4.2.

7 Stratégies d'utilisation des algorithmes

7.1 Stratégies d'initialisation

Il y a cinq façons différentes d'initialiser un algorithme dans MIXMOD. En dehors de la première qui est déterministe, il est recommandé de répéter plusieurs fois le tandem {stratégie d'initialisation/algorithme} afin de sélectionner la meilleure solution possible vis-à-vis du critère optimisé, la vraisemblance observée pour EM ou SEM, et la vraisemblance complétée pour CEM :

- les trois algorithmes peuvent être initialisés par une partition $\mathbf{z}^{u0}$ ou des valeurs des paramètres du mélange $\boldsymbol{\theta}^0$ spécifiés par l'utilisateur ;
- ils peuvent être initialisés par une valeur aléatoire de $\boldsymbol{\theta}^0$. Dans MIXMOD, ce départ aléatoire est obtenu en tirant au hasard la moyenne des composants parmi les observations, en fixant des proportions égales et en choisissant une matrice de variance commune et diagonale dont les éléments diagonaux sont égaux à la variance empirique de chaque variable. Cette stratégie, très utilisée, peut être considérée comme une stratégie de référence ;
- l'algorithme EM peut être initialisé par la position produisant la plus grande valeur de la vraisemblance complétée obtenue après plusieurs lancements aléatoires de l'algorithme CEM. Le nombre de réplications de CEM est *a priori* inconnu et dépend de la répartition entre les algorithmes du nombre total d'itérations disponibles fourni par l'utilisateur (voir Biernacki et al., 2003) ;
- l'algorithme EM peut être initialisé par la position produisant la plus grande vraisemblance obtenue après le lancement aléatoire de courtes et nombreuses exécutions de EM lui-même. Par exécution courte de EM, on entend que cet algorithme est

stoppé dès que $(L^q - L^{q-1})/(L^q - L^0) \leq 10^{-2}$, L^q étant la vraisemblance observée à la q^{e} itération. Ici 10^{-2} représente un seuil par défaut à choisir au jugé. Le nombre de réplications d'exécutions courtes de EM est *a priori* inconnu et dépend de la répartition entre les algorithmes du nombre total d'itérations disponibles fourni par l'utilisateur (voir Biernacki et al., 2003) ;
– l'algorithme EM peut être démarré par la position fournissant la plus grande vraisemblance dans une suite d'estimations produites par l'algorithme SEM initialisé au hasard avec toujours une répartition des itérations choisie par l'utilisateur (voir Biernacki et al., 2003).

7.2 Règles d'arrêt

Dans MIXMOD, il y a trois façons d'arrêter un algorithme :
– les algorithmes EM, SEM et CEM peuvent être arrêtés par un nombre maximal d'itérations ;
– un algorithme peut être arrêté par un seuil sur l'amélioration relative du critère en jeu (la logvraisemblance L (4) ou la logvraisemblance complétée CL (6)). Pour l'algorithme EM cette possibilité n'est pas recommandée car cet algorithme peut faire face à des situations de convergence lente. Il est recommandé d'arrêter l'algorithme CEM, qui converge en un nombre fini d'itérations, à sa position stationnaire ;
– un algorithme peut être arrêté dès que l'une des règles d'arrêt précédentes est remplie.

7.3 Chaînage d'algorithmes

Dans MIXMOD il est facile de combiner les algorithmes EM, SEM et CEM selon sa fantaisie. Cette possibilité peut produire des stratégies d'initialisation originales et efficaces, comme celles présentées dans Biernacki et al. (2003).

8 Sélection de modèles

Il est bien sûr du plus haut intérêt d'être capable de sélectionner automatiquement un modèle de mélange M et un nombre K de composants. Cependant, choisir un modèle de mélange raisonnable dépend beaucoup du but de la modélisation. C'est pourquoi, nous distinguons, dans ce qui suit, les points de vue estimation de densité, classification et analyse discriminante.

8.1 Les points de vue estimation de densité et classification

Dans MIXMOD trois critères sont disponibles dans un contexte non supervisé : BIC, ICL et NEC. Lorsqu'aucune information n'est disponible sur K, il est recommandé de faire varier K entre $K = 1$ et le plus petit entier plus grand que $n^{0.3}$ (voir Bozdogan, 1993).

Pour l'estimation de densité, BIC (*Bayesian Information Criterion*) doit être préféré. Notant $\nu_{M,K}$ le nombre de paramètres indépendants du modèle de mélange M avec K

composants, le critère BIC s'écrit comme un critère de vraisemblance pénalisée :

$$\mathrm{BIC}_{M,K} = -2L_{M,K} + \nu_{M,K} \ln(n). \tag{7}$$

Le couple (M, K) conduisant à la plus petite valeur de BIC est choisi. Bien que les conditions de régularité classiques justifiant BIC (Schwarz, 1978) ne sont pas remplies pour les mélanges, on peut prouver que pour de nombreux mélanges, BIC est un critère convergent (Kéribin, 2000). De plus BIC s'avère efficace en pratique (voir par exemple Fraley and Raftery, 1998).

Pour la classification, ICL et NEC peuvent choisir des modèles plus parcimonieux et robustes. Pour prendre en compte la capacité d'un modèle de mélange à révéler une structure en classes dans les données, on peut préférer au critère BIC le critère ICL (*Integrated Complete-data Likelihood*) (Biernacki et al., 2000) qui s'écrit :

$$\mathrm{ICL}_{M,K} = \mathrm{BIC}_{M,K} - 2 \sum_{i=m+1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \ln(t_{ik}), \tag{8}$$

où $t_{ik} = t_k(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_{M,K})$ (avec $\hat{\boldsymbol{\theta}}_{M,K}$ l'estimateur m.v. du paramètre pour le modèle M et le nombre de composants K) et où $\hat{\mathbf{z}} = \mathrm{MAP}(\hat{\boldsymbol{\theta}}_{M,K})$. Ce critère à minimiser est simplement le critère BIC pénalisé par un terme d'entropie qui mesure le degré d'imbrication des composants. Le critère NEC (*Normalized Entropy Criterion*) proposé par Celeux and Soromenho (1996) utilise un terme d'entropie similaire $E_K = -\sum_{i=m+1}^{n} \sum_{k=1}^{K} t_{ik} \ln(t_{ik})$, mais ce critère est principalement destiné à être utilisé pour déterminer le nombre de classes K, plutôt que la paramétrisation du modèle M (Biernacki and Govaert, 1999). Ce critère à minimiser s'écrit :

$$\mathrm{NEC}_K = \frac{E_K}{L_K - L_1}. \tag{9}$$

On peut remarquer que NEC_1 n'est pas défini. Biernacki et al. (1999) ont proposé la règle suivante, efficace pour lever cette indétermination : Soit $K^\star$ minimisant NEC_K $(2 \leq K \leq K_{\mathrm{sup}})$, K_{sup} étant un majorant du nombre de composants du mélange. On choisit $K^\star$ classes si $\mathrm{NEC}_{K^\star} \leq 1$, sinon on décide qu'il n'existe pas de structure en classes dans les données.

Exemple 5 (Départements français) *Cinq nombres de composants ($K = 1 - 5$) et trois mélanges gaussiens $[p_k \lambda_k DAD']$, $[p_k \lambda_k DA_k D']$ et $[p_k \lambda_k D_k A_k D_k]$ sont considérés. L'algorithme EM est utilisé pour chaque combinaison modèle–nombre de composants. Les figures 5 (a) et (b) donnent respectivement les valeurs de BIC pour chaque combinaison et la partition associée au choix de BIC. Les figures 5 (c) et (d) donnent les mêmes choses pour le critère ICL.*

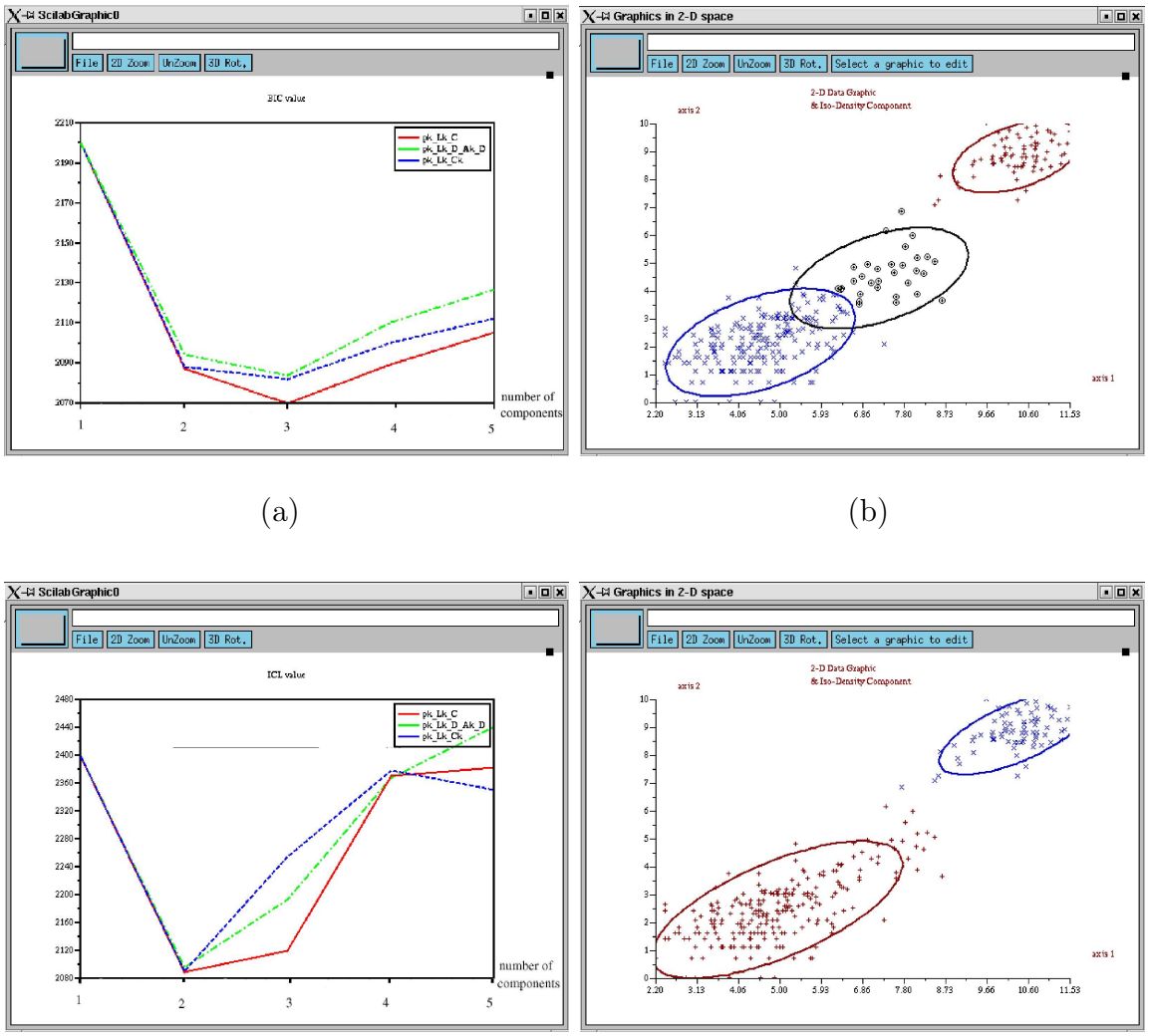


FIG. 5 – Sélection d'une combinaison modèle–nombre de composants pour les départements français : (a) valeurs de BIC ; (b) la partition optimale associée ; (c) valeurs de ICL ; (d) la partition optimale associée.

8.2 Le point de vue de l'analyse discriminante

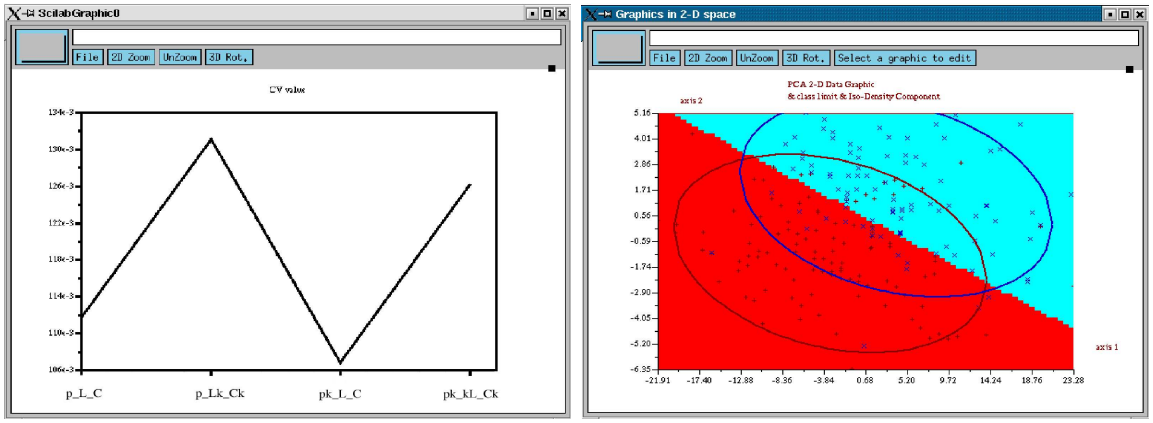
Dans ce cas, le modèle M doit être sélectionné, mais le nombre de composants du mélange est connu. Dans MIXMOD deux critères sont proposés dans un contexte supervisé : BIC et le taux d'erreur évalué par validation croisée (CV). Le critère CV est spécifique à la classification supervisée. Il est défini par :

$$\mathrm{CV}_M = \frac{1}{m}\sum_{i=1}^{m} \delta(\hat{\mathbf{z}}_i^{(i)}, \mathbf{z}_i) \tag{10}$$

où δ représente le coût 0-1 et $\hat{\mathbf{z}}_i^{(i)}$ le groupe d'affectation de $\boldsymbol{x}_i$ lorsque le classifieur est construit à partir de l'échantillon total $(\mathbf{x}, \mathbf{z})$ privé de $(\boldsymbol{x}_i, \mathbf{z}_i)$. Des estimations rapides des n règles de discrimination sont implantées dans le cas gaussien (Biernacki and Govaert, 1999).

Dans MIXMOD, selon une approche décrite dans Bensmail and Celeux (1996), il est possible de sélectionner l'un des quatorze mélanges gaussiens par minimisation du critère CV. On doit, cependant, signaler que ce critère fournit une estimation optimiste du vrai taux d'erreur. En effet, c'est une situation où la méthode inclut la sélection d'un modèle parmi plusieurs. Aussi le vrai taux d'erreur doit être évalué sur un échantillon indépendant. Typiquement, trois échantillons sont nécessaires : un échantillon d'*apprentissage* des modèles, un échantillon de *validation* pour choisir l'un des modèles et un échantillon *test* pour évaluer le vrai taux d'erreur de la méthode complète. Cela signifie que lorsque l'on utilise la validation croisée pour évaluer les performances d'un modèle, on doit effectuer une double validation croisée pour obtenir un estimateur sans biais du taux d'erreur. Cette procédure est implantée dans MIXMOD.

Exemple 6 (Oiseaux *borealis*) *Nous avons considéré quatre mélanges gaussiens* $[p\lambda DAD']$, $[p\lambda_k D_k A_k D_k']$, $[p_k \lambda DAD']$ *et* $[p_k \lambda_k D_k A_k D_k']$. *Les figures 6 (a) et (b) donnent respectivement les valeurs du critère CV pour chaque modèle et le classifieur associé au meilleur modèle pour ce critère.*



(a) (b)

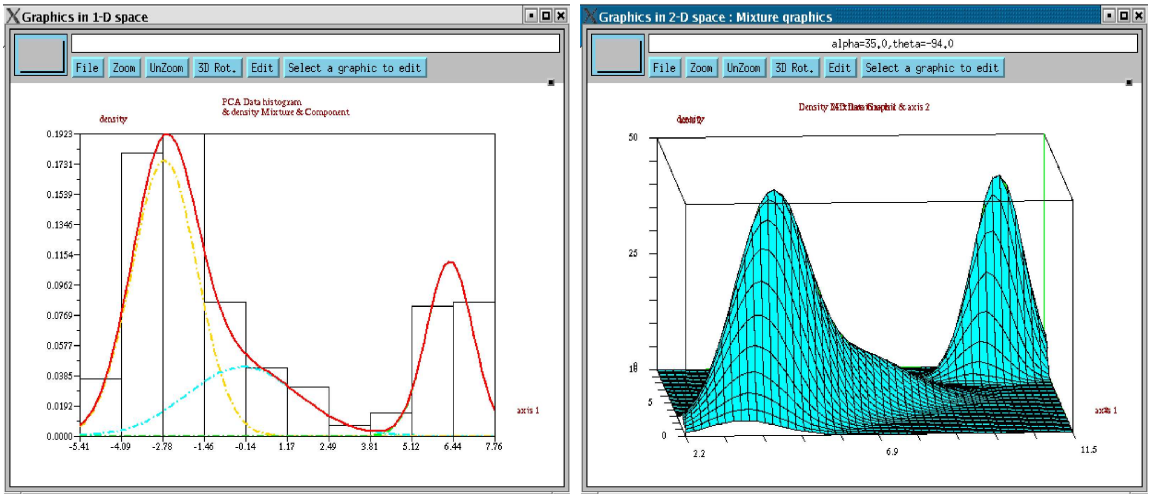
FIG. 6 – Sélection d'un mélange gaussien pour les oiseaux : (a) valeurs de CV et (b) règle de classement optimale associée.

9 Fonctions associées

Les environnements MATLAB et SCILAB fournissent des fonctions de haut niveau en particulier pour les représentations graphiques.

9.1 Représentation graphique des critères

L'une des sorties optionnelles de la fonction `mixmod` est un tableau de dimension quatre fournissant les valeurs de tous les critères demandés pour toutes les stratégies, tous les nombres de classes et tous les modèles demandés. À partir de ce tableau, des graphes de variation peuvent être dessinés dans MIXMOD. Des illustrations de cette possibilité sont données dans les figures 5 (a), (c) et 6 (a).



(a) (b)

FIG. 7 – Densité du mélange : (a) premier axe de l'ACP et (b) espace 2D d'origine.

9.2 La fonction MIXMODVIEW pour les graphiques

MIXMOD propose la fonction `mixmodView` de visualisation des résultats. Cette fonction permet de faire des graphiques générés à partir des sorties de la fonction `mixmod` (densités, isodensités, etc.) en dimension un, deux ou trois.

Les graphiques suivants sont disponibles (liste non exhaustive) :

- les isodensités, la représentation des densités des composants et du mélange sur le premier axe de l'ACP ;
- le tracé des limites de classes, les isodensités des composants et la représentation des individus dans le premier plan de l'ACP ;
- la densité du mélange dans le premier plan de l'ACP ;
- les individus et les labels dans le premier espace 3D de l'ACP.

Beaucoup de ces caractéristiques ont été illustrées dans les exemples précédents. L'exemple suivant montre des graphiques de densité.

Exemple 7 (Départements français) *Les figures 7 (a) et (b) donnent respectivement la densité du mélange sur le premier axe de l'ACP et dans l'espace 2D d'origine.*

9.3 La fonction PRINTMIXMOD pour des résumés

La fonction `printMixmod` peut être utilisée pour résumer les résultats de la fonction `mixmod`. Elle fournit un résumé synthétique des résultats (conditions d'entrée, valeur de critères, logvraisemblance, logvraisemblance complétée, estimation des paramètres, etc.)

9.4 La fonction INPUTMIXMOD pour les entrées

La fonction `inputMixmod` produit des structures SCILAB ou MATLAB qui peuvent être utilisées par la fonction `mixmod`. Elle permet de spécifier facilement les critères, les modèles de mélange, les algorithmes, leurs stratégies d'utilisation et leurs règles d'arrêt.

Références

Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49 :803–821.

Bensmail, H. and Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91(2) :1743–17448.

Biernacki, C., Beninel, F., and Bretagnolle, V. (2002). A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics*, 58(2) :387–397.

Biernacki, C., Celeux, G., and Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20 :267–272.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7) :719–725.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41 :561–575.

Biernacki, C. and Govaert, G. (1999). Choosing models in model-based clustering and discriminant analysis. *J. Statis. Comput. Simul.*, 64 :49–71.

Bock, H. (1986). Loglinear models and entropy clustering methods for qualitative data. In Gaull, W. and Schader, M., editors, *Classification as a tool of research*, pages 19–26. North-Holland, Amsterdam.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher information matrix. In Opitz, O., Lauritzen, B., and Klar, R., editors, *Information and Classification*, pages 40–54, Heidelberg. Springer-Verlag.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm : A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2 :73–82.

Celeux, G. and Govaert, G. (1991). Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8(2) :157–176.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3) :315–332.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5) :781–793.

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13 :195–212.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, B 39 :1–38.

Diday, E. and Govaert, G. (1974). Classification avec distance adaptative. *C. R. Acad. Sc. Paris, série A*, 278 :993–995.

Fraley, C. and Raftery, A. E. (1998). How many clusters ? Which clustering method ? Answers via model-based cluster analysis. *Computer Journal*, 41 :578–588.

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of American Statistical Association*, 62 :1159–1178.

Goodman, L. A. (1974). Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika*, 61 :215–231.

Kéribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyã, Series A*, 1 :49–66.

Lazarfield, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin Company, Boston.

Maronna, R. and Jacovkis, P. (1974). Multivariate clustering procedure with variable metrics. *Biometrics*, 30 :499–505.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

McLachlan, G. J. and Krishnan, K. (1997). *The EM Algorithm*. Wiley, New York.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Schroeder, A. (1976). Analyse d'un mélange de distributions de probabilité de même type. *Revue de Statistique Appliquée*, 24(1) :39–62.

Schwarz, G. (1978). Estimating the number of components in a finite mixture model. *Annals of Statistics*, 6 :461–464.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, 27 :387–397.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58 :236–244.