

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

Christophe Biernacki¹, Gilles Celeux², Anwuli Echenim²³, Gérard Govaert⁴, Florent Langrognet³

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](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