# Sélection de modèle pour la classification en présence d'une classification externe

Jean-Patrick Baudry, Gilles Celeux

# Sélection de modèle pour la classification en présence d'une classification externe

Jean-Patrick Baudry & Gilles Celeux

*Université Paris-Sud, département de mathématiques et Inria Saclay 91405 Orsay Cedex*

**Résumé** : En classification non supervisée de données, il est souvent utile d'interpréter les résultats de la classification cherchée en regard d'une partition des individus connue a priori et obtenue sur d'autres informations que les données disponibles. Nous proposons une approche fondée sur le modèle de mélange de lois qui permet de sélectionner un modèle de classification et un nombre de classes de sorte à produire une classification qui, à la fois, s'ajuste bien aux données et présente une bonne liaison avec la partition a priori. Cette approche utilise la vraisemblance intégrée jointe des données et des deux classifications en jeu. Il est à noter que l'obtention de la classification ne fait intervenir la partition a priori que dans la phase de sélection d'un modèle et non dans la phase de construction de la classification qui se fait de manière classique par maximum de vraisemblance. Des illustrations seront données et le fait de dissocier les étapes d'estimation et de sélection d'un modèle sera discuté.
**Mots-clés** : mélange de lois, vraisemblance intégrée, entropie, critère BIC

**Abstract** : In cluster analysis, it is often useful to interpret the obtained partition with respect to a known partition derived from alternative informations. An approach is proposed in the model-based clustering context to select a model and a number of clusters in order to get a partition which both provides a good fit with the data and closely related to the known partition. This approach makes use of the integrated joint likelihood of the data and the two partitions at hand. It is worth noticing that the known partition is only used to select a relevant mixture model. All the mixture models are estimated by the maximum likelihood methodology from the observed data. Numerical illustrations will be given and the strategy separating the estimation and the selection steps will be discussed.
**Key words** : mixture model, integrated likelihood, entropy, BIC criterion

## 1 Introduction

Le modèle de mélange de lois de probabilité fournit un cadre précis et souple pour construire une classification de données pertinente et dont on peut mesurer le bien-fondé et la stabilité avec des outils rigoureux (cf. par exemple Govaert, 2003). Ce modèle permet notamment de résoudre le délicat choix du nombre de classes avec objectivité. De plus, comme nous allons le voir dans cette communication, il permet facilement d'apporter des réponses judicieuses à des questions spécifiques qui se rencontrent fréquemment en classification.

Nous considérons le problème suivant. Il s'agit de réaliser la classification d'un ensemble de $n$ individus décrits par $d$ variables. Mais par ailleurs une partition d'intérêt $U$ en $K_u$ classes de ces individus, obtenue à partir d'autres informations est disponible. Dans ces conditions, il est souhaitable que la classification construite à partir des $d$ variables soit reliée à la partition $U$ pour faciliter l'exploitation de la classification cherchée. Le fait de se placer dans le cadre d'un modèle de mélange va permettre de répondre à cet objectif de manière naturelle et efficace.

# 2 Le modèle

Nous considérons $n$ individus décrits par $d$ variables $\mathbf{y}_1, \ldots, \mathbf{y}_n$ qu'il s'agit de classer en classes homogènes pour ces variables. La particularité du problème tient au fait que ces individus sont également caractérisés par une partition connue $\mathbf{u}$ obtenue sans la considération des $d$ variables descriptives. Il est alors souhaitable que la partition cherchée soit, autant que faire se peut, reliée à la partition $\mathbf{u}$.

Pour construire la partition $\mathbf{z}$ obtenue à partir des $\mathbf{y}$ nous considérons un modèle de mélange fini de distributions multivariées. Ce modèle sera par exemple un modèle de mélange gaussien si les données sont quantitatives ou un modèle de mélange multinomiale si les données sont qualitatives (cf. Biernacki *et al.* 2006).

Le problème le plus délicat est alors de choisir une forme de modèle et un nombre de classes pertinent. Pour nous aider à résoudre ce probème, nous allons tenir compte de la partition $\mathbf{u}$ et proposer un critère qui réalise un bon compromis entre la qualité d'ajustement du mélange de lois et l'adéquation entre la partitions $\mathbf{z}$ déduite du mélange et la partition donnée a priori $\mathbf{u}$.

# 3 Le critère de sélection de modèle

Dans l'idéal, on cherche une partition $\mathbf{z}$ telle que $\mathbf{y}$ et $\mathbf{u}$ soient *conditionnellement* indépendantes sachant $\mathbf{z}$ et maximisant la probablité de la loi jointe de $(\mathbf{y}, \mathbf{u}, \mathbf{z})$ pour le modèle de mélange choisi. Ainsi sous l'hypothèse que $\mathbf{y}$ et $\mathbf{u}$ sont conditionnellement indépendantes sachant $\mathbf{z}$, on cherche le modèle de mélange $m$ (caractérisée par sa forme et son nombre de composants) maximisant la vraisemblance intégrée complétée

$$p(\mathbf{y}, \mathbf{u}, \mathbf{z}|m) = \int p(\mathbf{y}, \mathbf{u}, \mathbf{z}|m, \theta_m)\pi(\theta_m)d\theta_m$$

$\theta_m$ étant le paramètre vectoriel caractérisant le modèle $m$ et $\pi(\theta_m)$ la loi a priori de ce paramètre. Grâce à l'hypothèse d'indépendance conditionelle, on peut écrire

$$p(\mathbf{y}, \mathbf{u}, \mathbf{z}|m) = p(\mathbf{y}, \mathbf{z}|m)p(\mathbf{u}|\mathbf{z}, m)$$

avec

$$p(\mathbf{y}, \mathbf{z}|m) = \int p(\mathbf{y}, \mathbf{z}|m, \theta_m)\pi(\theta_m)d\theta_m.$$

De plus, on a $p(\mathbf{u}|\mathbf{z}, m) = p(\mathbf{u}|\mathbf{z})$ puisque par définition la loi de $\mathbf{u}$ est indépendante du modèle $m$. L'estimation de la loi conditionnelle $p(\mathbf{u}|\mathbf{z})$ est obtenu directement à partir du tableau de contingence $(n_{kl})$ issu du croisement des partitions $\mathbf{u}$ et $\mathbf{z}$.

Maintenant, pour estimer $p(\mathbf{y}, \mathbf{z}|m)$, on utilise une approximation à la BIC comme il est fait dans Biernacki *et al.* (2000) pour définir le critère ICL qui permet de choisir un modèle de mélange dans un objectif de classification. Cela conduit au critère

$$SICL(m) = \mathrm{ICL}(m) + \sum_{l=1}^{U_{max}} \sum_{k=1}^{K_{max}} n_{kl} \log \frac{n_{kl}}{n_{k\cdot}},$$

le dernier terme à droite étant une pénalité supplémentaire au critère ICL d'autant plus faible que les partitions $\mathbf{u}$ et $\mathbf{z}$ sont liées.

# 4 Discussion

Nous présenterons des expérimentations illustrant l'intérêt de notre critère. De plus, nous commenterons notre choix de ne pas tenir compte de la partition a priori $\mathbf{u}$ dans la phase de construction de la partition $\mathbf{z}$. Nous verrons que d'autres choix sont possibles en remplaçant le contraste associé au maximum de vraisemblance par un autre contraste (Baudry *et al.* 2008) pour tenir compte de l'objectif de l'utilisateur dès la phase d'estimation et nous analyserons les avantages et les inconvénients de ces deux approches concurrentes.

## Bibliographie

[1] Baudry, J.-P., Celeux, G. and Marin, J.-M. (2008) Selecting models focussing on the modeller's purpose, *COMPSTAT 2008 : Proceedings in Computational Statistics* (P. Brito, Ed.), Physica-Verlag, Heidelberg, pp. 337-348

[2] Biernacki, C., Celeux, G., Govaert, G. (2000) Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719-725.

[3] Biernacki, C., Celeux, G., Govaert, G. et Langrognet, F. (2006) Le logiciel MIXMOD d'analyse de mélange pour la classification et l'analyse discriminante. *La Revue de Modulad*, 35, 25-44.

[4] Govaert, G. (2003). Classification et modèle de mélange. *Analyse de données*, Hermès, Paris, pp.263-292.