Pourquoi les modèles de mélange pour la classification ?

Christophe Biernacki

CNRS & Université de Lille 1, Villeneuve d'Ascq, France
biernack@math.univ-lille1.fr

Résumé Les modèles de mélange apportent une réponse rigoureuse, flexible et interprétable pour les multiples besoins de la classification : classification supervisée ou non, nature des données, choix du nombre de groupes, *etc.* Les domaines d'applications sont de plus en plus nombreux, aidés en cela par le développement de solutions logicielles adaptées.

Mots clés : classification supervisée, classification non supervisée, algorithme EM, choix de modèles.

Abstract Mixture models provide a mathematical-based, flexible and meaningful approach for the wide variety of classification requirements: Unsupervised or supervised classification, data features, number of classes selection, *etc.* Fields in which mixture models have been successfully applied are numerous and specific softwares are now available.

Keywords: Classification, clustering, EM algorithm, model selection.

1 Introduction

La classification a pris aujourd'hui une place importante en analyse des données *exploratoire* et *décisionnelle*, tant au niveau des domaines d'applications que des développements méthodologiques. L'objectif exploratoire est typiquement représenté par la classification dite *non supervisée*, ou classification automatique, qui vise à découvrir une partition hypothétique dans un ensemble d'objets. Le but décisionnel quant à lui se réfère plutôt à la classification *supervisée*, ou analyse discriminante, qui cherche généralement à affecter tout nouvel objet à des groupes préalablement définis. Pour cela, il est donc nécessaire de disposer de méthodes s'adaptant à la diversité des données à analyser comme l'abondance, ou au contraire la rareté, des objets (*individus*) disponibles ainsi que le nombre et/ou le type de descripteurs (*variables*) pour chacun d'entre eux.

Grâce à leur flexibilité, les mélanges finis de distributions de probabilité répondent à ces exigences (voir [36] et les nombreuses références associées). Ils sont devenus aujourd'hui un outil populaire et utilisé avec succès dans un nombre croissant de disciplines comme l'astronomie, la biologie, la génétique, l'économie, les sciences de l'ingénieur, le marketing, la reconnaissance d'images...En outre, le fait que des logiciels comme MIXMOD[1] [5] mettent à disposition ces méthodes sous forme de code performant et portable

[1]Site web du logiciel MIXMOD : http ://www-math.univ-fcomte.fr/mixmod/index.php

aide largement à leur diffusion. D'un point de vue théorique, les modèles de mélange permettent de bénéficier de l'ensemble des résultats de la statistique mathématique : lois multivariées paramétriques, estimation, choix de modèles. Ils permettent aussi de retrouver, voire de généraliser, de nombreuses méthodes de classification classiques non probabilistes à la base car plutôt géométriques. Ils font partie de la famille des modèles *génératifs* puisque l'ensemble du processus ayant généré les données (observées et manquantes) est totalement modélisé. Ce cadre formel apporte clairement un confort d'utilisation pour le praticien qui obtient sans effort supplémentaire un résumé exhaustif et généralement très interprétable de sa structure de classification au travers des paramètres du modèle. En outre, la rigidité apparente de cette modélisation est efficacement compensée par la mise en œuvre de méthodes de choix de modèles. Cependant, bien que cela soit possible, il serait dangereux de résumer la problématique de la classification à un simple sous-produit d'un modèle de mélange. En effet, il peut être efficace d'adjoindre aussi bien dans la phase d'estimation que de choix de modèles des informations supplémentaires en rapport avec l'objectif initial recherché. Cette optique est par ailleurs source de nombreux développements méthodologiques spécifiques récents.

Dans la suite de l'exposé, la Section 2 est dévolue à la présentation du modèle de mélange de lois de probabilités, à l'estimation de ses paramètres et aux méthodes habituelles de choix de modèles. La section suivante abordera la problématique de la classification non supervisée avec ce type de modèle. En particulier, on soulignera le lien avec des méthodes géométriques classiques dans le cas gaussien et on présentera aussi des critères de choix de modèles spécifiques. Une illustration numérique finalisera cette partie. La Section 4 reprend le même scénario pour la classification supervisée cette fois. Enfin, la Section 5 dresse un bilan de cette présentation ainsi que des nombreux nouveaux défis qui attendent les méthodes de classification (données en haute dimension, contexte semi-supervisé, choix de variables, ...) et les perspectives de réponses dans un cadre de modèle de mélange.

2 Modèle de mélange

2.1 Présentation générale

2.1.1 Définition

Une loi de mélange fini p sur un espace $\mathcal{X}$ est une loi de probabilité s'exprimant comme une combinaison linéaire de plusieurs lois de probabilité $\mathrm{p}_1, \ldots, \mathrm{p}_g$ sur $\mathcal{X}$. Autrement dit, il existe g coefficients $\pi_1, \ldots, \pi_g$ ($\pi_k > 0$ et $\sum_{k=1}^{g} \pi_k = 1$) tels que, pour tout $\mathbf{x}_1 \in \mathcal{X}$,

$$\mathrm{p}(\mathbf{x}_1) = \sum_{k=1}^{g} \pi_k \mathrm{p}_k(\mathbf{x}_1). \tag{1}$$

Les π_k et p_k sont respectivement appelées *proportions* et *composantes* du mélange.

2.1.2 Interprétation générative

La loi mélange p peut aussi s'interpréter comme la loi marginale de la variable aléatoire $\mathbf{X}_1$ obtenue à partir de la loi du couple de variables aléatoires $(\mathbf{X}_1, \mathbf{Z}_1)$ où

$$\mathbf{Z}_1 \sim \mathcal{M}_g(1, \pi_1, \ldots, \pi_g) \quad \text{et} \quad \mathbf{X}_1 | \mathbf{Z}_1 = \mathbf{z}_1 \sim \mathrm{p}_{\{k:z_{1k}=1\}}, \tag{2}$$

avec $\mathcal{M}_g(1, \pi_1, \ldots, \pi_g)$ la loi multinomiale de dimension g et d'ordre 1 de paramètre $(\pi_1, \ldots, \pi_g)$ et $\mathbf{z}_i = (z_{i1}, \ldots, z_{in})$ un vecteur binaire de dimension g. Cette interprétation permet de générer en deux étapes très simples un échantillon $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ tels que les $\mathbf{x}_i$ soient des réalisations i.i.d. de $\mathbf{X}_1$ $(i = 1, \ldots, n)$:

1. Générer un échantillon $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ tel que les $\mathbf{z}_i$ soient des réalisations i.i.d. de $\mathbf{Z}_1$;
2. Générer un échantillon $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ tel que les $\mathbf{x}_i$ soient des réalisations indépendantes de $\mathrm{p}_{\{k:z_{ik}=1\}}$.

Il est important de remarquer pour la suite que les vecteurs $\mathbf{z}_i$ identifient la composante p_k à partir de laquelle est ensuite généré chaque $\mathbf{x}_i$. On parlera parfois d'*étiquettes*.

2.1.3 L'exemple gaussien et multinomial

Généralement, on suppose en outre que chaque composante p_k appartient à une famille paramétrée $\mathrm{p}(.; \boldsymbol{\alpha}_k)$ et on note $\mathrm{p}(.; \boldsymbol{\theta})$ la loi mélange associée à ce paramétrage, $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ désignant le paramètre de ce modèle. Le choix se restreint à des familles $\mathrm{p}(.; \boldsymbol{\alpha}_k)$ conduisant à des lois mélanges identifiables, ou tout au moins qui le sont dans la plupart des situations d'intérêt. Il existe aussi des choix semi-paramétriques ou non paramétriques pour les p_k (voir [10]) mais nous ne les traiterons pas ici.

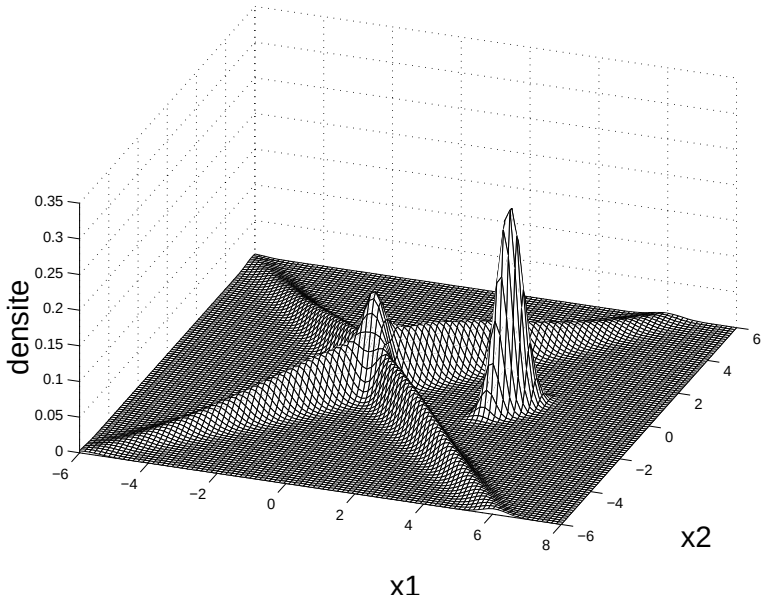


FIG. 1 – Un mélange gaussien dans $\mathbb{R}^2$ avec trois composantes.

Dans le cas continu où $\mathcal{X} = \mathbb{R}^d$, le modèle paramétrique le plus utilisé est la loi multinormale :

$$p(.; \boldsymbol{\alpha}_k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3}$$

avec $\boldsymbol{\alpha}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ désignant la moyenne de la composante k et $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$ la matrice de variance correspondante. La figure 1 donne une illustration d'un mélange gaussien dans $\mathbb{R}^2$ avec trois composantes.

Dans le cas qualitatif, $\mathcal{X} = \{1, \ldots, m_1\} \times \ldots \times \{1, \ldots, m_d\}$ correspond à d variables catégorielles chacune possédant m_j modalités ($j = 1, \ldots, d$) et les données $\mathbf{x}$ sont codées par le tableau *disjonctif complet* $n \times p$ suivant :

$$x_{ijh} = \begin{cases} 1 & \text{si l'individu } i \text{ prend la modalité } h \text{ pour la variable } j \\ 0 & \text{sinon,} \end{cases} \tag{4}$$

où $h = 1, \ldots, m_j$. Dans ce contexte, il est naturel de modéliser la loi de la variable qualitative $\mathbf{X}_{1j}$ de la composante k par une loi multinomiale $\mathcal{M}_{m_j}(1, \alpha_{kj1}, \ldots, \alpha_{kjm_j})$, où $\alpha_{kjh} = \mathrm{p}(X_{1jh} = 1 | Z_{1k} = 1)$. Le *modèle des classes latentes* [26] fait l'hypothèse supplémentaire que les variables qualitatives $\mathbf{X}_{1j}$ sont indépendantes conditionnellement à chaque composante k, d'où l'expression de la loi de chaque composante :

$$\mathrm{p}(\mathbf{x}_1; \boldsymbol{\alpha}_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_{kjh})^{x_{1jh}}, \tag{5}$$

avec $\boldsymbol{\alpha}_k = (\alpha_{kjh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$.

Notons que bien d'autres modélisations existent par ailleurs comme la loi de poisson, la loi exponentielle, la loi de Student multivariée, *etc.* (Voir [36].)

2.2 Estimation des paramètres par EM

2.2.1 Principe

En statistique, le paramètre $\boldsymbol{\theta}$ est inconnu et l'objectif est de l'estimer à partir de l'échantillon $\mathbf{x}$. Pour cela, on peut maximiser la log-vraisemblance donnée par :

$$L(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{g} \pi_k \mathrm{p}(\mathbf{x}_i; \boldsymbol{\alpha}_k) \right). \tag{6}$$

On parle aussi de log-vraisemblance *observée* car calculée uniquement sur les données observées $\mathbf{x}$.

Cependant, l'optimisation directe sur $\boldsymbol{\theta}$ est difficile : il n'y a généralement pas de solution analytique et les algorithmes généralistes d'optimisation de type Newton peuvent être compliqués à mettre en œuvre (calcul du hessien par exemple) et d'une efficacité toute relative dans ce cadre particulier. L'algorithme EM (*Espérance-Maximisation*) est un algorithme spécifique à la maximisation de la vraisemblance dans le cadre de données manquantes qui conduit à de bons résultats en pratique [21]. Sa mise en œuvre est très simple à partir du moment où la maximisation de la log-vraisemblance *complétée* qui serait calculée sur l'ensemble des données $(\mathbf{x}, \mathbf{z})$ issues du processus sous-jacent de génération aléatoire de $\mathbf{x}$ serait facile, ce qui est le cas pour nombre de lois paramétriques usuelles retenues pour les composantes $\mathrm{p}(.; \boldsymbol{\alpha}_k)$. On note cette log-vraisemblance par

$$L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \ln \left(\pi_k \mathrm{p}(\mathbf{x}_i; \boldsymbol{\alpha}_k) \right). \tag{7}$$

Ainsi, partant d'un paramètre initial $\boldsymbol{\theta}^-$ arbitraire, l'algorithme itère sur les étapes suivantes :

– Étape E (espérance) : calculer $t_{ik}^{+} = \mathrm{E}[Z_{ik}|\mathbf{x}_i; \boldsymbol{\theta}^{-}]$;
– Étape M (maximisation) : calculer $\boldsymbol{\theta}^{+} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{t}^{+})$ où $\{\mathbf{t}^{+}\}_{ik} = t_{ik}^{+}$.
L'algorithme s'arrête après un nombre prédéfini d'itérations ou bien à la stationarité du critère de log-vraisemblance observée. Cette seconde option découle de la propriété de croissance de $L(\boldsymbol{\theta}; \mathbf{x})$ à chaque itération.

2.2.2 Interprétation

Il est informatif de remarquer que l'espérance t_{ik} calculée à l'étape E peut aussi s'interpréter comme la probabilité conditionnelle que l'individu $\mathbf{x}_i$ ait été généré par la composante k :

$$t_{ik} = \mathrm{p}(Z_{ik} = 1|\mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\theta}) = \frac{\pi_k \mathrm{p}(\mathbf{x}_i; \boldsymbol{\alpha}_k)}{\mathrm{p}(\mathbf{x}_i; \boldsymbol{\theta})}. \tag{8}$$

Ainsi, les étiquettes manquantes $\mathbf{z}$ indiquant la composante d'origine des $\mathbf{x}_i$ ont simplement été remplacées par des probabilités de provenir de chacune de ces composantes. Cela explique aussi la simplicité de l'étape M en pratique car ces probabilités servent simplement de poids à chaque individu dans la phase d'estimation. Par exemple, les proportions du mélange obtenues à l'étape M sont données par les fréquences suivantes :

$$\pi_k^{+} = \frac{n_k^{+}}{n}, \tag{9}$$

avec $n_k^{+} = \sum_{i=1}^{n} t_{ik}^{+}$ l'effectif « flou » provenant de la composante k.

L'estimation des paramètres $\boldsymbol{\alpha}_k$ des composantes dépend quant à elle des lois retenues mais ne pose généralement pas plus de difficulté. Ainsi, dans le cas gaussien, on obtient

$$\boldsymbol{\mu}_k^{+} = \frac{1}{n_k^{+}} \sum_{i=1}^{n} t_{ik}^{+} \mathbf{x}_i \quad \text{et} \quad \boldsymbol{\Sigma}_k^{+} = \frac{1}{n_k^{+}} \sum_{i=1}^{n} t_{ik}^{+} (\mathbf{x}_i - \boldsymbol{\mu}_k^{+})(\mathbf{x}_i - \boldsymbol{\mu}_k^{+})', \tag{10}$$

et dans le cas multinomial d'indépendance conditionnelle

$$\alpha_{kjh}^{+} = \frac{1}{n_k^{+}} \sum_{i=1}^{n} t_{ik}^{+} x_{ijh}. \tag{11}$$

2.2.3 Propriétés

On a déjà signalé que chaque itération de EM fait croître la log-vraisemblance. Il converge en outre vers un point stationnaire de la vraisemblance qui peut être typiquement le maximum global, un maximum local ou un point selle. Pour une recherche plus efficace du maximum global, il est donc préférable de faire plusieurs essais en changeant les paramètres initiaux, puis de ne retenir que la solution donnant la plus grande vraisemblance. C'est aussi un algorithme qui converge linéairement, donc il peut souffrir d'une certaine lenteur au voisinage de la stationnarité.

On retrouvera nombre de propriétés plus détaillées sur EM dans [35] ou encore des propositions de stratégies pour bien initialiser l'algorithme [7]. Enfin, on pourra utilement se reporter à l'algorithme SEM [14], variante stochastique de EM, pour obtenir sans estimation de hessien une évaluation de la variabilité des estimateurs.

2.3 Choix de modèles

2.3.1 Problématique

Il est cependant fréquent que la structure de modélisation d'où sont issues les données $\mathbf{x}$ soit elle-même incertaine. Par exemple, on ignore le nombre g de composantes du modèle et ce nombre doit donc être lui aussi estimé. De façon générique, un modèle $\mathbf{m}$ représente un ensemble de contraintes sur l'espace du paramètre $\boldsymbol{\theta}$.

Il faut garder à l'esprit par ailleurs que le modèle de mélange a pu être simplement utilisé pour estimer une loi de probabilité inconnue grâce à sa grande flexibilité et donc qu'il n'y a pas de « vrai » modèle parmi ceux proposés. Cela oblige alors à définir autrement la notion même de « bon » modèle. C'est sur ce point que les approches divergent en général.

2.3.2 Deux critères de référence

Pour choisir parmi plusieurs modèles en compétition, on peut utiliser des critères très classiques en statistique mathématique qui s'expriment généralement comme une pénalisation de la log-vraisemblance maximale de la forme :

$$^*\mathrm{IC}_{\mathbf{m}} = L(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \mathbf{x}) - \nu_{\mathbf{m}}.f(n), \tag{12}$$

où $\hat{\boldsymbol{\theta}}_{\mathbf{m}}$ correspond à l'estimateur du maximum de la vraisemblance sous le modèle $\mathbf{m}$, $\nu_{\mathbf{m}}$ donne le nombre de paramètres à estimer et f est une fonction de n. On retient alors le modèle ayant obtenu la plus grande valeur du critère. Fondamentalement, ce type de critère tente de réaliser un compromis entre l'adéquation du modèle aux données mesuré par la log-vraisemblance maximale et la complexité de celui-ci, mesuré par sa dimension. Le point de vue adopté pour ajuster ce compromis permettra de définir la fonction f :

- Du point de vue fréquentiste, un « bon » modèle est celui qui réalise un compromis en terme de « biais-variance ». Dans ce cadre, le critère AIC (*An Information Criterion*) [1] est proposé avec $f(n) = 1$. La pénalité ne dépend donc pas de n.
- Du point de vue bayésien, un « bon » modèle est celui qui maximise la vraisemblance intégrée (sur les paramètres) lorsque chaque modèle en compétition est équiprobable *a priori*. Le critère BIC (*Bayesian Information Criterion*) [39] propose alors la pénalité $f(n) = 0.5 \ln(n)$, dépendant cette fois de n.

2.3.3 Propriétés

Tous deux sont des critères dits asymptotiques en le sens que les propriétés idéales dont ils sont issues sont atteintes seulement asymptotiquement. Ainsi, asymptotiquement, AIC retiendra le modèle minimisant l'écart de Kullback *moyen* avec la vraie loi inconnue tandis que BIC sélectionnera le modèle minimisant l'écart de Kullback avec la vraie loi. En ce sens, BIC est convergeant si le vrai modèle est dans la liste [31].

Il faut noter que les arguments de construction et de propriétés de AIC et BIC sont affaiblis dans le cas du choix de g pour les mélanges (voir une discussion détaillée à ce sujet dans [36], Chap. 6). En pratique, les deux critères donnent cependant de bons résultats, avec une préférence assez marqués pour BIC par les utilisateurs, AIC favorisant souvent des modèles trop complexes.

3 Classification non supervisée

3.1 Problématique

3.1.1 Données et objectif

L'objectif de la *classification non supervisée*, appelée aussi classification automatique, est de partitionner l'ensemble des n objets de $\mathbf{x}$ en g classes $G_1, \ldots, G_g$. D'autres structures que la partition peuvent aussi être recherchées, par exemple les hiérarchies qui sont des emboîtements de partitions, mais ces situations ne seront pas considérées ici. On peut également désigner la partition estimée sous forme d'indicatrices avec le tableau binaire $\hat{\mathbf{z}}$ et on a bien sûr l'équivalence suivante entre les notations :

$$\mathbf{x}_i \in G_k \Leftrightarrow \hat{z}_{ik} = 1. \tag{13}$$

L'objectif de la classification non supervisée est illustré sur la figure 2 dans le cas bivarié pour trois classes recherchées.

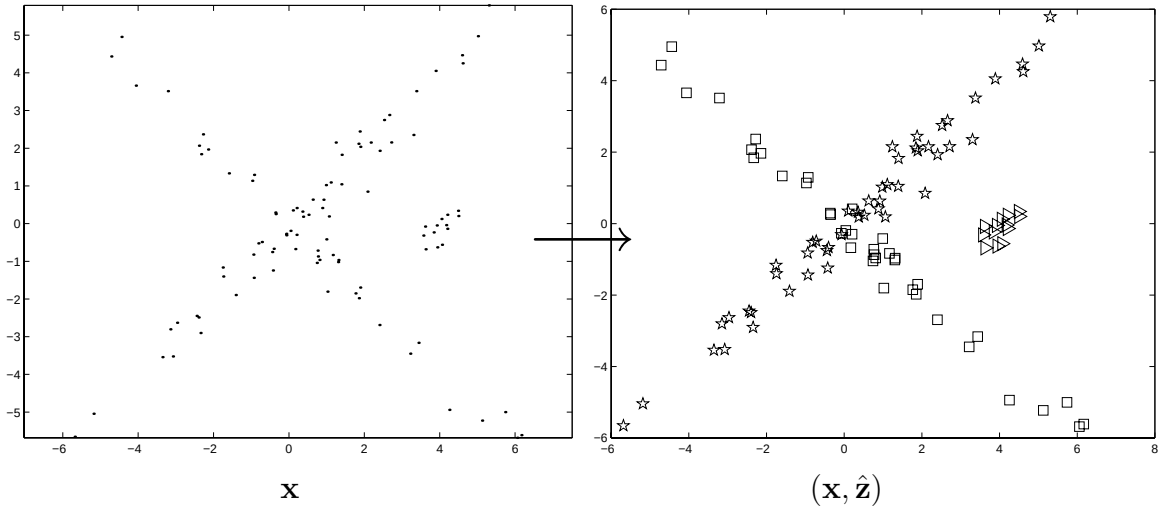


FIG. 2 – Classification non supervisée : illustration de l'objectif pour des données $\mathbf{x}$ de $\mathbb{R}^2$ et une partition estimée $\hat{\mathbf{z}}$ en trois classes.

Cet objectif n'est pas seulement formel. Il faut comprendre que l'intérêt sous-jacent de la démarche de classification est d'aider le praticien à analyser des données. Le regroupement en classes est pour lui une façon de synthétiser pour isoler l'information pertinente. En effet, il est difficile pour un humain d'appréhender directement cette information en présence de données parfois nombreuses, éventuellement décrites par de multiples dimensions dans des espaces eux-mêmes un peu complexes. La classification automatique est donc généralement une procédure mathématique qui propose une ou plusieurs partitions parmi la multitudes de possibilités offertes *a priori*.

Au-delà de cet objectif de base, toute aide supplémentaire permettant de faciliter l'interprétation des partitions elles-mêmes est la bienvenue pour le praticien. Il n'est pas difficile à imaginer en effet qu'il calculera des statistiques sur la partition obtenue pour mieux la comprendre. Par exemple, une simple moyenne par classe fera apparaître la classe « des grands » en opposition à la classe « des petits ». Ainsi, outre le fait d'obtenir

une partition, une modélisation plus abstraite de la notion de classe aidera certainement pour l'interprétation ultérieure. Cela fait partie des avantages qu'apportent les modèles de mélange comme nous allons le voir maintenant.

3.1.2 Modélisation par les mélanges

Dans sa présentation générale, la classification non supervisée est typiquement un problème *mal posé* car sa mise en œuvre repose implicitement sur l'évaluation de la « qualité » d'une partition $\hat{\mathbf{z}}$ et pourtant aucun critère objectif n'est spécifié au départ. La résolution du problème de classification passe donc par la définition d'un tel critère. Une définition précise en sera donnée dans la section 3.2 mais nous nous concentrons pour le moment sur la modélisation des classes.

Remarquons tout d'abord que, dans le processus d'acquisition des données, les objets étudiés sont généralement tirés aléatoirement dans une certaine population. Ainsi, il est naturel d'interpréter ces objets comme la réalisation d'une expérience probabiliste. Dans ce cadre, les modèles de mélange finis apportent alors une *hypothèse simplificatrice* sur la distribution de la population en faisant ressortir explicitement une structure en sous-populations, hypothèse proche de l'objectif de la classification. En effet, une définition précise est donnée de cette façon à la notion de classe puisque des individus appartiennent au même ensemble si et seulement si ils sont tirés dans la même sous-population [9].

Le problème de classification initial se ramène donc à estimer la partition $\mathbf{z}$ qui a permis la réalisation de l'échantillon $\mathbf{x}$ de façon i.i.d. suivant une loi mélange p. Ce pose alors maintenant le choix des composantes p_k associées.

Ce choix dépend d'une part de la nature des données (continues, catégorielles, ou autres). D'autre part, c'est aussi l'occasion de fournir au praticien une modélisation *interprétable* et *réduite* de la classe. Le cadre paramétrique atteint naturellement ces deux objectifs. En effet, pour les données continues, le choix d'une loi multinormale est particulièrement informatif pour l'utilisateur : la moyenne $\boldsymbol{\mu}_k$ définit la position centrale de la classe k et la matrice de variance $\boldsymbol{\Sigma}_k$ informe sur sa dispersion autour de ce centre. On peut ainsi facilement en déduire une zone (ellipsoïde) de confiance autour du centre. Donc, en peu de paramètres, la classe est résumée. Pour les données catégorielles, le modèle des classes latentes décrit explicitement les classes par la probabilité α_{kjh} de chacune des modalités, variable par variable. Dans tous les cas, les proportions π_k apportent de plus une information directement interprétable : l'effectif relatif de chaque classe.

3.1.3 Géométrie d'une classe : modèles parcimonieux gaussiens

On vient de voir l'intérêt de la modélisation paramétrique d'une classe pour aider à son interprétation. Dans le cas gaussien, des caractéristiques géométriques fines de la classe ellipsoïdale de centre $\boldsymbol{\mu}_k$ peuvent être en outre contrôlées grâce à une décomposition spectrale de la matrice de variance $\boldsymbol{\Sigma}_k$.

Suivant [3, 18], chaque matrice de variance des composantes du mélange peut s'écrire :

$$\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k' \tag{14}$$

avec $\lambda_k = |\boldsymbol{\Sigma}_k|^{1/d}$, $\mathbf{D}_k$ étant la matrice des vecteurs propres de $\boldsymbol{\Sigma}_k$ et $\mathbf{A}_k$ étant une matrice diagonale, telle que $|\mathbf{A}_k| = 1$, dont la diagonale est constituée des valeurs propres normalisées de $\boldsymbol{\Sigma}_k$ rangées en ordre décroissant. Le paramètre λ_k caractérise le *volume*

de la k^{e} classe, $\mathbf{D}_k$ son *orientation* et $\mathbf{A}_k$ sa *forme*. Une illustration dans le plan de la géométrie d'une classe est donnée sur la figure 3. En permettant à ces paramètres de varier ou non entre les classes, on obtient des modèles plus ou moins parcimonieux. En tout, [18] proposent quatorze modèles différents en permettant à tout ou partie des termes géométriques de varier ou non entre classes et en les associant aussi à des modèles très simples comme des matrices de variances diagonales (les $\mathbf{D}_k$ sont des matrices de permutation) ou encore sphériques ($\mathbf{A}_k = \mathbf{I}$).

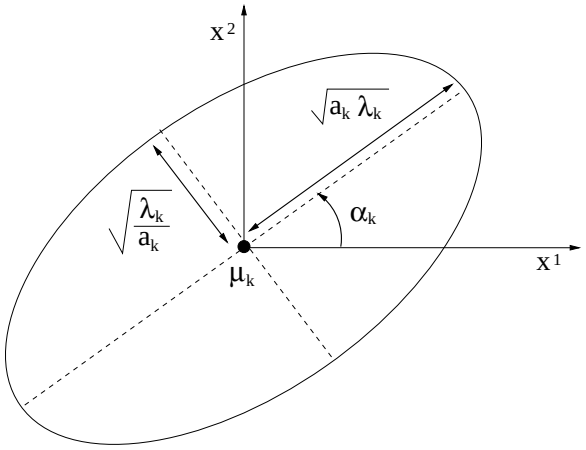


FIG. 3 – Illustration géométrique dans le plan du paramétrage de l'ellipse isodensité de la classe k par la décomposition spectrale de $\mathbf{\Sigma}_k$ avec $\mathbf{A}_k$ une matrice diagonale de termes diagonaux a et $1/a$, et $\mathbf{D}_k$ une matrice de rotation d'angle α.

Ainsi, chaque géométrie particulière, représentée par chaque modèle, permet au praticien de caractériser simplement non seulement les classes mais aussi leurs similitudes éventuelles, et ainsi d'aider à leur interprétation.

3.1.4 Centre et dispersion dans le cas multinomial

Les notions de centre et de dispersion d'une classe sont très intuitives pour appréhender rapidement les caractéristiques d'une classe et les différences entre plusieurs d'entre elles. Par analogie au cas gaussien, ces notions ont été introduites dans le cas Bernoulli pour données binaires [15] puis étendues au cas multinomial pour données qualitatives générales ([27], Chap. 9).

Le principe est de contraindre tous les vecteurs $(\alpha_{kj1}, \ldots, \alpha_{kjm_j})$ à prendre la forme

$$\left(\frac{\varepsilon_{kj}}{m_j - 1}, \frac{\varepsilon_{kj}}{m_j - 1}, .., \frac{\varepsilon_{kj}}{m_j - 1}, 1 - \varepsilon_{kj}, \frac{\varepsilon_{kj}}{m_j - 1}, ..., \frac{\varepsilon_{kj}}{m - 1}\right) \tag{15}$$

avec $\varepsilon_{kj} < \frac{m_j - 1}{m_j}$. Les vecteurs des probabilités sont alors simplement caractérisés par une modalité majoritaire (le mode ou le centre) et un terme de dispersion ε_k^j qui représente la probabilité d'en être différent.

Comme pour le modèle de mélange gaussien, il est possible d'imposer des contraintes supplémentaires sur la dispersion : elle peut être indépendante de la variable, de la classe, voire d'aucun des deux (voir de nouveau [27], Chap. 9).

3.2 Estimation d'une partition

3.2.1 Approche « mélange »

Ayant immergé la problématique de classification dans un modèle de mélange, il s'agit maintenant d'en tirer partie pour estimer $\mathbf{z}$. Dans un premier temps, l'algorithme EM permet d'obtenir une estimation $\hat{\boldsymbol{\theta}}$ du paramètre de mélange sous-jacent. Notons que l'étape M sera spécifique au modèle considéré, en particulier pour les modèles parcimonieux gaussiens et multinomiaux précédents tout en gardant en général une grande simplicité de calcul (voir [18] et [27], Chap.9). On en déduit ensuite une estimation $\hat{\mathbf{t}}$ des probabilités conditionnelles d'appartenance aux composantes du mélange, donc aux classes. Rappelons que $\{\hat{\mathbf{t}}\}_{ik} = \hat{t}_{ik}$ avec $\hat{t}_{ik} = \mathrm{p}(Z_{ik} = 1|\mathbf{X}_i = \mathbf{x}_i; \hat{\boldsymbol{\theta}})$. Une partition $\hat{\mathbf{z}}$ est alors obtenue par le principe du MAP (*Maximum A Posteriori*), c'est-à-dire en affectant chaque individu $\mathbf{x}_i$ à la classe de plus grande probabilité conditionnelle estimée :

$$\hat{\mathbf{z}} = \mathrm{MAP}(\hat{\mathbf{t}}) \quad \Leftrightarrow \quad \hat{z}_{ik} = \begin{cases} 1 & \text{si } k = \arg\max_{k'=1,\ldots,g} \hat{t}_{ik'} \\ 0 & \text{sinon.} \end{cases} \tag{16}$$

Remarquons, qu'au delà d'une estimation $\hat{\mathbf{z}}$ de la partition, le modèle fournit aussi deux informations supplémentaires intéressantes pour le praticien :

- une estimation $\hat{\boldsymbol{\theta}}$ des paramètres dont nous savons maintenant toute l'importance pour aider à l'interprétation des classes ;
- une estimation $\hat{\mathbf{t}}$ des probabilités conditionnelles d'appartenance aux classes, ce qui permet d'évaluer le *risque de classement* de chaque individu.

3.2.2 Approche « classification »

L'approche mélange précédente considère le problème de classification uniquement comme un sous-produit du modèle de mélange. Cependant, il n'est pas pris en considération explicitement l'objectif final de la classification qui est de faciliter l'analyse des données (voir sous-section 3.1.1). Par exemple, l'approche mélange peut tout à fait conduire à estimer des classes très imbriquées, c'est-à-dire dont les paramètres sont très similaires. Dans ce cas, le praticien aura du mal à interpréter ce qui caractérise des classes si imbriquées et aura le penchant naturel de les regrouper. Cela signifie que la partition qui lui a été fournie n'est pas complètement pertinente de son point de vue.

L'approche classification introduite par [40] peut être vue comme une méthode permettant d'obtenir explicitement des classes peu imbriquées comme nous le détaillons maintenant. Pour toute partition $\mathbf{z}$, on obtient la décomposition suivante de la log-vraisemblance en une log-vraisemblance complétée et un terme entropique [29] :

$$L(\boldsymbol{\theta}; \mathbf{x}) = L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) + E(\mathbf{t}, \mathbf{z}) \tag{17}$$

avec $E(\mathbf{t}, \mathbf{z}) = -\sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \ln t_{ik}$ un terme d'entropie de partition mesurant l'écart entre la partition $\mathbf{z}$ et les probabilités d'appartenance $\mathbf{t}$. Idéalement, obtenir des classes totalement non imbriquées revient à obtenir des probabilités d'appartenance $\mathbf{t}$ telles que $E(\mathbf{t}, \mathbf{z}) = 0$ pour une partition $\mathbf{z}$ particulière. Sous cette contrainte, la relation (17) indique qu'il est équivalent de maximiser la log-vraisemblance $L(\boldsymbol{\theta}; \mathbf{x})$ en $\boldsymbol{\theta}$ ou la log-vraisemblance complétée $L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})$ sur le couple $(\boldsymbol{\theta}, \mathbf{z})$. Passant outre la contrainte sur $\boldsymbol{\theta}$ imposée par $E(\mathbf{t}, \mathbf{z}) = 0$, l'approche classification se focalise sur la maximisation de la log-vraisemblance complétée $L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})$ pour le couple $(\boldsymbol{\theta}, \mathbf{z})$.

En pratique, $L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})$ est optimisée par l'agorithme CEM [16], sorte de version classification de l'algorithme EM. Partant d'un paramètre initial $\boldsymbol{\theta}^-$ arbitraire, l'algorithme CEM itère sur les étapes suivantes :

- Étape E (espérance) : calculer t_{ik}^+ comme EM ;
- Étape C (classification) : calculer la partition $\mathbf{z}^+ = \text{MAP}(\mathbf{t}^+)$;
- Étape M (maximisation) : calculer $\boldsymbol{\theta}^+ = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}^+)$.

L'étape M est très similaire à celle de EM puisqu'il s'agit en pratique de remplacer les $\mathbf{t}^+$ par les $\mathbf{z}^+$ dans les formules permettant d'obtenir $\boldsymbol{\theta}^+$. De plus, l'algorithme est stationnaire et fait croître à chaque itération la vraisemblance complétée. On peut donc l'arrêter à la stabilité de la partition. En pratique, sa convergence est beaucoup plus rapide que celle de EM ce qui peut être utile dans des situations où le temps d'estimation a une importance.

On ne s'étonnera pas que les paramètres $\tilde{\boldsymbol{\theta}}$ estimés par CEM soient biaisés, même asymptotiquement, ce biais étant d'autant plus fort que les composantes sous-jacentes du mélange sont imbriquées [13]. Cependant, l'approche classification peut donner de meilleurs résultats sur l'estimation du paramètre et aussi sur l'estimation de la partition dans le cas de petits échantillons. En effet, dans ce cas, rechercher explicitement une partition semble apporter une information utile pour réaliser le nécessaire compromis biais-variance [17].

Notons enfin, que cette approche ne prive pas l'utilisateur des sous-produits habituels d'une modélisation par mélange :

- des paramètres $\tilde{\boldsymbol{\theta}}$ pour aider à l'interprétation des classes ;
- des probabilités d'appartenance aux classes $\tilde{\mathbf{t}}$ pour évaluer le risque de classement ;
- une partition $\tilde{\mathbf{z}} = \text{MAP}(\tilde{\mathbf{t}})$.

3.2.3 Lien avec des approches à base de distances

Il est possible de faire des liens entre l'approche classification précédente et certaines méhodes de classification standards proposées antérieurement dans un cadre non probabiliste et s'appuyant plutôt sur des métriques. Ce type de rapprochement permet de révéler des hypothèses qui a l'origine n'étaient pas explicites. Il est alors envisageable d'étendre ces méthodes géométriques on utilisant toute la variété des modèles parcimonieux.

Ainsi, dans le cadre des données continues, la méthode des *centres mobiles* [43] sélectionne la partition $\mathbf{z}$ maximisant le critère d'inertie intraclasse trace $(\mathbf{W}(\mathbf{z}))$ où $\mathbf{W}(\mathbf{z})$ désigne la matrice d'inertie intraclasse donnée par

$$\mathbf{W}(\mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)' \tag{18}$$

avec $\bar{\mathbf{x}}_k$ la moyenne arithmétrique des individus de la classe G_k. Il s'agit précisément de la partition retenue par maximum de vraisemblance complétée avec proportions égales et un modèle gaussien sphérique de même volume entre composantes. De plus, l'algorithme CEM est exactement l'algorithme des centres mobiles. De la même façon, le critère $|\mathbf{W}(\mathbf{z})|$ [24] s'identifie au modèle à proportions égales et avec modèle gaussien homoscédastique ($\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_g$). D'autres correspondances sont encore possibles (voir [27], Chap. 9).

Dans le cadre des données catégorielles, [15] ont montré que l'algorithme CEM avec modèle des classes latentes maximise là aussi un critère d'information classique, proche d'une métrique du χ^2.

3.3 Choix de modèles

On dispose maintenant d'un ensemble de modèles à disposition grâce à des contraintes pleines de sens sur les paramètres du mélange. Cependant, il n'est pas rare que la structure la plus adéquate soit elle-même ignorée du praticien. De la même façon, le nombre de classes peut être inconnu, seule une borne supérieure g_{sup} étant généralement disponible. On est alors face à une problématique de choix d'un modèle $\mathbf{m}$, regroupant la structure des classes et/ou le nombre des classes.

3.3.1 Critères « mélange »

Les critères informationnels classiques comme BIC ou AIC sont disponibles pour aider à sélectionner $\mathbf{m}$. Cependant, ces critères souffrent du même défaut que celui évoqué en phase d'estimation à savoir la perte de l'objectif de classification. En effet, le critère BIC étant convergeant (au moins en pratique), il détectera toutes les composantes du mélange, même si certaines sont très imbriquées, et ce dès que la taille de l'échantillon sera suffisamment grande. La partition qui sera déduite du modèle retenu sera donc insatisfaisante pour le praticien car il y a moins de classes interprétables dans les données que de classes présentes dans le mélange.

Plus ennuyeux encore, l'hypothèse de mélange n'est probablement pas vraiment vérifiée dans la population sous-jacente aux données. Dans ce cas, on peut s'attendre à ce que les critères classiques sélectionnent un nombre très important de composantes pour le mélange afin de réaliser une bonne estimation de la loi inconnue de la population (voir des résultats sur l'estimation consistante de densité par BIC dans [38]).

On pourrait avancer que l'estimation par approche classification, s'il elle a été utilisée, a déjà produite des classes plutôt séparées. Cependant, le paramètre estimé par maximum de vraisemblance complétée ($\tilde{\boldsymbol{\theta}}$) n'est pas utilisable en toute rigueur à la place du paramètre estimé par maximum de vraisemblance ($\hat{\boldsymbol{\theta}}$) dans les expressions de BIC et AIC.

3.3.2 Critère « classification »

L'idée est de reporter l'objectif de classification de la méthode d'estimation vers la méthode de choix de modèle. En d'autres termes c'est à la méthode de choix de modèle de retenir un modèle produisant des classes bien séparées tout en respectant « au mieux » la distribution des données.

Dans cette perspective, [6] ont proposé le critère ICL (*Integrated Complete Likelihood*) qui pénalise la log-vraisemblance complétée calculée en $\hat{\boldsymbol{\theta}}$ (et non pas calculée en $\tilde{\boldsymbol{\theta}}$) par le même terme de complexité que le critère BIC. Ce critère, à maximiser, s'écrit sous les trois formes équivalentes suivantes :

$$\begin{aligned} \text{ICL}_{\mathbf{m}} &= L(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \mathbf{x}, \hat{\mathbf{z}}_{\mathbf{m}}) - 0.5\nu_{\mathbf{m}} \ln(n) && (19)\\ &= L(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \mathbf{x}) - E(\hat{\mathbf{t}}_{\mathbf{m}}, \hat{\mathbf{z}}_{\mathbf{m}}) - 0.5\nu_{\mathbf{m}} \ln(n) && (20)\\ &= \text{BIC}_{\mathbf{m}} - E(\hat{\mathbf{t}}_{\mathbf{m}}, \hat{\mathbf{z}}_{\mathbf{m}}). && (21) \end{aligned}$$

La seconde ligne utilise la relation (17) entre log-vraisemblance et log-vraisemblance complétée et permet ainsi d'interpréter ICL comme une pénalisation du maximum de log-vraisemblance par un terme de complexité du modèle et un terme d'imbrication des classes. La dernière ligne interprète ICL comme un critère BIC pénalisé par ce terme

d'imbrication. D'une part, on remarque donc que ICL est tout aussi simple à calculer que BIC. D'autre part, on déduit que ICL pénalisera les modèles produisant des classes trop imbriquées, ce que ne faisait pas BIC.

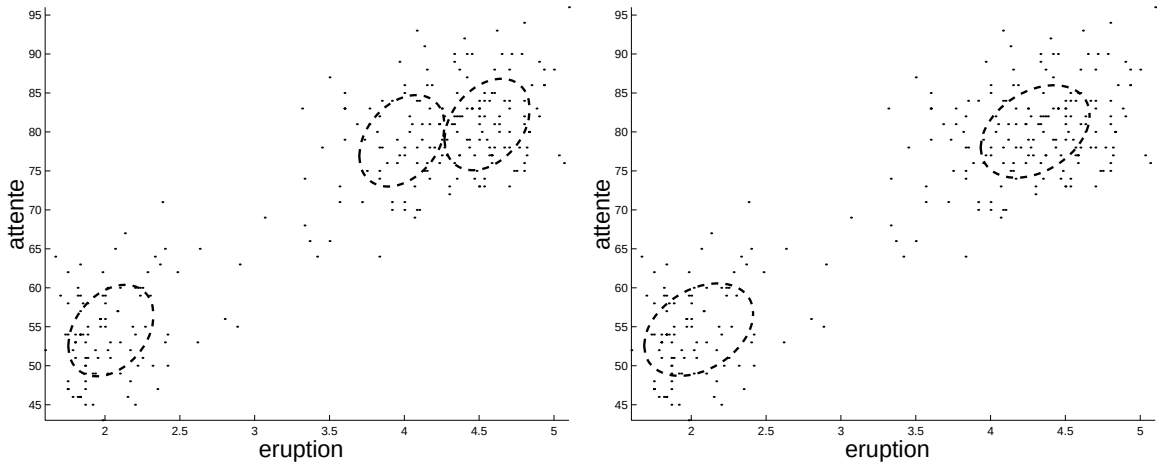
3.3.3 Exemples illustratifs

Il s'agit maintenant d'illustrer experimentalement les principales différences de comportement entre BIC et ICL. Plus généralement, c'est aussi l'occasion de montrer des exemples de mise en situation des mélanges gaussiens.

$\Delta\mu$	2.9		3.0		3.1		3.2		3.3	
n	BIC	ICL	BIC	ICL	BIC	ICL	BIC	ICL	BIC	ICL
100	94	23	96	31	97	44	95	45	97	60
400	100	9	100	21	100	48	100	70	100	85
700	100	8	100	15	100	39	100	72	100	96
1000	100	6	100	16	100	56	100	75	100	91

TAB. 1 – Nombre de sélections (parmi 100 essais) de deux composantes, au lieu d'une seule, dans un mélange gaussien univarié bimodal où $\pi_1 = \pi_2 = 0.5$, $|\mu_2 - \mu_1| = \Delta\mu$ et $\Sigma_1 = \Sigma_2 = 1$.

Dans un premier temps, les expériences suggèrent que ICL, au contraire de BIC, n'est pas nécessairement convergeant, la pénalité entropique d'ICL empêchant la sélection de classes trop imbriquées. En effet, ICL convergera vers le bon nombre de classes si ces dernières sont « suffisamment » séparées tandis qu'il sous-estimera ce nombre dans le cas contraire, même asymptotiquement. Ce comportement est illustré par des données simulées dans la table 1.



BIC ICL

FIG. 4 – Ellipses iso-densité des classes retenues par BIC et ICL sur les données du geyser.

Cependant, la pénalité entropique permet de robustifier le choix de modèle en faisant un compromis entre l'adéquation modèles/données et la séparation des classes. Ceci est illustré dans le cas du jeu de données réelles constitué de 272 eruptions de *the Old Faithful*

geyser du Yellowstone National Park (la version utilisée par [42]), chaque observation étant décrite par deux mesures : la durée de l'éruption et le temps d'attente entre deux éruptions (tous deux en minutes). En mettant en compétition les 28 modèles gaussiens combinés à un choix du nombre de classes entre 1 et $g_{\text{sup}} = 6$, BIC retient trois classes avec un modèle homoscédastique à proportions égales tandis que ICL sélectionne un modèle à deux classes à proportions libres et orientations égales. On remarque sur la figure 4 que la solution donnée par ICL fait alors ressortir la présence de deux groupes bien séparés tandis que la solution retenue par BIC illustre l'écart à la normalité de ses deux groupes plutôt que la présence d'une 3$^{\text{e}}$ classe d'intérêt pour la classification.

4 Classification supervisée

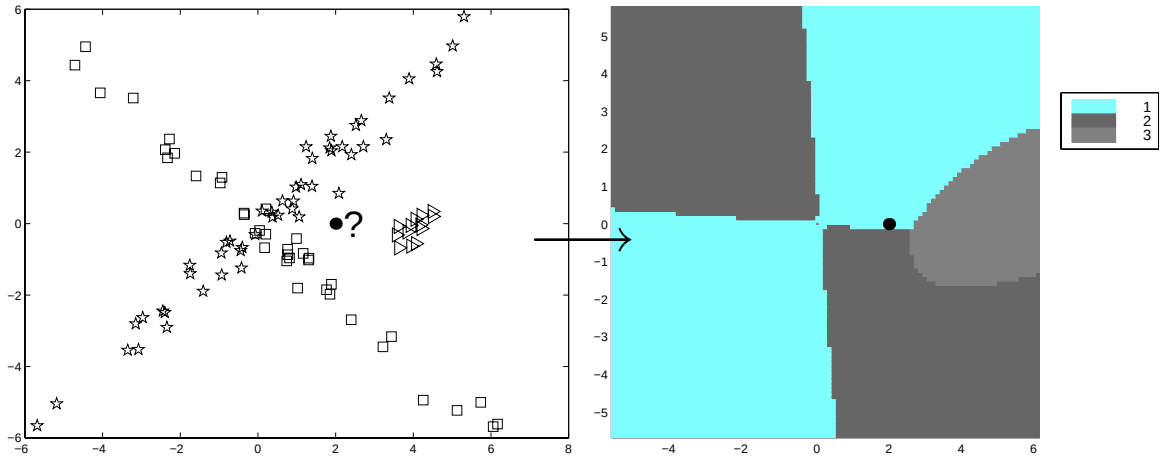
4.1 Problématique

4.1.1 Données et objectif

En *classification supervisée*, appelée aussi analyse discriminante, le couple $(\mathbf{x}, \mathbf{z})$ des n objets et de leur étiquette respective est connu (voir par exemple [34] ou [27], Chap. 7). L'objectif est d'estimer le groupe $\mathbf{z}_{n+1}$ de tout nouvel individu $\mathbf{x}_{n+1}$ de $\mathcal{X}$ arrivant ultérieurement à $(\mathbf{x}, \mathbf{z})$ et dont le groupe de provenance serait inconnu. En d'autres termes, il s'agit donc d'estimer une règle de classement r définie par

$$\begin{array}{rccc} r : & \mathcal{X} & \longrightarrow & \{1, \ldots, g\} \\ & \mathbf{x}_{n+1} & \longmapsto & r(\mathbf{x}_{n+1}). \end{array} \tag{22}$$

L'estimation s'appuiera sur l'ensemble des données disponibles $(\mathbf{x}, \mathbf{z})$, appelé souvent pour l'occasion *ensemble d'apprentissage*. L'objectif de la classification supervisée est illustré sur la figure 5 avec $\mathcal{X} = \mathbb{R}^2$ et trois groupes.



$(\mathbf{x}, \mathbf{z})$ et $\mathbf{x}_{n+1}$ $\qquad$ $\hat{r}$ et $\hat{\mathbf{z}}_{n+1}$

FIG. 5 – Classification supervisée : illustration de l'objectif pour des données $(\mathbf{x}, \mathbf{z})$ de $\mathbb{R}^2$ en trois classes. Le nouvel individu à classer $\mathbf{x}_{n+1}$ est représenté par le symbole « • ».

Au delà ce cet objectif *décisionnel*, on identifie parfois un objectif complémentaire,

voire alternatif, que l'on qualifie de *descriptif*. Dans ce cas, il s'agit plutôt de donner une description des groupes ou de la frontière qui les sépare.

4.1.2 Modélisation probabiliste générative

Le modèle probabiliste classique et naturel pour résoudre le problème de la classification supervisée s'appuie sur les proportions π_k des groupes et les lois conditionnelles p_k. Dans ce cadre, on retient la règle de classement r qui minimise l'erreur moyenne de classement donnée par

$$e(r) = 1 - E_{(\mathbf{X}_1,\mathbf{Z}_1)}[Z_{1r(\mathbf{X}_1)}]. \tag{23}$$

La règle de classement optimale r^*, dite *règle de Bayes*, est celle qui minimise $e(r)$. Elle correspond tout simplement à la règle du MAP déjà présentée en (16) pour $\mathbf{z}$ et que nous réexprimons ici pour r :

$$\forall \mathbf{x}_{n+1} \in \mathcal{X} \quad r(\mathbf{x}_{n+1}) = \arg \max_{k\in\{1,\ldots,g\}} \mathrm{p}(Z_{n+1k} = 1 | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}). \tag{24}$$

Bien entendu, le praticien peut prendre des libertés par rapport à la règle automatique du MAP en refusant de classer un individu $\mathbf{x}_{n+1}$ trop proche de la frontière de décision. On parle parfois dans ce cas de *rejet d'ambiguïté* [22]. Il est possible de prendre également en considération des coûts de mauvais classement, ce que nous ne ferons pas ici.

Dans le modèle probabiliste, la définition de la règle optimale de classement passe donc par la connaissance des probabilités conditionnelles d'appartenance aux groupes. Il est généralement efficace pour la prédiction et riche pour l'interprétation d'imposer une forme paramétrique à ces probabilités conditionnelles. On distingue alors deux approches pour la définir :

- *Approche prédictive* : les probabilités conditionnelles de groupes sont directement paramétrées sans définition précise de la loi p_k des groupes, c'est pourquoi on parle généralement de modèle *semi-paramétrique*. Le modèle de *régression logistique* [2] en fait partie.
- *Approche générative* : on paramétrise précisément cette fois la loi des groupes ($\mathrm{p}_k = \mathrm{p}(.;\boldsymbol{\alpha}_k)$) et on déduit la forme paramétrique des probabilités conditionnelles d'appartenance par la formule (8) combinant les $\mathrm{p}(\cdot;\boldsymbol{\alpha}_k)$ et les π_k. La *discrimination gaussienne* (voir [41] par exemple) et le *modèle d'indépendance conditionnelle* [25], que nous allons décrire tous deux dans la section suivante, en font partie.

Les deux approches permettent une caractérisation fine de la règle de classement. L'approche générative fournit en outre plusieurs avantages notables pour le praticien :

- Elle fournit une description précise et concise des groupes eux-mêmes grâce à la paramétrisation des p_k, facilitant ainsi leur interprétation.
- Elle permet d'identifier des individus $\mathbf{x}_{n+1}$ atypiques car très en dehors des zones de confiance associées aux groupes. L'individu peut par exemple provenir d'un groupe non modélisé. La règle du MAP aurait généralement classé sans hésitation cet individu dans une des classes existantes tandis que cette information supplémentaire incitera le praticien à plus de prudence. On parle alors de *rejet de distance* [22] si la décision est de ne pas classer l'individu en fin de compte.
- Enfin, elle permet de prendre en considération de façon naturelle pour estimer r l'information apportée par des individus $\mathbf{x}$ dont certaines appartenances $\mathbf{z}$ ne seraient pas connues. On parle de *classification semi-supervisée*, ce qui sort un peu de notre

contexte actuel mais mérite d'être signalé car cette problématique de classement se rencontre de plus en plus fréquemment [20].

Nous nous focalisons maintenant sur l'approche générative.

4.1.3 Géométrie des groupes et de la frontière de classement

Dans le cadre génératif, la discrimination gaussienne et le modèle d'indépendance conditionnelle sont certainement les plus utilisés. La discrimination gaussienne consiste simplement à modéliser la loi d'un groupe par la loi multinormale (3) et s'applique donc aux données continues. La discrimination avec modèle d'indépendance conditionnelle modélise la loi d'un groupe par (5) et concerne donc les données catégorielles. Dans les deux cas, on peut utiliser les propriétés géométriques décrites en 3.1.3 et 3.1.4 pour aider à l'interprétation des groupes et il est donc inutile d'en parler plus en détail ici.

On déduit aussi de ces modélisations des expressions paramétriques explicites et interprétables des règles de classement car la règle de classement est devenue elle-même une fonction de $\boldsymbol{\theta}$: $r = r(.;\boldsymbol{\theta})$. Par exemple, dans le cas gaussien homoscédastique à deux groupes ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$), on a, pour tout $\mathbf{x}_{n+1} \in \mathcal{X}$:

$$r(\mathbf{x}_{n+1};\boldsymbol{\theta}) = 1 \text{ si } \mathbf{x}_{n+1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \ln\frac{\pi_1}{\pi_2} > 0. \qquad (25)$$

La forme géométrique de la frontière de classement s'en déduit alors directement. Comme illustré sur la figure 6, un modèle homoscédastique gaussien correspond à une frontière linéaire entre deux groupes (un hyperplan de $\mathbb{R}^d$) tandis qu'un modèle à volume libre donne une frontière quadratique entre deux groupes (une quadrique de $\mathbb{R}^d$).

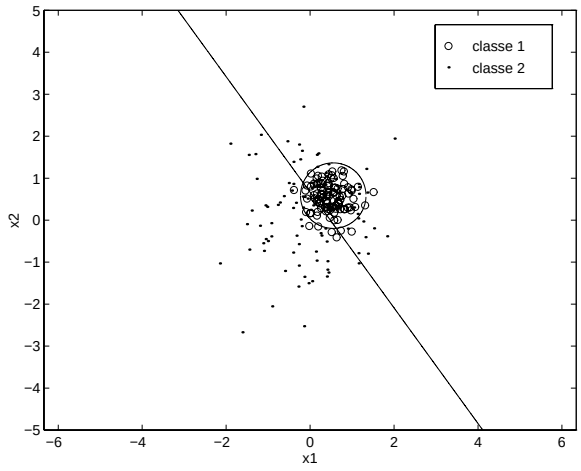


FIG. 6 – Illustration géométrique dans le plan de la forme de la frontière de classement dans le cas gaussien avec deux groupes : la frontière linéaire correspond au cas sphérique à volume égal et la frontière quadratique au cas sphérique à volume libre.

4.2 Estimation d'une règle de classement

4.2.1 Estimation par maximum de vraisemblance

Une des méthodes les plus simples pour estimer $r(.;\boldsymbol{\theta})$ est celle dite du *plug-in*. Il s'agit tout d'abord d'estimer $\boldsymbol{\theta}$ en maximisant la log-vraisemblance *observée*

$$L(\boldsymbol{\theta};\mathbf{x},\mathbf{z}) = \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik}\ln\left(\pi_k \mathrm{p}(\mathbf{x}_i;\boldsymbol{\alpha}_k)\right), \tag{26}$$

puis d'injecter l'estimateur $\hat{\boldsymbol{\theta}}$ obtenu dans la règle de classement :

$$\hat{r} = r(.;\hat{\boldsymbol{\theta}}). \tag{27}$$

La maximisation de la vraisemblance est immédiate car elle peut s'interpréter simplement comme une unique étape M de EM où les poids $\mathbf{t}^+$ ont été remplacés par les vraies appartenances $\mathbf{z}$ (voir 2.2.2). Cependant, il faut veiller au processus d'échantillonnage sous-jacent pour estimer convenablement les proportions π_k : si les données $\mathbf{x}$ proviennent bien de la loi mélange $\mathrm{p}(.;\boldsymbol{\theta})$, les proportions sont estimées comme indiquées auparavant tandis que dans le cas d'un tirage *cas-témoin* (en médecine typiquement), les proportions réelles de groupes doivent être connues par ailleurs (voir [19] par exemple). En outre, dans le cas gaussien, on peut aussi parfois remplacer les estimateurs du maximum de vraisemblance des matrices de variance par leurs expressions corrigées permettant de supprimer le biais d'estimation.

Au delà de la méthode du MAP, la connaissance de $\hat{\boldsymbol{\theta}}$ permet bien entendu de mettre en œuvre également des stratégies de rejet en ambiguïté ou en distance.

4.2.2 Lien avec d'autres méthodes

Tout comme en classification non supervisée, il est possible d'établir des connexions entre l'approche générative et certaines méthodes géométriques à base de distance. Ainsi, le cas homoscédastique avec deux groupes correspond à la discrimination linéaire de Fisher [23] s'appuyant sur la distance de Mahalanobis dans un espace euclidien.

Notons aussi que la discrimination gaussienne et la discrimination multinomiale avec indépendance conditionnelle sont aussi liées à la régression logistique linéaire (voir [34] par exemple).

4.3 Choix de modèles et de variables

On s'intéresse maintenant au choix d'un modèle $\mathbf{m}$. Il s'agit typiquement de sélectionner l'un des modèles gaussiens ou multinomiaux parcimonieux. Cependant, on peut étendre le choix de modèle à la mise en compétition d'un modèle génératif à un modèle non génératif, par exemple le modèle logistique. Enfin, il est aussi crucial en classification supervisée de s'intéresser au choix des variables les plus discriminantes.

4.3.1 Critères mélange

Dans le cas génératif paramétrique, on peut envisager des critères de vraisemblance pénalisée comme le critère BIC pour choisir $\mathbf{m}$. Cependant, comme c'était le cas en classification non-supervisée, BIC se focalise sur la forme de la distribution des données sans

prêter attention à l'objectif qui est de sélectionner ici une règle de classement d'erreur faible.

En ce qui concerne le choix entre un modèle génératif paramétrique et un modèle non génératif, le critère BIC n'est pas utilisable car les vraisemblances ne s'expriment pas sur les mêmes espaces probabilisés. Pour la même raison, BIC ne peut s'appliquer tel quel à la sélection de variables.

4.3.2 Critères liés à l'erreur de classement

Une technique universelle pour sélectionner un modèle génératif et/ou un modèle non génératif et/ou des variables en classification supervisée s'appuie sur une estimation du taux d'erreur de classement associé à la règle $\hat{r}$ estimée. Pour se faire, on peut s'appuyer sur des stratégies de rééchantillonnage comme la *validation croisée*. Le principe est de diviser aléatoirement l'ensemble $(\mathbf{x}, \mathbf{z})$ en v parties (approximativement) égales puis, notant $\hat{\boldsymbol{\theta}}^{\ell}_{\mathbf{m}}$ $(\ell = 1, \ldots, v)$ l'estimateur de $\boldsymbol{\theta}$ à partir de $(\mathbf{x}^{\ell}, \mathbf{z}^{\ell})$ correspondant à $(\mathbf{x}, \mathbf{z})$ privé de sa ℓ^{e} partie, le critère de validation croisée est défini par :

$$\mathrm{VC}_{\mathbf{m}} = 1 - \frac{1}{v} \sum_{\ell=1}^{v} \sum_{\{i:(\mathbf{x}_i,\mathbf{z}_i)\in(\mathbf{x}^{\ell},\mathbf{z}^{\ell})\}} Z_{ir(\mathbf{x}_i;\hat{\boldsymbol{\theta}}^{\ell}_{\mathbf{m}})}. \tag{28}$$

On retient alors le modèle et/ou les variables conduisant à la valeur du critère la plus faible. Lorsque n est petit, il est recommandé de prendre $v = n$, ce qui revient à la procédure dite du *leave-one-out* consistant à soustraire une unique observation à chaque fois. Lorsque l'échantillon est suffisamment grand, on peut cependant se contenter de la méthode de *l'échantillon test* consistant à diviser l'échantillon en deux parties seulement, la première servant à apprendre tandis que la seconde sert à évaluer la qualité de la règle.

Remarquons que la procédure peut être relativement coûteuse en temps de calcul car nécessitant de nombreuses estimations de la règle de classement. Cependant, il est parfois possible, en particulier dans le cas génératif, d'éviter de recalculer entièrement la règle de décision à chaque fois [8]. Signalons aussi qu'il est risqué d'utiliser ensuite $\mathrm{VC}_{\mathbf{m}}$ pour estimer le taux d'erreur associé au modèle retenu car il souffre d'un biais d'optimisme. On aura alors recours à une méthode de *double validation croisée* ([27], pp. 210–211).

4.3.3 Exemple illustratif

On va maintenant étudier un jeu de données issu d'une problématique marketing[2] pour illustrer l'utilisation du modèle multinomial d'indépendance conditionnelle et les méthodes de choix de modèles. Il est constitué de $n = 6876$ familles de la baie de San Francisco décrites par $d = 12$ variables catégorielles relatives au mode de vie et comportant deux variables ou plus : statut marital, classe d'âge, niveau d'éducation,... Les groupes sont au nombre de trois ($g = 3$) et représentent le niveau de revenu de la famille : revenus faibles (moins de 19 999$), moyens (entre 20 000$ et 39 999$) et forts (plus de 40 000$). La figure 7 représente les données dans le 1$^{\text{er}}$ plan de l'analyse en correspondance multiple (ACM).

On met en compétition deux modèles multinomiaux d'indépendance conditionnelle (le modèle général non contraint et le modèle à modalité majoritaire défini par (15)) et le

[2]Origine : Impact Resources, Inc., Columbus, OH (1987). (From questionnaires containg 502 questions filled out by shopping mall customers in the San Francisco Bay area.)

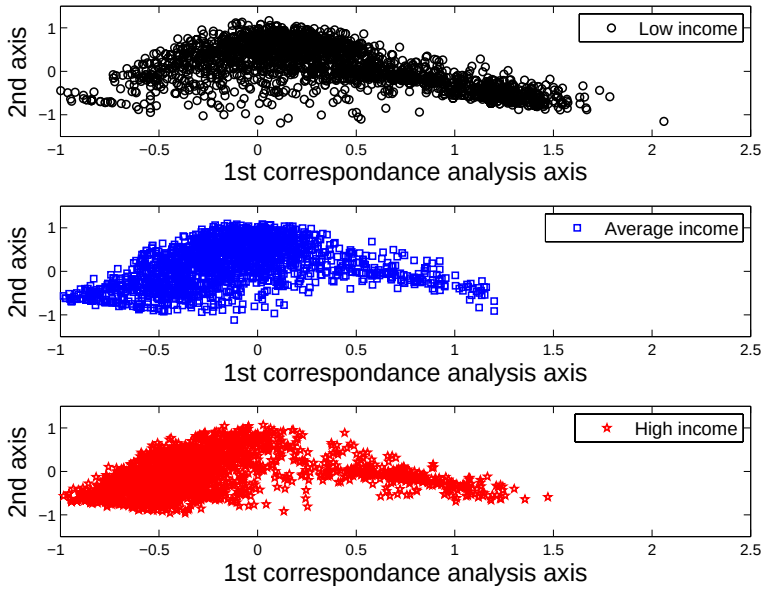


FIG. 7 – Les données marketing sur les deux premiers axes de l'ACM du nuage global. Les trois groupes sont représentés dans des sous-figures différentes pour faciliter leur comparaison.

modèle logistique. Comme l'échantillon est d'assez grande taille, on utilise la méthode de l'échantillon test en prenant un échantillon d'apprentissage de taille 4000 individus et on conservant les 2876 individus restant pour évaluer la règle de classement de chaque modèle. La table 2 donne l'estimation de l'erreur pour les trois modèles et la valeur du critère BIC pour les deux modèles génératifs. BIC sélectionne le modèle génératif le plus complexe, tout comme le critère d'erreur.

Modèle	$\widehat{e(\hat{r})}$ (%)	BIC
Multinomial général	**36.79**	**-56619**
Multinomial à modalité majoritaire	39.78	-62459
Logistique	39.36	(incomparable)

TAB. 2 – Classification supervisée sur les données marketing. L'erreur est estimée par la méthode de l'échantillon test et le critère BIC est calculé sur l'échantillon au complet.

En outre, il est possible de caractériser facilement chacun des groupes avec le modèle multinomial retenu en regardant de près les paramètres estimés : la table 3 donne les proportions π_k estimées et la table 4 donne les paramètres α_{kjh} estimés pour la variable « statut marital ». On peut faire de même pour les autres variables.

5 Bilan et perspectives

Les modèles de mélange apportent une réponse cohérente au problème de la classification : modélisation explicite et interprétable, outils de la statistique mathématique

Revenus	faibles	moyens	forts
π_k (%)	34.63	28.40	36.98

TAB. 3 – Proportions estimées des groupes avec le modèle multinomial d'indépendance conditionnelle général pour les données marketing.

Revenus \ modalités	marié	en couple	divorcé	veuf	célibataire
Faibles	10.90	7.15	9.10	3.90	68.95
Moyens	37.06	8.54	13.82	3.43	37.15
Forts	62.27	6.90	6.22	1.49	23.12

TAB. 4 – Paramètres α_{kjh} estimés (en %) du modèle multinomial d'indépendance conditionnelle général pour la variable « statut marital » avec les données marketing.

à disposition, généralisation de méthodes traditionnelles, mise en œuvre simple par des logiciels dédiés. Tout cela contribue à la diffusion croissante de ce type de méthode dans de nombreux domaines d'applications.

Comme partout, la classification est constamment confrontée à de nouveaux défis et les modèles de mélange apparaissent encore comme une approche suffisamment flexible et rigoureuse pour apporter des réponses pertinentes. Ainsi, du point de vue de la nature des données, les modèles de mélange gaussien peuvent être adaptés à des données de très haute dimension (plusieurs milliers de variables) [11, 12], à des données bruitées par l'introduction d'une « composante de bruit » [3] ou encore à des données partiellement étiquetées (classification semi-supervisée) [20]. Les objectifs peuvent aussi évoluer : classification croisée des individus et des variables [28], choix de variables en classification non supervisée (voir [37, 33] pour une réinterprétation comme un choix de modèles dans le cas gaussien), classification supervisée ou non de plusieurs échantillons à la fois par établissement de liens entre populations [4, 32, 30].

Références

[1] H. Akaike. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, AC-19 :716–723, 1974.

[2] J.A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59 :19–35, 1972.

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

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

[5] C. Biernacki, G. Celeux, A. Anwuli, G. Govaert, and F. Langrognet. Le logiciel mixmod d'analyse de mélange pour la classification et l'analyse discriminante. *La Revue de Modulad*, 35 :25–44, 2006.

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

[7] C. Biernacki, G. Celeux, and G. Govaert. 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, 2003.

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

[9] H. Bock. Probabilistic aspects in cluster analysis. In O. Opitz, editor, *Conceptual and numerical analysis of data*, pages 12–44. Springer-Verlag, Berlin, 1989.

[10] L. Bordes, S. Mottelet, and P. Vandekerkhove. Semiparametric estimation of a two components mixture model. *Annals of Statistics*, 34 :1204–1232, 2006.

[11] C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering. *Computational Statistics and Data Analysis*, 52(1) :502–519, 2007.

[12] C. Bouveyron, S. Girard, and C. Schmid. High dimensional discriminant analysis. *Communications in Statistics : Theory and Methods*, 36(14) :2607–2623, 2007.

[13] P.G. Bryant and J.A. Williamson. Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika*, 65 :273–281, 1978.

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

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

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

[17] G. Celeux and G. Govaert. Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and Simulation*, 47 :127–146, 1993.

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

[19] G. Celeux and J.P. Nakache. *Analyse discriminante sur variables qualitatives*. Polytechnica, 1994.

[20] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, 2006.

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

[22] B. Dubuisson. Decision with rejects options. In *Eusipco*, Barcelone, 1990.

[23] R. A. Fisher. Multiple measurements in taxinomie problems. *Annals of Eugenics*, 7(2) :179–188, 1936.

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

[25] M. Goldstein and W.R. Dillon. *Discrete discriminant analysis*. John Wiley & Sons, New York, 1978.

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

[27] G. Govaert. *Analyse des données*. Hermes, Paris, 2003.

[28] G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2) :463–473, 2003.

[29] R.J. Hathaway. Another interpretation of the em algorithm for mixture distributions. *Statistics & Probability Letters*, 4 :53–56, 1986.

[30] J. Jacques and C. Biernacki. Extension of model-based classification for binary data when training and test populations differ. *Journal of Applied Statistics*, (à paraître), 2009.

[31] E. Lebarbier and T. Mary-Huard. Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la SFdS*, (à paraître), 2006.

[32] A. Lourme and C. Biernacki. Gaussian model-based classification when training and test population differ : Estimating jointly related parameters. In *In First joint meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of SIS*, Caserta, Italy, 2008.

[33] C. Maugis, G. Celeux, and M.L. Martin-Magniette. Variable selection for clustering with gaussian mixture models. *Biometrics*, (à paraître), 2009.

[34] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, 2004.

[35] G. J. McLachlan and K. Krishnan. *The EM Algorithm*. Wiley, New York, 1997.

[36] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

[37] A.E. Raftery and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101 :168–178, 2006.

[38] K. Roeder and L. Wasserman. Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439) :894–902, 1997.

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

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

[41] R. Tomassone, M. Danzard, J.J. Daudin, and J.P. Masson. *Discrimination et classement*. Masson, 1988.

[42] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, 1994.

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