UNIVERSITÉ DE FRANCHE-COMTÉ

Mémoire d'Habilitation à Diriger des Recherches

Contribution à l'étude des mélanges de lois et à leurs applications

---

par

Christophe Biernacki

Enseignant-chercheur au laboratoire de Mathématiques de l'Université de Franche-Comté – UMR CNRS 6623

Soutenue le 9 décembre 2004 devant le jury composé de :

Rapporteurs :	
J.J. Daudin	Professeur, Institut National Paris-Grignon
M. Lavielle	Professeur, Université Paris 5
G.J. McLachlan	Professeur, University of Queensland
D.M. Titterington	Professeur, University of Glasgow
Examinateurs :	
G. Celeux	Directeur de Recherches, Inria Futurs
G. Govaert	Professeur, Université de Technologie de Compiègne
C. Stricker	Professeur, Université de Franche-Comté
M. Weil	Professeur honoraire, Université de Franche-Comté

Table des matières

Principales notations et abréviations

MV, MAP	maximum de vraisemblance, maximum *a posteriori*
v.a., i.i.d.	variable aléatoire, indépendamment et identiquement distribué
$\mathbf{I}_\nu$, $\mathbf{0}$	matrice identité d'ordre ν, matrice nulle d'ordre implicite
$\mathsf{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\mathsf{N}_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$	loi normale unidimensionnelle (d dimensionnelle) de moyenne $\boldsymbol{\mu}$ et de matrice de variance $\boldsymbol{\Sigma}$
$\mathsf{Mult}_K(1,\mathbf{p})$	loi multinomiale d'ordre un et de paramètre $\mathbf{p} \in [0,1]^K$
$\mathsf{tr}[\mathbf{A}]$, $\|\mathbf{A}\|$	trace et déterminant de la matrice carrée $\mathbf{A}$
$\mathsf{E}_{\mathbf{X}_1}[\mathbf{X}_1]$, $\mathsf{E}_{\boldsymbol{\theta}}[\mathbf{X}_1]$	espérance de $\mathbf{X}_1$ par rapport à $p(\mathbf{x}_1)$, à $p(\mathbf{x}_1;\boldsymbol{\theta})$
$\mathsf{E}_{\mathbf{Z}^u}[\mathbf{Z}^u\|\mathcal{D}]$, $\mathsf{E}_{\boldsymbol{\theta}}[\mathbf{Z}^u\|\mathcal{D}]$	espérance de $\mathbf{Z}^u$ par rapport à $p(\mathbf{z}^u\|\mathcal{D})$, à $p(\mathbf{z}^u\|\mathcal{D};\boldsymbol{\theta})$
$\mathsf{Var}_{\mathbf{X}_1}[\mathbf{X}_1]$, $\mathsf{Var}_{\boldsymbol{\theta}}[\mathbf{X}_1]$	matrice de variance de $\mathbf{X}_1$ par rapport à $p(\mathbf{x}_1)$, à $p(\mathbf{x}_1;\boldsymbol{\theta})$
$\xrightarrow{d}$, $\xrightarrow{p}$, $\xrightarrow{p.s.}$	convergence en distribution, en probabilité, presque sûre
$X_n = o_p(Y_n)$	(X_n) et (Y_n) deux suites de v.a. telles que $X_n/Y_n \xrightarrow{p} 0$ quand $n \to \infty$
$X_n = O_p(1)$	$\|X_n\|$ bornée en probabilité : $\forall \epsilon > 0$, $\exists c > 0$ tel que $p(\|X_n\| > c) < \epsilon$ quand $n \to \infty$
$X_n = O_p(Y_n)$	$X_n/Y_n = O_p(1)$
$x_n = o(y_n)$	(x_n) et (y_n) deux suites réelles telles que $x_n/y_n \to 0$ quand $n \to \infty$
$x_n = O(y_n)$	$\exists c > 0$ $\|x_n/y_n\| \leq c$ quand $n \to \infty$
$\mathbf{m}$, ν	modèle paramétrique, nombre de paramètres associé
$\mathbf{x}^l$, $\mathbf{x}^u$, $\mathbf{x}$	individus de labels connus, inconnus ; tous les individus $\mathbf{x} = (\mathbf{x}^l,\mathbf{x}^u)$
$\mathbf{z}^l$, $\mathbf{z}^u$, $\mathbf{z}$	labels connus, inconnus ; tous les labels $\mathbf{z} = (\mathbf{z}^l,\mathbf{z}^u)$
$\mathcal{D}$, $\mathcal{D}_l$, $\mathcal{D}_u$	données disponibles $\mathcal{D} = (\mathbf{x},\mathbf{z}^l) = (\mathcal{D}_l,\mathcal{D}_u)$ avec $\mathcal{D}_l = (\mathbf{x}^l,\mathbf{z}^l)$ et $\mathcal{D}_u = \mathbf{x}^u$
$\ell(\boldsymbol{\theta};\mathcal{D})$	log-vraisemblance de $\boldsymbol{\theta}$ avec les données $\mathcal{D}$
$\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}$, $\boldsymbol{\theta}^{(q)}$, $\boldsymbol{\theta}^*$	valeur de $\boldsymbol{\theta}$ maximisant la log-vraisemblance moyenne $\mathsf{E}_{\mathcal{D}}[\ln p(\mathcal{D};\boldsymbol{\theta})]$, maximisant la log-vraisemblance $\ell(\boldsymbol{\theta};\mathcal{D})$, à la $q$ème itération d'un algorithme d'optimisation de $\ell(\boldsymbol{\theta};\mathcal{D})$, à la stationnarité
K, G_k	nombre de groupes, $k$ème groupe
$\mathbf{t}$, $\mathbf{c}$	matrice des probabilités conditionnelles, de classification floue
r	règle de classement
β	proportion d'individus d'appartenance connue dans $\mathcal{D}$
$\mathbf{K}$, $\mathbf{K}_c$, $\mathbf{K}_\beta$	forme produit de la matrice d'information de Fisher calculée en $\boldsymbol{\theta}_0$ pour un individu : label inconnu, label connu, situation hybride
$\mathbf{J}$, $\mathbf{J}_c$, $\mathbf{J}_\beta$	forme hessienne de la matrice d'information de Fisher calculée en $\boldsymbol{\theta}_0$ pour un individu : label inconnu, label connu, situation hybride
$(\cdot)^+$	$\sup(\cdot,0)$

Introduction générale

Il est tout à fait possible de découvrir les mélanges finis de lois à partir d'une de leurs nombreuses applications. Dès le DEA, ce fut mon cas au travers de la problématique de classification automatique. C'est ainsi que l'on peut s'aventurer à expliquer la ligne directrice de mes recherches, partagées entre l'intérêt porté aux applications et le besoin d'élucider certains aspects fondamentaux plus spécifiques aux mélanges. En effet, les mélanges de distributions de probabilités fournissent une méthode de modélisation statistique extrêmement flexible qui débouche non seulement sur un vaste champ d'applications mais encore sur de nombreux développements théoriques. La littérature est certes abondante sur le sujet mais la spécificité des mélanges, parfois étroitement mêlée à un objectif applicatif, laisse ouvertes de multiples questions pourtant classiques en statistique comme l'estimation ou encore le choix de modèles.

Ce mémoire d'Habilitation à Diriger des Recherches dresse un panorama unifié, critique et relativement exhaustif des thèmes de recherche abordés pendant, et surtout postérieurement à, la thèse. Les aspects techniques sont succinctement développés en général, l'objectif étant plutôt de fournir une vision claire et simplifiée, presque pédagogique, des enjeux, des difficultés, des lignes directrices de résolution et des perspectives de chacun des sujets considérés. Ainsi, le lecteur est invité à retrouver le détail mathématique (énoncé et preuve précise de certains théorèmes par exemple) dans les travaux cités en référence. Néanmoins, des développements théoriques seront parfois donnés afin d'éclaircir des points particuliers de la bibliographie.

Un découpage en cinq parties a été adopté pour la présentation. Le premier des chapitres introduit les modèles de mélange comme une conséquence possible et naturelle de la problématique de classification, ce terme regroupant aussi bien la classification automatique que l'analyse discriminante. Les mélanges sont, en effet, un outil de choix pour modéliser l'hétérogénéité d'une population. Rapidement, le sujet se focalise sur les différentes méthodes ou critères permettant de choisir un modèle. Après avoir dressé un panorama des principaux critères statistiques généraux (typiquement les critères NIC et BIC) et de leur caractéristiques, l'accent est mis sur des réponses spécifiques à l'objectif de classification. En particulier, cette présentation aboutit au critère de classification ICL qui à la triple spécificité de se justifier d'une façon similaire à BIC, de s'écrire très simplement comme une pénalisation de BIC par une entropie de partition et, enfin et surtout, de montrer un comportement plus robuste vis à vis des violations de modèles. Ce chapitre est aussi l'occasion d'introduire nombre de notations et de définitions qui auront cours tout au long du mémoire.

Le chapitre suivant abandonne la problématique du choix de modèles et les aspects applicatifs pour se focaliser sur les problèmes d'estimation par la méthode du maximum de vraisemblance dans les mélanges, en particulier les mélanges gaussiens multivariés. Après un rappel des difficultés liées à cette technique d'estimation et sur les propriétés de l'algorithme d'optimisation EM, la question fondamentale de l'estimation est traitée sous trois aspects différents et complémentaires. Tout d'abord, des propositions d'initialisation de EM sont présentées, la plus notable étant certainement celle consistant à utiliser les propriétés de trajectoires de EM afin d'effectuer une recherche plus exhaustive de l'ensemble des solutions stationnaires de la vraisemblance. Ensuite, un test permettant de décider si une estimation candidate correspond ou non à celle du maximum de vraisemblance est décrit. Ce test, dépassant d'ailleurs le cadre des mélanges, montre une puissance supérieure à celle d'un autre test antérieur et similaire. Enfin, la question de la dégénérescence de la fonction vraisemblance dans les mélanges gaussiens est abordée. Sur ce thème délicat mais assez peu traité dans la bibliographie finalement, deux propositions ont été faites. Dans un premier temps, il a été établi que EM converge très rapidement vers une solution dégénérescente lorsqu'il en est à proximité, ce résultat permettant de confirmer que la démarche simple, pragmatique et largement répandue qui consiste à relancer l'algorithme lorsque la précision numérique est atteinte n'est pas infondée. Il est aussi exhibé une borne supérieure asymptotique de la vraisemblance qui permet ainsi d'apporter une réponse plus conceptuelle à ce problème de vraisemblance non bornée supérieurement.

Le chapitre qui suit s'intéresse à un type particulier de données que l'on retrouve naturellement, par exemple dès qu'un instrument de mesure à précision finie est utilisé. Il s'agit des données groupées, terme définissant typiquement des intervalles dans le cas continu univarié pour fixer les idées. Deux questions associées aux données groupées sont alors traitées. D'une part, il s'agit de découvrir si la problématique de la dégénérescence des mélanges gaussiens existe également avec ce type de données. Bien que la vraisemblance soit bornée supérieurement, il apparaît que le maximum de la vraisemblance peut être néanmoins obtenu pour des situations dégénérées. D'autres surprises nous attendent, en particulier avec l'étude de l'évolution de l'algorithme EM au voisinage de telles solutions. Il ressort par exemple que EM connaît une évolution dangereusement lente au voisinage d'un paramètre dégénéré et il devient donc important de proposer une règle d'arrêt spécifique pour EM, les règles *ad hoc* étant à proscrire. Notons aussi que la dégénérescence de EM dans le cas groupé est plus fréquente que le cas individuel étudié au chapitre précédent. D'autre part, ce chapitre se focalise sur le comportement du critère BIC pour les données groupées. En particulier, il est montré que BIC sélectionne généralement des modèles plus simples avec des données groupées qu'avec les données individuelles dont ces dernières sont issues.

Le quatrième chapitre s'intéresse à une évolution de l'analyse discriminante classique en violant l'hypothèse fondamentale que les populations d'apprentissage et de test sont identiques. Cette démarche est motivée par des besoins applicatifs, notamment en biologie ou l'exemple « classique » peut se présenter comme la question de la détermination du sexe d'oiseaux juvéniles à partir d'une règle de classement déjà établie pour des oiseaux adultes de la même espèce. Considérée dans le cadre des mélanges gaussiens, cette discrimination généralisée est construite à partir d'un certain nombre d'hypothèses réalistes sur les deux populations, ces hypothèses conduisant à la paramétrisation d'un lien stochastique entre les deux populations. Les résultats encourageants obtenus dans le contexte biologique ayant initié cette étude laissent espérer de nombreux prolongements à cette direction de recherche.

Enfin, le dernier chapitre du mémoire relate des expériences diverses en transfert de connaissances ou encore en applications dans le milieu des entreprises. Tout d'abord, étant donné l'importance des modèles de mélange au sein du contexte applicatif et aussi les nombreuses évolutions sur le sujet, il a semblé opportun de mettre à disposition de la communauté scientifique, du monde de l'enseignement et du milieu économique un logiciel gratuit offrant des fonctionnalités certes scientifiques, mais aussi informatiques, de bon niveau. Ainsi, sous la responsabilité scientifique de Gilles Celeux (Directeur de Recherches, INRIA Grenoble et Paris-Sud), de Gérard Govaert (Professeur, Université de Technologie de Compiègne) et de moi-même, le développement d'un logiciel dénommé MIXMOD a été entrepris grâce au concours d'informaticiens comme Florent Langrognet (Ingénieur de Recherches CNRS, Université de Besançon) ou encore d'ingénieurs experts (Gregory Noulin, Yann Vernaz) recrutés spécialement par l'INRIA grâce à une aide spécifique pour le développement logiciel. Un second point important de ce chapitre concerne l'utilisation des modèles de mélange gaussiens pour l'indexation d'images, domaine en forte extension. Les mélanges ont alors été mis à contribution soit pour estimer la distribution des caractéristiques d'une image comme la couleur (cas de l'indexation d'images), soit pour mettre en œuvre l'analyse discriminante afin de retrouver typiquement des acteurs dans un film (cas de l'indexation de vidéo). L'indexation de vidéo étant le sujet de la thèse de Riad Hammoud, un étudiant du laboratoire Gravir-Movi où j'étais à ce moment, une partie de sa thèse s'est donc déroulée en co-responsabilité avec son directeur officiel. La dernière partie de ce chapitre s'attache à de véritables transferts vers le monde des entreprises comme l'utilisation de la classification pour aider à réduire l'absentéisme dans les collectivités locales (contrat avec une compagnie d'assurances filiale de DEXIA), ou encore un problème de fiabilité bayésienne dans les tuyauteries de centrales nucléaires (contrat avec EDF), cette dernière problématique étant nettement indépendante des mélanges.

Chapitre 1

Choix de modèles en classification

Sommaire

Publications associées à ce chapitre

Les publications marquées par un astérisque « ∗ » désignent des travaux issus de la thèse.

Revues avec comité de lecture

[R1] C. Biernacki, G. Celeux and G. Govaert (2000). *Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood.* IEEE Transactions on Pattern Analysis and Machine Intelligence, **22**(7), 719–725.

[R2] C. Biernacki, G. Celeux and G. Govaert (1999). *An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture Model.* Pattern Recognition Letters, **20**(3), 267–272.

∗ [R3] C. Biernacki and G. Govaert (1999). *Choosing Models in Model-based Clustering and Discriminant Analysis.* Journal of Statistical Computation and Simulation, **64**, 49–71.

∗ [R4] C. Biernacki (1999). *Précision sur les données et coude de la vraisemblance pour trouver le nombre de classes dans un mélange.* Revue de Statistique Appliquée, **47**(1), 47–62.

Conférences avec comité de lecture

∗ [C5] C. Biernacki (1998). *Choosing the Number of Clusters from the Data Precision.* COMPSTAT'98, Bristol, Angleterre, 24–28 août.

∗ [C6] C. Biernacki and G. Govaert (1997). *Choosing Gaussian Models in Discriminant Analysis.* IV International Meeting of Multidimentional Data Analysis, Bilbao, Spain, 10–12 septembre.

∗ [C7] G. Celeux, C. Biernacki and G. Govaert (1997). *Choosing Models in Model-based Clustering and Discriminant Analysis.* Joint Statistical Meetings of the American Statistical Association (ASA), Anaheim/Orange County, USA, 10–14 août. Invited paper.

∗ [C8] C. Biernacki and G. Govaert (1997). *Using the Classification Likelihood to Choose the Number of Clusters.* Computing Science and Statistics, **29**, 2, 451–457.

Conférences sans comité de lecture

∗ [c9] C. Biernacki (1998). *Nombre de classes en classification automatique.* 30ièmes Journées de Statistique, Rennes, France, 25–29 Mai.

∗ [c10] C. Biernacki and G. Govaert (1996). *Comparaison de critères en classification pour la détection de modèles gaussiens parcimonieux.* Quatrièmes journées de la Société Francophone de Classification (SFC), Vannes, France, 19–20 septembre.

∗ [c11] C. Biernacki and G. Govaert (1995). *Detecting the Number of Clusters using Parsimonious Models with a Penalized Likelihood in Clustering.* Congrès International sur les Mélanges de Distribution, Aussois, France, 17–21 septembre.

Rapports

∗ [r12] C. Biernacki (1997). *Choix de modèles en classification.* Thèse de doctorat, Université de Technologie de Compiègne.

∗ [r13] C. Biernacki (1993). *Classification à partir de gaussiennes à liberté restreinte.* Rapport de DEA, Université de Technologie de Compiègne.

Séminaires

[s14] C. Biernacki (2000). *Évaluation d'un modèle de mélange en classification par la vraisemblance complétée intégrée.* Workshop Valclass II, Paris, France, 21-22 septembre.

∗ [s15] C. Biernacki and G. Govaert (1998). *Choix du nombre de classes par la vraisemblance classifiante.* Séminaire Statistique et Modélisation Stochastique de l'Université Joseph Fourier et de l'INPG, Grenoble, France, 19 mars.

∗ [s16] C. Biernacki and G. Govaert (1996). *Choix du nombre de classes par la vraisemblance classifiante.* Séminaire du CREST-INSEE, Paris, France, 16 décembre.

Seul ce chapitre est en étroite liaison avec le sujet de la thèse soutenue en septembre 1997 à l'Université de Technologie de Compiègne [r12] ; il en reprend d'ailleurs le titre. Certaines publications s'identifient aux résultats de la thèse ([R3] , [R4] , [C∗] , [c∗] , [s15] , [s16]). D'autres correspondent à des travaux menés par la suite lors de mes deux années post-doctorales à l'INRIA Rhônes-Alpes ([R1] , [R2] , [s14]). En particulier, [R1] concrétise une perspective évoquée en conclusion du travail de thèse.

Trois parties ponctuent ce chapitre. La section 1.1 situe tout d'abord la classification automatique et l'analyse discriminante dans le cadre des modèles de mélange en insistant sur les différentes hypothèses relatives à notre cadre de travail (formulation paramétrique, méthode d'estimation, etc.). Mettant en évidence l'aspect incontournable du choix de modèles, la section 1.3 présente, avec les aspects mathématiques qui y conduisent, des critères généraux et classiques de sélection de modèles tels que AIC, NIC ou BIC. Enfin, la section 1.3 exhibe des critères spécifiques de choix de modèles comme ICL, NEC ou le critère de validation croisée. Ces critères, contrairement aux critères généraux précédents, prennent pleinement en considération l'objectif attendu de la méthode de classification retenue, permettant alors de retenir des modèles plus adaptés au réel besoin de l'utilisateur.

1.1 Classification et modèle de mélange

1.1.1 Problématiques de la classification

Nous entendons par classification, deux problèmes généralement distincts : la classification supervisée dite aussi analyse discriminante et la classification non supervisée dite aussi classification automatique.

En analyse discriminante, les données sont constituées d'une part de n individus $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$ appartenant à un espace $\mathcal{X}$ de dimension d, et d'autre part d'une partition de ceux-ci en K groupes $G_1,\ldots,G_K$. Cette partition est notée $\mathbf{z} = (\mathbf{z}_1,\ldots,\mathbf{z}_n)$, où $\mathbf{z}_i = (z_{i1},\ldots,z_{iK})'$ est un vecteur de $\{0,1\}^K$ tel que $z_{ik} = 1$ si l'individu $\mathbf{x}_i$ appartient au $k$ème groupe G_k, et $z_{ik} = 0$ sinon ($i = 1,\ldots,n$, $k = 1,\ldots,K$). L'ensemble des données est donc constitué de tous les couples $\mathcal{D} = (\mathbf{x},\mathbf{z}) = ((\mathbf{x}_1,\mathbf{z}_1),\ldots,(\mathbf{x}_n,\mathbf{z}_n))$. Il est qualifié d'ensemble d'apprentissage. L'objectif est d'estimer le groupe $\mathbf{z}_{n+1}$ de tout nouvel individu $\mathbf{x}_{n+1}$ de $\mathcal{X}$ dont la classe est inconnue. En d'autres termes, il s'agit d'estimer une règle classement r à partir de $\mathcal{D}$ et définie comme suit :

$$\begin{array}{cccc} r: & \mathcal{X} & \longrightarrow & \{1,\ldots,K\} \\ & \mathbf{x}_{n+1} & \longmapsto & r(\mathbf{x}_{n+1}). \end{array} \tag{1.1}$$

Notons que l'espace des individus $\mathcal{X}$ correspond généralement à $\mathbb{R}^d$ dans le cas continu ou encore à $\{0,1\}^d$ dans le cas binaire.

En classification automatique, seuls les individus sont connus et donc les données correspondent à l'ensemble $\mathcal{D} = \mathbf{x}$. L'objectif est alors d'estimer une partition $\mathbf{z}$ de $\mathbf{x}$. Dans sa version la plus générale et la plus difficile, le nombre de groupes K doit également être estimé par $\hat{K}$.

Nous proposons de présenter les deux problématiques dans un cadre légèrement plus général où l'ensemble des données est composé de m individus ($0 \leq m \leq n$) $\mathbf{x}^l = (\mathbf{x}_1,\ldots,\mathbf{x}_m)$ dont les appartenances aux groupes $\mathbf{z}^l = (\mathbf{z}_1,\ldots,\mathbf{z}_m)$ sont connues, tandis que les $n - m$ individus restant $\mathbf{x}^u = (\mathbf{x}_{m+1},\ldots,\mathbf{x}_n)$ sont de labels inconnus $\mathbf{z}^u = (\mathbf{z}_{m+1},\ldots,\mathbf{z}_n)$. Nous écrirons alors $\mathcal{D} = (\mathcal{D}_l,\mathcal{D}_u)$ avec $\mathcal{D}_l = (\mathbf{x}^l,\mathbf{z}^l)$ et $\mathcal{D}_u = \mathbf{x}^u$. Ainsi, en classification automatique, seuls les labels $\mathbf{z}_{m+1}$ à $\mathbf{z}_n$ doivent être estimés et en analyse discriminante même les points non classés peuvent servir à l'apprentissage d'une règle de classement sur tout l'espace des individus (McLachlan 1992 [39] pp.37–43, [r12] p. 31).

1.1.2 Modélisation par des mélanges de lois

Un point clé de la résolution des problèmes d'analyse discriminante et de classification automatique repose sur la définition précise de la notion de groupe. Intuitivement, un groupe rassemble des éléments qui se ressemblent. Dans un cadre probabiliste, le point commun ou bien encore la ressemblance entre les éléments d'un même groupe peut se traduire par le fait qu'ils proviennent d'une même loi de probabilité. La superposition des lois associées à chaque groupe conduit alors naturellement à un mélange de lois.

Ainsi, dire qu'un individu $\mathbf{x}_1 \in \mathcal{X}$ appartient au groupe G_k est équivalent à dire que cet individu est réalisation d'une variable aléatoire $\mathbf{X}_1 \in \mathcal{X}$ conditionnellement au fait que $\{\mathbf{Z}_{1k} = 1\}$, où $\mathbf{Z}_1 = (Z_{11},\ldots,Z_{1K})'$ est un vecteur de $\{0,1\}^K$ indiquant le groupe d'appartenance de $\mathbf{X}_1$ avec la convention déjà introduite, à savoir que $Z_{ik} = 1$ si l'individu $\mathbf{X}_1$ appartient au $k$ème groupe G_k, et $Z_{1k} = 0$ sinon ($k = 1,\ldots,K$). Nous noterons comme suit la loi de $\mathbf{X}_1$ conditionnelle au groupe G_k, c.-à-d. la loi de la v.a. $\mathbf{X}_{1|Z_{1k}=1}$,

$$\mathbf{X}_{1|Z_{1k}=1} \sim p_k(\mathbf{x}_1). \tag{1.2}$$

La loi de $\mathbf{Z}_1$ est définie en outre par une loi multinomiale d'ordre 1

$$\mathbf{Z}_1 \sim \mathsf{Mult}_K(1,\mathbf{p}), \tag{1.3}$$

avec $\mathbf{p} = (\pi_1,\ldots,\pi_K)$ où π_k ($k = 1,\ldots,K$) représente la proportion de la composante k dans le mélange où encore la probabilité *a priori* qu'un individu provienne de cette composante, c'est-à-dire ($\sum_{k=1}^K \pi_k = 1$ et $\pi_k \geq 0$):

$$\pi_k = p(Z_{1k} = 1). \tag{1.4}$$

Cela revient ainsi à considérer que chaque groupe $G_1,\ldots,G_K$ est présent en proportions $\pi_1,\ldots,\pi_K$ respectivement. La loi jointe du couple $(\mathbf{X}_1,\mathbf{Z}_1)$ s'écrit alors

$$(\mathbf{X}_1,\mathbf{Z}_1) \sim \prod_{k=1}^{K} \left[\pi_k p_k(\mathbf{x}_1)\right]^{z_{1k}}, \tag{1.5}$$

et on déduit la loi marginale de $\mathbf{X}_1$, $p(\mathbf{x}_1)$, appelée aussi loi mélange :

$$\mathbf{X}_1 \sim p(\mathbf{x}_1) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_1). \tag{1.6}$$

Adoptant cette modélisation, la loi de $\mathbf{Z}_1$ conditionnelle à $\{\mathbf{X}_1 = \mathbf{x}_1\}$, c'est-à-dire la loi de la variable $\mathbf{Z}_{1|\mathbf{X}_1=\mathbf{x}_1}$, est donnée par

$$\mathbf{Z}_{1|\mathbf{X}_1=\mathbf{x}_1} \sim \mathsf{Mult}_K(1,\mathbf{t}_1), \tag{1.7}$$

où $\mathbf{t}_1 = (t_{11},\ldots,t_{1K})$ et t_{1k} ($k = 1,\ldots,K$) est une probabilité conditionnelle déduite simplement par le théorème de Bayes

$$\begin{aligned} t_{1k} &= p(Z_{1k} = 1|\mathbf{X}_1 = \mathbf{x}_1) \\ &= \frac{\pi_k p_k(\mathbf{x}_1)}{p(\mathbf{x}_1)}. \end{aligned} \tag{1.8}$$

Grâce à ces probabilités conditionnelles, il est possible de proposer une règle de classement r pour tout individu $\mathbf{x}_1$ de $\mathcal{X}$ par la méthode du *maximum a posteriori* (nous la noterons MAP) et ainsi de répondre du même mouvement aux deux problématiques de l'analyse discriminante et de la classification automatique. Cette règle consiste simplement à classer un individu dans le groupe de plus forte probabilité :

$$\forall \mathbf{x}_1 \in \mathcal{X} \quad r(\mathbf{x}_1) = k \quad \text{si } t_{1k} \geq t_{1h} \quad \text{pour } h = 1,\ldots,K. \tag{1.9}$$

Au delà de l'aspect intuitif de cette procédure de classement, se cache une notion plus subtile. Considérant égaux les coûts de mauvais classement pour chacun des groupes (ce qui est souvent le cas), utiliser la règle du MAP revient en fait à minimiser la probabilité d'erreur de classement $e(r)$ associée à toute règle r et définie par

$$\begin{aligned} e(r) &= \sum_{k=1}^{K} \pi_k \sum_{h=1,h\neq k}^{K} p(r(\mathbf{X}_1) = h|Z_{ik} = 1) \qquad (1.10)\\ &= 1 - \mathsf{E}_{(\mathbf{X}_1,\mathbf{Z}_1)}[Z_{1r(\mathbf{X}_1)}]. \qquad (1.11) \end{aligned}$$

Cette règle optimale est généralement qualifiée de *règle de Bayes* en théorie de la décision. On trouvera les développements nécessaires dans de nombreuses références comme MacLachlan (1992) [39] chap. 1, Flury (1997)[27] chap. 7 ou encore [r12] chap. 1.

1.1.3 Formulation paramétrique

Considérées sous l'angle des mélanges, l'analyse discriminante et la classification automatique se résument au calcul des probabilités conditionnelles puis à l'utilisation de la règle optimale du MAP. Les probabilités conditionnelles s'exprimant en fonction des proportions $\pi_1,\ldots,\pi_K$ et des lois conditionnelles $p_1(\mathbf{x}_1),\ldots,p_K(\mathbf{x}_1)$, ces quantités doivent être estimées non seulement à partir des données disponibles $\mathcal{D}$ mais aussi à l'aide d'hypothèses plus ou moins réalistes, en tout cas souvent simplificatrices, dont on dispose sur ce modèle de mélange.

Une première hypothèse concerne le type d'échantillonnage. Les couples individus-labels $(\mathbf{x}_1,\mathbf{z}_1),\ldots,(\mathbf{x}_n,\mathbf{z}_n)$ sont supposés être des réalisations i.i.d. (indépendamment et identiquement distribuées) de n couples de variables aléatoires $(\mathbf{X}_1,\mathbf{Z}_1),\ldots,(\mathbf{X}_n,\mathbf{Z}_n)$ suivant la même loi que $(\mathbf{X}_1,\mathbf{Z}_1)$, loi définie par (1.5). Cette hypothèse est faite tout autant en classification automatique qu'en analyse discriminante même si, il est vrai, les labels ne sont pas observés dans ce premier cas. Notons que cette hypothèse d'indépendance peut être remise en cause au besoin comme dans les modèles de Markov cachés où l'indépendance entre les variables conditionnelles $\mathbf{X}_{1|Z_{1k}=1},\ldots,\mathbf{X}_{n|Z_{nk}=1}$ est conservée tandis qu'elle est relaxée entre les variables $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$ (voir par exemple Besag 1986 [6], MacLachlan et Peel 2000 [38] chap. 13).

Une seconde hypothèse concerne les lois conditionnelles $p_1(\mathbf{x}_1),\ldots,p_K(\mathbf{x}_1)$. Délaissant les estimations non paramétriques (Silverman 1986 [54], McLachlan 1992 [39] chap. 9), nous ferons l'hypothèse que $p_k(\mathbf{x}_1)$ est totalement définie par un paramètre éventuellement vectoriel $\boldsymbol{\alpha}_k$, ce que nous noterons $(k=1,\ldots,K)$

$$p_k(\mathbf{x}_1) = p(\mathbf{x}_1;\boldsymbol{\alpha}_k). \tag{1.12}$$

Notant $\boldsymbol{\theta} = (\mathbf{p},\mathbf{a})$ le paramètre du mélange avec $\mathbf{p} = (\pi_1,\ldots,\pi_K)$ et $\mathbf{a} = (\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_K)$, la loi mélange est alors donnée par

$$\begin{aligned} p(\mathbf{x}_1) &= p(\mathbf{x}_1;\boldsymbol{\theta}) \\ &= \sum_{k=1}^{K} \pi_k p(\mathbf{x}_1;\boldsymbol{\alpha}_k), \end{aligned} \tag{1.13}$$

et la probabilité conditionnelle est alors paramétrée aussi par $\boldsymbol{\theta}$: $t_{1k} = t_{1k}(\boldsymbol{\theta})$. Ainsi, le couple constitué par la loi paramétrique $p(\boldsymbol{x}_1;\boldsymbol{\theta})$ et un espace Θ dans lequel évolue ce paramètre définit ce que nous appellerons maintenant un *modèle* et que nous noterons de façon générique $\mathbf{m}$:

$$\mathbf{m} = \{p(\mathbf{x}_1;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}. \tag{1.14}$$

Nous appellerons aussi $\nu = \dim(\Theta)$ le nombre de paramètres dans le modèle.

Dans la suite, nous supposerons que la famille de mélange de densités considérée est identifiable, c'est-à-dire que deux paramètres différents, même avec permutation des labels, conduisent à deux densités différentes, et réciproquement (voir McLachlan et Peel 2000 [38] sec. 1.14).

1.1.4 Estimation de la règle optimale

À ce stade, la méthode d'estimation des probabilités conditionnelles la plus simple et certainement la plus répandue est celle du *plug-in*. Cette méthode consiste à estimer le paramètre $\boldsymbol{\theta}$, sous les contraintes d'un modèle $\mathbf{m}$, puis à utiliser directement cette estimation dans l'expression des probabilités conditionnelles. Nous délaissons donc des méthodes alternatives qui estiment directement les paramètres d'intérêt que sont les probabilités conditionnelles : en analyse discriminante il s'agit soit de la méthode prédictive bayésienne (voir Ripley 1996 p. 45–55 [48] ou [r12] p. 18), soit de la régression logistique (voir Ripley 1996 p. 43–45 [48]); en classification automatique il s'agit de la classification bayésienne de Binder (1978) [8] (voir aussi [r12] p. 20). Considérant donc le principe du *plug-in*, Pearson (1894) [46] utilisa à l'origine la méthode des moments pour estimer le paramètre de mélange dans le cas d'un mélange gaussien bimodal univarié. Malgré quelque regain de popularité de cette approche (voir par exemple Monfrini (2003) [42] ou encore les références dans McLachlan et Peel (2000) [38] chap. 1) cette méthode est peu utilisée aujourd'hui. Délaissant aussi les techniques bayésiennes pour estimer $\boldsymbol{\theta}$ (voir par exemple Robert (1992) [50] ou une discussion dans [r12] pp. 38–43) nous retiendrons la méthode du maximum de vraisemblance

(MV) pour sa popularité, sa simplicité de mise en œuvre et les bonnes propriétés de ses estimateurs sous des conditions assez générales, en particulier que le maximum global existe et soit unique (voir Lehmann (1983) [36] chap. 6). Nous reportons au chapitre suivant la discussion liée aux problèmes de non existence ou de non unicité du maximum. Notant $\ell(\boldsymbol{\theta};\mathcal{D}) = p(\mathcal{D};\boldsymbol{\theta})$ la log-vraisemblance de $\boldsymbol{\theta}$, l'estimateur du maximum de vraisemblance est donné par

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta};\mathcal{D}). \tag{1.15}$$

La log-vraisemblance s'exprime facilement grâce à l'hypothèse d'indépendance des données. Elle s'écrit

$$\begin{aligned} \ell(\boldsymbol{\theta};\mathcal{D}) &= \ell(\boldsymbol{\theta};\mathcal{D}_l) + \ell(\boldsymbol{\theta};\mathcal{D}_u) && (1.16)\\ &= \sum_{i=1}^{m}\sum_{k=1}^{K} z_{ik} \ln\left(\pi_k p(\mathbf{x}_i;\boldsymbol{\alpha}_k)\right) + \sum_{i=m+1}^{n} \ln\left(p(\mathbf{x}_i;\boldsymbol{\theta})\right). && (1.17) \end{aligned}$$

La log-vraisemblance $\ell(\boldsymbol{\theta};\mathbf{x},\mathbf{z}) = \ell(\boldsymbol{\theta};\mathcal{D},\mathbf{z}^u)$ est qualifiée de « log-vraisemblance complétée » car elle s'exprime avec les données complètes $\mathbf{x}$ et $\mathbf{z}$. Elle est plus facile à optimiser que la log-vraisemblance $\ell(\boldsymbol{\theta};\mathcal{D})$ (si bien sûr $\mathcal{D}_u \neq \emptyset$), la connaissance de la partition $\mathbf{z}$ supprimant la difficulté inhérente au problème de mélange. Pour optimiser $\ell(\boldsymbol{\theta};\mathcal{D})$ dans le cas général, on a souvent recours à l'algorithme EM de Dempster *at al.* (1977) [22], algorithme de données incomplètes connu pour sa facilité de mise en œuvre et ses bonnes propriétés théoriques (voir McLachlan et Krishnan 1997 [40] pour une revue détaillée). Des variantes sont possibles comme une version stochastique de EM, l'algorithme SEM de Celeux et Diebolt (1985) [16] qui ne converge pas ponctuellement vers $\hat{\boldsymbol{\theta}}$ mais dont la moyenne de la loi stationnaire approche cet optimum global. Tout comme les questions d'existence et d'unicité du maximum de vraisemblance, les procédures maximisant la vraisemblance seront discutées plus en détail dans le chapitre suivant.

Notons enfin, qu'en classification automatique, il existe une méthode d'estimation spécifique qualifiée d'*approche classification* par opposition à celle du maximum de vraisemblance qualifiée d'*approche mélange* (Celeux et Govaert 1993 [19]). Elle consiste à maximiser la log-vraisemblance complétée $\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{z}^u)$ sur le couple $(\boldsymbol{\theta},\mathbf{z}^u)$:

$$(\hat{\boldsymbol{\theta}}_c,\hat{\mathbf{z}}_c^u) = \arg\max_{\boldsymbol{\theta}\in\Theta,\mathbf{z}^u} \ell(\boldsymbol{\theta};\mathcal{D},\mathbf{z}^u). \tag{1.18}$$

L'intérêt de cette approche est de tenir compte de l'objectif de classification sans pour autant sacrifier à la simplicité du principe de *plug-in*. Ainsi, lorsque la taille de l'échantillon est faible, on constate que l'estimation de la partition est meilleure avec l'approche classification qu'avec l'approche mélange (voir par exemple [r12] p. 52). Par contre, l'estimateur du maximum de vraisemblance complétée $\hat{\boldsymbol{\theta}}_c$ est biaisé, même asymptotiquement, en particulier si les groupes sont faiblement séparés (Bryant et Williamson (1978) [15]). Cependant, un autre apport intéressant de l'approche classification est de retrouver des techniques classiques et non probabilistes en classification automatique (Celeux et Govaert 1993 [19]). Nous en donnerons des exemples dans le cas gaussien ci-dessous. L'optimisation de la vraisemblance complétée peut se faire avec l'algorithme CEM qui est une version classification de EM (Celeux et Govaert (1992) [18]).

1.1.5 Modèles gaussiens

Il s'agit maintenant de présenter des familles de mélange classiquement utilisées, ainsi que les différentes hypothèses qui sont souvent faites sur l'espace Θ du paramètre de mélange. Le modèle de mélange gaussien multivarié est certainement le plus connu et le plus employé. Il a été largement utilisé aussi bien en classification automatique (voir par exemple Wolfe (1970) [62], Bock (1981) [9]) qu'en analyse discriminante (voir les références dans McLachlan (1992) [39]). Dans le contexte gaussien, les données $\mathbf{x}_i$ $(i = 1,\dots,n)$ sont des variables continues de $\mathbb{R}^d$ et la densité conditionnelle des composantes s'écrit $(k = 1,\dots,K)$

$$p(\mathbf{x}_i;\boldsymbol{\alpha}_k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\right), \tag{1.19}$$

avec $\boldsymbol{\alpha}_k = (\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ la moyenne de la composante et $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$ sa matrice de variance-covariance. À ce stade, il est tout à fait courant d'imposer des restrictions sur le paramètre $\boldsymbol{\theta}$ au travers de l'espace Θ, et ce, pour deux raisons essentielles : soit on dispose d'une information *a priori* que l'on prend ainsi en considération, soit la taille de l'échantillon est trop petite pour retenir le modèle général. L'objectif recherché est bien entendu d'obtenir une meilleure estimation de $\boldsymbol{\theta}$, donc, par ricochet, des probabilités conditionnelles.

Au delà de contraintes simples que l'on peut envisager sur les proportions (typiquement fixer les proportions égales ou bien les estimer toutes), des contraintes spécifiques au cas gaussien ont été développées pour les matrices de variance. Reprenant l'approche de Banfield et Raftery (1993) [5], Celeux et Govaert (1995) [20] proposent une décomposition spectrale des matrices de variance qui permet une interprétation géométrique simple. Chaque matrice est décomposée sous la forme $\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k$, avec $\lambda_k = |\mathbf{\Sigma}_k|^{1/d}$ représentant le volume de la composante k, $\mathbf{D}_k$ la matrice orthogonale des vecteur propres de $\mathbf{\Sigma}_k$ qui s'interprète comme l'orientation de cette composante et $\mathbf{A}_k$ la matrice diagonale des valeurs propres normalisées classées par ordre décroissant sur la diagonale et de déterminant un qui s'interprète comme la forme de cette composante. En permettant à certains paramètres, et non nécessairement à tous, de varier ou non entre composantes, Celeux et Govaert obtiennent quatorze modèles différents qu'ils regroupent en trois familles : la famille sphérique où la forme est l'identité et donc seul le volume à un rôle, la famille diagonale où la matrice de variance est diagonale, et enfin la famille générale qui regroupe tous les autres cas (par exemple le cas homoscédastique des matrices égales ou encore le cas le plus général des matrices sans aucune contrainte). Leur article détaille l'étape M des algorithmes de type EM (EM, CEM et SEM) pour chacun de ces modèles.

En classification automatique, Celeux et Govaert (1992) [18] ont montré que l'approche classification permettait de retrouver, en fonction du modèle considéré, certains critères classiques construits sur des notions de distances. Ainsi, dans le cas d'un modèle à proportions égales, le critère des K-means de Ward (1963) [59] revient à considérer un modèle sphérique avec composantes de même volume, le critère de Friedman et Rubin (1967) [28] correspond à un modèle homoscédastique et le critère de Scott et Symons (1971) [53] se retrouve par le modèle le plus général.

De la même façon, en analyse discriminante, la règle de discrimination linéaire de Fisher (1936) [26] utilisant la distance de Mahalanobis entre deux groupes peut s'interpréter comme un mélange gaussien avec modèle homoscédastique. Notons que Friedman (1989) [29] proposa une méthode de régularisation afin de réaliser un compromis entre une règle linéaire (modèle homoscédastique) et une règle quadratique générale (modèle sans aucune contrainte). Néanmoins, contrairement aux quatorze modèles gaussiens précédents, cette régularisation ne fournit pas des règles de classement facilement interprétables.

1.1.6 Modèles des classes latentes

L'utilisation de variables binaires est très courante en statistique : on peut citer par exemple, les données de présence-absence en écologie ou le codage en noir et blanc des pixels en traitement d'images. Les données $\mathbf{x}_i$ $(i = 1,\ldots,n)$ sont des vecteurs $(x_{i1},\ldots,x_{id})'$ de valeurs x_{ij} appartenant à l'ensemble $\{0,1\}$. Précisons que l'approche que nous allons décrire peut être étendue à des variables qualitatives générales, généralisant ainsi les données binaires à plus de deux modalités. Dans le cas binaire, le modèle des classes latentes (Everitt 1984 [24], Celeux et Govaert 1991 [17]) suppose que chaque variable $X_{ij|Z_{ik}=1}$ suit une loi de Bernoulli

$$p(x_{ij};\alpha_{kj}) = (\alpha_{kj})^{x_{ij}}(1-\alpha_{kj})^{1-x_{ij}}, \tag{1.20}$$

où $\alpha_{kj} = p(X_{ij} = 1|Z_{ik} = 1)$. Faisant l'hypothèse d'indépendance conditionnelle des coordonnées de $\mathbf{X}_{i|Z_{ik}=1}$, la loi de la composante k s'écrit alors comme le produit des termes précédents :

$$p(\mathbf{x}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{d}(\alpha_{kj})^{x_{ij}}(1-\alpha_{kj})^{1-x_{ij}}, \tag{1.21}$$

avec $\boldsymbol{\alpha}_k = (\alpha_{k1},\ldots,\alpha_{kd})'$. Afin de proposer des situations parcimonieuses, le modèle peut être réécrit de la façon suivante (Celeux et Govaert 1991 [17]) :

$$p(\mathbf{x}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{d}(\varepsilon_{kj})^{|x_{ij}-\delta_{kj}|}(1-\varepsilon_{kj})^{1-|x_{ij}-\delta_{kj}|} \tag{1.22}$$

où $(\delta_{kj},\varepsilon_{kj}) = (0,\alpha_{kj})$ si $\alpha_{kj} < 1/2$ et $(\delta_{kj},\varepsilon_{kj}) = (1,1-\alpha_{kj})$ sinon. Ainsi les paramètres $\boldsymbol{\alpha}_k$ sont maintenant définis par $\boldsymbol{\alpha}_k = (\boldsymbol{\delta}_k,\boldsymbol{\varepsilon}_k)$ avec d'une part $\boldsymbol{\delta}_k = (\delta_{k1},\ldots,\delta_{kd})'$ un vecteur binaire de dimension

d qui joue le rôle de centre du groupe et qui donne pour chaque variable la valeur binaire la plus fréquente, et d'autre part $\boldsymbol{\varepsilon}_k = (\varepsilon_{k1}, \ldots, \varepsilon_{kd})'$ un vecteur appartenant à l'ensemble $]0,1/2[^d$ qui caractérise la dispersion du composant et qui est pour chaque variable la probabilité d'avoir une valeur différente du centre. On retrouve ainsi le paramétrage utilisé par Aitchinson et Aitken (1976) [1] en discrimination non paramétrique sur variables qualitatives par la méthode des noyaux.

À partir de cette formulation, il est possible d'envisager des situations parcimonieuses en imposant des contraintes sur le modèle le plus général. Ainsi, trois autres modèles sont proposés : le modèle le plus simple ne dépend ni du composant, ni de la variable; un autre modèle impose de ne dépendre que du composant; enfin le dernier modèle impose de ne dépendre que de la variable. Dans chaque situation, Celeux et Govaert explicitent l'étape M des algorithmes de type EM.

Notons, qu'en classification automatique, l'approche classification avec le modèle le plus simple et des proportions égales s'identifie à une critère de classification binaire établi par ailleurs dans un cadre non probabiliste (voir par exemple Gower 1974 [31]).

1.2 Critères généraux de choix de modèles

1.2.1 Nécessité de sélectionner un modèle

Le fait d'avoir décidé d'un modèle paramétrique $\mathbf{m} = \{p(\mathbf{x}_1;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ pour la loi mélange $p(\mathbf{x}_1)$ a permis de se placer dans un cadre simplifié où les outils puissants de l'inférence paramétrique sont disponibles. L'hypothèse paramétrique est néanmoins contraignante car la vraie loi peut être bien différente de la loi candidate paramétrée : par exemple, les densités des groupes ne sont pas gaussiennes ou encore le nombre réel de composants est supérieur à celui pris en charge par le modèle. On imagine alors aisément les conséquences sur l'estimation de la vraie loi $p(\mathbf{x}_1)$. Il faut aussi comprendre par mauvais modèle, la notion plus subtile de modèle surparamétré. Par exemple, l'utilisation d'un modèle gaussien général alors que les « vraies » composantes sont des gaussiennes sphériques conduira, pour des petites tailles d'échantillon, à des estimations de qualité inférieure à celles données par un modèle faisant cette hypothèse sphérique. La cause essentielle est la variabilité de l'estimation quand trop de paramètres sont inconnus. On aurait tout autant ce problème avec un nombre de composants dans le modèle supérieur au vrai nombre de groupes.

Afin de limiter l'aspect subjectif du modèle paramétrique on peut alors envisager de remettre en cause ce modèle, ce qui permet d'introduire tout de même une certaine flexibilité dans la modélisation. En général, on dispose d'une famille $\mathbf{m}_1, \ldots, \mathbf{m}_q$ de modèles, et il s'agit de choisir l'un d'entre eux. Dans ce cadre, les tests d'hypothèses, comme le test du rapport des vraisemblances maximales, sont certes disponibles mais finalement peu adaptés à la sélection pour deux raisons essentielles. D'une part, un enchaînement de tests entraîne inévitablement une incertitude sur le risque associé au test global. D'autre part, le test du rapport des vraisemblances maximales ne permet que de tester des modèles « emboîtés », c'est-à-dire des modèles dont les espaces de paramètres ont une relation d'inclusion (voir plus de détails sur ce type de tests dans [r12] pp. 60–63 par exemple).

Par ailleurs, dans le cas d'un test entre deux nombres distincts de composants, la loi limite du rapport des vraisemblances maximales ne suit plus nécessairement une distribution du χ^2 avec le nombre habituel de degrés de libertés malgré l'emboîtement des modèles (Aitkin et Rubin (1985) [2], Everitt (1981) [25]). En effet, la validité de la distribution asymptotique classique du rapport des vraisemblances se fonde sur un développement de Taylor de la log-vraisemblance, développement contestable s'il s'effectue au voisinage d'un paramètre qui se trouve sur la frontière de l'espace des paramètres comme c'est le cas si certaines proportions π_k sont nulles pour indiquer le fait que des composantes n'existent pas. En fait, des propositions existent pour contourner cette difficulté : Wolfe (1971) [62] conjecture une autre distribution de la statistique de test en remarquant qu'il y a incertitude sur le nombre de degrés de liberté dans la loi du χ^2, Aitkin et Rubin (1985) [2] évitent de tester sur les proportions en les manipulant comme des paramètres de nuisance, ou encore McLachlan (1987) [41] propose un bootstrap paramétrique qui permet d'estimer la loi de la statistique. Nous verrons cependant que la difficulté spécifique apportée par le choix du nombre de composants n'est pas propre aux tests car nous la retrouverons dans nombre de critères que nous allons présenter.

Notons que les praticiens, alternativement au test du rapport des vraisemblances maximales, utilisent parfois l'approche empirique qui consiste à détecter un coude dans le courbe du MV, afin de choisir un modèle dans le cas de modèles emboîtés. Par exemple, notant $\ell_K = \ell(\hat{\boldsymbol{\theta}}_K;\mathcal{D})$ l'estimateur du MV pour K composantes, Cutler et Windham (1993) [21] proposent le critère EL (*Elbow Likelihood*) qui, procédant dans l'ordre croissant du nombre de groupes à partir de $K = 1$, sélectionne K groupes si $\ell^-_{K+1} \leq \ell_K$ où $\ell^-_{K+1} = \ell_{K+1} - 0.01|\ell_1|$. Afin de lever une part d'empirisme de cette procédure, une autre définition de ℓ^-_{K+1}, s'appuyant cette fois sur la précision des données, à été introduite par [R4] et [C5] dans le cas homoscédastique. L'idée de base consiste à remplacer les données $\mathbf{x}_1,\ldots,\mathbf{x}_n$ par $\mathbf{x}_1+\mathbf{U}_1,\ldots,\mathbf{x}_n+\mathbf{U}_n$ où $\mathbf{U}=(\mathbf{U}_1,\ldots,\mathbf{U}_n)$ regroupe des vecteurs aléatoires i.i.d. uniformément dans un hyperparallélépipède rectangle dont la longueur des demi côtés représente la précision sur les données. Cette précision traduit aussi bien une incertitude de mesure sur les données qu'un ordre de grandeur jugé suffisant par l'expert du domaine (même si les données sont beaucoup plus précises en réalité). Notant $\hat{\boldsymbol{\theta}}_K(\mathbf{U})$ l'estimateur du MV obtenu à partir de $\mathcal{D}(\mathbf{U})$, où les individus $\mathbf{x}_1,\ldots,\mathbf{x}_n$ ont été remplacés par $\mathbf{x}_1+\mathbf{U}_1,\ldots,\mathbf{x}_n+\mathbf{U}_n$, on définit finalement ℓ^-_K par $\ell^-_K = \mathsf{E}_{\mathbf{U}'}[\ell(\mathsf{E}_{\mathbf{U}}[\hat{\boldsymbol{\theta}}_K(\mathbf{U})]);\mathcal{D}(\mathbf{U}')]$. Si les probabilités conditionnelles d'appartenance aux groupes calculés par $\hat{\boldsymbol{\theta}}_K(\mathbf{U})$ sont peu différentes de celles calculées par $\hat{\boldsymbol{\theta}}_K$ (condition raisonnable si l'imprécision n'est pas trop forte), on peut montrer que $\ell^-_K \leq \ell_K$. Dans ce cas, le calcul de $\mathsf{E}_{\mathbf{U}}[\hat{\boldsymbol{\theta}}_K(\mathbf{U})]$ est explicite et l'estimation de l'espérance de la log-vraisemblance dans l'expression de ℓ^-_K se fait par une méthode de Monte-Carlo.

Les critères présentés ici s'affranchissent de toute notion de test statistique. Ils se contentent de sélectionner le modèle, parmi l'ensemble $\mathbf{m}_1,\ldots,\mathbf{m}_q$, qui est le plus proche des données en un certain sens. On peut les qualifier de critères généraux car ils dépassent le cadre de la classification par modèle de mélange, s'intéressant à tout type de modèle paramétré au sens large, même si à certains moments nous ferons des commentaires spécifiques au cas des mélanges. Ils s'expriment souvent comme une pénalisation du maximum de log-vraisemblance par une mesure de la complexité du modèle considéré. Les critères décrits ici sont les principaux critères de ce genre mais il peut en exister quelques autres comme le critère ICOMP (*Informational Complexity*) de Bozdogan (1990, 1993, 1995) [12, 13, 14] dont la complexité, dite de van Emden, est décrite par une fonction du nombre de paramètres du modèle et de la matrice d'information de Fisher (voir par exemple [r12] pp. 73–74)).

1.2.2 Approche fréquentiste et déviance

Dans un contexte fréquentiste, il semble raisonnable de retenir le modèle $\mathbf{m}$ minimisant la *déviance moyenne* D de l'échantillon et qui s'exprime par la double espérance suivante

$$D = 2\mathsf{E}_{\mathcal{D}\mathcal{D}'}[\ln p(\mathcal{D}') - \ln p(\mathcal{D}';\hat{\boldsymbol{\theta}})], \tag{1.23}$$

où $\hat{\boldsymbol{\theta}}$ correspond à l'estimateur du maximum de vraisemblance qui est donc une fonction de l'échantillon $\mathcal{D}$ ($\mathcal{D}$ est considéré dans sa version aléatoire) et aussi du modèle $\mathbf{m}$, $\mathcal{D}'$ est un échantillon indépendant de $\mathcal{D}$ mais de même loi, et les espérances sur $\mathcal{D}$ et $\mathcal{D}'$ sont prises respectivement par rapport à la vraie loi jointe $p(\mathcal{D})$ et $p(\mathcal{D}')$. On peut interpréter D comme le double de la moyenne (sur $\mathcal{D}$) de la *divergence de Kullback-Leibler* entre la vraie loi jointe des données $p(\mathcal{D}')$ (loi de référence) et la loi jointe des données $p(\mathcal{D}';\hat{\boldsymbol{\theta}})$ amenée par le modèle.

En vue de donner une estimation de D, nous devons tout d'abord énoncer et démontrer deux propositions qui généralisent respectivement les propositions 2.2 et 2.3 de Ripley (1996) [48] pp. 32-34. Cette étude particulière est rendue nécessaire par l'hétérogénéité de notre ensemble de données $\mathcal{D}$ qui peut éventuellement être composé de couples de données complètes $(\mathbf{X}_i,\mathbf{Z}_i)$ et aussi incomplètes $\mathbf{X}_i$. Nous noterons dans la suite $\boldsymbol{\theta}_0$ la valeur de $\boldsymbol{\theta}$ qui minimise la divergence de Kullback-Leibler entre $p(\mathcal{D}')$ et $p(\mathcal{D}';\boldsymbol{\theta})$:

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathsf{E}_{\mathcal{D}'}[\ln p(\mathcal{D}') - \ln p(\mathcal{D}';\boldsymbol{\theta})]. \tag{1.24}$$

Sous certaines conditions de régularité, on a $\hat{\boldsymbol{\theta}} \xrightarrow{p.s.} \boldsymbol{\theta}_0$ (voir par exemple White 1982 [60]). Si la vraie loi est dans la famille paramétrique, nous aurons alors bien sûr $p(\mathcal{D}') = p(\mathcal{D}';\boldsymbol{\theta}_0)$.

Proposition 1 *Sous certaines conditions de régularité*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathsf{N}_\nu\left(\mathbf{0},\mathbf{J}_\beta{}^{-1}\mathbf{K}_\beta\mathbf{J}_\beta{}^{-1}\right) \tag{1.25}$$

avec ν le nombre de paramètres à estimer dans le modèle $\mathbf{m}$, $N_\nu(\mathbf{0},\mathbf{V})$ la loi normale ν dimensionnelle de moyenne nulle et de matrice de variance $\mathbf{V}$, $\beta = m/n \in [0,1]$ la proportion d'individus d'appartenance connue dans $\mathcal{D}$, $\mathbf{J}_\beta = (\beta\mathbf{J}_c + (1-\beta)\mathbf{J})$ et $\mathbf{K}_\beta = (\beta\mathbf{K}_c + (1-\beta)\mathbf{K})$ où

$$\mathbf{J}_c = -E_{(\mathbf{X}_1,\mathbf{Z}_1)}\nabla^2 \ln p(\mathbf{X}_1,\mathbf{Z}_1;\boldsymbol{\theta}_0), \qquad \mathbf{J} = -E_{\mathbf{X}_1}\nabla^2 \ln p(\mathbf{X}_1;\boldsymbol{\theta}_0), \tag{1.26}$$

$$\mathbf{K}_c = Var_{(\mathbf{X}_1,\mathbf{Z}_1)}\nabla \ln p(\mathbf{X}_1,\mathbf{Z}_1;\boldsymbol{\theta}_0), \qquad \mathbf{K} = Var_{\mathbf{X}_1}\nabla \ln p(\mathbf{X}_1;\boldsymbol{\theta}_0). \tag{1.27}$$

L'espérance est prise par rapport à la vraie loi jointe $p(\mathbf{x}_1,\mathbf{z}_1)$ pour $\mathbf{J}_c$ et $\mathbf{K}_c$, et par rapport à la vraie loi marginale $p(\mathbf{x}_1)$ pour $\mathbf{J}$ et $\mathbf{K}$. Les dérivées premières et secondes portent sur $\boldsymbol{\theta}$. Si la vraie loi est dans la famille paramétrique, alors $\mathbf{J}_c = \mathbf{K}_c$ et $\mathbf{J} = \mathbf{K}$.

Preuve L'estimateur du maximum de vraisemblance vérifie $\nabla\ell(\hat{\boldsymbol{\theta}};\mathcal{D}) = \mathbf{0}$. Un développement de Taylor à l'ordre un permet d'ecrire

$$\mathbf{0} = \nabla\ell(\hat{\boldsymbol{\theta}};\mathcal{D}) = \nabla\ell(\boldsymbol{\theta}_0;\mathcal{D}) + \nabla^2\ell(\tilde{\boldsymbol{\theta}};\mathcal{D})\ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{1.28}$$

avec $\tilde{\boldsymbol{\theta}}$ un vecteur entre $\hat{\boldsymbol{\theta}}$ et $\boldsymbol{\theta}_0$ au sens multidimensionnel. Utilisant le théorème central limite, la loi des grands nombres et faisant l'hypothèse que le rapport m/n reste égal à β quand $n \to \infty$, on obtient

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \Big[-\frac{1}{n}\nabla^2\ell(\tilde{\boldsymbol{\theta}};\mathcal{D}_l) - \frac{1}{n}\nabla^2\ell(\tilde{\boldsymbol{\theta}};\mathcal{D}_u)\Big]^{-1}\Big[\frac{1}{\sqrt{n}}\nabla\ell(\boldsymbol{\theta}_0;\mathcal{D}_l) + \frac{1}{\sqrt{n}}\nabla\ell(\boldsymbol{\theta}_0;\mathcal{D}_u)\Big] \tag{1.29}$$

$$\xrightarrow{d} \Big[\beta\mathbf{J}_c + (1-\beta)\mathbf{J}\Big]^{-1}\mathsf{N}_\nu\Big(\mathbf{0},\Big[\beta\mathbf{K}_c + (1-\beta)\mathbf{K}\Big]\Big) = \mathsf{N}_\nu\left(\mathbf{0},\mathbf{J}_\beta{}^{-1}\mathbf{K}_\beta\mathbf{J}_\beta{}^{-1}\right). \tag{1.30}$$

Le fait que $\mathbf{J}_c = \mathbf{K}_c$ et $\mathbf{J} = \mathbf{K}$ quand le modèle est exact est une propriété déjà connue (voir par exemple Lehmann 1983 [36] p. 118). □

Proposition 2 *Notant $\nu^* = tr[\mathbf{K}_\beta\mathbf{J}_\beta^{-1}]$, on peut exprimer la moyenne D de la déviance par*

$$D = 2\{\ln p(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\} + 2\nu^* + O_p(\sqrt{n}). \tag{1.31}$$

Si la vraie loi est dans la famille paramétrique, alors $\nu^* = \nu$, le nombre de paramètres.

Preuve Commençons par un développement de Taylor à l'ordre deux en $\boldsymbol{\theta}_0$ de la log-vraisemblance $2\ell(\hat{\boldsymbol{\theta}};\mathcal{D}')$:

$$2\ell(\hat{\boldsymbol{\theta}};\mathcal{D}') \approx 2\ell(\boldsymbol{\theta}_0;\mathcal{D}') + 2(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\nabla\ell(\boldsymbol{\theta}_0;\mathcal{D}') + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\nabla^2\ell(\boldsymbol{\theta}_0;\mathcal{D}')(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{1.32}$$

$$= 2\ell(\boldsymbol{\theta}_0;\mathcal{D}') + 2(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\nabla\ell(\boldsymbol{\theta}_0;\mathcal{D}') + \mathsf{tr}[\nabla^2\ell(\boldsymbol{\theta}_0;\mathcal{D}')(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)']. \tag{1.33}$$

Ce résultat associé au fait que $\mathsf{E}_{\mathcal{D}'}\nabla\ell(\boldsymbol{\theta}_0;\mathcal{D}') = \mathbf{0}$ puis à l'indépendance entre $\mathcal{D}$ et $\mathcal{D}'$, permet d'écrire

$$D \approx 2\mathsf{E}_{\mathcal{D}'}[\ln p(\mathcal{D}') - \ell(\boldsymbol{\theta}_0;\mathcal{D}')] - 2\mathsf{E}_{\mathcal{D}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathsf{E}_{\mathcal{D}'}\nabla\ell(\boldsymbol{\theta}_0;\mathcal{D}') - \mathsf{tr}[\mathsf{E}_{\mathcal{D}'}\nabla^2\ell(\boldsymbol{\theta}_0;\mathcal{D}')\mathsf{E}_{\mathcal{D}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'] \tag{1.34}$$

$$\approx 2\mathsf{E}_{\mathcal{D}'}[\ln p(\mathcal{D}') - \ell(\boldsymbol{\theta}_0;\mathcal{D}')] - \mathsf{tr}[\{\mathsf{E}_{\mathcal{D}'_l}\nabla^2\ell(\boldsymbol{\theta}_0;\mathcal{D}'_l) + \mathsf{E}_{\mathcal{D}'_u}\nabla^2\ell(\boldsymbol{\theta}_0;\mathcal{D}'_u)\}\mathsf{Var}_{\mathcal{D}}\hat{\boldsymbol{\theta}}] \tag{1.35}$$

$$= 2\mathsf{E}_{\mathcal{D}'}[\ln p(\mathcal{D}') - \ell(\boldsymbol{\theta}_0;\mathcal{D}')] - \mathsf{tr}[(-n\mathbf{J}_\beta)(\mathbf{J}_\beta{}^{-1}\mathbf{K}_\beta\mathbf{J}_\beta{}^{-1}/n)] \tag{1.36}$$

$$= 2\mathsf{E}_{\mathcal{D}'}[\ln p(\mathcal{D}') - \ell(\boldsymbol{\theta}_0;\mathcal{D}')] + \mathsf{tr}[\mathbf{K}_\beta\mathbf{J}_\beta{}^{-1}]. \tag{1.37}$$

L'erreur dans cette expression est d'ordre $O(1/\sqrt{n})$. Il reste à estimer le premier terme à partir de l'échantillon $\mathcal{D}$ disponible pour conclure :

$$2\mathsf{E}_{\mathcal{D}'}[\ln p(\mathcal{D}') - \ell(\boldsymbol{\theta}_0;\mathcal{D}')] \approx 2\{\ln p(\mathcal{D}) - \ell(\boldsymbol{\theta}_0;\mathcal{D})\} \tag{1.38}$$

$$\approx 2\{\ln p(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\} - 2(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\nabla\ell(\hat{\boldsymbol{\theta}};\mathcal{D}) - \mathsf{tr}[\nabla^2\ell(\hat{\boldsymbol{\theta}};\mathcal{D})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'] \tag{1.39}$$

$$= 2\{\ln p(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\} - \mathsf{tr}[\{\nabla^2\ell(\hat{\boldsymbol{\theta}};\mathcal{D}_l) + \nabla^2\ell(\hat{\boldsymbol{\theta}};\mathcal{D}_u)\}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'] \tag{1.40}$$

$$\approx 2\{\ln p(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\} - \mathsf{tr}[\{-n\mathbf{J}_\beta\}\mathsf{Var}_{\mathcal{D}}\hat{\boldsymbol{\theta}}] = 2\{\ln p(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\} + \nu^*. \tag{1.41}$$

L'erreur dans cette dernière approximation est d'ordre $O_p(\sqrt{n})$, ce qui devient donc l'ordre global de l'erreur sur D. Notant $\mathbf{I}_\nu$ la matrice identité $\nu \times \nu$, on remarque aussi que si la vraie loi est dans la famille paramétrée, alors $\nu^* = \mathsf{tr}[\mathbf{K}_\beta\mathbf{K}_\beta^{-1}] = \mathsf{tr}\mathbf{I}_\nu = \nu$. □

Ainsi, la déviance moyenne D s'exprime en fonction de la déviance observée $2\{\ln p(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\}$ pénalisée par une mesure de la complexité du modèle, ν^*. On en déduit le critère NIC (*Network Information Criterion*) de Murata *et al.* (1991, 1993, 1994) [43, 44, 45] :

$$\mathrm{NIC} = \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\ - \nu^*, \tag{1.42}$$

et on retient le modèle $\mathbf{m}$ qui a la plus grande valeur de NIC. Lorsque la vraie loi est incluse dans la famille paramétrée candidate, on retrouve le critère AIC (*An Information Criterion*) de Akaike (1973 [3], 1974 [4]) :

$$\text{AIC} = \ell(\hat{\boldsymbol{\theta}}; \mathcal{D}) \ - \nu. \tag{1.43}$$

L'utilisation de NIC en pratique est assez limitée car il est difficile d'estimer les matrices $\mathbf{J}_\beta$ et $\mathbf{K}_\beta$. On peut toujours opter pour sa version simplifiée AIC mais au prix de l'hypothèse douteuse que la vraie loi est la loi paramétrique candidate. Cela impose d'ailleurs de ne choisir que parmi des déclinaisons de modèles emboîtés de cette loi générique. Il est possible aussi de diviser l'échantillon complet en deux parties, un échantillon d'apprentissage et un échantillon test, pour estimer soit les matrices $\mathbf{J}_\beta$ et $\mathbf{K}_\beta$, soit directement la quantité D. L'indépendance des deux types d'échantillon garantit l'absence de biais de l'estimateur obtenu. L'indépendance n'étant pas nécessaire pour justifier l'absence de biais d'une moyenne, une utilisation plus efficace de l'approche *apprentissage-test*, car diminuant la variance de l'estimateur, consiste à moyenner les estimateurs obtenus à partir de divisions *apprentissage-test* différentes du même jeu de données de départ. Dans cette optique, Stone (1977) [56] avait obtenu le critère NIC bien avant Murata *et al.* en établissant un lien entre le critère AIC et le critère de validation croisée. Ainsi, une autre possibilité pour estimer D consiste à employer le critère de validation croisée CV (*Cross Validation*) défini par

$$\text{CV} = \sum_{i=1}^{m} \ln p(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_{\{i\}}) + \sum_{i=m+1}^{n} \ln p(\mathbf{x}_i, \mathbf{z}_i; \hat{\boldsymbol{\theta}}_{\{i\}}), \tag{1.44}$$

où $\hat{\boldsymbol{\theta}}_{\{i\}}$ est l'estimateur du maximum de vraisemblance obtenu à partir des données $\mathcal{D}$ privé de la $i$ème d'entre elles. Ce critère converge asymptotiquement vers NIC. Notons que Smyth (2000) [55] suggère une validation croisée « à gros grain » en itérant avec des échantillons tests de plus d'un individu. On peut penser que cette approche est équivalente à CV si la taille de l'échantillon test est asymptotiquement négligeable par rapport à celle de l'échantillon complet (résultat obtenu dans un autre contexte, celui de la régression linéaire, par Zhang (1993) [63]). Si tel n'est pas le cas, par exemple un échantillon test représentant la moitié de l'échantillon complet, cela devient un nouveau critère concurrent de CV dont il serait intéressant d'étudier théoriquement les différences.

Il est prouvé que NIC est malheureusement un critère inconsistant de choix de modèles car il sélectionne un modèle trop complexe avec une probabilité non nulle, même asymptotiquement. Prenons par exemple deux modèles emboîtés $\mathbf{m}_1$ et $\mathbf{m}_2$ avec $\Delta\nu = \nu_2 - \nu_1 > 0$ et faisons l'hypothèse que le modèle le plus parcimonieux des deux corresponde à la vraie loi. Alors, le développement suivant permet d'établir que l'on peut choisir à tord le modèle le plus complexe pour des grandes tailles d'échantillon

$$2(\text{AIC}_2 - \text{AIC}_1) = 2(\ell(\hat{\boldsymbol{\theta}}_2; \mathcal{D}) - \ell(\hat{\boldsymbol{\theta}}_1; \mathcal{D})) - 2\Delta\nu \xrightarrow{d} \chi^2_{\Delta\nu} - 2\Delta\nu, \tag{1.45}$$

puisque $p(\chi^2_{\Delta\nu} > 2\Delta\nu) > 0$. En fait, dans le cas du choix du nombre de composants d'un mélange, on ne connaît plus la loi du rapport des vraisemblances maximales comme on l'a évoqué dans la section précédente et, par conséquent, on ne sait plus rien sur la non-consistance de AIC dans cette situation. Le choix du nombre de groupes remet même en cause la validité du critère NIC puisque certains développements de Taylor nécessaires pour l'obtenir ne sont plus justifiés. C'est pour cela qu'il existe des variantes de AIC, comme le critère AIC3 de Bozdogan (1981,1983) [10, 11] qui, s'appuyant sur la conjecture de Wolfe (1971) [62] déjà évoquée, propose une pénalisation légèrement plus forte que celle de AIC (voir aussi plus de détails sur ce critère dans [r12] annexe B) :

$$\text{AIC3} = \ell(\hat{\boldsymbol{\theta}}; \mathcal{D}) \ - 1.5\nu. \tag{1.46}$$

Néanmoins, ce nouveau critère ne résout en rien l'inconsistance de AIC même s'il sélectionnera des modèles plus parcimonieux que AIC du fait de sa plus forte pénalisation.

1.2.3 Approche bayésienne et vraisemblance intégrée

Dans un contexte bayésien, le principe est naturellement de retenir le modèle $\mathbf{m}$ dont la probabilité *a posteriori* est la plus grande. Cette probabilité s'exprime par

$$p(\mathbf{m}|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{m})p(\mathbf{m}). \tag{1.47}$$

Si tous les modèles sont *a priori* équiprobables, cela revient à sélectionner le modèle maximisant $p(\mathcal{D}|\mathbf{m})$. Cette quantité, généralement appelée *vraisemblance intégrée* ou encore *vraisemblance marginale*, est le point clé de la méthode bayésienne et s'écrit

$$p(\mathcal{D}|\mathbf{m}) = \int_\Theta p(\mathcal{D}|\boldsymbol{\theta},\mathbf{m})p(\boldsymbol{\theta}|\mathbf{m})d\boldsymbol{\theta}, \tag{1.48}$$

où, avec les notations adoptées jusqu'ici, $p(\mathcal{D}|\boldsymbol{\theta},\mathbf{m}) = p(\mathcal{D};\boldsymbol{\theta})$. Nous noterons aussi, pour simplifier les notations, $p(\boldsymbol{\theta}|\mathbf{m}) = p(\boldsymbol{\theta})$, ne laissant donc apparaître le modèle que dans l'expression de la vraisemblance intégrée $p(\mathcal{D}|\mathbf{m})$. Pour calculer cette probabilité, outre la définition d'une loi *a priori* sur $\boldsymbol{\theta}$, l'évaluation de l'intégrale est nécessaire. L'évaluation analytique étant rarement possible, diverses directions d'approximation existent (voir par exemple Kass et Raftery 1995 [33]) : les méthodes numériques mais elles sont assez instables en grande dimension, les méthodes de Monte-Carlo comme l'échantillonneur de Gibbs ou de Metroplis-Hastings, ou encore l'approximation de Laplace-Metropolis faisant une approximation de Taylor à l'ordre deux de l'intégrale. Une méthode alternative consiste à approcher la log-vraisemblance intégrée par le maximum de log-vraisemblance pénalisé par une fonction du nombre de paramètres ν et aussi de la taille de l'échantillon n. Ainsi, on obtient une expression simple qui permet d'ailleurs de s'affranchir de définir la loi *a priori*. La proposition suivante énonce cette importante propriété (voir par exemple Kass et Wasserman 1995 [34], Raftery 1995 [47] pp. 130–133 ou encore Ripley 1996 [48] pp. 62-65), en la démontrant ici dans le cas général d'un modèle pouvant éventuellement ne pas inclure la vraie loi.

Proposition 3 *Sous certaines conditions de régularité, on a*

$$\ln p(\mathcal{D}|\mathbf{m}) = \ell(\hat{\boldsymbol{\theta}};\mathcal{D}) - \frac{\nu}{2}\ln(n) + O_p(1/\sqrt{n}). \tag{1.49}$$

Preuve La loi *a posteriori* $p(\boldsymbol{\theta}|\mathcal{D})$ de $\boldsymbol{\theta}$ est supposée être approximativement une gaussienne $\mathsf{N}_\nu(\tilde{\boldsymbol{\theta}},\mathbf{V})$. Dans ce cas, la moyenne correspond au mode de cette loi, c.-à-d. $\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} p(\boldsymbol{\theta}|\mathcal{D})$, et la matrice de variance s'identifie à l'inverse du hessien de $-\ln p(\tilde{\boldsymbol{\theta}}|\mathcal{D})$, c.-à-d. $\mathbf{V} = [-\nabla^2 \ln p(\tilde{\boldsymbol{\theta}}|\mathcal{D})]^{-1}$ (il suffit d'effectuer le calcul pour s'en convaincre). Lorsque l'échantillon est suffisamment grand, la loi $p(\boldsymbol{\theta}|\mathcal{D})$ se concentre autour de son mode, donc $g(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta};\mathcal{D}) - \ln p(\boldsymbol{\theta})$ est également concentré autour de $\tilde{\boldsymbol{\theta}}$ puisque $p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp\{-g(\boldsymbol{\theta})\}$. Ainsi, le développement de Taylor à l'ordre deux de $g(\boldsymbol{\theta})$ au voisinage de $\tilde{\boldsymbol{\theta}}$ est valide pour des grands échantillons, ce qui permet d'écrire en remarquant que $\nabla g(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$

$$\begin{aligned} p(\mathcal{D}|\mathbf{m}) &= \int_\Theta p(\mathcal{D};\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_\Theta \exp\{-g(\boldsymbol{\theta})\}d\boldsymbol{\theta} && (1.50)\\ &\approx \exp\{-g(\tilde{\boldsymbol{\theta}})\}\int_\Theta \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})'\mathbf{V}^{-1}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})\right\}d\boldsymbol{\theta} && (1.51)\\ &= \exp\{-g(\tilde{\boldsymbol{\theta}})\}(2\pi)^{\nu/2}|\mathbf{V}|^{1/2}. && (1.52)\end{aligned}$$

La dernière équation provient du fait que l'intégrale est égale au terme normalisateur d'une loi gaussienne $\mathsf{N}_\nu(\tilde{\boldsymbol{\theta}},\mathbf{V})$. L'erreur étant d'ordre $O_p(1/n)$ (voir Tierney et Kadane 1986 [57]), on obtient

$$\ln p(\mathcal{D}|\mathbf{m}) = \ell(\tilde{\boldsymbol{\theta}};\mathcal{D}) + \ln p(\tilde{\boldsymbol{\theta}}) + \frac{\nu}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{V}| + O_p(1/n). \tag{1.53}$$

Pour les grands échantillons, nous faisons les deux approximations $\tilde{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}$ et $\mathbf{V} \approx \frac{1}{n}\hat{\mathbf{J}}_\beta^{-1}\hat{\mathbf{K}}_\beta\hat{\mathbf{J}}_\beta^{-1}$ où $\hat{\mathbf{J}}_\beta$ et $\hat{\mathbf{K}}_\beta$ représentent respectivement $\mathbf{J}_\beta$ et $\mathbf{K}_\beta$ en remplaçant $\boldsymbol{\theta}_0$ par $\hat{\boldsymbol{\theta}}$ à l'intérieur des espérances (voir la proposition 1), et ainsi $|\mathbf{V}| \approx n^{-\nu}|\hat{\mathbf{J}}_\beta^{-1}\hat{\mathbf{K}}_\beta\hat{\mathbf{J}}_\beta^{-1}|$. Une erreur d'ordre $O_p(1/\sqrt{n})$ étant introduite par ces approximations, on obtient

$$\ln p(\mathcal{D}|\mathbf{m}) = \ell(\hat{\boldsymbol{\theta}};\mathcal{D}) + \ln p(\hat{\boldsymbol{\theta}}) + \frac{\nu}{2}\ln(2\pi) - \frac{\nu}{2}\ln(n) + \frac{1}{2}\ln|\hat{\mathbf{J}}_\beta^{-1}\hat{\mathbf{K}}_\beta\hat{\mathbf{J}}_\beta^{-1}| + O_p(1/\sqrt{n}). \tag{1.54}$$

Dans cette équation, le premier terme est d'ordre $O_p(n)$, le quatrième d'ordre $O_p(\ln(n))$ et tous les autres d'ordre inférieur ou égal à $O_p(1)$. Supprimant les termes d'ordre inférieur ou égal $O_p(1)$, cela donne alors :

$$\ln p(\mathcal{D}|\mathbf{m}) = \ell(\hat{\boldsymbol{\theta}};\mathcal{D}) - \frac{\nu}{2}\ln(n) + O_p(1). \tag{1.55}$$

Cette approximation est assez mauvaise car d'ordre élevé. Raftery (1995) [47] pp. 130-133 propose de retenir une loi *a priori* particulière, $p(\boldsymbol{\theta}) = \mathsf{N}_\nu(\hat{\boldsymbol{\theta}},\hat{\mathbf{J}}_\beta^{-1}\hat{\mathbf{K}}_\beta\hat{\mathbf{J}}_\beta^{-1})$, qui apporte, en moyenne, autant d'information qu'une seule observation. Ainsi,

$$\ln p(\hat{\boldsymbol{\theta}}) = -\frac{\nu}{2}\ln(2\pi) - \frac{1}{2}\ln|\hat{\mathbf{J}}_\beta^{-1}\hat{\mathbf{K}}_\beta\hat{\mathbf{J}}_\beta^{-1}| \tag{1.56}$$

et, combinant avec l'équation (1.54), des termes s'éliminent et l'ordre le plus élevé est $O_p(1/\sqrt{n})$. □

Cette approximation conduit à maximiser le critère BIC (*Bayesian Information Criterion*) de Schwarz (1978) [52] :

$$\mathrm{BIC} = \ell(\hat{\boldsymbol{\theta}};\mathcal{D}) - \frac{\nu}{2}\ln(n). \tag{1.57}$$

Contrairement aux critères NIC et AIC, la pénalisation de BIC s'exprime simplement en fonction du nombre de paramètres que le modèle soit ou non adapté : la difficulté d'estimation de ν^* présente dans NIC n'existe donc pas. Ce critère a aussi été proposé en théorie du codage par Rissanen (1989) [49] sous le nom de MDL pour *Minimum Description Length*. Comme la pénalisation de BIC est bien plus forte que celle de AIC pour toute taille raisonnable d'échantillon ($\ln(n) > 2$ donc $n > 8$), on peut s'attendre à ce qu'il choisisse des modèles plus parcimonieux. En fait, on peut même montrer que BIC est un critère consistant. Par exemple, dans le cas des modèles $\mathbf{m}_1$ et $\mathbf{m}_2$ emboîtés considérés à la fin de la section sur AIC, on a, de la même façon que l'équation (1.45),

$$2(\mathrm{BIC}_2 - \mathrm{BIC}_1) = 2(\ell(\hat{\boldsymbol{\theta}}_2;\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}}_1;\mathcal{D})) - \Delta\nu\ln(n) \xrightarrow{d} \chi^2_{\Delta\nu} - \Delta\nu\ln(n). \tag{1.58}$$

Notant $\mu = \Delta\nu$ et $\sigma^2 = 2\Delta\nu$ respectivement la moyenne et la variance de la variable $\chi^2_{\Delta\nu}$ et utilisant l'inégalité de Chebyschev, on peut écrire

$$p(\chi^2_{\Delta\nu} > \Delta\nu\ln(n)) \leq p(|\chi^2_{\Delta\nu} - \mu| > \Delta\nu\ln(n) - \mu) \leq \frac{\sigma^2}{(\Delta\nu\ln(n) - \mu)^2} \xrightarrow{n\to\infty} 0, \tag{1.59}$$

ce qui signifie, qu'asymptotiquement, BIC sélectionnera le modèle le plus simple, donc le vrai modèle $\mathbf{m}_1$. On peut montrer aussi que BIC ne sous estime pas l'ordre du modèle, c.-à-d. que si le modèle le plus complexe $\mathbf{m}_2$ est le vrai cette fois, il sera retenu asymptotiquement avec probabilité un. La preuve se fait de même façon en s'appuyant sur la loi du rapport des vraisemblances maximales qui est une loi du χ^2 décentrée (voir [r12] pp. 74–75).

Bien entendu, ces preuves de consistance reposent toujours sur le fait que le paramètre ne soit pas sur le bord de l'espace Θ, ce qui pose de nouveau le problème de la validité de ces résultats dans le cadre du choix du nombre de composantes d'un mélange de lois. Des travaux spécifiques à ce problème existent. Leroux (1992) [37] a montré que BIC ne sous-estimait pas le vrai nombre de groupes, asymptotiquement. Roeder et Wasserman (1997) [51] ont montré, dans le cadre des mélanges gaussiens pour estimer une densité de façon non paramétrique, que l'usage de BIC pour sélectionner le nombre de gaussiennes conduisait à un estimateur de densité consistant. Keribin (2000) [35] généralise les résultats de Leroux en prouvant, sous certaines conditions et en utilisant une reparamétrisation localement conique pour rendre valable les développements de Taylor, que BIC ne surestime pas non plus le nombre de groupes asymptotiquement.

1.3 Critères spécifiques de choix de modèles

1.3.1 Limitation des critères généraux

Il est essentiel de faire le constat que, généralement, tout modèle est faux. Comme on ne peut donc pas retenir le « vrai » modèle, au moins est-on en droit de retenir le modèle le moins mauvais, notion que l'on peut traduire par un modèle *utile*. Dans cette optique, il faut alors définir l'utilité d'un modèle en fonction de l'objectif que l'on fixe au processus de modélisation. Trois objectifs sont clairement identifiés dans notre contexte de mélange :

- retenir le modèle qui produit la meilleure estimation de la densité mélange : c'est l'approche *mélange* ;
- retenir le modèle qui produit la meilleure règle de classement : c'est l'approche *discrimination* ;
- retenir le modèle qui produit une partition permettant de représenter une structure sous-jacente présente dans les données : c'est l'approche *classification automatique*.

Les critères précédents mesurent la qualité d'un modèle au sens de l'estimation d'une densité. Leur généralité fait qu'ils peuvent s'employer dans le cadre d'une loi de mélange, *modulo* quelques réserves théoriques parfois dans le cas du nombre de composantes. On peut donc les classer dans l'approche mélange définie ci-dessus. Bien entendu, la discrimination et la classification automatique pouvant s'interpréter dans notre exposé comme un sous-produit de l'estimation de la loi mélange, ces critères sont utilisables dans ces contextes spécifiques.

Néanmoins, retenir un modèle qui conduit à une mauvaise estimation de la vraie loi n'a pas nécessairement les mêmes conséquences en classification qu'en mélange. En effet, un mauvais modèle au sens de la densité peut avoir, dans certains cas, des répercutions négatives limitées sur l'estimation soit de la partition, soit de la règle de classement (en fonction de la problématique) si ce modèle retenu affecte peu les valeurs des probabilités conditionnelles. C'est le cas par exemple lorsque les groupes sont bien séparés ou encore lorsque le modèle permet tout de même de bien estimer les probabilités conditionnelles d'appartenance aux groupes dans les zones de forte confusion (voir une illustration dans Ripley (1996) [48] p. 27).

L'objectif de cette section est alors de présenter des critères concurrents aux critères généraux NIC ou BIC et prenant pleinement en considération l'utilité recherchée. Ainsi, en discrimination, on s'appuiera sur un critère lié à l'erreur de classement. En classification automatique, on présentera des critères liés à la notion d'entropie de la partition et favorisant ainsi la séparation des groupes.

1.3.2 Analyse discriminante et erreur de classement

L'objectif premier de l'analyse discriminante est d'estimer une règle de discrimination conduisant à l'erreur de classement la plus faible. Idéalement, on retiendrait alors le modèle dont la règle associée $\hat{r}$ obtenue à partir de $\mathcal{D}$ produit le moins d'erreur en moyenne, critère e qui s'exprime par :

$$e = \mathsf{E}_{\mathcal{D}}[e(\hat{r})] = 1 - \mathsf{E}_{\mathcal{D}(\mathbf{X}_1',\mathbf{Z}_1')}[Z'_{1\hat{r}(\mathbf{X}_1')}]. \tag{1.60}$$

Différentes façons d'estimer e existent. La plus simple d'entre elles est l'erreur *apparente* de classement $\hat{e}^{\text{apparent}}$ définie par

$$\hat{e}^{\text{apparent}} = 1 - \frac{1}{m}\sum_{i=1}^{m} Z_{i\hat{r}(\mathbf{X}_i)}. \tag{1.61}$$

C'est un estimateur convergent de e mais bien connu pour son biais optimiste puisque l'utilisation du même échantillon pour apprendre la règle et la tester conduit naturellement à une sous-estimation de l'erreur en moyenne : $\mathsf{E}_{\mathcal{D}}[\hat{e}^{\text{apparent}}] \leq e$. L'estimation de l'erreur par la méthode de la *partition* est plus pertinente car elle divise l'échantillon complet en deux échantillons distincts. Le premier, noté $\mathcal{D}_t$, est composé de $\mathcal{D}$ privé des individus d'indices t ($t \subset \{1,\ldots,m\}$) et sert à *apprendre* une règle notée $\hat{r}_t$. Le second est composé des individus restant, c.-à-d. les n_t individus avec labels $\{(\mathbf{x}_i,\mathbf{z}_i)\}_{i\in t} \in \mathcal{D}_l$, et sert à *tester* cette règle $\hat{r}_t$. On a

$$\hat{e}_t^{\text{partition}} = 1 - \frac{1}{n_t}\sum_{i\in t} Z_{i\hat{r}_t(\mathbf{X}_i)}, \tag{1.62}$$

ce qui produit un estimateur sans biais de $\mathsf{E}_{\mathcal{D}_t}[e(\hat{r}_t)]$ puisqu'il est facile de vérifier que

$$\mathsf{E}_{\mathcal{D}}[\hat{e}_t^{\text{partition}}] = 1 - \mathsf{E}_{\mathcal{D}_t(\mathbf{X}_1',\mathbf{Z}_1')}[Z'_{1\hat{r}_t(\mathbf{X}_1')}] = \mathsf{E}_{\mathcal{D}_t}[e(\hat{r}_t)]. \tag{1.63}$$

On remarque tout de suite que $\mathsf{E}_{\mathcal{D}_t}[e(\hat{r}_t)] \simeq e$ uniquement si l'ensemble d'apprentissage est assez grand. Il vient alors logiquement l'idée de choisir un ensemble de test composé d'un unique individu $t = \{i\} \in \{1,\ldots,m\}$ pour agrandir au maximum l'ensemble d'apprentissage, ce qui conduit à l'estimateur $\hat{e}_{\{i\}}^{\text{partition}}$. Comme cet estimateur à trop de variance à cause de la taille excessivement faible de son ensemble de test, l'estimateur par *validation croisée* propose de le moyenner sur tous les individus avec labels :

$$\hat{e}^{\text{cv}} = \frac{1}{m}\sum_{i=1}^{m} \hat{e}_{\{i\}}^{\text{partition}}. \tag{1.64}$$

Les variables $\hat{e}_{\{1\}}^{\text{partition}},\ldots,\hat{e}_{\{m\}}^{\text{partition}}$ étant de même loi mais non indépendantes, on vérifie alors que cet estimateur reste bien sans biais pour $\mathsf{E}_{\mathcal{D}_{\{1\}}}[e(\hat{r}_{\{1\}})]$

$$\mathsf{E}_{\mathcal{D}}[\hat{e}^{\text{cv}}] = \mathsf{E}_{\mathcal{D}}[\hat{e}_{\{1\}}^{\text{partition}}] = \mathsf{E}_{\mathcal{D}_{\{1\}}}[e(\hat{r}_{\{1\}})], \tag{1.65}$$

et aussi que sa variance est bien inférieure à celle de $\hat{e}_{\{1\}}^{\text{partition}}$

$$\mathsf{Var}_{\mathcal{D}}[\hat{e}^{\text{cv}}] = \frac{1}{m^2}\mathsf{Var}_{\mathcal{D}}\left[\sum_{i=1}^{m}\hat{e}_{\{i\}}^{\text{partition}}\right] < \mathsf{Var}_{\mathcal{D}_{\{1\}}}[\hat{e}_{\{1\}}^{\text{partition}}]. \tag{1.66}$$

Cette dernière inégalité provient du fait que deux v.a. Y_1 et Y_2 de même loi vérifient $\mathsf{Var}[Y_1 + Y_2] = 2\mathsf{Var}[Y_1] + 2\mathsf{Cov}[Y_1,Y_2]$ et que, si de plus il n'existe pas de relation fonctionnelle affine entre elles, alors $\mathsf{Cov}[Y_1,Y_2] < \mathsf{Var}[Y_1]$.

Les principaux concurrents à l'estimateur obtenu par validation croisée sont l'estimateur *Jacknife* (Tuckey 1958 [58]) et l'estimation par *bootstrap* (Efron 1983 [23]) mais l'estimateur par validation croisée conduit à de bons résultats au prix d'une mise en oeuvre assez simple, ce qui le fait généralement préféré aux autres.

L'estimateur par validation croisée nécessite le calcul de m règles de classement $\hat{r}_{\{i\}}$ $(i = 1,\ldots,m)$ ce qui peut s'avérer coûteux en temps de calcul. Dans le cas gaussien où l'échantillon est complet, c'est-à-dire quand $m = n$, il est possible de mettre à jour avec peu de calculs la règle sur l'échantillon complet $\hat{r}$ pour obtenir $\hat{r}_{\{i\}}$. En effet, huit des quatorze modèles gaussiens reposent sur le calcul de certains des éléments suivants qui peuvent être aisément actualisés ([R3] annexe A) : effectif de chaque groupe, proportions, centres, matrice d'inertie intra-classe ainsi que son déterminant, son inverse et sa trace, matrice d'inertie totale ainsi que son déterminant, son inverse et sa trace.

Dans un contexte de données simulées, des comparaisons numériques ont été menées entre les critères AIC, AIC3, BIC et $\hat{e}^{cv}$. Il ressort que la validation croisée est le critère le plus adapté dans le cas de petits échantillons et aussi dans le cas où aucun des modèles n'est le véritable modèle [R3] [C6] [C7]. Le critère $\hat{e}^{cv}$ est donc à recommander pour sa robustesse par rapport aux critères généraux classiques.

1.3.3 Classification automatique et entropie

La classification automatique a pour but essentiel de produire une partition d'un ensemble de données en différents groupes. L'utilité de ces groupes est ensuite fonction des besoins : ils apportent généralement une compréhension simplifiée des données initiales et parfois ils sont aussi utilisés pour définir une fonction de décision afin d'attribuer un nouvel objet à l'un d'entre eux. Que cela soit pour l'interprétation facile des groupes ou encore pour la définition d'une règle de classement efficace ensuite, il est indispensable que les groupes obtenus soit assez séparés entre eux.

Certains critères se contentent alors de retenir le modèle conduisant à la meilleure séparabilité des classes. C'est le cas du critère PC (*Partition Coefficient*) de Bezdek (1981) [7] sommant le carré de toutes les probabilités condionnelles, où encore du critère MIR (*Minimum Information Ratio*) de Windham et Cutler (1992) [61], et de ses variantes, critères plus évolués qui s'appuient sur un rapport des matrices d'information de Fisher de l'échantillon classifié et non classifié pour mesurer la capacité des données à être séparées (voir plus de détails dans [r12] pp. 84–85). Ces critères donnent des résultats médiocres et souffrent de ne pouvoir décider entre l'absence et la présence de groupes puisqu'ils sont optimaux pour un seul groupe. L'idée fondamentale des critères que nous allons présenter consiste à faire un compromis entre la proximité du modèle aux données et la séparabilité des groupes obtenus. Ils ne se cantonnent donc pas à mesurer la seule séparabilité comme les critères précédents.

Les critères proposés ci-dessous s'appuient tous, à un moment ou à un autre, sur une décomposition remarquable de la log-vraisemblance en un terme de type log-vraisemblance complétée et en un terme de type entropie. Cette relation s'énonce

$$\ell(\boldsymbol{\theta};\mathcal{D}) = \ell(\boldsymbol{\theta};\mathcal{D},\mathbf{c}^u) + \xi(\boldsymbol{\theta};\mathbf{c}^u), \tag{1.67}$$

où $\mathbf{c}^u = (c_{ik})$ $(k = 1,\ldots,K,\ i = m+1,\ldots,n)$ est une matrice stochastique, c'est-à-dire que $c_{ik} \in [0,1]$ avec $\sum_{k=1}^K c_{ik} = 1$. Cette matrice $\mathbf{c}^u$ peut s'interpréter comme une matrice de *classification floue* d'appartenance des individus sans labels $\mathbf{x}^u$ aux différents groupes. Ainsi, le terme $\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{c}^u)$ s'interprète comme la log-vraisemblance de $\boldsymbol{\theta}$ portant sur l'échantillon complété $(\mathbf{x},\mathbf{z})$ où les labels inconnus $\mathbf{z}^u$ ont été remplacés par les valeurs floues $\mathbf{c}^u$

$$\begin{aligned}
\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{c}^u) &= \ell(\boldsymbol{\theta};\mathcal{D}_l) + \ell(\boldsymbol{\theta};\mathcal{D}_u,\mathbf{c}^u) && (1.68)\\
&= \sum_{i=1}^m \sum_{k=1}^K z_{ik} \ln\left(\pi_k p(\mathbf{x}_i;\boldsymbol{\alpha}_k)\right) + \sum_{i=m+1}^n \sum_{k=1}^K c_{ik} \ln\left(\pi_k p(\mathbf{x}_i;\boldsymbol{\alpha}_k)\right). && (1.69)
\end{aligned}$$

Le deuxième terme, que nous commenterons plus en détail en fonction des choix de $\mathbf{c}^u$, $\xi(\boldsymbol{\theta};\mathbf{c}^u)$, s'exprime par

$$\xi(\boldsymbol{\theta};\mathbf{c}^u) = -\sum_{i=m+1}^n \sum_{k=1}^K c_{ik} \ln(t_{ik}) \geq 0, \tag{1.70}$$

où il a été noté, rappelons-le, $t_{ik} = t_{ik}(\boldsymbol{\theta})$. Ce terme varie entre 0 et $m \ln(K)$.

Dans le cas où $\mathbf{c}^u = \mathbf{t}^u$, avec $\mathbf{t}^u = (t_{ik})$ $(k = 1,\ldots,K,\ i = m+1,\ldots,n)$ la matrice des probabilités conditionnelles pour les individus $\mathbf{x}^u$, le terme $\xi(\boldsymbol{\theta};\mathbf{t}^u)$ correspond à l'*entropie* de la partition floue $\mathbf{t}^u$. L'entropie est une mesure du chevauchement des classes : sa plus petite valeur indique des groupes très fortement séparés, sa plus grande des groupes fortement imbriqués. Dans ce contexte, la relation fondamentale (1.67) est identique à la relation d'Hathaway (1986) [32] liant la log-vraisemblance et l'entropie. Le critère NEC (*Normalized Entropy Criterion*) de Celeux et Soromenho (1996) [30] est établi à partir de cette relation mais notons qu'il n'est valable que pour un échantillon dépourvu de labels connus, donc le cas où $\mathcal{D} = \mathbf{x}$. Ce critère s'exprime comme une normalisation de l'entropie par des termes de log-vraisemblance :

$$\mathrm{NEC}_K = \frac{\xi_K}{\ell_K - \ell_1} \text{ si } K > 1, \tag{1.71}$$

avec $\ell_k = \ell(\hat{\boldsymbol{\theta}}_k;\mathcal{D})$ et $\xi_k = \xi(\hat{\boldsymbol{\theta}}_k;\hat{\mathbf{t}}_k)$ où $\hat{\boldsymbol{\theta}}_k$ est l'estimateur du maximum de vraisemblance pour k groupes et $\hat{\mathbf{t}}_k = \mathbf{t}(\hat{\boldsymbol{\theta}}_k)$. Notons que $\hat{\boldsymbol{\theta}}_K$ et $\hat{\boldsymbol{\theta}}_1$ doivent être obtenus avec les mêmes modèles de contraintes sur les paramètres (par exemple, sous l'hypothèse gaussienne, un modèle sphérique pour les deux nombres de groupes). On retient alors le modèle $\mathbf{m}$ (modèle de contraintes et/ou nombre de groupes) qui possède la valeur de NEC la plus faible. Notons que NEC_1 conduit à une forme indéterminée puisque $\xi_1 = 0$. À l'origine, Celeux et Soromenho (1996) [30] ont proposé une heuristique pour lever cette indétermination mais cette procédure est peu efficace dans certaines situations et également limitée au cas gaussien. Pour répondre à ces deux limitations, une nouvelle procédure a été proposée dans [R2] . Notant le terme de vraisemblance complétée pour k classes $\ell c_k = \ell(\hat{\boldsymbol{\theta}}_k;\mathcal{D},\hat{\mathbf{t}}_k)$, l'idée part de la remarque que l'existence réelle de K groupes $(K > 1)$ bien séparés est équivalente à avoir $\xi_K \simeq 0$ ce qui équivaut aussi à $\ell c_K \simeq \ell_K$. Comme on a nécessairement $\ell_K \geq \ell_1$ (en effet, il y a emboîtement de l'espace des paramètres) on peut alors espérer avoir $\ell c_K \geq \ell_1$ si la présence de K groupes est évidente. Notant enfin que $\ell_1 = \ell c_1$ et qu'on a la relation $1 = (\ell c_K - \ell c_1)/(\ell_K - \ell_1) + \mathrm{NEC}_K$ d'après (1.67), on déduit que si K groupes bien séparés existent $(K > 1)$ on a alors $\mathrm{NEC}_K \leq 1$. La nouvelle procédure revient ainsi à retenir le nombre K, où $2 \leq K \leq K_{sup}$ avec K_{sup} une borne supérieure donnée pour K, minimisant NEC_K pourvu que $\mathrm{NEC}_K \leq 1$. Si aucun des nombres de groupes considérés ne permet d'obtenir $\mathrm{NEC}_K \leq 1$, on décide qu'il n'existe pas de structure de classification, c'est-à-dire qu'il n'y a qu'un seul groupe. Cette procédure est strictement équivalente à poser $\mathrm{NEC}_1 = 1$ puis à retenir la valeur de NEC la plus faible. L'intérêt de cette nouvelle stratégie d'utilisation de NEC est double. Non seulement, la nouvelle procédure a gagné en stabilité par rapport à l'ancienne et est utilisable pour des mélanges non gaussiens (voir des comparaisons avec l'ancienne procédure ainsi qu'un essai avec des données binaires dans [R2]). Mais encore, la valeur de NEC apparaît porteuse de sens par elle-même sur la qualité de la classification car les valeurs inférieures à 1 indiquent la présence d'une structure qui mérite d'être regardée tout de même avant de décider définitivement du nombre de composantes. La figure 3 de [R2] illustre bien tout l'intérêt de ce nouvel usage de NEC.

Considérons maintenant un choix différent pour $\mathbf{c}^u$. En choisissant $\mathbf{c}^u = \mathbf{z}^u$ avec $\mathbf{z}^u$ la partition de $\mathbf{x}^u$ obtenue par MAP de $\mathbf{t}^u$, le terme entropique correspond à l'opposé du logarithme de la probabilité conditionnelle de la partition $\mathbf{z}^u$, puisque cette probabilité s'écrit $p(\mathbf{z}^u|\mathbf{x}^u) = \prod_{i=m+1}^{n}\prod_{k=1}^{K} t_{ik}^{z_{ik}}$. Le terme $\xi(\boldsymbol{\theta};\mathbf{z}^u)$ mesure donc un écart entre les probabilités conditionnelles $\mathbf{t}^u$ et la partition $\mathbf{z}^u$ la plus proche en un certain sens. Le critère retenu dans [C8] est alors simplement la vraisemblance complétée $\ell(\hat{\boldsymbol{\theta}};\mathcal{D},\hat{\mathbf{z}}^u)$, $\hat{\mathbf{z}}^u$ étant le MAP de $\hat{\mathbf{t}}^u$, critère que nous appelons CL (*Completed Likelihood*) :

$$\mathrm{CL} = \ell(\hat{\boldsymbol{\theta}};\mathcal{D},\hat{\mathbf{z}}^u) = \ell(\hat{\boldsymbol{\theta}};\mathcal{D}) - \xi(\hat{\boldsymbol{\theta}};\hat{\mathbf{z}}^u). \tag{1.72}$$

Le modèle retenu est celui qui produit la plus grande valeur de CL. Ce critère s'interprète comme une pénalisation de la log-vraisemblance maximale, non plus par une complexité du modèle comme AIC ou BIC, mais par une pénalité mesurant le degré de chevauchement des groupes.

La dernière approche que nous présentons dans cette section est bayésienne. L'idée est de prendre comme critère la vraisemblance complétée intégrée $p(\mathcal{D},\mathbf{z}^u|\mathbf{m})$, démarche à rapprocher de celle utilisant la vraisemblance intégrée $p(\mathcal{D}|\mathbf{m})$ et conduisant à BIC. Supposant pour le moment connus les labels $\mathbf{z}^u$ et notant $\hat{\boldsymbol{\theta}}^*$ l'estimateur du maximum de vraisemblance associé à l'échantillon complet $(\mathcal{D},\mathbf{z}^u)$, la proposition 3 permet d'écrire

$$\ln p(\mathcal{D},\mathbf{z}^u|\mathbf{m}) \approx \ell(\hat{\boldsymbol{\theta}}^*;\mathcal{D},\mathbf{z}^u) - \frac{\nu}{2}\ln(n). \tag{1.73}$$

Ce critère est inutilisable en l'état puisque $\mathbf{z}^u$ est inconnu (c'est précisément ce que l'on recherche !) et, par conséquent, $\hat{\boldsymbol{\theta}}^*$ est indisponible. Comme $\hat{\boldsymbol{\theta}}^*$ et $\hat{\boldsymbol{\theta}}$ (ce dernier est, rappelons le, l'estimateur du maximum de vraisemblance associé à l'échantillon incomplet $\mathcal{D}$) sont tous deux convergents, on peut tout d'abord faire l'approximation $\hat{\boldsymbol{\theta}}^* \approx \hat{\boldsymbol{\theta}}$ pour n assez grand. Ensuite, il suffit d'estimer $\mathbf{z}^u$ par $\hat{\mathbf{z}}^u$ le MAP associé à $\hat{\boldsymbol{\theta}}$. Cela conduit au critère ICL (*Integrated Completed Likelihood*) [R1] :

$$\begin{aligned} \text{ICL} &= \ell(\hat{\boldsymbol{\theta}};\mathcal{D},\hat{\mathbf{z}}^u) - \frac{\nu}{2}\ln(n) = \ell(\hat{\boldsymbol{\theta}};\mathcal{D}) - \xi(\hat{\boldsymbol{\theta}};\hat{\mathbf{z}}^u) - \frac{\nu}{2}\ln(n) \qquad (1.74)\\ &= \text{CL} - \frac{\nu}{2}\ln(n) = \text{BIC} - \xi(\hat{\boldsymbol{\theta}};\hat{\mathbf{z}}^u) = \text{CL} + \text{BIC} - \ell(\hat{\boldsymbol{\theta}};\mathcal{D}). \qquad (1.75)\end{aligned}$$

On retiendra le modèle ayant la plus grande valeur de ICL. Ce critère peut s'interpréter comme le maximum de log-vraisemblance pénalisé par deux termes de nature différente : un terme pénalisant les groupes peu séparés (comme le critère CL) et un autre pénalisant le trop grand nombre de paramètres (comme le critère BIC). Ces relations avec CL et BIC ont d'ailleurs été mis explicitement dans les équations précédentes. Notons que le critère AWE (*Approximate Weight of Evidence*) proposé aussi dans un cadre bayésien par Banfield et Raftery (1993) [5], est très ressemblant à ICL car il pénalise aussi la vraisemblance complétée par une fonction du nombre de paramètres. Néanmoins, l'estimateur utilisé est celui de la vraisemblance complétée décrit par (1.18) et de plus la pénalisation de la complexité du modèle est plus forte que celle de ICL.

Nous proposons maintenant d'énoncer les principaux résultats expérimentaux concernant les différents critères présentés. Pour le choix d'un des 14 modèles gaussiens (à nombre fixé de composantes), il a été établi dans [R3] [c9] [c10] que NEC et CL sont moins adaptés que BIC, AIC ou encore AIC3 car ils choisissent généralement des modèles conduisant à une plus grande erreur de classement. Pour le choix du nombre de composants (modèle gaussien fixé), l'apport de NEC et CL sur les critères généraux est beaucoup plus clair car ils sont nettement plus robustes à un écart sur l'hypothèse de normalité, pourvu que les groupes soient suffisamment séparés [C7] [C8] [s15]. En effet, même si les composantes sont bien séparées, le fait qu'elles soient non gaussiennes induit une forte surestimation du nombre de composantes par BIC, AIC ou encore AIC3, au contraire de NEC et CL. Néanmoins, si les proportions sont libres entre composantes, CL souffre d'une surestimation du nombre de composants, ce qui n'est pas le cas de NEC [C8] [s16]. Finalement, le critère ICL utilisé dans le cadre le plus général (nombre de composantes et/ou modèle de contrainte) produit une synthèse intéressante de toutes les bonnes propriétés de robustesse de NEC et CL sans pour autant avoir les inconvénients décrits ci-dessus [R1] [s14].

1.4 Bilan et perspectives

Ce chapitre regroupe des aspects aussi bien expérimentaux que théoriques en ce qui concerne les méthodes de choix de modèles. En effet, au delà de la justification mathématique des critères, leur réelle efficacité dans certains contextes spécifiques comme celui du choix du nombre de composantes ne peut s'évaluer que par l'expérimentation. Il ressort que, au même titre qu'il n'existe pas de « vrai » modèle, il n'existe pas non plus de critère universel même si le critère BIC se rapproche quelque peu de cette utopie. Il faut retenir en effet qu'il existe autant de critères que d'objectifs que l'on attend par la suite du modèle sélectionné.

Un certain nombre d'extensions ou de nouvelles voies dans les méthodes de sélection de modèles peuvent être esquissées à ce stade de la présentation. C'est l'objectif des paragraphes suivants qui terminent ainsi ce chapitre.

Une première piste de recherche concerne une adaptation du critère ICL à certains mélanges de lois spécifiques. Il est exact que ICL est un critère suffisamment général pour s'appliquer à tous types de mélanges de lois, mais ICL n'en est pas moins un critère asymptotique au même titre que BIC d'ailleurs. Par contre, contrairement à BIC, ICL passe par l'intégration d'une vraisemblance complétée ce qui peut vraisemblablement permettre une intégration explicite si la loi *a priori* est bien choisie, par exemple une loi conjuguée (voir par exemple Robert 1994 [50]) avec un choix des hyperparamètres rendant cette loi « peu informative ». Cette démarche permet d'espérer un critère ICL encore plus efficace pour des petites tailles d'échantillons dans ces situations particulières. Le cas des classes latentes avec variables binaires ou qualitatives est particulièrement prometteur car la loi non informative de Jeffreys correspond dans ce cas à une loi conjuguée et donc le problème de choix des hyperparamètres, étape souvent subjective, peut être évité.

Une autre direction d'investigation concerne les propriétés de convergence du critère ICL. Des expérimentations mettent facilement en évidence que, contrairement à BIC, ICL peut sous-estimer le vrai nombre de composantes si celles-ci sont « trop peu » séparées, et cela même asymptotiquement. Néanmoins, à partir d'un certain niveau de séparation, ICL semble redevenir, au même titre que BIC, un critère convergent. La clarification des propriétés théoriques d'ICL est indispensable pour compléter les propriétés expérimentales encourageantes que nous avons décrites dans le chapitre. En particulier, s'il est confirmé qu'un seuil de séparation des composantes est nécessaire à sa convergence, une discussion sur la signification profonde de ce seuil peut être des plus enrichissantes.

Actuellement, les critères présentés peuvent se diviser en deux grands groupes : ceux dont la valeur à une signification en elle-même comme les critères NEC et d'erreur de classement et ceux qui n'ont qu'une valeur relative par rapport à celle des autres modèles. Autrement dit, bien qu'un certain modèle soit retenu au détriment d'autres par, disons, l'erreur de classement, il est tout de même possible de le refuser si l'erreur qui lui est associée est jugée trop forte. De même, bien que dans une moindre de mesure, le critère NEC exprime une sorte de normalisation en rendant possible sa comparaison par rapport à une valeur de référence qui est un. Il serait alors intéressant de permettre une certaine normalisation des autres critères que sont BIC, NIC ou ICL pour permettre de refuser tout de même le meilleur des modèles proposé si besoin.

Bibliographie

[1] J. Aitchinson and C. G. G. Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63:413–420, 1976.

[2] M. Aitkin and D. B. Rubin. Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B*, 47:67–75, 1985.

[3] H. Akaike. Information Theory as an Extension of the Maximum Likelihood Principle. In B. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, Akademiai Kiado, 1973.

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

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

[6] J. Besag. On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society B*, 48:259–302, 1986.

[7] J.D. Bezdeck. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York, 1981.

[8] S. Binder. Bayesian cluster analysis. *Biometrika*, 65:31–38, 1978.

[9] H.H. Bock. Statistical Testing and Evaluation Methods in Cluster Analysis. In *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions*, pages 116–146, Calcutta, December 16-19 1981.

[10] H. Bozdogan. *Multi-Sample Cluster Analysis and Approaches to Validity Studies in Clustering Individuals*. PhD thesis, Department of Mathematics, University of Illinois, Chicago, IL 60680, 1981.

[11] H. Bozdogan. Determining the number of clusters in the standart multivariate normal mixture model using model-selection criteria. Technical report no. uic/dqm/a83-1, Quantitative Methods Department, University of Illinois, Chicago, IL 60680, June 16 1983. ARO contract DAAGL90820K-0155.

[12] H. Bozdogan. On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics, Theory and Methods*, 19:221–278, 1990.

[13] H. Bozdogan. Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In *Information and Classification*. Springer-Verlag, Heidelberg, 1993.

[14] H. Bozdogan. Mixture-Model Cluster Analysis using Model Selection Criteria and a New Informational Measure of Complexity. In *Congrès International sur les mélanges de distributions*, Aussois, France, 17-21 septembre 1995.

[15] P.G. Bryant and J.A. Williamson. Asymptotic Behaviour of Classification Maximum Likelihood Estimates. *Biometrika*, 65:273–281, 1978.

[16] 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(1):73–92, 1985.

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

[18] G. Celeux and G. Govaert. A classification EM algorithm and two stochastic versions. *Computational Statistics & Data Analysis*, 14:315–332, 1992.

[19] 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.

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

[21] A. Cutler and M. P. Windham. Information-based validity functionals for mixture analysis. In Kluwer H. Bozdogan, editor, *Proceedings of the first US-Japan Conference on the Frontiers of Statistical Modeling*, pages 149–170, Amsterdam, 1993.

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

[23] B. Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983.

[24] B. S. Everitt. *An Introduction to Latent Variables Models*. London: Chapman & Hall, 1984.

[25] B.S. Everitt. A Monte-Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, 16:171–180, 1981.

[26] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. Pt. II.

[27] B. Flury. *A first course in multivariate statistics*. Springer, New York, 1997.

[28] H.P. Friedman and J. Rubin. On some Invariant Criteria for Grouping Data. *Journal of the American Statistical Association*, 62:1159–1178, 1967.

[29] J. Friedman. Regularized Discriminant Analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

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

[31] J. C. Gower. Maximal predictive classification. *Biometrics*, 30:643–654, 1974.

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

[33] R. E. Kass and A. E. Raftery. Bayes factors and model uncertainty. *Journal of the American Statistical Association*, 90:773–795, 1995.

[34] R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90:928–934, 1995.

[35] C. Keribin. Consistent estimation of the order of mixture models. *Sankhyã, Series A*, 62(1):49–66, 2000.

[36] E.L. Lehmann. *Theory of Point Estimation*. Wiley, New-York, 1983.

[37] B. G. Leroux. Consistent estimation of a mixing proportion. *Annals of Statistics*, 20:1350–1360, 1992.

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

[39] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

[40] G. J. McLachlan and T. Krishnam. *The EM algorithm and Extensions*. Wiley, New York, 1997.

[41] G.J. Mclachlan. On Bootstraping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture. *Applied Statistics*, 36:318–324, 1987.

[42] E. Monfrini. Unicité dans la méthode des moments pour les mélanges de deux distributions normales. *C. R. Acad. Sci. Paris*, 336(Série I):89–94, 2003.

[43] N. Murata, S. Yoshizawa, and S. Amari. A criterion for determining the number of parameters in an artificial neural network model. In T. Hohonen, K. Mäkisara, O. Simula, and J. Kangas, editors, *Artificial Neural Networks. Proceesings of ICANN-91*, volume 1, pages 9–14, Amsterdam: North Holland, 1991.

[44] N. Murata, S. Yoshizawa, and S. Amari. Learning curves, model selection and complexity of neural networks. In *NIPS5*, pages 607–614, 1993.

[45] N. Murata, S. Yoshizawa, and S. Amari. Network information criterion – determining the number of hidden units for artificial neural network models. *IEEE Transactions on Neural Networks*, 5:865–872, 1994.

[46] K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society*, 185:71–110, 1894.

[47] A. E Raftery. Bayesian model selection in social research (with discussion). In P. V. Marsden, editor, *Sociological Methodology*, pages 111–195. Cambridge, Mass.: Blackwells, 1995.

[48] B. D. Ripley. *Neural Networks and Pattern Recognition*. Cambridge University Press, New York, 1996.

[49] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Teaneck, New Jersey, 1989.

[50] C. P. Robert. *The Bayesian Choice : a Decision-Theoretic Motivation*. New York: Springer-Verlag, 1994.

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

[52] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

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

[54] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[55] P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10:63–72, 2000.

[56] M. Stone. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of Royal Statistical Society, Series B*, 39:44–47, 1977.

[57] L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal distributions. *Journal of the American Statistical Association*, 81:82–86, 1986.

[58] J. Tukey. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29:164, 1958.

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

[60] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

[61] M.P. Windham and A. Cutler. Information Ratios for Validating Cluster Analysis. *Journal of the American Statistical Association*, 87:1188–1192, 1992.

[62] J. H. Wolfe. A monte carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Technical Bulletin STB 72-2, US Naval Personnel Research Activity, San Diego, California, 1971.

[63] P. Zhang. Model selection via multifold cross validation. *The Annals of Statistics*, 21:299–313, 1993.

Chapitre 2

Contributions à la maximisation de la vraisemblance

Sommaire

Publications associées à ce chapitre

Revues avec comité de lecture

[R1] C. Biernacki. *Testing for a Global Maximum of the Likelihood.* Journal of Computational and Graphical Statistics, à paraître.

[R2] C. Biernacki. *An Asymptotic Upper Bound of the Likelihood to Prevent Gaussian Mixtures from Degenerating.* Journal of Statistical Planning and Inference, en révision.

[R3] C. Biernacki (2004). *Initializing EM Using the Properties of its Trajectories in Gaussian Mixtures.* Statistics and Computing, **14**, 3, 267-279.

[R4] C. Biernacki, G. Celeux and G. Govaert (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.

[R5] C. Biernacki and S. Chrétien (2003). *Degeneracy in the Maximum Likelihood Estimation of Univariate Gaussian Mixtures with EM.* Statistics & Probability Letters, **61**, 373–382.

Conférences avec comité de lecture

[C6] C. Biernacki and C. Chrétien (2001). *Degeneracy in the likelihood approach to Gaussian mixtures estimation.* ASMDA'2001, Xth International Symposium on Applied Stochastic Models and Data Analysis, Compiègne, France, 12–15 juin.

Conférences sans comité de lecture

[c7] C. Biernacki (2003). *Un test pour le maximum global de vraisemblance.* SFdS'2003, 35ièmes journées de statistiques, Lyon, France, 2–6 juin.

[c8] C. Biernacki (2003). *Initialisation de EM en utilisant ses propriétés de trajectoires dans les mélanges gaussiens.* SFdS'2003, 35ièmes journées de statistiques, Lyon, France, 2–6 juin.

[c9] C. Biernacki, G. Celeux and G. Govaert (2000). *Stratégies algorithmiques pour maximiser la vraisemblance dans les modèles de mélange.* SFdS'2000, 32ièmes journées de statistiques, Fès, Maroc, 15–19 mai.

Dans le chapitre précédent, nous avons essentiellement considéré l'estimation du paramètre d'intérêt $\boldsymbol{\theta}$ sous l'angle du principe du maximum de vraisemblance. Par contre, nous n'avons que peu évoqué les difficultés liées à cette technique, typiquement les questions relatives à l'existence de multiples points stationnaires ou encore aux procédures d'optimisation utilisées. Il est évident que les choix opérés pour la mise en œuvre de la méthode du maximum de vraisemblance (racine retenue, méthode d'optimisation employée) auront un impact direct sur la qualité de l'estimateur proposé. Par ailleurs, il est important de noter que nombre de critères de choix de modèles s'appuient sur une pénalisation du maximum de log-vraisemblance et, ainsi, une mauvaise recherche de cet optimum peut naturellement conduire à retenir un modèle inadéquate. Une illustration de ce phénomène a été donnée dans l'exemple introductif de [R4] où, dans un cas simple de deux composantes gaussiennes, le choix de l'un ou l'autre des deux maxima répertoriés conduisait à un choix différent du nombre de composantes par le critère AIC.

Ce chapitre se compose de quatre parties principales. La section 2.1 propose de rappeler certaines difficultés liées à la technique du maximum de vraisemblance, en balayant plusieurs aspects statistiques et algorithmiques. La section 2.2 est consacrée à l'algorithme EM comme technique renommée d'optimisation de la vraisemblance ainsi qu'à ses propriétés et certaines de ses variantes. Il sera essentiellement discuté du problème important relatif à l'initialisation de cet algorithme, en particulier de l'idée de tirer les paramètres de départ au hasard sur les trajectoires de EM. Bien que portant généralement dans le cas gaussien, des extensions de ces techniques sont également envisageables dans d'autres types de mélange. La section 2.3 se place d'emblée dans un cadre général, dépassant même le cas des mélanges. Il s'agit de proposer un test statistique permettant de décider si une racine de la vraisemblance peut être ou non considérée comme celle du maximum de la vraisemblance. Enfin, la section 2.4 se penche sur le problème de la dégénérescence de la vraisemblance dans le cas d'un mélange gaussien hétéroscédastique. Ce problème est abordé sous l'aspect algorithmique et également d'un point de vue conceptuel.

2.1 Problématique du maximum de la vraisemblance

2.1.1 Racines multiples et choix de la plus grande racine

Sous le nom générique de méthode du « maximum de la vraisemblance » se dissimulent un certain nombre de problèmes, propres ou non d'ailleurs au cas particulier des mélanges. En effet, dans de nombreuses applications où cette technique d'estimation est utilisée, les statisticiens sont souvent confrontés à l'existence de multiples racines de l'équation de la log-vraisemblance, c'est-à-dire aux différentes valeurs de $\boldsymbol{\theta}$ qui vérifient l'équation

$$\nabla \ell(\boldsymbol{\theta}; \mathcal{D}) = \mathbf{0}. \tag{2.1}$$

Bien entendu, sous certaines conditions de régularité, la théorie affirme l'existence d'une unique racine convergente de cette équation (voir par exemple Cramér 1946 [12] ou encore sa généralisation dans le cas multivarié par Tarone et Gruenhage 1965 [38]), mais généralement peu d'indications sont données sur le choix de la racine convergente dans le cas de plusieurs racines. Le papier bibliographique de Small *et al.* (2000) [36] discute d'approches variées pour sélectionner une des racines (voir aussi une discussion antérieure dans Lehmann 1983 Chap. 6 [23]), incluant par exemple une méthode itérant à partir d'estimateurs convergents, l'emploi d'une technique de bootstrap ou encore l'examen des propriétés asymptotiques des racines lorsque leur formulation est explicite. Une autre possibilité consiste simplement à sélectionner la racine associée à la valeur maximum de la vraisemblance puisque Wald (1949) [41] a établi la convergence du maximum global de la vraisemblance sous certaines conditions. Les propriétés de Wald concernant cet estimateur du maximum de vraisemblance ont été ensuite généralisées par White (1982) [42] dans le cas plus réaliste d'un modèle éventuellement mal spécifié (voir à ce propos la section 1.2.2 et en particulier la proposition 1 du chapitre 1). Ainsi, la stratégie qui consiste à retenir la valeur maximum de la vraisemblance est souvent adoptée.

2.1.2 Cas pathologiques associés à la méthode

Néanmoins, il existe des situations pour lesquelles ce maximum global est inconsistant et on pourra en trouver des exemples dans Neyman et Scott (1948) [31], Ferguson (1982) [16] ou bien encore Stefanski et Carroll (1987) [37]. Dans le cadre des mélanges gaussiens, retenir l'optimum global de la vraisemblance peut aussi conduire à différentes situations « pathologiques », essentiellement dans le cas hétéroscédastique. Tout d'abord, le maximum global de la vraisemblance n'existe pas puisque la vraisemblance n'est pas bornée comme l'ont noté en premier Kiefer et Wolfowitz (1956) [22] (notons que ce maximum n'est pas une racine de 2.1). Pour s'en convaincre, il suffit de placer une gaussienne dégénérée (c'est-à-dire un Dirac) en un des points de l'échantillon, tout en imposant que la variance généralisée (i.e. le déterminant de la matrice de variance) d'au moins une des autres gaussiennes ne soit pas nulle. Par ailleurs, parmi les autres maxima locaux de la vraisemblance, certains d'entre eux peuvent correspondre à des *spurious maximizers* comme dénommés par McLachlan et Peel (2000) [27] sec. 3.10, c'est-à-dire à des solutions où une ou plusieurs matrices de variance sont proches de la matrice nulle, solutions conduisant parfois à une assez grande vraisemblance mais ne représentant pourtant aucune réalité par rapport au paramètre recherché.

2.1.3 Difficultés techniques pour obtenir une racine adéquate

On pourra se reporter à Redner et Walker (1984) [34] sec. 2.2 ou McLachlan et Peel (2000) [27] sec. 1.18 pour une revue détaillée sur l'historique des méthodes visant à maximiser la vraisemblance dans les mélanges de distributions.

Dans ce cadre des mélanges, la résolution de l'équation (2.1) conduit à des équations hautement non linéaires, donc généralement impossibles à résoudre sous forme explicite. L'évolution des moyens de calculs, grâce à l'informatique, a permis de surmonter graduellement cette difficulté. Dans un premier temps, certains cas simples de mélanges ont été résolus avec succès par des méthodes itératives. Par exemple, Rao (1948) [33] utilisa la méthode du *scoring* de Fisher pour étudier un mélange de deux gaussiennes homoscédastiques univariées, Mendenhall et Hader (1958) [30] utilisèrent une méthode de Newton pour une situation encore plus simple puisque ayant une unique inconnue. Par la suite, Day (1969) [13] dans le cas d'un mélange multivarié de deux gaussiennes homoscédastiques, et Wolfe (1971) [43] (ainsi que d'autres références antérieures de ce même auteur) avec un nombre quelconque de composantes hétéroscédastiques, ont employé à des périodes similaires des méthodes d'optimisation annonçant déjà l'algorithme EM de Dempster *et al.* (1977) [15]. L'algorithme EM, voire certaines de ses nombreuses variantes depuis les travaux de Dempster *et al.* (1977) [15], est certainement aujourd'hui la méthode d'estimation de la vraisemblance la plus utilisée aujourd'hui.

Bien que de tels algorithmes permettent ainsi d'apporter des solutions relativement simples et satisfaisantes pour maximiser la vraisemblance, ils sont tout de même confrontés, en général, aux problèmes théoriques inhérents à la méthode du maximum de vraisemblance : multiplicité des points stationnaires, dégénérescence de la vraisemblance et solutions *spurious* dans les mélanges gaussiens. S'agrègent parfois aux problèmes de la méthode du maximum de la vraisemblance des difficultés propres à la méthode d'optimisation retenue, comme une certaine lenteur de convergence comme cela peut être le cas pour EM. Ainsi, l'ensemble de ces questions ouvre la voie à un certain nombre de recherches dont certaines seront débattues dans la suite de ce chapitre.

2.2 Stratégies d'initialisation de l'algorithme EM

2.2.1 L'algorithme EM et ses propriétés

L'algorithme EM de Dempster *et al.* (1977) [15] est une procédure générale pour maximiser la vraisemblance. Elle est adaptée à de nombreuses situations décrites sous forme de problèmes avec *données incomplètes*. Depuis cet article de référence, de nombreux auteurs ont décrit cet algorithme, ses propriétés et parfois ses variantes, par exemple McLachlan et Krishnan (1997) [28] ou encore, dans le cas des mélanges de lois, Redner et Walker (1984) [34]. Dans le cadre des modèles de mélange que nous avons présentés dans le chapitre précédent, les données manquantes correspondent aux labels inconnus $\mathbf{z}^u$. Partant d'un paramètre initial $\boldsymbol{\theta}^{(0)}$, cet algorithme procède en deux étapes successives, l'étape E (pour *Expectation*) et l'étape M (pour *Maximization*). Notant

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) = \mathsf{E}_{\boldsymbol{\theta}^{(q)}}[\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{Z}^u)|\mathcal{D}] \tag{2.2}$$

l'espérance de la log-vraisemblance complétée $\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{Z}^u)$ par rapport à la loi conditionnelle $p(\mathbf{z}^u|\mathcal{D};\boldsymbol{\theta}^{(q)})$, ces deux étapes s'énoncent :

Étape E Calculer $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$;

Étape M Choisir $\boldsymbol{\theta}^{(q+1)} \in \Theta$ tel que $\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$.

S'il existe plusieurs valeurs de $\boldsymbol{\theta}^{(q+1)}$ possibles à l'étape M, il suffit de retenir l'une d'entre elles. Puis l'algorithme s'arrête dès que la log-vraisemblance devient stationnaire, c'est-à-dire dès que

$$|\ell(\boldsymbol{\theta}^{(q+1)};\mathcal{D}) - \ell(\boldsymbol{\theta}^{(q)};\mathcal{D})| \leq \varepsilon, \tag{2.3}$$

avec ε une valeur arbitrairement petite.

Nous proposons maintenant d'énoncer certaines propriétés importantes de EM. En embrassant ainsi les forces et faiblesses de cet algorithme, nous comprendrons mieux les différentes extensions et pistes de recherche qui y sont liées.

Une première propriété importante de EM concerne la monotonie que nous pouvions déjà soupçonner à la vue du critère d'arrêt cité ci-dessus. En effet, la vraisemblance augmente de manière monotone à chaque itération de EM, c'est-à-dire $\ell(\boldsymbol{\theta}^{(q+1)};\mathcal{D}) \geq \ell(\boldsymbol{\theta}^{(q)};\mathcal{D})$ pour $q \geq 0$. Afin de montrer ce point (voir par exemple McLachlan et Krishnan (1997) Chap. 3 [28]), on peut s'appuyer sur la décomposition de la log-vraisemblance en un terme de vraisemblance complète et un terme entropique (voir la relation (1.67) énoncée au chapitre précédent) :

$$\ell(\boldsymbol{\theta};\mathcal{D}) = \ell(\boldsymbol{\theta};\mathcal{D},\mathbf{z}^u) + \xi(\boldsymbol{\theta};\mathbf{z}^u). \tag{2.4}$$

En prenant l'espérance par rapport à $p(\mathbf{z}^u|\mathcal{D};\boldsymbol{\theta}^{(q)})$ des deux membres de cette dernière équation, on obtient alors

$$\begin{aligned}
\ell(\boldsymbol{\theta};\mathcal{D}) &= \mathsf{E}_{\boldsymbol{\theta}^{(q)}}[\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{Z}^u)|\mathcal{D}] + \mathsf{E}_{\boldsymbol{\theta}^{(q)}}[\xi(\boldsymbol{\theta};\mathbf{Z}^u)|\mathcal{D}] \qquad (2.5)\\
&= Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) + \xi(\boldsymbol{\theta};\mathbf{t}^u(\boldsymbol{\theta}^{(q)})). \qquad (2.6)
\end{aligned}$$

La transformation du terme entropique est due au fait que $t_{ik}(\boldsymbol{\theta}^{(q)}) = p(Z_{ik} = 1|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta}^{(q)}) = \mathsf{E}_{\boldsymbol{\theta}^{(q)}}[Z_{ik}|\mathbf{X}_i = \mathbf{x}_i]$ $(i = m+1,\ldots,n,\ k = 1,\ldots,K)$. Il est alors déduit que

$$\ell(\boldsymbol{\theta}^{(q+1)};\mathcal{D}) - \ell(\boldsymbol{\theta}^{(q)};\mathcal{D}) = \{Q(\boldsymbol{\theta}^{(q+1)};\boldsymbol{\theta}^{(q)}) - Q(\boldsymbol{\theta}^{(q)};\boldsymbol{\theta}^{(q)})\} + \{\xi(\boldsymbol{\theta}^{(q+1)};\mathbf{t}^u(\boldsymbol{\theta}^{(q)})) - \xi(\boldsymbol{\theta}^{(q)};\mathbf{t}^u(\boldsymbol{\theta}^{(q)}))\}. \tag{2.7}$$

Le premier terme du second membre de cette équation est positif comme conséquence directe de la définition de l'étape M. On conclut alors en remarquant que le second terme est lui aussi positif puisque

$$\xi(\boldsymbol{\theta}^{(q+1)};\mathbf{t}^u(\boldsymbol{\theta}^{(q)})) - \xi(\boldsymbol{\theta}^{(q)};\mathbf{t}^u(\boldsymbol{\theta}^{(q)})) = \sum_{i=m+1}^{n}\left\{\sum_{k=1}^{K} t_{ik}(\boldsymbol{\theta}^{(q)}) \ln\left(\frac{t_{ik}(\boldsymbol{\theta}^{(q)})}{t_{ik}(\boldsymbol{\theta}^{(q+1)})}\right)\right\} \geq 0. \tag{2.8}$$

En effet, on reconnaît, pour chaque $i = m+1,\ldots,n$, la divergence de Kullback-Leibler entre les distributions $\{t_{ik}(\boldsymbol{\theta}^{(q)})\}_{k=1,\ldots,K}$ et $\{t_{ik}(\boldsymbol{\theta}^{(q+1)})\}_{k=1,\ldots,K}$.

Une seconde propriété de EM concerne sa vitesse de convergence vers un point stationnaire. En effet, EM est souvent considéré comme un algorithme convergeant assez lentement. Nous allons voir que ce taux de convergence est linéaire au voisinage d'un point stationnaire $\boldsymbol{\theta}^*$ de la vraisemblance (voir McLachlan et Krishnan (1997) Chap. 3.9 [28]), contrairement à des méthodes de type Newton qui bénéficient d'une convergence quadratique localement. Chaque itération de EM correspond à une application g de Θ dans Θ tel que $\boldsymbol{\theta}^{(q+1)} = g(\boldsymbol{\theta}^{(q)})$. Si $\boldsymbol{\theta}^{(q)}$ converge vers un point $\boldsymbol{\theta}^*$ et que g est une application continue, alors $\boldsymbol{\theta}^* = g(\boldsymbol{\theta}^*)$. Un développement de Taylor de $g(\boldsymbol{\theta}^{(q)})$ au voisinage de $\boldsymbol{\theta}^*$ permet d'écrire

$$\boldsymbol{\theta}^{(q+1)} - \boldsymbol{\theta}^* \approx \mathbf{H}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}^{(q)} - \boldsymbol{\theta}^*), \tag{2.9}$$

avec $\mathbf{H}(\boldsymbol{\theta}^*)$ la matrice jacobienne $\nu \times \nu$ de $g(\boldsymbol{\theta})$. Ainsi, une itération de EM est quasiment linéaire au voisinage de la convergence avec matrice de convergence $\mathbf{H}(\boldsymbol{\theta}^*)$. Comme le taux global de convergence est généralement donné par

$$\gamma = \lim_{q\to\infty} \frac{\|\boldsymbol{\theta}^{(q+1)} - \boldsymbol{\theta}^*\|}{\|\boldsymbol{\theta}^{(q)} - \boldsymbol{\theta}^*\|} \tag{2.10}$$

pour n'importe quelle norme $\|\cdot\|$ de $\mathbb{R}^\nu$, il correspond ainsi à la plus grande valeur propre de $\mathbf{H}(\boldsymbol{\theta}^*)$. La vitesse de convergence de EM sera donc dépendante de la valeur de γ, une grande valeur imposant une convergence lente.

Au delà de ses propriétés théoriques, EM est largement apprécié pour sa simplicité d'implantation, ses itérations généralement peu gourmandes en temps de calcul, le peu de mémoire nécessaire pour le faire fonctionner (il nécessite peu de stockage) et enfin son principe assez naturel heuristiquement. Ces divers points apparaissent lorsqu'on passe en revue chacune de ses deux étapes. L'étape E, dans le cas de mélange que nous considérons, consiste finalement à calculer les probabilités conditionnelles $t_{ik}(\boldsymbol{\theta}^{(q)})$ ($i = m+1, \ldots, n$, $k = 1, \ldots, K$) puisque la log-vraisemblance complétée $\ell(\boldsymbol{\theta}; \mathcal{D}, \mathbf{z}^u)$ est linéaire vis-à-vis des données manquantes $\mathbf{z}^u$. En d'autres termes, nous avons l'identité $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}) = \ell(\boldsymbol{\theta}; \mathcal{D}, \mathbf{t}^u(\boldsymbol{\theta}^{(q)}))$. L'étape M, quant à elle, permet souvent de trouver le paramètre $\boldsymbol{\theta}^{(q+1)}$ de façon explicite. En effet, il est généralement facile de trouver le maximum de vraisemblance dans le cas de la vraisemblance complète et l'étape M consiste finalement à maximiser la vraisemblance complète où les données manquantes d'appartenance ont été remplacées par leur espérance. Les proportions sont données par

$$\pi_k^{(q+1)} = \frac{n_k^{(q)}}{n}, \tag{2.11}$$

où $n_k^{(q)} = \sum_{i=1}^m z_{ik} + \sum_{i=m+1}^n t_{ik}(\boldsymbol{\theta}^{(q)})$ correspond à la population « floue » de la composante k. Les autres paramètres dépendent du modèle paramétrique considéré. Par exemple, dans le cas gaussien général (appelé aussi cas *hétéroscédastique*), nous retrouvons des formules familières pour les centres et les matrices de variances ($k = 1, \ldots, K$) :

$$\boldsymbol{\mu}_k^{(q+1)} = \frac{1}{n_k^{(q)}} \left(\sum_{i=1}^m z_{ik}\mathbf{x}_i + \sum_{i=m+1}^n t_{ik}(\boldsymbol{\theta}^{(q)})\mathbf{x}_i \right) \tag{2.12}$$

$$\boldsymbol{\Sigma}_k^{(q+1)} = \frac{1}{n_k^{(q)}} \left(\sum_{i=1}^m z_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})' + \sum_{i=m+1}^n t_{ik}(\boldsymbol{\theta}^{(q)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})' \right). \tag{2.13}$$

Dans le cas *homoscédastique* (même matrice de variance $\boldsymbol{\Sigma}^{(q+1)}$ entre groupes), on a, à partir de l'expression précédente,

$$\boldsymbol{\Sigma}^{(q+1)} = \frac{1}{n} \sum_{k=1}^K n_k^{(q)} \boldsymbol{\Sigma}_k^{(q+1)}. \tag{2.14}$$

Celeux et Govaert (1995) [9] ont décrit les étapes M pour chacun des 14 modèles gaussiens présentés à la section 1.1.5 du chapitre précédent. Dans le cas des classes latentes, les différents modèles considérés à la section 1.1.6 conduisent à des étapes M différentes détaillées par Celeux et Govaert (1991) [6].

2.2.2 Aperçu des propositions pour répondre aux limitations de EM

Comme l'algorithme EM peut s'avérer assez lent dans certaines situations, de nombreux auteurs ont récemment proposé des versions modifiées de celui-ci pour accélérer sa convergence tout en préservant la simplicité de ses itérations. Dans ce contexte, Liu et Sun (1997) [25] considèrent, dans le cadre des mélanges, l'algorithme ECME (*Expectation Conditional Maximization of Either*) de Liu et Rubin (1994) [24]. Dans ECME, l'étape E de EM est inchangée mais l'étape M de EM est remplacée par l'étape CM (*Conditional Maximization*) qui maximise, au choix en fonction des paramètres, soit l'espérance de la log-vraisemblance complète comme c'est déjà le cas dans EM, soit directement la log-vraisemblance. Modifiant cette fois l'étape E, Ueda et Nakano (1998) [39] proposent une version déterministe de type *recuit simulé* de EM, l'algorithme DAEM (*Deterministic Annealing EM*), dans le but de résoudre les problèmes de maxima locaux. Pour cela, à l'étape E, les probabilités conditionnelles d'appartenance aux groupes sont élevées à une certaine puissance qui peut s'apparenter à une température, cette valeur tendant vers un quand le nombre d'itérations augmente. Pilla et Lindsay (2001) [32] ont suggéré une nouvelle définition des données manquantes afin de réduire leur nombre. Ainsi, le taux de convergence est amélioré dans certaines directions paramétriques qui peuvent d'ailleurs être changées au fil des itérations de EM en modifiant à volonté la formulation des données manquantes. Dans le but également d'explorer en profondeur l'espace des paramètres dans différentes directions, Celeux *et al.* (2001) [3] ont considéré un algorithme EM qui met à jour séquentiellement les paramètres des différentes composantes.

Au delà de la vitesse de convergence de EM, la qualité de l'estimateur produit par l'algorithme dépend directement de la valeur d'initialisation fournie à la procédure. En effet, EM converge vers un point stationnaire de la vraisemblance qui peut tout autant se présenter sous la forme du maximum global, d'un maximum local ou encore d'un point selle. Notons bien que la possibilité que EM se fasse piéger dans des points stationnaires peu intéressants est une caractéristique partagée par bien d'autres algorithmes d'optimisation et est plutôt la conséquence de la forme parfois assez chaotique de la courbe de la vraisemblance. Pour éviter de n'avoir qu'un seul point stationnaire dont on ne connaît pas l'intérêt, il semble indispensable d'associer, d'une façon ou d'une autre, un certain caractère aléatoire à l'algorithme EM. Dans cet objectif, l'algorithme SEM (*Stochastic EM*) de Celeux et Diebolt (1985) [4] incorpore une étape aléatoire S (*Stochastic*) entre les étapes E et M. Cette étape consiste à tirer les appartenances suivant une loi multinomiale d'ordre un avec paramètres les probabilités conditionnelles d'appartenance aux groupes, plutôt que de prendre l'espérance de ces appartenances comme le fait EM. Partant de $\boldsymbol{\theta}^{(0)}$, il s'énonce alors

Étape E Calculer $\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{t}^u(\boldsymbol{\theta}^{(q)}))$;
Étape S Pour chaque $i = m+1,\dots,n$ tirer $\mathbf{z}_i(\boldsymbol{\theta}^{(q)})$ suivant la loi $\mathsf{Mult}_K(1,\mathbf{t}_i(\boldsymbol{\theta}^{(q)}))$;
Étape M Choisir $\boldsymbol{\theta}^{(q+1)} \in \Theta$ tel que $\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta};\mathcal{D},\mathbf{z}^u(\boldsymbol{\theta}^{(q)}))$.

Comme la suite des paramètres $(\boldsymbol{\theta}^{(q)})$ engendrée par SEM ne converge pas ponctuellement à cause des perturbations aléatoires introduites à l'étape S, l'algorithme s'arrête après un nombre prédéterminé d'itérations. En effet, cette suite converge en loi vers l'unique distribution stationnaire. Asymptotiquement, la moyenne de cette loi fournit un optimum local significatif de la vraisemblance. Ainsi, SEM permet théoriquement de s'affranchir de la valeur initiale $\boldsymbol{\theta}^{(0)}$ si un nombre « suffisant » d'itérations est effectué par l'algorithme. Notons qu'il existe une version de type recuit simulé de SEM, l'algorithme SAEM (*Simulated Algorithm EM*) de Celeux et Diebolt (1990) [5], qui permet de débuter avec SEM puis de terminer avec EM grâce à une contrôle de température et ainsi de converger ponctuellement tout en s'affranchissant relativement, comme SEM, de la position initiale.

Alternativement à des variantes de EM qui introduisent de l'aléatoire dans la procédure elle-même, de l'aléatoire peut aussi bien être injecté au niveau du choix des valeurs initiales de EM. Idéalement, le choix d'un paramètre initial judicieux peut permettre de pallier simultanément les deux difficultés essentielles de EM qui sont la lenteur de la convergence et la convergence vers des points stationnaires inintéressants. En pratique, il est couramment admis que plusieurs essais de EM à partir de valeurs initiales différentes sont bienvenus pour ensuite ne sélectionner que la valeur associée à la plus grande vraisemblance. Se pose alors la question du choix de ses valeurs initiales. Dans ce contexte, Coleman et Woodruff (2000) [11] utilisent une méthode de classification qui est lancée à partir d'une partition aléatoire d'un sous-échantillon et qui produit une partition de ce sous-échantillon, cette partition étant ensuite utilisée pour initialiser EM. McLachlan et Peel (2000) [27] proposent, dans le cas gaussien, de fixer les proportions du départ égales entre composantes, de générer les K centres suivant une loi normale multivariée de moyenne égale à la moyenne de l'échantillon complet et de matrice de variance égale à la matrice de variance de l'échantillon complet. Enfin, les matrices de variance de chaque composante sont initialisées à cette matrice de variance particulière. Par ailleurs, Markatou *et al.* (1998) [26] ont proposé une méthode d'initialisation basée sur une technique de bootstrap afin de construire automatiquement une région raisonnable de recherche.

Mais, comme souligné par Meila et Heckerman (2001) [29], le choix d'un paramètre initial est finalement une affaire de compromis entre sa pertinence et son coût de calcul. Les deux sections suivantes présentent des procédures originales d'initialisation qui prennent pleinement en considération cette remarque essentielle. Elles considèrent la maximisation de la vraisemblance principalement dans le cadre des modèles de mélange de lois normales à nombre de classes et modèle sous-jacent connus, même si des extensions à des autres lois sont aussi envisageables.

Notons que généralement l'ensemble des méthodes décrites (ci-dessus et ci-dessous) se placent dans une situation ou aucun label n'est connu. En effet, si tel n'est pas le cas, il suffit d'utiliser les individus dont les labels sont connus pour initialiser EM. Nous nous plaçons donc dans la suite dans la situation où une telle information n'est pas disponible.

2.2.3 Enchaînement d'algorithmes

Pour éviter les situations de dégénérescence de la vraisemblance ainsi que des maxima locaux correspondant à des *spurious maximizers*, cette section est restreinte aux modèles avec volumes égaux entre composantes gaussiennes. Rappelons que le volume d'une gaussienne de matrice de variance $\boldsymbol{\Sigma}$ en dimension d est défini par $|\boldsymbol{\Sigma}|^{1/d}$ comme nous l'avons vu à la section 1.1.5 du chapitre 1. Ces modèles sont les seuls qui garantissent l'absence des situations pathologiques décrites. Notons que le modèle homoscédastique est un cas particulier de ce modèle puisque nous supposons que les volumes sont égaux, ce qui n'impose donc pas nécessairement l'égalité des matrices de variances dans leur totalité.

À l'origine, Celeux et Govaert (1993) [8] ont constaté expérimentalement qu'un enchaînement de l'algorithme de classification CEM (Celeux et Govaert 1992 [7]) suivi de l'algorithme EM initialisé avec les paramètres obtenus par CEM atténuait la dépendance de EM vis-à-vis des paramètres de départ. L'idée qui en découle alors est d'examiner si des enchaînements bien choisis d'algorithmes ne conduisent pas à une stratégie statistiquement meilleure. Cette recherche est facilitée par le logiciel MIXMOD qui permet des enchaînements d'algorithmes très aisément (voir le chapitre 5). Comme précédemment dit, le temps alloué à la recherche de la solution est bien sûr un paramètre primordial à prendre en considération lors de cette optimisation, c'est pourquoi l'étude des stratégies en compétition intègre cet aspect.

Les stratégies étudiées s'appuient sur les trois algorithmes EM, CEM et SEM. L'algorithme CEM (pour *Classification* EM) consiste à ajouter une étape C dite de classification entre les étapes E et M de EM. Cette étape se résume à estimer les labels inconnus par un MAP des probabilités conditionnelles estimées à l'étape E. Cet algorithme s'énonce :

Étape E Calculer $\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{t}^u(\boldsymbol{\theta}^{(q)}))$;
Étape C Pour chaque $i=m+1,\dots,n$ et $k=1,\dots,K$, $\mathbf{z}_{ik}(\boldsymbol{\theta}^{(q)})=1$ si $k=\arg\max_{h=1,\dots,K} t_{ih}(\boldsymbol{\theta}^{(q)})$;
Étape M Choisir $\boldsymbol{\theta}^{(q+1)}\in\Theta$ tel que $\boldsymbol{\theta}^{(q+1)}=\arg\max_{\boldsymbol{\theta}\in\Theta}\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{z}^u(\boldsymbol{\theta}^{(q)}))$.

Rappelons que CEM n'optimise pas la vraisemblance mais la vraisemblance complétée $\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{z}^u)$ sur le couple d'inconnues $(\boldsymbol{\theta},\mathbf{z}^u)$.

Dans [R4] [c9], les stratégies proposées s'appuient sur une procédure en trois étapes que l'on peut qualifiée de « recherche/iteration/sélection ». Cette stratégie est décrite comme suit :

Recherche Une méthode de recherche fournit s valeurs initiales pour EM ;
Itération EM est lancé à partir de chacune de ces positions initiales ;
Sélection Parmi les s essais précédents de EM, celui qui produit la vraisemblance la plus grande est retenu.

L'originalité du travail réside dans la définition de l'étape « recherche ». Cette étape est composée soit de l'algorithme CEM, soit de l'algorithme SEM ou encore de l'algorithme EM lui-même, utilisé sous forme de petites séquences. L'arrêt des algorithmes dans la partie « recherche » se fait par stabilité du critère de vraisemblance pour EM, du critère de vraisemblance complétée pour CEM et uniquement par le nombre d'itérations pour SEM. Si le nombre alloué d'itérations n'est pas encore atteint, une nouvelle séquence de CEM ou de EM est alors relancée. Chaque stratégie est initialisée au tout début de la procédure avec des proportions égales, des centres au hasard parmi l'échantillon et des matrices de variance diagonales égales dont la diagonale est composée des variances de chaque variable. La mesure du temps est effectuée par le nombre *total* d'itérations de chaque stratégie. La répartition des itérations au sein des essais et des algorithmes d'une stratégie est uniforme. Dans le cas de l'utilisation de l'algorithme CEM, si celui-ci converge avant son nombre alloué d'itérations, on le redémarre en lui allouant le nombre restant d'itérations et ainsi de suite, puis EM est initialisé avec les paramètres donnant la vraisemblance complétée la plus grande.

Les stratégies sont évaluées au travers de simulations de Monte-Carlo diverses et aussi de jeux de données réelles. Dans le choix des jeux de données, nous nous sommes attachés à une grande variété de situations susceptibles d'être favorables à l'une ou l'autre des stratégies testées de sorte de tirer un bilan sur des bases équitables. Il ressort que la stratégie de recherche consistant en un ensemble de petites séquences de EM semble légèrement plus performante que les autres, mais il faut surtout retenir des expériences que, pour avoir des chances d'obtenir le maximum de la vraisemblance, un nombre conséquent d'itérations doit être alloué à la procédure d'estimation.

2.2.4 Tirage sur les trajectoires

Cette partie présente une stratégie originale et non heuristique pour initialiser l'algorithme EM dans le cas des mélanges gaussiens multivariés [R3] [c8]. L'idée est de s'appuyer sur une distribution des paramètres initiaux définie sur un espace incluant toutes les trajectoires possibles de EM. Cet espace s'exprime simplement comme une relation entre les deux premiers moments empiriques et les paramètres du mélange à n'importe quelle itération de EM. Notant $\boldsymbol{\theta}^{(q)}$ le paramètre à l'itération q de EM et $\bar{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i/n$ et $\bar{\boldsymbol{\chi}} = \sum_{i=1}^{n} \boldsymbol{\chi}_i/n$ respectivement les premiers et second moments empiriques avec $\boldsymbol{\chi}_i = \mathbf{x}_i\mathbf{x}_i'$, un calcul direct montre que

$$\bar{\mathbf{x}} = \sum_{k=1}^{K} \pi_k^{(q)} \boldsymbol{\mu}_k^{(q)} \tag{2.15}$$

$$\bar{\boldsymbol{\chi}} = \sum_{k=1}^{K} \pi_k^{(q)} (\boldsymbol{\Sigma}_k^{(q)} + \boldsymbol{\mu}_k^{(q)} \boldsymbol{\mu}_k^{(q)\prime}). \tag{2.16}$$

En conséquence, toute trajectoire de EM est incluse dans l'espace défini par ces deux equations et il en va de même pour le maximum de vraisemblance.

À partir de cette remarque, l'idée est alors de définir une distribution peu informative $p(\boldsymbol{\theta}^{(0)})$ sur le paramètre initial qui vérifie les deux équations (2.15) et (2.16). Nous espérons ainsi parcourir efficacement l'espace des paramètres possibles de EM. Notant $\mathbf{p} = (\pi_1, \ldots, \pi_K)$, $\mathbf{M} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ et $\mathbf{S} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$, nous définissons tout d'abord une loi uniforme $p(\mathbf{p}^{(0)})$ sur l'espace des proportions. Puis, nous construisons une loi conditionnelle $p(\mathbf{M}^{(0)}|\mathbf{p}^{(0)})$ sur les centres en les tirant au hasard dans l'échantillon, sauf l'un d'entre eux qui est déterminé par la première relation (2.15). Ensuite, une loi sur les variances $p(\mathbf{S}_{var}^{(0)}|\mathbf{p}^{(0)}, \mathbf{M}^{(0)})$ est choisie, où $\mathbf{S}_{var}^{(0)} = (\{\boldsymbol{\Sigma}_k^{(0)}\}_{jj}, 1 \leq k \leq K, 1 \leq j \leq d)$ est l'ensemble des variances, en tirant uniformément toutes les variances dans l'espace déterminé par la seconde relation (2.16). Enfin, une loi $p(\mathbf{S}_{cov}^{(0)}|\mathbf{p}^{(0)}, \mathbf{M}^{(0)}, \mathbf{S}_{var}^{(0)})$ sur les covariances, où $\mathbf{S}_{cov}^{(0)} = (\{\boldsymbol{\Sigma}_k^{(0)}\}_{j\ell}, 1 \leq k \leq K, 1 \leq j < \ell \leq d)$ est l'ensemble des covariances, est définie dans l'espace des matrices diagonalement dominantes sauf les covariances de la dernière composante qui sont déterminées par la seconde relation (2.16).

Cette nouvelle procédure d'initialisation est ensuite comparée avec trois stratégies concurrentes. La première est identique à celle sur les trajectoires excepté le fait que les contraintes (2.15) et (2.16) sont totalement inexistantes dans le tirage. Le but est donc de montrer l'apport spécifique de ces contraintes. Ensuite, nous reprenons la stratégie de McLachlan et Peel (2000) [27] chap. 2 dont nous avons d'ailleurs parlé précédemment et qui consiste à fixer les proportions égales, à tirer les centres de façon indépendante suivant la loi $\mathsf{N}_d(\bar{\mathbf{x}}, \mathbf{V})$ où $\mathbf{V}$ est la matrice de variance empirique de l'échantillon complet, et enfin à fixer toutes les matrices de variances initiales égales à $\mathbf{V}$. Si les centres sont maintenant tirés au hasard dans l'échantillon, nous obtenons une version modifiée qui correspond à notre troisième stratégie concurrente.

Notons qu'une stratégie consistant à choisir les paramètres initiaux à partir d'une partition aléatoire de l'échantillon, comme la procédure EMRS de Coleman et Woodruff (2000) [11], n'est pas considérée dans nos essais. En effet, bien que cette stratégie produise des paramètres initiaux vérifiant pourtant les équations (2.15) et (2.16), cette procédure produit systématiquement des centres et des matrices de variances qui sont similaires entre les différentes composantes. On peut alors s'attendre à une recherche assez pauvre de l'espace des valeurs initiales possibles.

Les expériences sont menées au travers de simulations de Monte-Carlo d'une part et de jeux de données réelles d'autre part. Dans les deux cas, différentes dimensions et diverses tailles d'échantillon sont étudiées. Les mesures de performance sont ensuite présentées sous deux formes complémentaires. Tout d'abord, la fréquence d'obtention du maximum de vraisemblance (à une précision numérique près) est calculée à chaque itération de EM, permettant ainsi de juger de la fréquence et de la rapidité de convergence de EM. Ensuite, pour chaque itération de EM de nouveau, la proportion de maxima locaux (parmi tous les maxima identifiés) détectée par chaque stratégie est calculée, permettant ainsi de mesurer le degré d'exploration de l'espace de la vraisemblance induit par chaque méthode. Par ailleurs, le temps nécessaire pour initialiser EM avec chaque stratégie est mesuré.

Il ressort ainsi que la nouvelle procédure détecte systématiquement, à chaque itération de EM, une plus grande proportion de maxima locaux que toutes les autres stratégies. En pratique, aucune des stratégies ne dispense d'effectuer de multiples tirages suivis d'une routine EM. Par contre, parmi l'ensemble des valeurs données par EM après un nombre fixé d'itérations, la nouvelle procédure propose un plus large éventail de maxima, incluant non seulement les maxima faciles à détecter et que toute autre stratégie fournie aussi, mais également des maxima locaux relativement « étroits » difficiles à détecter car vraisemblablement peu de trajectoires de EM y conduisent. Comme nous avons considéré pour le moment le seul cas hétéroscédastique, certains de ces maxima locaux peuvent bien entendu correspondre à des situations pathologiques comme des *spurious maximizers* mais il suffit alors de les écarter et de choisir un maximum plus adapté parmi l'ensemble des solutions fournies. Si le maximum local intéressant est difficile à détecter, comme cela est susceptible d'être le cas dans de nombreuses situations, la stratégie proposée, permet, à chaque itération de EM, de détecter cette valeur avec une plus grande fréquence que les autres stratégies.

Concernant le temps nécessaire à l'obtention d'un paramètre initial, la nouvelle stratégie requiert un temps supérieur aux autres à cause de nombreux rejets dans la procédure de tirage des paramètres. Néanmoins, ce temps d'initialisation reste négligeable par rapport au gain apporté dans de nombreux cas, sauf si l'espace des paramètres devient grand et que, conjointement, la taille de l'échantillon diminue.

Notons que la procédure est facilement adaptable au cas homoscédastique, puisque l'équation (2.16) devient dans ce cas

$$\bar{\chi} = \mathbf{\Sigma}^{(q)} + \sum_{k=1}^{K} \pi_k^{(q)} \boldsymbol{\mu}_k^{(q)} \boldsymbol{\mu}_k^{(q)\prime}, \tag{2.17}$$

ce qui consiste à choisir la matrice de variance commune initiale $\mathbf{\Sigma}^{(0)}$ simplement de la façon suivante :

$$\mathbf{\Sigma}^{(0)} = \bar{\chi} - \sum_{k=1}^{K} \pi_k^{(0)} \boldsymbol{\mu}_k^{(0)} \boldsymbol{\mu}_k^{(0)\prime}. \tag{2.18}$$

Il semble acquis que la situation homoscédastique permettra une obtention encore plus rapide des paramètres initiaux du fait de sa grande simplicité.

En tirant sur l'espace des trajectoires, l'espace des paramètres possibles de EM semble donc parcouru plus efficacement qu'avec les méthodes classiques. Suite à ces résultats encourageants, plusieurs extensions de ce travail son envisageables : adaptation à d'autres modèles gaussiens définis dans Celeux et Govaert (1995) [9], étude de mélanges non gaussiens comme Poisson, exponentiel ou encore Bernoulli.

2.3 Validation d'une estimation

2.3.1 Problématique

Nous avons vu que, pour de nombreuses raisons d'ordre théorique et pratique, la méthode d'estimation par maximum de la vraisemblance est très populaire. En général, cette méthode consiste à retenir comme estimateur le paramètre qui conduit à la plus grande valeur de la vraisemblance. En effet, comme discuté à la section 2.1.1, ce paramètre correspond souvent à l'unique racine convergente de l'équation de la vraisemblance, même si des cas pathologiques existent dans certaines situations (voir la section 2.1.2). Délaissant de telles situations dans la présente section, il reste tout de même la difficulté d'identifier la valeur du maximum de vraisemblance, car la fonction vraisemblance est bien connue aussi pour souvent comporter de nombreux points stationnaires. La section précédente a alors non seulement rappelé mais aussi proposé des méthodes pour détecter le plus grand nombre de ces points stationnaires. Il ne reste plus, ensuite, qu'à retenir la racine adéquate. Généralement, ces méthodes étaient liées à l'algorithme EM.

Néanmoins, en pratique, la recherche de toutes les racines de la vraisemblance peut prendre un temps considérable, sans pourtant aucune garantie que tous les maxima locaux soient identifiés en un temps fini, même si le nombre de racines est borné (on pourra trouver un exemple comportant un nombre infini de racines dans Barnett 1966 [2]). Finalement, au delà de cette stratégie basique consistant à rechercher toutes les racines, peu de propositions existent. Par exemple, De Haan (1981) [14] a proposé un intervalle de confiance sur la valeur du maximum de la vraisemblance à partir de la théorie asymptotique des valeurs extrêmes. Comme mis en évidence par Veall (1991) [40] dans un contexte économétrique, cette approche devient impraticable à cause du nombre de calculs nécessaires quand le support de l'espace des paramètres est grand ou encore quand cet espace est multidimensionnel. Une approche alternative peut consister à construire un test statistique sur la convergence d'une racine de l'équation de la vraisemblance. Cette démarche est justifier par le fait que, comme nous l'avons rappelé ci-dessus, la racine maximisant la vraisemblance est généralement la seule racine convergente. En d'autres mots, une telle méthode permet de décider si une racine $\hat{\boldsymbol{\theta}}$ donnée peut ou non être adoptée comme le maximum global de la vraisemblance au travers des hypothèses suivantes :

$$\begin{cases} H_0 : & \hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0 \\ H_1 : & \hat{\boldsymbol{\theta}} \text{ ne converge pas vers } \boldsymbol{\theta}_0, \end{cases}$$

où $\boldsymbol{\theta}_0$ est le paramètre à estimer défini par l'équation 1.24 du chapitre 1. Ainsi, il devient possible de rechercher une nouvelle racine (par exemple en utilisant les stratégies associées à EM et proposées à la section précédente), puis de la tester, jusqu'à ce que la racine courante ne soit plus rejetée par le test.

Nous présentons une bibliographie de tests de ce type puis nous nous focalisons sur une nouvelle proposition dans ce contexte. Notons que ces résultats dépassent le cadre des mélanges et peuvent donc s'appliquer à des situations très générales qui mettent en œuvre le principe du maximum de vraisemblance.

2.3.2 Aperçu de l'existant

Dans ce contexte, Heyde (1997) [19] et Heyde et Morton (1998) [20] ont proposé de construire un test soit en employant un critère d'adéquation permettant de sélectionner la meilleure racine, soit en retenant la racine pour laquelle le hessien de la log-vraisemblance a un comportant asymptotique similaire à son espérance évaluée avec la présente racine. Dans le même esprit, Gan et Jiang (1999) [17] ont choisi une statistique de décision qui est basée sur la différence entre une estimation de la forme produit $\mathbf{K}_\beta(\boldsymbol{\theta})$ et une estimation de la forme hessienne $\mathbf{J}_\beta(\boldsymbol{\theta})$ de la matrice d'information de Fisher calculée en $\boldsymbol{\theta}$, avec $\beta = m/n \in [0,1]$ la proportion d'individus d'appartenance connue dans $\mathcal{D}$, $\mathbf{J}_\beta(\boldsymbol{\theta}) = (\beta\mathbf{J}_c(\boldsymbol{\theta})+(1-\beta)\mathbf{J}(\boldsymbol{\theta}))$ et $\mathbf{K}_\beta(\boldsymbol{\theta}) = (\beta\mathbf{K}_c(\boldsymbol{\theta}) + (1-\beta)\mathbf{K}(\boldsymbol{\theta}))$ où [1]

$$\mathbf{J}_c(\boldsymbol{\theta}) = -\mathsf{E}_{\boldsymbol{\theta}_0}\nabla^2 \ln p(\mathbf{X}_1,\mathbf{Z}_1;\boldsymbol{\theta}), \qquad \mathbf{J}(\boldsymbol{\theta}) = -\mathsf{E}_{\boldsymbol{\theta}_0}\nabla^2 \ln p(\mathbf{X}_1;\boldsymbol{\theta}), \tag{2.19}$$

$$\mathbf{K}_c(\boldsymbol{\theta}) = \mathsf{Var}_{\boldsymbol{\theta}_0}\nabla \ln p(\mathbf{X}_1,\mathbf{Z}_1;\boldsymbol{\theta}), \qquad \mathbf{K}(\boldsymbol{\theta}) = \mathsf{Var}_{\boldsymbol{\theta}_0}\nabla \ln p(\mathbf{X}_1;\boldsymbol{\theta}). \tag{2.20}$$

Notons que les espérances sont prises par rapport à la loi paramétrée par $\boldsymbol{\theta}_0$ car le test que nous présentons maintenant repose sur l'hypothèse que le modèle est exact. Remarquant que

$$\mathsf{E}_{\boldsymbol{\theta}_0}\nabla\ell(\boldsymbol{\theta}_0;\mathcal{D}) = \mathbf{0} \tag{2.21}$$

et que

$$\mathbf{K}_\beta(\boldsymbol{\theta}_0) - \mathbf{J}_\beta(\boldsymbol{\theta}_0) = \mathbf{0}, \tag{2.22}$$

le test de Gan et Jiang consiste intuitivement à retenir le paramètre $\hat{\boldsymbol{\theta}}$ qui vérifie les deux équations suivantes :

$$\nabla\ell(\hat{\boldsymbol{\theta}};\mathcal{D}) = \mathbf{0} \tag{2.23}$$

et

$$\hat{\mathbf{K}}_\beta(\hat{\boldsymbol{\theta}}) - \hat{\mathbf{J}}_\beta(\hat{\boldsymbol{\theta}}) \approx \mathbf{0}, \tag{2.24}$$

où $\hat{\mathbf{K}}_\beta(\cdot)$ et $\hat{\mathbf{J}}_\beta(\cdot)$ sont respectivement des estimateurs de $\mathbf{K}_\beta(\cdot)$ et $\mathbf{J}_\beta(\cdot)$, les espérances étant estimées par méthode de Monte-Carlo à partir des données $\mathcal{D}$. S'appuyant sur certaines conditions de régularité et sur la conjecture suivante

$$\mathsf{E}_{\boldsymbol{\theta}_0}\nabla\ell(\boldsymbol{\theta};\mathcal{D}) = \mathbf{0} \text{ et } \mathbf{K}_\beta(\boldsymbol{\theta}) - \mathbf{J}_\beta(\boldsymbol{\theta}) = \mathbf{0} \Longrightarrow \boldsymbol{\theta} = \boldsymbol{\theta}_0, \tag{2.25}$$

ces auteurs établissent que $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ est équivalent à $\hat{\mathbf{K}}_\beta(\hat{\boldsymbol{\theta}}) - \hat{\mathbf{J}}_\beta(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{0}$. Obtenant la loi asymptotique de la statistique de décision $\hat{\mathbf{K}}_\beta(\hat{\boldsymbol{\theta}})-\hat{\mathbf{J}}_\beta(\hat{\boldsymbol{\theta}})$, il est alors aisé de construire le test bilatéral associé. Dans leur travaux, se focalisant dans le cas d'un paramètre univarié, sa distribution asymptotique sous l'hypothèse nulle est donnée par

$$\frac{\hat{\mathbf{K}}_\beta(\hat{\theta}) - \hat{\mathbf{J}}_\beta(\hat{\theta})}{\sqrt{\mathsf{Var}_{\boldsymbol{\theta}_0}[\hat{\mathbf{K}}_\beta(\hat{\theta}) - \hat{\mathbf{J}}_\beta(\hat{\theta})]}} \xrightarrow{p} \mathsf{N}(0,1), \tag{2.26}$$

la variance du dénominateur étant approchée également par une méthode de Monte-Carlo. Malheureusement, ce test révèle une faible puissance pour des tailles modérées d'échantillon dans des situations pourtant limitées à l'estimation d'un unique paramètre réel.

1. Avec les notations de (1.26) et (1.27) du chapitre 1 on a $\mathbf{J}_\beta = \mathbf{J}_\beta(\boldsymbol{\theta}_0)$ et $\mathbf{K}_\beta = \mathbf{K}_\beta(\boldsymbol{\theta}_0)$ s'il s'agit du vrai modèle.

2.3.3 Un nouveau test du maximum de la vraisemblance

Nous présentons ici une procédure similaire à celle de Gan et Jiang (1999) [17] pour tester la convergence d'une solution stationnaire de la log-vraisemblance, le changement intervenant uniquement dans la statistique de test. Notant

$$\delta(\boldsymbol{\theta};\mathcal{D}) = n^{-1}\{[\ell(\boldsymbol{\theta};\mathcal{D})] - [\mathsf{E}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\mathcal{D})]\}, \tag{2.27}$$

l'idée générale reste très simple [R1] [c7] : si le modèle est exact, une solution convergente $\hat{\boldsymbol{\theta}}$ devra vérifier, en plus de l'équation (2.23), l'équation suivante (qui remplace donc (2.24))

$$\delta(\hat{\boldsymbol{\theta}};\mathcal{D}) \approx 0, \tag{2.28}$$

car les termes $n^{-1}\{[\ell(\boldsymbol{\theta};\mathcal{D})]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\}$ et $n^{-1}\{[\mathsf{E}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\mathcal{D})]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\}$ convergent tout deux vers la même valeur $\beta\mathsf{E}_{\boldsymbol{\theta}_0}\ln p(\mathbf{X}_1,\mathbf{Z}_1;\boldsymbol{\theta}_0) + (1-\beta)\mathsf{E}_{\boldsymbol{\theta}_0}\ln p(\mathbf{X}_1;\boldsymbol{\theta}_0)$. S'appuyant sur certaines conditions de régularité et sur la conjecture suivante

$$\mathsf{E}_{\boldsymbol{\theta}_0}\nabla\ell(\boldsymbol{\theta};\mathcal{D}) = \mathbf{0} \text{ et } \mathsf{E}_{\boldsymbol{\theta}_0}\ell(\boldsymbol{\theta};\mathcal{D}) - \mathsf{E}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\mathcal{D}) = 0 \Longrightarrow \boldsymbol{\theta} = \boldsymbol{\theta}_0, \tag{2.29}$$

il est alors possible d'établir de nouveau que $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ est équivalent à $\delta(\hat{\boldsymbol{\theta}};\mathcal{D}) \xrightarrow{p} 0$. La distribution de cette statistique sous l'hypothèse nulle est donnée par

$$\delta(\hat{\boldsymbol{\theta}};\mathcal{D}) \xrightarrow{d} \mathsf{N}\left(0, \frac{\beta\mathsf{Var}_{\boldsymbol{\theta}_0}\ln p(\mathbf{X_1},\mathbf{Z_1};\boldsymbol{\theta}_0) + (1-\beta)\mathsf{Var}_{\boldsymbol{\theta}_0}\ln p(\mathbf{X_1};\boldsymbol{\theta}_0)}{n}\right). \tag{2.30}$$

Les variances inconnues sont simplement estimées par $[\beta\mathsf{Var}_{\boldsymbol{\theta}}\ln p(\mathbf{X}_1,\mathbf{Z}_1;\boldsymbol{\theta}) + (1-\beta)\mathsf{Var}_{\boldsymbol{\theta}}\ln p(\mathbf{X}_1;\boldsymbol{\theta})]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$. Un test bilatéral est alors construit. Ce test est bilatéral car on peut trouver des valeurs de $E_{\boldsymbol{\theta}_0}\ell(\boldsymbol{\theta};\mathcal{D}) - E_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\mathcal{D})$ aussi bien positives que négatives.

L'implantation de cette nouvelle méthode est ainsi fort simple à mettre en œuvre quelle que soit la dimension du paramètre. De plus, des expérimentations avec des données simulées et des données réelles ont montré que la puissance obtenue est bien supérieure à celle du précédent test permettant ainsi de considérer cette fois des paramètres multidimensionnels. La conjecture (2.29) a été vérifiée, à chaque fois que cela était possible, dans les différents essais considérés.

Notons que, bien que l'estimateur du maximum de vraisemblance soit invariant (ou équivariant) lors d'une transformation de l'espace des données, la log-vraisemblance ne possède pas cette propriété, même après standardisation. En conséquence, le test proposé peut accepter une racine dans un système de coordonnées tout en rejetant cette même racine avec une transformation du système de coordonnées. Néanmoins, pour un niveau de signification donné, le test prendra asymptotiquement la même décision, indépendamment du système de coordonnées.

Il est important de remarquer que l'applicabilité du test suppose que le modèle soit bien spécifié. En effet, si tel n'est pas le cas, on a sous l'hypothèse nulle,

$$\delta(\hat{\boldsymbol{\theta}};\mathcal{D}) \xrightarrow{d} \mathsf{N}\left(\delta_0, \frac{\beta\mathsf{Var}_{(\mathbf{X}_1,\mathbf{Z}_1)}\ln p(\mathbf{X}_1,\mathbf{Z}_1;\boldsymbol{\theta}_0) + (1-\beta)\mathsf{Var}_{\mathbf{X}_1}\ln p(\mathbf{X}_1;\boldsymbol{\theta}_0)}{n}\right), \tag{2.31}$$

avec cette fois $|\delta_0| > 0$. On ne peut construire un test dans cette situation car la valeur de δ_0 est naturellement inconnue. Nous sommes alors face à un paradoxe car, d'une part le test du maximum de vraisemblance requiert d'avoir le bon modèle et, d'autre part, nombre de critères de choix de modèles comme AIC (Akaike 1973 [1]) ou BIC (Schwarz 1978 [35]) nécessitent de connaître cette valeur du maximum de vraisemblance. Afin de lever ce problème, nous avons expérimentalement remarqué que, si le modèle n'est pas adapté mais tout de même relativement « flexible » au sens que δ_0 n'est pas trop éloigné de zéro, le test restait tout de même opérationnel pour des tailles d'échantillon relativement petites. Par exemple, le cas d'un mauvais nombre de composantes a été considéré avec succès dans les expérimentations.

2.4 Dégénérescence dans les mélanges gaussiens

2.4.1 Problématique et aperçu de l'existant

Comme nous l'avons déjà évoqué, il est bien connu que la fonction vraisemblance n'est pas bornée supérieurement dans le cas des modèles de mélange gaussiens hétéroscédastiques (Kiefer et Wolfowitz

1956 [22], Day 1969 [13]). En effet, si au moins l'un des centres gaussiens coïncide avec un individu parmi $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ et si le déterminant de la matrice de variance correspondante tend vers zéro, alors la fonction vraisemblance augmente sans limite (elle croit d'ailleurs vers l'infini à la même vitesse que l'inverse de la racine carrée de la variance généralisée du fait de l'expression de la densité gaussienne). À défaut d'un maximum global borné pour ce type de mélange, la méthode du maximum de vraisemblance consiste alors à retenir un maximum local de la vraisemblance. Par contre, des méthodes d'optimisation générales comme celle de Newton ou des méthodes d'optimisation plus spécialisées comme celle de l'algorithme EM (Dempster *et al.* 1977 [15]) sont susceptibles de diverger vers ces solutions de dégénérescence. Il est donc nécessaire de pouvoir se prémunir de telles dérives.

Une façon très simple d'appréhender la dégénérescence consiste à se procurer des données labellées comme cela a été remarqué dans le cas univarié par Hosmer (1973) [21]. À défaut de cette information supplémentaire souvent difficile à obtenir, les approches alternatives ayant pour but d'éviter la dégénérescence proposent essentiellement soit de transformer, soit de contraindre la fonction de vraisemblance initiale. Une approche naturelle consiste ainsi à contraindre les variances généralisées $|\mathbf{\Sigma}_1|,\ldots,|\mathbf{\Sigma}_K|$ (ou de façon équivalente le volume des composantes, voir la section 1.1.5 du chapitre précédent) à être plus grandes qu'une valeur *a priori* $\alpha > 0$: $\forall k = 1,\ldots,K$, $|\mathbf{\Sigma}_k| > \alpha$. Grâce à cette contrainte, le maximum de la vraisemblance est borné. Un autre façon de procéder est d'imposer une contrainte relative entre les variances généralisées du type $\min_k |\mathbf{\Sigma}_k| \geq \beta \max_k |\mathbf{\Sigma}_k|$, avec $0 < \beta \leq 1$ une constante donnée. Cette procédure a été proposée en premier par Hathaway (1985) [18]. Dans le contexte bayésien, le principe de maximisation de la vraisemblance est remplacé par la maximisation d'une fonction agrégeant la vraisemblance et une densité *a priori* $p(\boldsymbol{\theta})$ sur le paramètre de mélange. Dans tous les cas, des petites valeurs de variances généralisées sont lourdement pénalisées, c'est-à-dire que cette loi est choisie afin que la limite de la nouvelle fonction $\ell(\boldsymbol{\theta};\mathcal{D}) + \ln p(\boldsymbol{\theta})$ soit égale à $-\infty$ quand au moins une des variances généralisées tend vers zéro (voir Ciuperca *et al.* 2003 [10] pour une mise en oeuvre dans le cas univarié).

Ainsi, toutes ces propositions nécessitent une information supplémentaire aux données qui change le problème initial et, par conséquent, risque d'affecter l'estimation de $\boldsymbol{\theta}$. Dans le cas bayésien, Ciuperca *et al.* (2003) [10] ont prouvé la convergence de leur estimateur. Néanmoins, pour des tailles modérées d'échantillon, la densité *a priori* peut fortement affecter la convergence de celui-ci. Dans l'approche où un minimum est défini pour la variance généralisée, le résultat de l'estimation est évidemment dépendant du choix de cette borne. En effet, à la place de dégénérer, la vraisemblance risque de converger vers une variance généralisée arbitrairement choisie qui n'offre finalement pas plus d'interprétation concrète que la solution de dégénérescence elle-même (voir [C6] pour un exemple univarié). Dans la situation des contraintes relatives entre variances généralisées, le fait d'imposer les variances généralisées égales est un cas intéressant obtenu avec $\beta = 1$. Par exemple, cette situation avait été considérée à la section 2.2.3 pour éviter la dégénérescence dans nos expériences. Un cas particulier de l'égalité des variances généralisées consiste en l'égalité des matrices de variances dans leur intégralité. En effet, l'homoscédaticité peut être une information *a priori* naturelle dans certains cas, comme en biologie où cette propriété est souvent vérifiée quand des variables biométriques de mâles et de femelles sont étudiées. Cependant, quand aucune information *a priori* n'est disponible, toutes les solutions présentées restent subjectives pour éviter la dégénérescence.

Dans les deux sections suivantes, nous discutons de techniques objectives afin de répondre au problème de la dégénérescence de la fonction de vraisemblance. La première relève plus d'une approche pragmatique en remarquant que, au moins dans le cas univarié, l'algorithme EM converge extrêmement rapidement vers la dégénérescence et donc que la précision de l'ordinateur sera suffisante pour détecter ces solutions indésirables. La seconde section s'attache plutôt à définir, dans un cas multivarié cette fois, une borne asymptotique pour le maximum non dégénéré de la vraisemblance.

2.4.2 Aspect pratique : comportement de EM

En pratique, de nombreux utilisateurs de l'algorithme EM (Dempster *et al.* 1977 [15]) dans le cadre des mélanges gaussiens ont remarqué que la procédure pouvait s'arrêter de façon incidentelle. Cet arrêt imprévu est généralement du au fait que certaines matrices de variance sont excessivement mal conditionnées, au point que la précision numérique de l'ordinateur soit atteinte pour calculer leur déterminant ou les inverser. Cet événement est typiquement un indice de convergence vers une solution de dégénérescence. Plusieurs remarques importantes doivent être associées à cette situation, toutes ayant été illustrées au travers de l'exemple univarié présenté en introduction de [R5]. Tout d'abord, les utilisateurs constatent que la dégénérescence peut être atteinte même lorsque l'algorithme EM est initialisé avec des grandes variances généralisées, c'est-à-dire quand une situation de dégénérescence est éloignée. Deuxièmement, avant d'atteindre une situation dégénérée, il est tout à fait possible que EM passe par des étapes de quasi-stabilité de la vraisemblance ou des paramètres pendant de nombreuses itérations, au point de simuler la convergence vers un maximum local. Ainsi, le passage bien connu de EM par des points selles n'est pas exclu avant que EM ne dégénère. Enfin, la convergence vers des points de dégénérescence est extrêmement rapide lorsque la procédure se trouve dans un voisinage d'une de ces solutions limites. Cette dernière constatation incite finalement les praticiens à gérer la dégénérescence de EM de façon assez pragmatique puisqu'ils se contentent souvent de relancer l'algorithme à partir d'une position différente lors d'une situation de dépassement de la précision numérique.

Les travaux [R5] [C6] apportent une justification théorique, au moins dans le cas univarié, à ce comportement pragmatique fort répandu. Cette justification se fait en deux temps. Tout d'abord, il est démontré qu'il existe un domaine d'attraction autour de toute solution dégénérée (théorème 1 de [R5]). Ainsi, si EM est assez proche d'une solution dégénérée, alors EM y convergera et son point fixe sera dégénéré. Dans un second temps il est établi que, si EM converge vers une dégénérescence, alors la variance de la composante qui est la cause de la dégénérescence tend vers zéro avec une vitesse exponentielle (théorème 2 de [R5]). Comme la vraisemblance tend vers l'infini aussi vite que l'inverse de l'écart-type, la divergence de la vraisemblance est également exponentielle. Cette vitesse de convergence est paramétrée par une constante multiplicative de valeur inconnue qui peut néanmoins diminuer la vitesse de convergence au début si elle est trop grande, le caractère exponentiel reprenant de toute façon très rapidement sa domination au bout d'un certain nombre d'itérations. En général, on constate cependant que deux ou trois itérations suffisent pour atteindre la limite de l'ordinateur.

L'ensemble de ces résultats ont été établis dans le cas univarié pour EM mais, les expériences montrent qu'il est raisonnable d'espérer des propriétés similaires dans le cas multivarié ainsi que pour d'autres types d'algorithmes comme une procédure de Newton et un algorithme de gradient.

2.4.3 Aspect théorique : une borne supérieure de la vraisemblance

La section précédente conduit aux deux remarques contradictoires suivantes. Tout d'abord, la procédure consistant à laisser la précision numérique de l'ordinateur jouer un rôle de détection de la dégénérescence est tout de même arbitraire puisque cela revient finalement à fixer une variance minimale égale à cette précision. Par contre, la convergence extrêmement rapide de la variance vers zéro semble indiquer qu'il sera difficile d'établir un test visant à détecter la dégénérescence avant que la limite numérique ne soit atteinte.

Conscient de la difficulté de détecter en pratique la dégénérescence de EM ou d'autres algorithmes avant le franchissement fatidique de la précision informatique, [R2] propose néanmoins de répondre au caractère subjectif de cette borne fixée par les limites numériques de l'ordinateur. Il s'agit d'établir une borne supérieure pour la valeur de la vraisemblance non dégénérée dans le cas multivarié, la validité de cette borne n'étant valable qu'asymptotiquement. Le théorème 1 de [R2] démontre que

$$\ell_{sup} = n \left[\ln K + \left(-\frac{1}{2} \ln |\mathbf{V}^*| \right)^+ \right] \tag{2.32}$$

est une borne supérieure asymptotique de $\ell(\hat{\boldsymbol{\theta}}; \mathcal{D})$, c'est-à-dire que $p(\ell(\hat{\boldsymbol{\theta}}; \mathcal{D}) \leq \ell_{sup}) \longrightarrow 1$ quand $n \to \infty$, avec $\mathbf{V}^*$ la matrice empirique de variance de plus petit déterminant obtenue avec $d+1$ individus parmi toutes les combinaisons possibles de $d+1$ individus dans l'échantillon. Cette borne est valable que le modèle soit ou non correct.

Afin de mesurer la validité à taille finie de cette borne, deux approches ont été menées. Tout d'abord, il a été établi que, si les composantes sont bien séparées, cette borne est valable pour des tailles d'échantillon très réduites (théorème 2 de [R2]). Ensuite, des expérimentations ont montré que, même si les composantes sont peu séparées, la borne semble tout de même valide pour des échantillons de taille réduite.

Il semble néanmoins établi que cette borne n'est pas suffisamment petite pour prévenir la dégénérescence effective de la vraisemblance avec l'algorithme EM car la variance généralisée atteint sa limite numérique généralement avant que la borne ℓ_{sup} ne soit atteinte. Cette borne apporte alors plutôt une réponse théorique à la dégénérescence qu'une méthode pratique réellement efficace, à moins de disposer d'un ordinateur de précision extrêmement élevée. Par ailleurs, il est à noter que la complexité de calcul de ℓ_{sup} ne dépend pas du nombre de composantes mais elle évolue de façon combinatoire en fonction du nombre d'individus et de la dimension de l'espace du fait du calcul de $|\mathbf{V}^*|$. En effet, ce calcul peut s'avérer coûteux en temps.

2.5 Bilan et perspectives

Nous avons rappelé que la méthode du maximum de vraisemblance, en particulier dans le cas des mélanges gaussiens, posait un certain nombre de difficultés théoriques et algorithmiques. Le problème de l'existence de plusieurs maxima locaux a conduit a proposer des réponses de différentes natures. Tout d'abord, en amont, des stratégies d'initialisation de l'algorithme EM ont été discutées, la plus notable étant certainement celle consistant à tirer les paramètres initiaux de EM sur un sous-espace des paramètres incluant toutes les trajectoires de cet algorithme. Ensuite, en aval, une procédure de test permettant de décider si une racine de l'équation de la vraisemblance correspond au maximum a été proposée. Ce test est d'une portée plus large que les modèles de mélange. Par ailleurs, le problème de la dégénérescence de la vraisemblance dans les mélanges gaussiens a été abordé. En premier lieu, il a été établi que l'algorithme EM convergeait extrêmement rapidement vers de telles solutions en leur voisinage, ce qui tendait ainsi à justifier la pratique répandue de la relance de EM à chaque échec du à la dégénérescence. Ensuite, une borne supérieure asymptotique du maximum de vraisemblance non dégénéré a été proposée pour apporter une réponse plus théorique à cette difficulté des mélanges gaussiens. Diverses extensions et perspectives peuvent alors être envisagées suite aux travaux précédents.

Une première extension consiste à poursuivre l'approche innovante et prometteuse liée au tirage des paramètres initiaux de EM sur ses trajectoires. Actuellement, seuls les modèles gaussiens hétéroscédastiques et homoscédastiques ont été considérés et il serait utile d'adapter des lois de tirage aux douze autres modèles gaussiens de Celeux et Govaert (1995) [9]. D'ailleurs, dans le cas hétéroscédastique, nous avons remarqué une lenteur de la procédure de tirage due à de nombreux rejets lorsque la dimension de l'espace et le nombre de composantes augmentent. La proposition faite dans le cas hétéroscédastique pourra donc certainement être elle-même améliorée. Il semble également raisonnable d'étendre ces stratégies de tirages à des mélanges non-gaussiens, comme les mélanges de Poissons, exponentiels ou encore des classes latentes. Enfin, les stratégies des trajectoires pourraient être considérées dans le cas de labels partiellement connus, ou encore adaptées pour des algorithmes autres que EM. Citons notamment des variantes de EM comme l'algorithme de classification CEM ou encore des procédures indépendantes comme une méthode de Newton. Une application directe à Newton peut surprendre mais, en fait, il faut garder à l'esprit que le maximum de vraisemblance vérifie lui-même les équations de trajectoires et qu'il est ainsi raisonnable d'initialiser tout algorithme en utilisant cette information. Peut être est-il aussi envisageable de maximiser la vraisemblance sous les contraintes de trajectoires dans ce type d'algorithme. En effet, EM peut être interprété comme une méthode de maximisation de la vraisemblance sous contraintes, bien que ces contraintes soit naturelles avec EM puisqu'elles correspondent aux caractéristiques de ses propres trajectoires.

Par ailleurs, le nouveau test du maximum de la vraisemblance a montré une meilleure puissance que le principal test existant. Néanmoins, comme ce dernier, sa justification repose partiellement sur la validité d'une conjecture qui correspond à l'équation (2.29) dans notre proposition. Un contre-exemple a été exhibé dans le cas du test concurrent mais sa portée était assez réduite du fait de son caractère plutôt artificiel. L'objectif est alors d'identifier assez précisément le domaine de validité de notre propre conjecture.

Concernant l'étude de la convergence de EM vers des solutions dégénérescentes, une extension au cas multidimensionnel semble intéressante dans un premier temps. Une étude similaire pourrait être aussi menée pour d'autres algorithmes que EM comme un gradient ou encore une méthode de Newton. Au chapitre 3, une première extension a été proposée pour les mélanges gaussiens univariés avec données groupées.

Enfin, une perspective importante concerne le cas des *spurious solutions*. Certes, McLachlan et Peel (2000) [27] proposent d'éliminer ces solutions en examinant les variances généralisées ou encore les valeurs propres des matrices de variance. Cependant, cette procédure reste assez subjective et l'on peut même affirmer que la détection rigoureuse de ces racines particulières de la vraisemblance est une question ouverte. Deux pistes de recherche sont pour le moment proposées. Tout d'abord, le test du maximum de la vraisemblance pourrait être employé en l'etat puisqu'il consiste finalement à tester si une racine de l'équation de la vraisemblance est ou non convergente. En effet, une solution *spurious* n'est pas une solution convergente. Deuxièmement, il pourrait être établit une borne asymptotique pour le maximum de la vraisemblance associé à la racine convergente. La borne asymptotique destinée à prévenir d'une solution dégénérée est une bonne base de départ car elle est construite en supposant la convergence de la racine de la vraisemblance. Elle devra cependant être améliorée pour pouvoir être efficace dans cette nouvelle situation.

Bibliographie

[1] H. Akaike. Information Theory as an Extension of the Maximum Likelihood Principle. In B. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, Akademiai Kiado, 1973.

[2] V. D. Barnett. Evaluation of the maximum likelihood estimator where the likelihood equation has multiple roots. *Biometrika*, 53:151–166, 1966.

[3] G. Celeux, S. Chrétien, F. Forbes, and A. Mkhadri. A component-wise EM algorithm for mixtures. *Journal of Computational and Graphical Statistics*, 10:699–712, 2001.

[4] 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(1):73–92, 1985.

[5] G. Celeux and J. Diebolt. Une version de type recuit simulé de l'algorithme EM. *Notes aux Comptes Rendus de l'Académie des Sciences*, 310:119–124, 1990.

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

[7] G. Celeux and G. Govaert. A classification EM algorithm and two stochastic versions. *Computational Statistics & Data Analysis*, 14:315–332, 1992.

[8] 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.

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

[10] G. Ciuperca, A. Ridolfi, , and J. Idier. Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, 30(1):45–59, 2003.

[11] D. A. Coleman and D. L. Woodruff. Cluster analysis for large datasets: An effective algorithm for maximizing the mixture likelihood. *Journal of Computational and Graphical Statistics*, 9(4):672–688, 2000.

[12] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ, 1946.

[13] N. E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56:463–474, 1969.

[14] L. De Haan. Estimation of the minimum of a function using order statistics. *Journal of the American Statistical Association*, 76:467–469, 1981.

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

[16] T. S. Ferguson. An inconsistent maximum likelihood estimate. *Journal of the American Statistical Association*, 77(380):831–834, 1982.

[17] L. Gan and J. Jiang. A test for global maximum. *Journal of the American Statistical Association*, 94(447):847–854, 1999.

[18] R.J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal distributions. *Annals of Statistics*, 13:795–800, 1985.

[19] C. C. Heyde. *Quasi-Likelihood and Its Application*. Springer, New-York, 1997.

[20] C. C. Heyde and R. Morton. Multiple roots in general estimating equations. *Biometrika*, 85:954–959, 1998.

[21] D. W. Hosmer, Jr. A comparison of iterative maximum-likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, 29:761–770, 1973.

[22] J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27:887–906, 1956.

[23] E.L. Lehmann. *Theory of Point Estimation.* Wiley, New-York, 1983.

[24] C. Liu and D. B. Rubin. The ecme algorithm: a simple extension of em and ecm with faster monotone convergence. *Biometrika*, 81:633–648, 1994.

[25] C. Liu and D. X. Sun. Acceleration of EM algorithm for mixtures models using ecme. In *ASA Proceedings of The Stat. Comp. Session*, volume 5, pages 109–114, 1997.

[26] M. Markatou, A. Basu, and B. G. Lindsay. Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, 93(442):740–750, 1998.

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

[28] G. J. McLachlan and T. Krishnam. *The EM algorithm and Extensions.* Wiley, New York, 1997.

[29] M. Meila and D. Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42:9–29, 2001.

[30] W. Mendenhall and R. J. Hader. Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika*, 45:504–520, 1958.

[31] J. Neyman and E. Scott. Consistent estimators based on partially consistent observations. *Econometrica*, 16:1–32, 1948.

[32] R. S. Pilla and B. G. Lindsay. Alternative EM methods for nonparametric finite mixture models. *Biometrika*, 88:535–550, 2001.

[33] C. R. Rao. The utilization of multiple measurements in problems of biological classification (with discussion). *Journal of the Royal Statistical Society, Series B*, 10:159–203, 1948.

[34] R.A. Redner and H.F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, 1984.

[35] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[36] C. G. Small, J. Wang, and Z. Yang. Eliminating multiple root problems in estimation. *Statistical Science*, 15(4):313–341, 2000.

[37] L. A. Stefanski and R. J. Carroll. Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74(4):703–716, 1987.

[38] R. D. Tarone and G. Gruenhage. A note on the uniqueness of roots of the likelihood equations for vector-valued parameters. *Journal of the American Statistical Association*, 70:903–904, 1975.

[39] N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11:271–282, 1998.

[40] M. R. Veall. Testing for a global maximum in a econometric context. *Econometrica*, 56:1959–1965, 1991.

[41] H. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of the Mathematical Statistics*, 20:595–601, 1949.

[42] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

[43] J. H. Wolfe. A monte carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Technical Bulletin STB 72-2, US Naval Personnel Research Activity, San Diego, California, 1971.

Chapitre 3

Spécificité des données groupées

Sommaire

Publications associées à ce chapitre

Revues avec comité de lecture

[R1] C. Biernacki. *Influence of the Bin Dimension on Selecting a Model by the BIC Criterion in Gaussian Mixtures with Grouped Data*. En préparation.

[R2] C. Biernacki. *Degeneracy in the Maximum Likelihood Estimation of Univariate Gaussian Mixtures for Grouped Data and Behaviour of the EM Algorithm*. Scandinavian Journal of Statistics, en révision.

Dans de nombreuses situations, les données collectées se présentent sous une forme dite groupée, c'est-à-dire que la position exacte des données est inconnue à l'intérieur d'un *bin*, terme désignant un sous-ensemble de l'espace des données (typiquement un intervalle dans le cas unidimensionnel). On parle d'ailleurs parfois de *binned data* (voir par exemple McLachlan et Peel 2000 [6] chap. 9). Ce type de données apparaît systématiquement lors de l'utilisation d'un instrument de mesure avec une résolution finie, mais les données groupées peuvent aussi être obtenues intentionnellement, par exemple dans le but de simplifier un grand tableau de données. Tout autant que dans le cas des données individuelles (terme utilisé par opposition aux données groupées et désignant les données « classiques » que nous avons vus jusqu'à présent), des besoins en classification automatique, en analyse discriminante en encore en estimation de densité peuvent être naturellement associées à ce nouveau type de données. Nous avons vu dans les chapitres précédents que les modèles de mélange apportent des solutions intéressantes à ces diverses problématiques et il est donc judicieux de disposer de ces outils dans le cas groupé également. Bien entendu, les méthodes d'estimation et de choix de modèles qui ont été décrites dans les deux chapitres précédents sont suffisamment générales pour s'appliquer aux *binned data*, moyennant quelques adaptations techniques. Par contre, des questions spécifiques aux données groupées doivent être adressées : il s'agit typiquement de l'influence de la dimension des *bins* sur les méthodes d'estimation et de choix de modèles.

Ce chapitre se compose de deux parties principales. La section 3.1 traite tout d'abord des indispensables adaptations techniques pour passer des données individuelles aux données groupées. Ainsi, les modèles de mélange et l'algorithme EM sont décrits dans ce nouveau contexte. Ensuite, la section 3.2 discute dans un premier temps des problèmes de dégénérescence dans les mélanges gaussiens, aussi bien au niveau de la valeur de la fonction vraisemblance que du comportement de l'algorithme EM. Des différences remarquables avec le cas individuel sont mises en évidence et l'influence majeure de la dimension des *bins* sur ce comportement est identifiée. Dans un second temps, l'influence de cette dimension sur le comportement du critère de choix de modèles BIC est discutée. Cette étude, bien qu'illustrée dans le cadre des mélanges, dépasse en fait le cadre des mélanges (gaussiens ou autre) cette fois.

3.1 Modèles de mélange et algorithme EM

3.1.1 Modélisation

On considère une partition de l'espace $\mathcal{X}$ des individus $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ en H sous-espaces (ou *bins*) $\Omega_1, \ldots, \Omega_H$. Dans le contexte des données groupées, la valeur de chaque individu $\mathbf{x}_i$ ($i = 1, \ldots, n$) est ignorée, seul le *bin* Ω_h ($h = 1, \ldots, H$) auquel appartenait cet individu étant conservée dans un vecteur binaire $\mathbf{y}_i = (y_{i1}, \ldots, y_{iH})'$ avec $y_{ih} = 1$ si $\mathbf{x}_i \in \Omega_h$ et $y_{ih} = 0$ sinon. Ainsi, l'ensemble des valeurs individuelles initiales $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ est maintenant réduit à une matrice binaire $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$. Comme les individus $\mathbf{x}_1, \ldots, \mathbf{x}_n$ étaient supposés être des réalisations i.i.d. de la loi mélange $p(\mathbf{x}_1; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k p(\mathbf{x}_1; \boldsymbol{\alpha}_k)$, les nouvelles données groupées $\mathbf{y}_1, \ldots, \mathbf{y}_n$ sont elles-même n réalisations i.i.d. d'un vecteur aléatoire $\mathbf{Y}_1$ de distribution

$$p(Y_{1h} = 1; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k a_h(\boldsymbol{\alpha}_k) \tag{3.1}$$

$$= A_h(\boldsymbol{\theta}) \tag{3.2}$$

où

$$a_h(\boldsymbol{\alpha}_k) = \int_{\Omega_h} p(\boldsymbol{\lambda}; \boldsymbol{\alpha}_k) d\boldsymbol{\lambda}. \tag{3.3}$$

Souvenons nous alors que seuls les labels des m premiers individus étaient connus pour déduire que le nouvel ensemble de données disponible devient $\mathcal{D} = (\mathcal{D}_l, \mathcal{D}_u)$ avec $\mathcal{D}_l = (\mathbf{y}^l, \mathbf{z}^l)$ et $\mathcal{D}_u = \mathbf{y}^u$, où, sans surprise, $\mathbf{y}^l = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$ et $\mathbf{y}^u = (\mathbf{y}_{m+1}, \ldots, \mathbf{y}_n)$. Dans ce contexte, le log-vraisemblance de $\boldsymbol{\theta}$ s'écrit

$$\ell(\boldsymbol{\theta}; \mathcal{D}) = \sum_{h=1}^H \sum_{k=1}^K \psi_{kh} \ln(\pi_k a_h(\boldsymbol{\alpha}_k)) + \sum_{h=1}^H \Psi_h \ln(A_h(\boldsymbol{\theta})), \tag{3.4}$$

où, d'une part, $\psi_{hk} = \sum_{i=1}^{m} y_{ih} z_{ik}$ désigne le nombre d'individus appartenant au groupe G_k et qui sont de plus dans le *bin* Ω_h ($k = 1,\ldots,K$, $h = 1,\ldots,H$, $\sum_{h=1}^{H}\sum_{k=1}^{K}\psi_{kh} = m$) et, d'autre part, $\Psi_h = \sum_{i=m+1}^{n} y_{ih}$ désigne le nombre d'individus de label inconnu qui sont dans le *bin* Ω_h ($h = 1,\ldots,H$, $\sum_{h=1}^{H}\Psi_h = n-m$).

En fait, nous pouvons remarquer que $\mathbf{y}$ n'est pas réellement nécessaire dans l'expression de la vraisemblance puisque uniquement les effectifs ψ_{kh} et Ψ_h ($k = 1,\ldots,K$, $h = 1,\ldots,H$) sont utilisés. C'est pour cette raison que McLachlan et Jones (1988) [7] ont considéré que les données groupées sont ψ_{kh} et Ψ_h ($k = 1,\ldots,K$, $h = 1,\ldots,H$) plutôt que $\mathbf{y}$. Cette nouvelle formulation des données oblige à utiliser une distribution multinomiale qui, finalement, conduit à une log-vraisemblance identique à (3.4) à une constante additive près.

3.1.2 Adaptation de l'algorithme EM

L'algorithme EM peut bien entendu être utilisé pour maximiser $\ell(\boldsymbol{\theta};\mathcal{D})$ et obtenir l'estimateur $\hat{\boldsymbol{\theta}}$ du MV. Dans le cadre des données groupées, les données manquantes ne sont plus seulement les labels inconnus $\mathbf{z}^u$ mais aussi les données individuelles $\mathbf{x}$. La log-vraisemblance complétée est donc inchangée par rapport aux cas non groupé, c'est-à-dire qu'elle s'exprime par $\ell(\boldsymbol{\theta};\mathbf{x},\mathbf{z})$.

L'étape E de l'algorithme EM (Dempster *et al.* 1997 [4]) à l'itération $q+1$ consiste alors à calculer l'espérance de cette log-vraisemblance complétée par rapport à la loi conditionnelle $p(\mathbf{x},\mathbf{z}^u|\mathcal{D};\boldsymbol{\theta}^{(q)})$, espérance qui s'écrit $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) = \mathsf{E}_{\boldsymbol{\theta}^{(q)}}[\ell(\boldsymbol{\theta};\mathbf{X},\mathbf{Z})|\mathcal{D}]$. Quelques manipulations permettent alors d'exprimer ce terme sous la forme suivante :

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) = \sum_{h=1}^{H}\sum_{k=1}^{K}\left\{\psi_{kh}\mathsf{E}_h^{(q)}[\ln\{\pi_k p(\mathbf{X}_1;\boldsymbol{\alpha}_k)\}] + \Psi_h\mathsf{E}_h^{(q)}[\ln\{\pi_k p(\mathbf{X}_1;\boldsymbol{\alpha}_k)\}t_{1k}(\boldsymbol{\theta}^{(q)})]\right\} \tag{3.5}$$

où $\mathsf{E}_h^{(q)}$ se rapporte à une espérance relative à la loi conditionnelle $p(\mathbf{x}_1;\boldsymbol{\theta}^{(q)})/A_h(\boldsymbol{\theta}^{(q)})$.

Dans la situation où aucun label n'est connu ($m = 0$), Cadez *et al.* (2002) [3] ont donné la formulation de $\pi_k^{(q+1)}$, $\boldsymbol{\mu}_k^{(q+1)}$ et $\boldsymbol{\Sigma}_k^{(q+1)}$ de l'étape M dans le cas gaussien le plus général ($k = 1,\ldots,K$). Notant

$$n_k^{(q)} = \sum_{h=1}^{H}\left\{\psi_{kh} + \Psi_h\mathsf{E}_h^{(q)}[t_{1k}(\boldsymbol{\theta}^{(q)})]\right\} \tag{3.6}$$

et

$$\mathbf{W}_{kh}^{(q)} = \psi_{kh}\mathsf{E}_h^{(q)}[(\mathbf{X}_1 - \boldsymbol{\mu}^{(q+1)})(\mathbf{X}_1 - \boldsymbol{\mu}^{(q+1)})'] + \Psi_h\mathsf{E}_h^{(q)}[(\mathbf{X}_1 - \boldsymbol{\mu}^{(q+1)})(\mathbf{X}_1 - \boldsymbol{\mu}^{(q+1)})'t_{1k}(\boldsymbol{\theta}^{(q)})], \tag{3.7}$$

l'adaptation des termes donnés par Cadez *et al.* (2002) dans le cas hybride de données partiellement classées donne

$$\pi_k^{(q+1)} = \frac{n_k^{(q)}}{n}, \quad \boldsymbol{\mu}_k^{(q+1)} = \frac{\sum_{h=1}^{K}\left(\psi_{kh}\mathsf{E}_h^{(q)}[\mathbf{X}_1] + \Psi_h\mathsf{E}_h^{(q)}[\mathbf{X}_1 t_{1k}(\boldsymbol{\theta}^{(q)})]\right)}{n_k^{(q)}} \quad \text{et} \quad \boldsymbol{\Sigma}_k^{(q+1)} = \frac{\sum_{h=1}^{H}\mathbf{W}_{kh}^{(q)}}{n_k^{(q)}}. \tag{3.8}$$

Cadez *et al.* (2002) ont proposé un certain nombre de méthodes numériques pour évaluer, dans le cas multivarié général, les intégrales présentes dans les espérances $\mathsf{E}_h^{(q)}$. Les modèles diagonaux, par contre, ne posent pas de problèmes numériques de ce genre. Notant $\text{diag}(\mathbf{W}_{kh}^{(q)})$ la matrice diagonale composée des termes diagonaux de $\mathbf{W}_{kh}^{(q)}$, les modèles diagonaux hétéroscédastiques et homoscédastiques donnent respectivement

$$\boldsymbol{\Sigma}_k^{(q+1)} = \frac{\sum_{h=1}^{H}\text{diag}(\mathbf{W}_{kh}^{(q)})}{n_k^{(q)}} \quad \text{et} \quad \boldsymbol{\Sigma}^{(q+1)} = \frac{\sum_{k=1}^{K}\sum_{h=1}^{H}\text{diag}(\mathbf{W}_{kh}^{(q)})}{n}. \tag{3.9}$$

Considérant en outre que les *bins* sont des hyper-rectangles de l'espace $\mathbb{R}^d$ (typiquement, la partition de $\mathbb{R}^d$ correspond à une grille), il est possible d'exprimer chacun des termes du mélange en utilisant la fonction de répartition de la loi $\mathsf{N}(0,1)$. Ces expressions ont été données dans l'annexe A de [R1] dans le cas où aucun label n'est connu. Ces résultats généralisent ceux obtenus par McLachlan et Jones (1988) [7] dans le cas univarié. L'extension au cas partiellement classé ne pose pas de problème particulier.

3.2 Influence de la dimension des *bins*

3.2.1 Dégénérescence dans les mélanges gaussiens

Nous nous plaçons ici dans le cas des modèles de mélanges gaussiens univariés et dont aucune donnée n'est labellée. La dégénérescence des mélanges gaussiens avec données individuelles à déjà été traitée dans la section 2.4 du chapitre 2. Deux types de résultats avaient alors été mis en évidence. D'une part, la fonction vraisemblance n'est pas bornée puisqu'elle tend vers l'infini lorsque l'un des centres gaussiens coïncide avec un individu tandis que la variance correspondante tend vers zéro. D'autre part, lorsqu'une itération de l'algorithme EM est « proche » d'une solution dégénérée, alors les itérés successifs convergent extrêmement rapidement vers cette solution. Ainsi, des difficultés aussi bien pratiques que théoriques de premier ordre sont liées à l'utilisation des mélanges gaussiens avec données individuelles. Il est donc crucial de savoir si ces problèmes sont transposés aux données groupées. C'est la problématique engagée dans cette section. Nous nous limiterons par ailleurs au cas le plus courant où les *bins* sont simplement des intervalles de $\mathbb{R}$.

Précisons tout d'abord qu'une solution dégénérée dans le cas groupé est définie, par analogie avec le cas individuel, par une gaussienne dont la variance est nulle et dont la moyenne est à l'intérieur d'un *bin*, sans autre précision sur sa position dans ce *bin*. Cette précision faite, on peut remarquer que la vraisemblance associée au cas groupé est bornée quelque soit la valeur du paramètre $\boldsymbol{\theta}$, solutions dégénérées incluses. En effet, le fait que les termes $a_h(\boldsymbol{\alpha}_k)$ de (3.3) soient des probabilités implique que $a_h(\boldsymbol{\alpha}_k) \leq 1$ ($h = 1,\ldots,H$, $k = 1,\ldots,K$), et donc $\ell(\boldsymbol{\theta};\mathcal{D}) \leq 0$. Malheureusement, cette propriété quelque peu rassurante n'est qu'une illusion pour une éventuelle vaccination contre la dégénérescence. Le théorème 1 de [R2] établit en effet que, si la dimension de tous les *bins* non vides est « suffisamment » petite, alors le maximum global de la vraisemblance est obtenu avec une solution dégénérée. Remarquons que si le *bin* est vide, il est établit par ailleurs que la solution dégénérée correspondante ne peut jamais être le maximum global et ce quelque soit la dimension des différents *bins*. Ces résultats indiquent donc que le phénomène de dégénérescence peut exister dans le cas groupé et que sa présence dépend de la dimension des *bins* non vides.

Cette propriété d'existence conditionnelle établie, il devient intéressant de connaître le comportement de l'algorithme EM au voisinage d'une solution dégénérée. Le théorème 2 de [R2] indique que le comportement d'une composante $k \in \{1,\ldots,K\}$ qui se trouve dans le voisinage d'une solution dégénérée à l'intérieur de l'intervalle non vide Ω_h ($h \in \{1,\ldots,H\}$) à l'itération q de EM dépend de la valeur du terme $\rho^{(q)}$ suivant

$$\rho^{(q)} = 1 - \frac{\Psi_{h'(q)}}{\Psi_h} R^{(q)}_{h'(q)} \qquad \text{avec} \qquad R^{(q)}_{h'(q)} = \frac{\pi_k^{(q)} + \sum_{l\neq k} \pi_l^{(q)} a_h(\boldsymbol{\alpha}_l^{(q)})}{\sum_{l\neq k} \pi_l^{(q)} a_{h'(q)}(\boldsymbol{\alpha}_l^{(q)})}, \tag{3.10}$$

où on a noté aussi $h^{'(q)}$ le numéro du *bin* à gauche ou a droite de Ω_h (celui dont un des bords est le plus proche de la moyenne de la composante k). Il est prouvé que le signe de $\rho^{(q)}$ régit le fait que EM s'approche ou non de la solution dégénérée considérée. Si $\rho^{(q)} > 0$, alors la gaussienne k à l'itération suivante $q+1$ se rapprochera encore de la dégénérescence. Par contre, si $\rho^{(q)} < 0$, cette gaussienne s'en éloignera. Le cas $\rho^{(q)} = 0$ est indéterminé dans l'étude. Notons qu'il n'est pas exclut que le signe de $\rho^{(q)}$ puisse changer au cours des itérations de EM et le fait que $\rho^{(q)} > 0$ à toutes les itérations n'est pas une condition suffisante pour dégénérer à coup sûr. En effet, cela indique seulement que la gaussienne est « piégée » autour de la dégénérescence mais sans garantie de convergence. Il est essentiel de remarquer que le signe de $\rho^{(q)}$ est partiellement contrôlé par le rapport $\Psi_{h'(q)}/\Psi_h$ qui correspond au rapport de l'effectif Ψ_h du *bin* Ω_h où la dégénérescence est étudiée et de l'effectif $\Psi_{h'(q)}$ du *bin* $\Omega_{h'(q)}$ qui est « le plus proche » du *bin* Ω_h. En effet, un faible ratio favorise $\rho^{(q)} > 0$ tandis qu'un grand ratio favorise les situations où $\rho^{(q)} < 0$. Un cas particulier important correspond à un *bin* $\Omega_{h'(q)}$ vide (i.e. $\Psi_{h'(q)} = 0$). On obtient alors $\rho^{(q)} = 1$ pour toute itération q et on peut ainsi établir avec certitude que EM convergera vers la situation dégénérée étudiée.

[R2] fait ressortir un autre résultat contrastant nettement avec la dégénérescence dans le cas des données individuelles. En effet, au voisinage d'une situation dégénérée, l'algorithme EM évolue *extrêmement* lentement puisqu'on montre que le taux global de convergence γ de EM (terme défini à l'équation (2.10) du chapitre 2) est égal à 1 en cet endroit. Comme conséquence, le praticien ne peut plus attendre patiemment que la précision numérique de l'ordinateur arrête EM. Dans le même esprit, il devient dangereux de tester directement la valeur de la variance $\sigma_k^{2(q)}$ de la composante k avec une valeur arbitrairement faible puisque, la lenteur rédhibitoire de la convergence ne laisse espérer atteindre une telle borne qu'en un temps immensément grand. Bien entendu, il est aussi exclu de tester directement l'évolution relative de la variance entre deux itérations de EM avec un seuil prédéfini car alors cet test confondrait la dégénérescence et la convergence. Néanmoins, il est important de pouvoir tout de même détecter une situation de dégénérescence pour les raisons suivantes. Tout d'abord, nous avons vu que la dégénérescence peut parfois conduire au maximum global de la vraisemblance et donc il indispensable d'identifier cet événement pour l'éliminer. Ensuite, l'arrêt de EM à l'approche d'un point dégénéré (cas où $\rho^{(q)} > 0$) permet d'économiser du temps de calcul en évitant de nombreuses itérations inutiles. Enfin, quand EM s'éloigne lentement d'une dégénérescence (cas $\rho^{(q)} < 0$), détecter cette situation empêche d'arrêter EM de façon illégitime en lui permettant de poursuivre son évolution vers d'autre solutions. Afin de répondre à ces besoins, les développements de [R2] permettent de proposer le critère suivant pour détecter la dégénérescence : on identifie une solution dégénérée de EM dès que

$$|\sigma_k^{2(q+1)} - \sigma_k^{2(q)} + \eta^{(q)}\rho^{(q)}u^{(q)}| < \epsilon \tag{3.11}$$

avec $\eta^{(q)} = |\lambda^{(q)} - \mu_k^{(q)}|$, $\lambda^{(q)}$ étant le bord de $\Omega_{h'^{(q)}}$ le plus proche de $\mu_k^{(q)}$, $u^{(q)} = \sigma_k^{2(q)} p(\lambda^{(q)}; \boldsymbol{\alpha}_k^{(q)})$ et ϵ une valeur arbitrairement petite, par exemple $\epsilon = 10^{-6}$. Bien entendu, le signe de $\rho^{(q)}$ permet d'identifier si on s'éloigne ou si on se rapproche de la solution de dégénérescence ainsi détectée. Notons aussi qu'il a été prouvé que EM ne peut pas dégénérer à l'intérieur d'un *bin* Ω_h vide.

Ces résultats ont été confirmé par des expériences numériques. Ces expériences ont montré par ailleurs que l'augmentation de la dimension des *bins* accroît les fréquences de dégénérescence. Comme situation limite à cette remarque, on a remarqué que les données groupées conduisent toujours à plus de dégénérescence que dans le cas des données individuelles. Ainsi, la règle d'arrêt (3.11) est d'une grande importance dans l'étude des mélanges gaussiens avec données groupées même si elle reste quelque peu arbitraire par la définition du seuil ϵ.

3.2.2 Comportement du critère BIC

Nous nous replaçons ici dans le cas multivarié avec données groupées. Cadez *et al.* (2002) [3] ont étudié expérimentalement l'influence de la dimension des *bins* sur la divergence de Kullback-Leibler entre l'estimateur du MV et la vraie valeur du paramètre. Il ressort que l'augmentation de la taille des *bins* décroît sans surprise la qualité de l'estimateur mais qu'il faut néanmoins des *bins* relativement larges pour constater des dégradations substantielles. Nous proposons d'étendre cette étude à l'effet de la dimension des *bins* sur des critères de sélection de modèles comme BIC (Schwarz 1978 [8]) qui est l'un des critères les plus populaires. Il est naturel de s'attendre aussi à une dégradation de la qualité du modèle sélectionné quand la taille des *bins* croît mais il s'agit de qualifier, et autant que possible de quantifier, cette dégradation.

Dans un premier temps, il est établit dans [R1] d'un point de vue théorique que le modèle sélectionné par le critère BIC avec des *bins* « suffisamment » petits est le même que le modèle sélectionné avec les données individuelles. C'est un résultat qui n'a bien entendu rien de surprenant mais qui aura son importance un peu plus loin. La partie intéressante de l'étude réside plutôt dans le cas des *bins* de plus grande dimension. La proposition 1 [R1] établit que, lorsque la taille des *bins* augmente, le critère BIC sélectionne des modèles de plus en plus simples en termes de nombre de paramètres. C'est un comportement qui peut apparaître raisonnable du point de vue d'un critère de parcimonie. En effet, l'augmentation de la taille des *bins* conduit finalement à une perte progressive de l'information présente initialement dans les données individuelles et il semble raisonnable de choisir un modèle de plus en plus simple quand l'information s'appauvrit. Comme, d'une part, des *bins* assez petits conduisent BIC à sélectionner le même modèle que dans le cas individuel et que, d'autre part, l'agrandissement des *bins* simplifie les modèles retenus par BIC, on déduit que BIC sélectionne généralement des modèles plus simples avec les données groupées qu'il ne l'aurait fait dans le cas individuel.

Des expériences numériques permettent de quantifier un peu le comportement de BIC [R1]. Il apparaît qu'il est nécessaire d'avoir des *bins* d'assez grande dimension pour noter des différences notables entre les modèles retenus par BIC dans le cas groupé et dans le cas individuel. Ce résultat corrobore donc celui de Cadez *et al.* (2002) [3] obtenu pour l'estimation des paramètres. Dans la pratique, il n'est pas rare d'avoir des données entachées d'une grande incertitude comme le prouve un exemple en astronomie. Dans cet exemple, nous montrons que le modèle retenu par BIC peut tout à fait être remis en cause si dans l'avenir la précision des instruments de mesure permet de réduire l'incertitude d'aujourd'hui. De façon plus générale, le fait de diminuer l'imprécision augmente la chance de choisir des modèles plus complexes.

En fait, l'ensemble de ces résultats sont valides dans des situations plus globales. D'une part, ils ne se limitent pas aux mélanges gaussiens car ils peuvent s'appliquer à tout type de modèle discret ou continu (modèle de mélange ou pas) du moment qu'il s'agit de données groupées. D'autre part, le comportement de BIC se transpose sans problème à des critères bâtis sur le même patron comme le critère AIC (Akaike 1974 [1]).

3.3 Bilan et perspectives

Nous avons vu qu'il était nécessaire de mener une étude spécifique aux données groupées pour les raisons que nous rappelons succinctement. Dans un premier temps, la possibilité de dégénérescence dans les mélanges gaussiens univariés a été mise en évidence bien que la vraisemblance soit bornée. Plus précisément, la possibilité que le maximum de vraisemblance corresponde à une solution dégénérée devient plus présente quand la dimension des *bins* non vides diminue. Du point de vue de l'algorithme EM, la dégénérescence dans le cas groupé est plus complexe à gérer que dans le cas individuel : il peut y avoir attraction ou répulsion dans un voisinage de la dégénérescence, la fréquence de dégénérescence est plus élevée (et elle augmente avec la dimension des *bins*), sa vitesse est extrêmement lente au point de devoir proposer une règle d'arrêt particulière. Cet antagonisme entre fréquence des solutions dégénérées dans EM (les *bins* doivent être grands) et la possibilité d'un maximum de vraisemblance associé à une solution dégénérée (les *bins* doivent être petits) est malheureusement difficile à quantifier. Dans un second temps, nous avons aussi remarqué l'influence de la dimension des *bins* sur le comportement du critère BIC. Il ressort typiquement que BIC appliqué aux données groupées sélectionne en général un modèle plus simple qu'avec les données individuelles (ou avec des données groupées de *bins* plus petits).

Les extensions possibles de l'étude dans le cadre de la dégénérescence des mélanges gaussiens concernent essentiellement l'extension au cas multivarié et l'analyse d'autres algorithmes comme Newton. Notons qu'en pratique le cas multivarié semble se comporter comme le cas unidimensionnel avec EM. Il s'agira donc en particulier de proposer une règle d'arrêt pertinente pour cet algorithme.

L'étude de l'influence de la dimension des *bins* sur le critère BIC pourrait par ailleurs être étendue à des critères de choix de modèles structurés différemment comme le critère de validation croisée utilisé en analyse discriminante ou encore les critères NEC (Celeux et Soromenho 1996 [5]) ou ICL (Biernacki *et al.* 2000 [2]) adaptés à la classification automatique. On pourrait aussi envisager une étude sur le comportement du test du rapport des vraisemblance maximales car des procédures variées de choix de modèles partent de ce principe général (voir par exemple McLachlan et Peel 2000 [6] chap. 6).

Enfin, puisque la question de l'initialisation de l'algorithme EM reste bien entendu d'actualité dans les mélanges avec données groupées, il serait pertinent d'adapter la méthode générique de tirage des paramètres initiaux sur les trajectoires présentée dans le chapitre 2. Des difficultés spécifiques au cas groupé seront probablement à traiter, en particulier celles liées à l'expression relativement complexe de l'étape M par rapport au cas individuel.

Bibliographie

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

[2] 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.

[3] I.V. Cadez, P. Smyth, G.J McLachlan, and C.E McLaren. Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. *Machine Learning*, 47(1):7–34, 2002.

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

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

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

[7] G.J. McLachlan and P.N. Jones. Finite mixture models to grouped and truncated data via the em algorithm. *Biometrics*, 44:571–578, 1988.

[8] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

Chapitre 4

Une extension de l'analyse discriminante : la discrimination généralisée

Sommaire

Publications associées à ce chapitre

Revues avec comité de lecture

[R1] C. Biernacki, F. Beninel and V. Bretagnolle (2002). *A Generalized Discriminant Rule when Training Population and Test Population Differ on their Descriptive Parameters.* Biometrics, **58**, 2, 387–397.

Conférences avec comité de lecture

[C2] C. Biernacki, F. Beninel and V. Bretagnolle (2000). *Model on a Population and Prediction on Another One: a Generalized Discriminant Rule*, COMPSTAT'2000, Utrecht, Pays-Bas, 21–25 août.

Conférences sans comité de lecture

[c3] C. Biernacki, F. Beninel and V. Bretagnolle (2000). *Modèle sur une population et prédiction sur une autre : fonction discriminante généralisée.* SFdS'2000, 32ièmes journées de statistiques, Fès, Maroc, 15–19 mai.

[c4] C. Biernacki, F. Beninel and V. Bretagnolle (1999). *Model on a Population and Prediction on Another One: Generalized Allocation Rule.* TMBM'99 (Theory and Mathematics in Biology and Medicine), Amsterdam, 29 juin–3 juillet.

Séminaires

[s5] C. Biernacki (2001). *Apprentissage sur une population et prédiction sur une autre : fonction discriminante généralisée.* Séminaire de l'IRDQ, Besançon, France, 22 janvier.

[s6] C. Biernacki, F. Beninel and V. Bretagnolle (2000). *Apprentissage sur une population et prédiction sur une autre : fonction discriminante généralisée.* Séminaire Agropolis, Montpellier, France, 6 novembre.

[s7] C. Biernacki, F. Beninel and V. Bretagnolle (2000). *Apprentissage sur une population et prédiction sur une autre : fonction discriminante généralisée.* Séminaire Écologie, Environnement et Méthodes Statistiques de l'IUT de Niort, Niort, France, 27 avril.

Ce chapitre propose une généralisation de l'analyse discriminante classique. L'évolution porte sur les hypothèses relatives aux deux échantillons mis en œuvre en discrimination. En effet, la population dont est issu l'échantillon d'apprentissage et celle dont est issu l'échantillon à classer sont en général supposées identiques tandis que nous proposons maintenant de relaxer cette condition. L'objectif est ainsi d'appliquer une règle de classement à un échantillon provenant d'une population différente de la population d'apprentissage. Le fait que les populations soient différentes ne signifie évidemment pas qu'elles soient sans liaison entre elles. En effet, il s'agit d'identifier l'information que l'on doit retenir de la population d'apprentissage pour l'appliquer à la nouvelle population à classer. L'enjeu majeur est alors certainement de trouver des liens entre les différentes populations. La portée de cette évolution semble vaste aussi bien d'un point de vue applicatif que d'un point de vue théorique, c'est pourquoi nous avons réservé un chapitre entier aux premiers pas dans cette direction.

La première partie de cette présentation donne un aperçu des évolutions successives de l'analyse discriminante afin de mettre en perspective les besoins en discrimination généralisée, ce qui conduit naturellement à la présentation d'une proposition assez empirique issue de la littérature destinée à mettre en œuvre ce principe. La seconde partie développe une proposition nouvelle et plus formalisée dans le contexte des mélanges gaussiens multivariés. Cette approche permet d'ailleurs de faire des liens avec l'approche empirique précédente. Une application en biologie permettant d'estimer le sexe d'oiseaux de mer à partir d'une règle apprise à partir d'une espèce identique mais différant soit de par son âge, soit de par sa provenance géographique des oiseaux précédents, a permis alors d'évaluer l'efficacité de la proposition dans une situation réelle. Un certain nombre d'extensions possibles à ce travail sont ensuite esquissées dans la dernière section.

Notons que [R1] correspond à la publication centrale de ce travail. L'ensemble des autres communications correspondent soit à des présentations lors de conférences [C2], [c∗] ou encore à des invitations à divers séminaires [s∗].

4.1 Évolutions de l'analyse discriminante

4.1.1 Aperçu des évolutions successives

Comme nous l'avons rappelé au premier chapitre, l'analyse discriminante procède traditionnellement de la façon suivante (Lachenbruch et Goldstein 1979 [15], Gnanaddesikan 1989 [13], McLachlan 1992 [16]) : un échantillon provient d'une population et une partition en plusieurs groupes de cet échantillon est connue. Chaque individu de la population est décrit par un ensemble de caractéristiques regroupées dans un vecteur appartenant à un certain espace $\mathcal{X}$. À partir de ces variables, une règle de classement est établie dans le but de classer tout autre élément de cette population dont le groupe serait inconnu.

Dans ce contexte, Fisher (1936) [10] a proposé une discrimination linéaire entre deux classes au moyen de la distance de Mahalanobis lorsque le vecteur de caractéristiques évolue dans l'espace euclidien, mais, depuis ce travail précurseur, de nombreuses extensions ont été proposées. Rao (1948) [17] a étendu l'approche de Fisher à plus de deux groupes (voir aussi Anderson 1958 [3]). Dans un cadre plus général le modèle de mélange gaussien multivarié permet d'obtenir des règles de classement quadratiques (voir par exemple Tomassone *et al.* 1988 [22]) et des modèles parcimonieux entre les possibilités linéaires et quadratiques sont concevables, comme les modèles introduits par Banfield et Raftery (1993) [4] et revisités par Celeux et Govaert (1995) [6]. Des méthodes non paramétriques ont aussi été explorées , utilisant des méthodes de noyaux (Silverman 1986 [19]), des techniques de k plus proches voisins (Fix et Hodges 1951 [11]) ou encore d'autres travaux de *projection pursuit* (Friedman et Stuetzle 1981 [12]). La discrimination logistique, à l'origine proposée par Anderson 1972 [2], est une autre direction qui peut être interprétée comme une classification semi-paramétrique. Une adaptation de certaines des méthodes précédentes ainsi que de nouvelles propositions ont été aussi considérées dans le cas de données qualitatives (Celeux et Nakache 1994 [7]). Un autre sujet important de recherche est la sélection de modèle, citons par exemple la méthode de validation croisée (voir Hand 1986 [14] pour un aperçu général) ou encore des critères de pénalisation du maximum de vraisemblance comme le critère AIC (Akaike 1974 [1]).

4.1.2 Une nouvelle direction à explorer

Une hypothèse sous-jacente à l'ensemble des procédures classiques de discrimination que nous venons de rappeler est que l'échantillon d'apprentissage soit représentatif de la population à classer, c'est-à-dire que la règle de classement estimée soit statistiquement peu différente de celle de la population. Il peut néanmoins être intéressant de considérer le problème d'analyse discriminante statistique lorsque l'échantillon d'apprentissage provient d'une population différente des éléments à classer. Une illustration concrète est fournie en biologie par des oiseaux adultes sur lesquels plusieurs variables biométriques sont mesurées. Le sexe de ces oiseaux est connu et le but est de déterminer celui de jeunes oiseaux impubères dont le sexe est impossible à obtenir sans procéder à une dissection. Cependant, dans certains cas, les oiseaux adultes n'ont pas les mêmes dimensions que les jeunes : dans certaines espèces d'oiseaux, la croissance se termine plusieurs années après l'envol (Bretagnolle *et al.* 1998 [5]). Ce genre de problème se pose aussi sur des différences ne portant par sur l'âge mais par exemple sur l'éloignement géographique : on connaît le sexe des oiseaux vivant sur une certaine île et on voudrait déterminer le sexe d'une espèce similaire mais généralement de taille différente et vivant sur une autre île.

Il est bien connu des biologistes que la plupart des variables biométriques sont distribuées suivant une loi gaussienne conditionnellement au sexe. En conséquence, les variables de chacune des deux populations (les jeunes et les vieux par exemple) sont distribuées selon un mélange de deux lois normales généralement multidimensionnelles, la distribution des variables de chaque sexe étant ainsi décrite par les paramètres suivants : une proportion, un centre et une matrice de variance. De plus, le mélange est souvent homoscédastique et donc la règle de classement entre mâle et femelle est linéaire. Notant $\mathbf{\Sigma}$ la matrice de variance commune pour la population d'apprentissage, π_1 et π_2 les proportions et $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$ les centres respectifs du groupe des femelles et des mâles de cette population, la règle optimale de classement correspondant au cas linéaire consiste classiquement à décider qu'un individu $\mathbf{x}_1$ est un mâle si

$$\mathbf{x}_1'\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > c \quad \text{avec} \quad c = 1/2(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \ln(\pi_2/\pi_1). \tag{4.1}$$

Dans le contexte d'espèces d'oiseaux de mer différant par leur habitat géographique, Van Franeker et Ter Brack (1993) [23] ont alors proposé une des premières techniques de discrimination généralisée. Ils ont simplement proposé de garder la même règle de classement (4.1) pour la population à classer en changeant uniquement le seuil c en un nouveau seuil c^*. En d'autres termes, ils décident de classer comme mâle un individu $\mathbf{x}_1^*$ de cette nouvelle population si

$$\mathbf{x}_1^{*\prime}\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > c^*. \tag{4.2}$$

Cette méthode est attrayante et simple à mettre en œuvre mais le nouveau seuil proposé est défini de façon relativement empirique dans leurs travaux.

La problématique d'apprentissage et de test sur des populations différentes a ainsi montré son intérêt dans une situation concrète. Néanmoins, les outils de résolution sont balbutiants et nécessitent des travaux de recherche adaptés, en particulier dans le cas général des mélanges gaussiens multivariés.

4.2 Discrimination généralisée dans le cas gaussien

Nous proposons, dans cette section, d'étendre l'analyse discriminante gaussienne multivariée à la situation où les deux populations d'apprentissage et de test sont différentes. Dans un premier temps, un certain nombre d'hypothèses sur les liaisons entre les deux populations sont établies, conduisant ainsi à une paramétrisation du lien. Ensuite, différents modèles de liaisons sont proposés afin de restreindre au besoin le nombre de paramètres à estimer présents dans le lien. Cela conduit à plusieurs modèles de liaison entre populations dont l'interprétation est assez aisée. Il s'agit alors de proposer un estimateur spécifique à chacune de ces situations. Enfin, un exemple réel de biologie a permis tout d'abord de vérifier les hypothèses avancées, puis de constater l'efficacité de la nouvelle règle de classement obtenue.

4.2.1 Les hypothèses

Dans le contexte de la discrimination gaussienne multivariée, la population d'apprentissage P (i.e. la population avec labels) et la population test P^* (i.e. la population à classer) sont toutes deux modélisées par des mélanges de lois normales de paramètres respectifs $\boldsymbol{\theta}$ et $\boldsymbol{\theta}^*$. En discrimination généralisée, ces deux paramètres peuvent être différents, mais nous supposerons tout de même que l'espace des données est identique entre populations, c'est-à-dire d'une part que le nombre de groupes K est identique et chaque groupe est de même nature et, d'autre part, que les variables mesurées dans chacune des populations sont de même nature dans le même espace $\mathbb{R}^d$. Nous disposons d'un échantillon $\mathcal{D} = (\mathbf{x},\mathbf{z}^l)$ de taille n provenant de façon i.i.d. de P. Éventuellement, certains labels peuvent donc être inconnus dans P. L'échantillon issu de façon i.i.d de P^* est de taille n^* et éventuellement certains labels peuvent être connus, notant ainsi cet échantillon $\mathcal{D}^* = (\mathbf{x}^*,\mathbf{z}^{l*})$. Fidèles à notre démarche de *plug-in*, l'objectif est naturellement d'estimer $\boldsymbol{\theta}^*$ afin de déduire une estimation de la règle r^* et donc une estimation de $\mathbf{z}^{u*}$.

L'approche proposée pour la discrimination généralisée consiste à établir une application $\boldsymbol{\phi}_k$ de $\mathbb{R}^d$ dans $\mathbb{R}^d$ liant en loi les vecteurs aléatoires du groupe k ($k = 1,\ldots,K$) des populations P et P^*. Ainsi

$$\mathbf{X}^*_{1|Z^*_{1k}=1} \sim \boldsymbol{\phi}_k(\mathbf{X}_{1|Z_{1k}=1}) = [\phi_{k1}(\mathbf{X}_{1|Z_{1k}=1}),\ldots,\phi_{kd}(\mathbf{X}_{1|Z_{1k}=1})]', \tag{4.3}$$

avec ϕ_{kj} une application de $\mathbb{R}^d$ dans $\mathbb{R}$ ($j = 1,\ldots,d$). Trois hypothèses essentielles sont alors faites à ce stade sur l'application $\boldsymbol{\phi}_k$. Tout d'abord, il est supposé que la $j$ème composante $\phi_{kj}(\mathbf{X}_{1|Z_{1k}=1})$ de $\boldsymbol{\phi}_k(\mathbf{X}_{1|Z_{1k}=1})$ ne dépend que de la $j$ème composante $X_{1j|Z_{1k}=1}$ de $\mathbf{X}_{1|Z_{1k}=1}$. Cela revient à admettre que ϕ_{kj} est une application de $\mathbb{R}$ dans $\mathbb{R}$ (par simplicité, la notation ϕ_{kj} est conservée), ce qui s'écrit aussi

$$\boldsymbol{\phi}_k(\mathbf{X}_{1|Z_{1k}=1}) = [\phi_{k1}(X_{11|Z_{1k}=1}),\ldots,\phi_{kd}(X_{1d|Z_{1k}=1})]'. \tag{4.4}$$

En second lieu, cette application ϕ_{kj} est supposée être de classe C^1. Ces deux premières hypothèses suffisent à établir que la fonction ϕ_{kj} est nécessairement affine (résultat provenant de De Meyer *et al.* 2000 [9]), ce qui conduit aux K relations suivantes

$$\mathbf{X}^*_{1|Z^*_{1k}=1} \sim \boldsymbol{\Delta}_k\mathbf{X}_{1|Z_{1k}=1} + \mathbf{b}_k, \tag{4.5}$$

avec $\boldsymbol{\Delta}_k$ une matrice diagonale de $\mathbb{R}^{d\times d}$ et $\mathbf{b}_k$ un vecteur de $\mathbb{R}^d$. Enfin, la dernière hypothèse restreint l'hypothèse affine au cas linéaire, c'est-à-dire que $\mathbf{b}_k = \mathbf{0}$. Ainsi, les deux populations sont liées en loi et conditionnellement aux groupes par la relation suivante : $\mathbf{X}^*_{1|Z^*_{1k}=1} \sim \boldsymbol{\Delta}_k\mathbf{X}_{1|Z_{1k}=1}$. Dans le cas où les matrices diagonales $\boldsymbol{\Delta}_1$ à $\boldsymbol{\Delta}_K$ sont connues, la linéarité du lien permet d'exprimer facilement les moyennes et les matrices de variances de P^* en fonction de celles de P puisque $\boldsymbol{\mu}^*_k = \boldsymbol{\Delta}_k\boldsymbol{\mu}_k$ et $\boldsymbol{\Sigma}^*_k = \boldsymbol{\Delta}_k\boldsymbol{\Sigma}_k\boldsymbol{\Delta}_k$.

4.2.2 Modèles de liaison entre populations

Afin d'expliciter certaines connaissances *a priori* qui seraient disponibles sur la liaison entre les deux populations P et P^*, des modèles de contraintes sur la matrice de transformation $\boldsymbol{\Delta}_k$ sont considérés. Ces modèles, au nombre de cinq et notés $\mathbf{m}_1$ à $\mathbf{m}_5$, sont définis comme suit :

($\mathbf{m}_1$) $\boldsymbol{\Delta}_k = \mathbf{I}_d$: les deux populations sont identiques ;
($\mathbf{m}_2$) $\boldsymbol{\Delta}_k = \alpha\mathbf{I}_d$: la transformation est indépendante du groupe et de la variable ;
($\mathbf{m}_3$) $\boldsymbol{\Delta}_k = \boldsymbol{\Delta}$: la transformation est uniquement indépendante du groupe ;
($\mathbf{m}_4$) $\boldsymbol{\Delta}_k = \alpha_k\mathbf{I}_d$: la transformation est uniquement indépendante de la variable ;
($\mathbf{m}_5$) $\boldsymbol{\Delta}_k$ est sans contrainte : il s'agit de la situation la plus générale.

Il est très important de noter que ces cinq modèles sont totalement différents de ceux proposés par Banfield et Raftery (1993) [4] ou encore Celeux et Govaert (1995) [6]. En effet, ces auteurs ont établi, comme nous l'avons vu à la section 1.1.5 du premier chapitre, des liens entre les matrices de variances des groupes à l'intérieur d'une unique population P. Par contre, les modèles présentés en discrimination généralisée correspondent à des contraintes sur les centres et les matrices de variances entre deux populations différentes P et P^*.

Certains modèles appellent des commentaires intéressants. Le modèle $\mathbf{m}_1$ correspond à l'analyse discriminante classique. Si la population P est homoscédastique, les modèles $\mathbf{m}_2$ et $\mathbf{m}_3$ préservent l'homoscédasticité dans la population P^* (si P est homoscédastique) et donc conservent la linéarité de la règle de classement. De plus, les modèles $\mathbf{m}_4$ et $\mathbf{m}_5$ peuvent transformer une règle linéaire de P en une règle quadratique pour P^*, ceci avec très peu de paramètres à estimer. Notons enfin que le modèle $\mathbf{m}_2$ permet de retrouver l'approche de Van Franeker et Ter Brack (1993) [23] présentée à la section 4.1.2. En effet, si P est composée de deux groupes homoscédastiques, on peut montrer que le modèle $\mathbf{m}_2$ conduit à la règle de classement (4.2) avec $c^* = \alpha(1/2(\boldsymbol{\mu_1} + \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) + \ln(\pi^*_2/\pi^*_1))$. Ce nouveau seuil est différent de celui proposé par Van Franeker et Ter Brack mais s'appuie cette fois sur une démarche plus rigoureuse.

Chacun des cinq modèle peut être couplé à une hypothèse sur les proportions $\pi_1^*,\ldots,\pi_K^*$ de P^*. En effet, soit on suppose qu'elles sont préservées entre les deux populations, soit on n'admet pas cette hypothèse et il s'agit de les estimer. Prenant ainsi en considération ces deux possibilités, cela conduit finalement à dix modèles différents par combinaison.

Nous avons retenu une procédure de *plug-in* pour estimer les matrices $\mathbf{\Delta}_1,\ldots,\mathbf{\Delta}_K$, c'est-à-dire que ces matrices de transformation seront exprimées en fonction du vrai paramètre $\boldsymbol{\theta}$ de la population P. Comme le paramètre de mélange $\boldsymbol{\theta}$ est généralement lui-même estimé à partir de $\mathcal{D}$ (on pourra se reporter aux chapitres précédents), il sera simplement remplacé par son estimateur dans les expressions de $\mathbf{\Delta}_1,\ldots,\mathbf{\Delta}_K$. Ainsi, des estimateurs convergents sont obtenus avec des calculs relativement simples.

L'estimation des matrices $\mathbf{\Delta}_k$ dépend du modèle utilisé. Pour les modèles $\mathbf{m}_2$ et $\mathbf{m}_3$, un estimateur des moindres carrés est disponible puisque la transformation est indépendante du groupe. L'avantage de cet estimateur est d'être obtenu par des calculs élémentaires. En effet, partant de la relation globale entre les espérances des deux populations $\mathsf{E}_{\boldsymbol{\theta}^*}\mathbf{X}_1^* = \mathbf{\Delta}\mathsf{E}_{\boldsymbol{\theta}}\mathbf{X}_1$, on déduit l'estimateur des moindres carrés suivant pour chacun des deux modèles concernés :

$$\text{pour } \mathbf{m}_2 : \hat{\alpha} = \frac{\bar{\mathbf{X}}^{*\prime}\bar{\mathbf{X}}}{\bar{\mathbf{X}}'\bar{\mathbf{X}}} \quad , \quad \text{pour } \mathbf{m}_3 : \{\hat{\mathbf{\Delta}}\}_{jj} = \frac{\{\bar{\mathbf{X}}^*\}_j}{\{\bar{\mathbf{X}}\}_j} \quad (j = 1,\ldots,d). \tag{4.6}$$

Cependant, l'estimation par moindres carrées n'est pas applicable si les proportions doivent être estimées pour $\mathbf{m}_2$ et $\mathbf{m}_3$. Elle n'est pas applicable non plus pour l'ensemble des autres modèles, que les proportions soient ou non estimées. Ainsi, le principe du maximum de vraisemblance à été mis en œuvre pour chacun des dix modèles (y compris $\mathbf{m}_2$ et $\mathbf{m}_3$ lorsque l'estimateur des moindres carrés est disponible). Dans chaque situation, une adaptation de l'étape M de l'algorithme EM (Dempster *et al.* 1977 [8]) a été proposée.

Les dix modèles étant maintenant disponibles, se pose le problème du choix de l'un d'entre eux. Comme les modèles semblent facilement compréhensibles par un praticien, celui-ci peut effectuer son choix grâce à sa propre connaissance du domaine étudié. À défaut, une sélection automatique du modèle peut s'opérer par un critère comme BIC (Schwarz 1978 [18]) qui pénalise le maximum de la vraisemblance par le nombre de paramètres à estimer. Notons que ce nombre de paramètres énumère uniquement les paramètres estimés dans la liaison entre P et P^*, i.e. $\mathbf{\Delta}_1,\ldots,\mathbf{\Delta}_K$ et éventuellement $\pi_1^*,\ldots,\pi_K^*$, et non les paramètres estimés dans P. En effet, comme nous l'avons précisé, l'estimation de la liaison se fait conditionnellement à la connaissance du paramètre de P.

4.2.3 Application en biologie

Un test de validité de notre procédure à été effectué dans un contexte biologique similaire à celui de Van Franeker et Ter Brack (1993) [23]. L'espèce d'oiseau de mer considérée dans notre exemple, l'espèce *Calanectris diomedea* (voir Thibault et Bretagnolle 1997 [21] pour un examen des caractéristiques biométriques de cette espèce), est présente en Méditerranée et dans l'Atlantique Nord, ces conditions océanographiques contrastées étant présumées à l'origine de l'existence de différences notables en taille, en couleur et en comportement entre ces différentes sous-espèces. Au total, un échantillon de trois sous-espèces différentes a été considéré et cinq variables morphologiques ont été mesurées sur chaque individu (Thibault *et al.* 1997 [20]). Chaque échantillon comporte une partie de mâles et de femelles et chaque individu est de sexe connu par dissection.

L'ensemble des combinaisons par paires d'échantillons a été considéré en faisant jouer tour à tour un rôle d'échantillon d'apprentissage et un rôle d'échantillon test à chacun des échantillons. Au total, six combinaisons différentes ont donc été obtenues. Dans un premier temps, les trois hypothèses de la section 4.2.1 ont été vérifiées avec succès. Ensuite, chacun des dix modèles ainsi qu'une méthode de classification automatique avec modèle homoscédastique (cela correspond au modèle optimal pour ces oiseaux) ont été appliqués à chacune des six combinaisons précédentes. Le critère de comparaison des méthodes à été l'erreur de classement dans la population test puisque, comme nous l'avons dit, l'ensemble des labels étaient connus en réalité. Il ressort que les techniques d'analyse discriminante classique et de classification automatique conduisent toutes deux à des erreurs de classement très élevées. Par contre, tous les modèles de transition entre populations améliorent notablement la qualité de l'affectation pourvu qu'ils prennent en compte une transformation des variables entre populations. La transformation semble par contre relativement indépendante du groupe. Ainsi, le modèle $\mathbf{m}_3$ (avec ou sans proportions à estimer) semble bien adapté à notre situation. Il est remarquable que même le modèle le plus complexe $\mathbf{m}_5$ (avec ou sans proportions à estimer également), conduit aussi à de très bons résultats car le nombre de paramètres dans ce modèle reste tout de même assez faible. Notons que le critère BIC a aussi montré son applicabilité dans ce contexte. Enfin, des situations avec un échantillon test partiellement labellé ont aussi été considérées avec succès.

4.3 Bilan et perspectives

Nous avons émis l'idée de considérer le principe de l'analyse discriminante lorsque les deux échantillons d'apprentissage et de test ne proviennent plus nécessairement d'une même population. L'analyse discriminante généralisée a été proposée comme nom de baptême de ce principe général. Une mise en œuvre spécifique au cas gaussien multivarié a été établie en proposant, conditionnellement aux groupes, une liaison linéaire en loi entre les deux populations. Ainsi, une règle d'allocation a été obtenue avec peu de paramètres à estimer. Ce nombre varie en fonction des hypothèses de contraintes représentées par différents modèles de liaison. Un exemple en biologie a permis de confirmer la pertinence de cette application de la discrimination généralisée, notamment en conduisant à de meilleurs classements que la discrimination standard ou encore que la classification automatique. Différentes extensions à ce travail peuvent être entrevues.

Tout d'abord, dans le contexte de la discrimination gaussienne généralisée, une méthode de maximisation de la vraisemblance par *plug-in* a été expérimentée. En effet, les paramètres de la population d'apprentissage ont été estimés par maximum de vraisemblance à partir du seul échantillon d'apprentissage dans un premier temps, puis ces estimateurs ont été introduits en l'état dans la vraisemblance servant à estimer les paramètres de liaison entre populations. Alternativement à cette procédure simple mettant en œuvre le principe du maximum de la vraisemblance, il est envisageable d'estimer les paramètres de la population d'apprentissage et de la transition entre populations au même moment en intégrant ces deux types de paramètres dans la même vraisemblance. Cette vraisemblance globale posera certainement des difficultés d'optimisation mais une comparaison de cette approche avec la méthode du *plug-in* présente un intérêt de premier ordre, la nouvelle approche présentant un caractère plus « optimal ».

Une conséquence logique de cette proposition est de conduire à la *classification généralisée*. Nous savons qu'en classification dans le cadre des modèles de mélange, une technique de base très employée consiste à estimer le paramètre de la population, puis à déduire soit une règle de classement, soit une partition, ceci en fonction de l'objectif. Cette estimation est naturellement d'autant plus précise que la taille de l'échantillon est grande. Comme il est parfois difficile d'augmenter la taille de l'échantillon provenant d'une population P d'intérêt, peut être est-il possible, dans certains cas, d'obtenir un échantillon complémentaire provenant d'une population P^* et dont un modèle de liaison existe avec la précédente population P. Ainsi, il semble raisonnable de penser que la taille réunie des deux échantillons considérés doit compenser avantageusement l'augmentation du nombre de paramètres à estimer (ceux de P et ceux de liaison entre P et P^*), en particulier si le modèle de liaison est parcimonieux. Par exemple, dans notre illustration en biologie, on pourra certainement déterminer le sexe de deux espèces d'oiseaux avec plus de précision en « simultané » qu'individuellement. Cette idée générale peut s'étendre sans difficulté à plus de deux populations ou encore à des échantillons partiellement labellés.

Par ailleurs, il serait bien entendu intéressant d'appliquer la discrimination généralisée à d'autres méthodes classiques de discrimination citées dans la section 4.1.1. En effet, au delà de l'hypothèse gaussienne pour laquelle une proposition a été développée, il peut être considéré des données qualitatives, de la discrimination non paramétrique ou encore de la discrimination logistique. Dans chaque nouvelle situation, le principal défit sera très certainement de faire apparaître une relation réaliste entre les différentes populations afin d'estimer plus efficacement la règle de la population sans labels. Nombre de ces évolutions seront vraisemblablement guidées par des besoins de praticiens de divers horizons comme cela a été le cas dans notre contexte avec la biologie. Par exemple, en marketing, on peut transposer le comportement de consommateurs d'un marché connu vers un marché prospectif afin d'estimer le potentiel de ce nouveau marché avec une étude spécifique très réduite.

Bibliographie

[1] H. Akaike. A new look at the 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] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley, New York, 1958.

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

[5] V. Bretagnolle, F. Genevois, and F. Mougeot. Intra- and intersexual function in the call of a non-passerine bird. *Behaviour*, 135:1161–1184, 1998.

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

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

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

[9] B. De Meyer, B. Roynette, P. Vallois, and M. Yor. On independent times and positions for Brownian motion. Technical Report 1, Les prépublications de l'Institut Élie Cartan, Institut Elie Cartan, Vandœuvre lès Nancy, France, 2000.

[10] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. Pt. II.

[11] E. Fix and J. L. Hodges. Discriminatory analysis - nonparametric discrimination: Consistency properties. Technical report, Report of the U.S.A.F. School of Aviation Medicine, Agrawala (1977), 1951.

[12] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

[13] R. Gnanaddesikan. Discriminant analysis and clustering, panel of experts. *Statistical Science*, 4(1):34–69, 1989.

[14] D. J. Hand. Recent advances in error-rate estimation. *Pattern Recognition letters*, 4:335–346, 1986.

[15] P. A. Lachenbruch and M. Goldstein. Discriminant analysis. *Biometrics*, 35:68–85, 1979.

[16] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York, 1992.

[17] C. R. Rao. The utilization of multiple measurements in problems of biological classification (with discussion). *Journal of the Royal Statistical Society, Series B*, 10:159–203, 1948.

[18] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[19] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

[20] J-. C. Thibault, V. Bretagnolle, and C. Rabouam. Cory's shearwater calonectris diomedea. *Birds of Western Paleartic Update*, 1:75–98, 1997.

[21] J-. C. Thibault and V. Bretagnolle. A mediterranean breeding population of cory's shearwater which shows behavioral and biometrical characters of the atlantic subspecies. *Ibis*, 140:523–528, 1997.

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

[23] J. A. Van Franeker and C. J. F. Ter Brack. A generalized discriminant for sexing fulmarine petrels from external measurements. *The Auk*, 110(3):492–502, 1993.

Chapitre 5

Applications et transferts

Sommaire

Publications associées à ce chapitre

Revues avec comité de lecture

[R1] C. Biernacki, G. Celeux, G. Govaert and F. Langrognet. *Model-Based Cluster and Discriminant Analysis with the* MIXMOD *Software.* Soumis à « Computational Statistics and Data Analysis »..

Conférences avec comité de lecture

[C2] A. Guérin-Dugué, C. Biernacki and J. Hérault (2001). *Statistical Modelling for Image Retrieval using a Biological Model of the Perceptive Colour Space.* ICIP'2001, IEEE International Conference on Image Processing, Thessaloniki, Greece, 7–10 octobre.

[C3] C. Biernacki and R. Mohr (1999). *Indexation et appariement d'images par modèle de mélange gaussien des couleurs.* GRETSI'99, 17ème colloque GRETSI sur le traitement du signal et des images, Vannes, 13–17 septembre.

[C4] R. Hammoud, R. Mohr and C. Biernacki (1999). *Robustification des signatures de couleurs par modélisation de leur variabilité intra-plan vidéo.* GRETSI'99, 17ème colloque GRETSI sur le traitement du signal et des images, Vannes, 13–17 septembre.

[C5] B. Villain, B. Vérité, C. Biernacki and G. Celeux (1999). *A Practical Approach of Expert Elicitation for Bayesian Reliability Analysis of Ageing.* ESREL'99 (European Safety and Reliability Conference), TUM Munich – Garching, Allemagne, 13–17 septembre.

Conférences sans comité de lecture

[c6] C. Biernacki, G. Celeux, G. Govaert, F. Langrognet, G. Noulin and Y. Vernaz (2003). *MIXMOD : un logiciel pour les modèles de mélange en classification et en analyse discriminante.* SFdS'2003, 35ièmes journées de statistiques, Lyon, France, 2–6 juin.

[c7] C. Biernacki (1999). *Le logiciel XEMgaus pour la classification et la discrimination gaussienne.* SFdS'99, 31ièmes journées de statistiques, Grenoble, 17–21 mai.

Rapports

[r8] M. Bricler et C. Biernacki (2001). *Projet d'élaboration d'une méthode aidant les collectivités locales et les établissements hospitaliers à réduire l'absentéisme* . Rapport final sur l'étude de faisabilité DEXIA-SOFAXIS/IRDQ.

[r9] C. Biernacki, G. Celeux, B. Villain et B. Vérité (1998). *Utilisation des opinions d'experts pour l'analyse de la dégradation des structures passives.* Rapport final de convention de recherche Inria-EDF.

Séminaires

[s10] C. Biernacki (2001). *Classification, discrimination et panorama des logiciels d'identification des mélanges.* Journée didactique IS2 sur les mélanges de lois de probabilité à l'Inria Rhône-Alpes, Grenoble, France, 9 mars.

Un certain nombre de travaux appliqués faisant appel à des méthodes de classification ou simplement d'estimation de densité, mais mettant généralement en œuvre les modèles de mélange gaussiens, ont été menés dans les années postérieures à la thèse. Ainsi, le logiciel MIXMOD, présenté à la section 5.1 ci-dessous, a été développé en collaboration avec d'autres collègues statisticiens ou informaticiens afin de permettre une large utilisation de certains résultats précédents sur la plupart des plateformes informatiques. Dans un contexte d'indexation d'images ou de vidéos, le second post-doctorat dans le laboratoire GRAVIR et le projet MOVI de l'INRIA Rhône-Alpes a été l'occasion d'introduire ce type d'outils statistiques dans un domaine traditionnellement tourné vers une approche plutôt informatique des problèmes (section 5.2). En particulier, cette période a permis de co-encadrer un étudiant en thèse au sein de l'équipe d'accueil. Des travaux se sont ensuite poursuivis après mon recrutement à Besançon avec d'une part l'équipe IS2 de l'INRIA et d'autre part l'équipe LIS de l'INPG. Enfin, des problématiques purement industrielles ont été abordées (voir la section 5.3). Ainsi, au travers d'un contrat avec l'électricien EDF qui s'est déroulé lors du premier post-doctorat, une étude particulière à été menée dans un cadre de fiabilité bayésienne afin d'aider à définir des lois *a priori* sur les paramètres d'un modèle de Weibull. Un logiciel spécifique a alors été développé pour mettre en œuvre la méthode proposée. Par ailleurs, le contrat avec la société de courtage en assurances SOFAXIS implantée à Bourges, filiale des assurances DEXIA, visant à aider à la réduction de l'absentéisme dans les collectivités locales à permis de trouver une voie de résolution originale au travers d'une classification des types d'absentéisme.

5.1 Le logiciel MIXMOD

5.1.1 Motivation

Comme nous l'avons vu au cours des chapitres précédents, les modèles de mélange finis sont un puissant outil pour l'estimation de densité, la classification automatique et l'analyse discriminante. MIXMOD (*MIXture MODelling*) [R1] [c6] est un logiciel générique qui prend en considération ces trois centres d'intérêt avec leurs évolutions les plus récentes dans un cadre multivarié et se propose de les mettre à disposition gratuitement pour un large public. Dans sa version actuelle, MIXMOD se résume aux modèles de mélange gaussiens multivariés, mais d'autres types de mélange de distributions sont planifiés. Ainsi, une version de MIXMOD incluant les mélanges de lois multinomiales (modèle des classes latentes) est actuellement en fin de développement.

Au cours de la thèse, un logiciel (XEM) a été écrit en MATLAB et en SPLUS. Il proposait déjà l'ensemble des quatorze modèles gaussiens, l'algorithme EM et certaines de ses variantes, ainsi que plusieurs critères de choix de modèles. Il a ensuite évolué vers une version plus modulaire, l'algorithme XEMgaus [c7], version écrite exclusivement en MATLAB et plus largement diffusée. Aujourd'hui, une nouvelle évolution, MIXMOD, a été conjointement développée par l'Inria Rhône-Alpes (projet IS2), le laboratoire de mathématiques de Besançon (Université de Franche-Comté - UMR CNRS 6623) et le laboratoire Heudiasyc de Compiègne (Université de Technologie de Compiègne - UMR CNRS 6599). C'est un logiciel orienté objet construit comme une collection de classes C++. Cela préserve totalement les possibilités, les performances, la précision et la faible consommation en mémoire de C, tout en offrant l'avantage de l'environnement de programmation du langage orienté objet C++. Une interface SCILAB ou MATLAB est disponible, mais pour les spécialistes, un mode expert donne accès à toutes les fonctionnalités du programme, permettant typiquement une mise en œuvre d'études par simulations. Ce mode expert se présente soit sous sa forme la plus simple (lancement direct du programme C++ à partir de fichiers d'entrée), soit sous une forme plus conviviale en tant que fonction intégrée MATLAB, SCILAB. De cette manière, MIXMOD se présente de façon transparente dans des environnements évolués et largement diffusés.

Ce logiciel est destiné aux scientifiques et aux ingénieurs qui ont besoin de résoudre des problèmes de classification ou de discrimination. Dans un contexte universitaire, il s'agit d'un outil d'enseignement pour des cours d'introduction ou avancés dans ces domaines. Dans l'industrie, MIXMOD est un instrument de choix pour la recherche, l'analyse et en général la reconnaissance statistique des formes.

MIXMOD est distribué gratuitement sur différentes plateformes (Linux, Unix, Windows) en « open source » via un site internet [1]. En conséquence, chacun peut le redistribuer et/ou le modifier en respectant les termes de « the GNU General Public License » comme publié par « the Free Software Foundation »

1. MIXMOD est disponible avec sa documentation : `http://www-math.univ-fcomte.fr/MIXMOD/index.htm`.

soit pour la version 2 de la licence, soit (au choix) pour toute version ultérieure. La documentation est constituée d'un guide de l'utilisateur, d'une documentation statistique et d'un manuel de référence sur la programmation elle-même. Les supports fournis sont disponibles sous forme de fichiers postscript ou pdf (*acrobat reader*).

5.1.2 Aperçu des logiciels concurrents

D'autres logiciels permettent d'analyser des données multivariées à l'aide de modèles comme les mélanges gaussiens par exemple mais pas seulement. Il s'agit en particulier des logiciels suivants dont la liste n'est pas exhaustive (pour une description plus détaillée ainsi que d'autres logiciels, se reporter à [s10] et McLachlan et Peel 2000 [23] pp. 343–348).

EMMIX

- auteurs : McLachlan *et al.* (1999) [24]
- disponibilité : `http://www.maths.uq.edu.au/~gjm/emmix/emmix.html`
- langage/systèmes : fortran/unix, windows
- caractéristiques : EM et SEM, certains modèles gaussiens multivariés, modèles de Student multivariés, critères de choix de modèles (BIC, AIC, AWE, P-value, bootstrap paramétrique), estimation de l'écart-type des estimateurs par bootstrap

MCLUST

- auteurs : Fraley et Raftery (1999) [16]
- disponibilité : `http://www.stat.washington.edu/fraley/software.html`
- langage/systèmes : fortran interfacé avec le logiciel SPLUS (ou R)/unix, windows
- caractéristiques : EM, certains modèles gaussiens multivariés avec possibilité d'ajouter un bruit poissonnien, classification hiérarchique avec certains modèles gaussiens multivariés, critère BIC

AUTOCLASS

- auteurs : Cheeseman et Stutz (1996) [11]
- disponibilité : `http://ic-www.arc.nasa.gov/ic/projects/bayes-group/autoclass/`
- langage/systèmes : C/unix, windows
- caractéristiques : approche bayésienne, certains modèles gaussiens, de Bernoulli ou de Poisson avec possibilité de combiner certains d'entre eux, critère bayésien de choix du nombre de groupes, gestion des valeurs manquantes

SNOB

- auteurs : Wallace et Dowe (1994) [37]
- disponibilité : `http://www.csse.monash.edu.au/~dld/Snob.html`
- langage/systèmes : FORTRAN 77
- caractéristiques : mélanges de différentes natures (classes latentes, gaussiennes diagonales, Poisson, von Mises), gestion des données manquantes et choix du nombre de groupes

NORMIX

- auteurs : Wolfe (1967) [38]
- disponibilité : `http://www.alumni.caltech.edu/~wolfe/normix.htm`
- langage/systèmes : FORTRAN 77/windows
- caractéristiques : mélange de gaussiennes et de Bernoulli

5.1.3 Spécificités scientifiques de MIXMOD

Comme de nombreux logiciels déjà disponibles, différents algorithmes d'estimation des paramètres du mélange sont proposés, principalement EM (Dempster *et al.* 1977 [14]), CEM (Celeux et Govaert 1992 [9]) et SEM (Celeux et Diebolt 1985 [8]). Par ailleurs, des fonctionnalités spécifiques sont disponibles. Ainsi, il est possible d'enchaîner ces différents algorithmes pour obtenir des stratégies susceptibles de fournir un optimum pertinent de la vraisemblance, comme cela a été décrit à la section 2.2.3 du chapitre 2 (voir aussi Biernacki *et al.* 2003 [4]). De plus, plusieurs critères d'information pour choisir un modèle parcimonieux (le nombre de composants du mélange notamment) sont également inclus comme BIC (Schwarz 1978 [33]), certains d'entre eux favorisant le parti pris de la classification comme le critère NEC (Celeux et Soromenho 1996 [17]) ou encore ICL (Biernacki *et al.* 1999 [3]), ou bien le parti pris de la discrimination comme le critère de validation croisée. Dans le cadre des mélanges gaussiens multivariés, les quatorze modèles gaussiens de Celeux et Govaert (1995) [10] sont disponibles, permettant ainsi de prendre en considération différentes hypothèses sur les matrices de variance au travers de leur décomposition spectrale. Il est aussi possible que la partition soit connue en totalité (cas de l'analyse discriminante classique) ou bien partiellement connue (situation valable aussi bien classification automatique qu'en discrimination).

5.2 Modèles de mélange et indexation d'images

5.2.1 Problématique de l'indexation

L'utilisation croissante de l'imagerie numérique au travers de divers supports, dont le web, a mis en évidence la nécessité de méthodes automatiques d'indexation offrant une meilleure accessibilité en recherche d'images. Les enjeux directement liés à la problématique de l'indexation sont de deux ordres. Il s'agit tout d'abord d'extraire une signature, c'est-à-dire une information discriminante, de l'ensemble des images de la base numérique. Ensuite, lors de la présentation d'une image requête, la procédure d'indexation doit être capable de ressortir de la base d'images l'ensemble des individus relatifs à cette requête. Outre le besoin de définition du caractère de ressemblance entre images, se pose des difficultés bien connues en imagerie et liées à une grande variabilité comme les effets d'éclairage et les prises de vues.

Depuis l'étude initiale de Swain et Ballard (1991) [36], l'information de couleur est devenue une des principales signature utilisée pour la segmentation d'images et pour l'indexation (Healey et Luong 1998 [21]). En général, la distribution des couleurs est modélisée par l'histogramme empirique (normalisé ou non selon les cas) calculé à partir de l'image. L'avantage de cette technique est sa simplicité et sa rapidité, mais elle souffre d'une sensibilité à la taille des cases utilisée. Par ailleurs, cette approche est employée pour décrire des caractéristiques globales de l'image car il y a perte de l'information spatiale. La phase d'appariement entre images est ensuite généralement effectuée par la distance du χ^2, la divergence de Jeffrey ou encore l'intersection d'histogrammes (Puzicha *et al.* 1999 [29]).

Notons que, au delà de la distribution des couleurs, se pose aussi le problème du choix de la caractérisation des couleurs (Wyszecki et Stiles 1982 [39]). Les principaux modèles de couleurs sont alors l'espace « classique » RGB, l'espace linéaire YCbCr bien adapté aux images en compression JPEG, l'espace LAB proche de la perception de la rétine des hommes et des primates et l'espace LAcBc qui, non seulement est proche de la perception, mais encore prend en compte des non-linéarités dans les photorécepteurs (Alleysson et Hérault 1999 [1]).

5.2.2 Propositions en indexation d'images

L'idée essentielle de cette section est de modéliser la distribution d'une caractéristique fondamentale $\mathbf{x}$ de l'image (comme la couleur) par un mélange de gaussiennes éventuellement multivariées. Ainsi, la signature de l'image correspond au paramètre estimé du mélange, donc $\hat{\boldsymbol{\theta}}$. L'appariement entre l'image requête de caractéristique $\mathbf{x}^R$ et les B images indexées dans la base par les paramètres $\hat{\boldsymbol{\theta}}^1,\ldots,\hat{\boldsymbol{\theta}}^B$ se fait ensuite en classant chacune de ces images par valeur décroissante de la log-vraisemblance $\ell(\hat{\boldsymbol{\theta}}^b;\mathbf{x}^R)$ de leur signature ($b=1,\ldots,B$). Comparée à la méthode des histogrammes, cette technique de modélisation permet de réduire le nombre de paramètres, à un effet de régularisation pour des données bruitées et la flexibilité offerte permet des stratégies plus élaborées. Il ressort néanmoins que les mélanges gaussiens sont considérablement plus gourmands en temps d'indexation que les histogrammes, la phase d'appariement quant à elle étant généralement du même ordre entre les deux méthodes. Cependant, la rapidité de la phase d'appariement est un point plus essentiel lors d'une requête d'images que celle associée à la phase d'indexation, cette dernière pouvant en effet généralement être déroulée en tâche de fond.

Considérant des caractéristiques combinant les coordonnées spatiales, la texture et la couleur de l'image, Carson *et al.* (1998) [7] ont proposé une modélisation par des mélanges gaussiens en dimension huit. Dans cette approche, toutes les caractéristiques sont utilisées au même niveau pour la segmentation en régions homogènes. L'intégration de l'information spatiale dans ce vecteur de caractéristiques produit des erreurs de classification dans des grandes régions de l'image qui sont alors arbitrairement découpées en régions plus petites. Pour répondre à cette difficulté, Greenspan *et al.* (2000) [20] ont préféré intégrer l'information spatiale après une segmentation d'image en des régions de couleur homogène. La distribution spatio-chromatique de chaque zone segmentée est ensuite modélisée par une unique gaussienne en dimension cinq intégrant les couleurs et une information spatiale. Utilisant une segmentation déjà établie, [C3] proposa d'établir la signature gaussienne à partir de différentes caractéristiques combinées de couleurs et de gradient (norme, orientation). L'usage de critères de choix du nombre de composantes comme ICL (Biernacki *et al.* 2000 [3]) a montré son intérêt pour retenir un nombre parcimonieux de composantes gaussiennes.

Faisant suite à ces premières approches, [C2] proposa certaines modifications, en particulier sur la méthode d'implantation, sur l'utilisation de l'espace de couleurs LAcBc sans l'information de luminance et sur les contraintes concernant la matrice de variance pour l'information spatio-chromatique. Ainsi, une stratégie d'indexation en trois étapes à été retenue : (1) la distribution des couleurs est modélisée par un mélange gaussien bivarié (deux informations de couleur) estimé par l'algorithme EM et dont le nombre de composantes est donné par le critère ICL, (2) une classification spatiale classique en imagerie est effectuée afin de transformer ces composantes chromatiques en régions spatiales homogènes et (3) chaque région ainsi obtenue est modélisée par un mélange de gaussiennes en dimension quatre (deux couleurs et deux informations géographiques) dont le nombre de composantes est de nouveau déterminé par ICL et dont la corrélation entre information spatiale et de couleur est supposée nulle. La signature de l'image correspond donc à l'ensemble des paramètres de mélange de toutes les zones identifiées. Mis en œuvre sur une base d'images de grande taille constituée de thèmes variés, l'appariement obtenu avec cette nouvelle méthode d'indexation a montré une nette domination en terme de pertinence des images renvoyées par la requête par rapport à la méthode des histogrammes utilisant différentes stratégies d'appariement (divergence de Jeffrey, intersection, méthode du χ^2).

5.2.3 Propositions en indexation de vidéo

L'évolution rapide de la vidéo numérique comme un moyen informatif riche et largement intégré dans de nombreuses applications a mis en lumière l'intérêt de méthodes automatisées de structuration et d'indexation afin de permettre une meilleure accessibilité en navigation et en recherche (Smoliar *et al.* 1995 [34]). L'élaboration de résumés vidéo, en particulier de résumés des objets d'intérêt désignés par l'utilisateur comme les acteurs, représente une des applications récentes et très demandées dans le domaine de la vidéo (Benayoun *et al.* 1998 [2]). Ainsi, l'utilisateur a la possibilité de ne sélectionner dans un film que les scènes d'un acteur particulier par exemple. Pour réaliser ce genre d'application qualifiée d'*hypervidéo*, des techniques de reconnaissance d'objets et de classification sont nécessaires. Cependant, l'indexation des objets vidéo semble être une tâche délicate principalement pour deux raisons. D'une part, l'apparence intra et inter-plan de l'objet d'intérêt (un plan est une suite d'images) est soumise à de nombreux facteurs de variabilité tels que les prises de vue, les effets d'éclairage, les mouvements de la caméra, les rotations et les occultations partielles. Ainsi, différentes techniques d'appariement (Murase et Nayar 1995 [27], Rowe *et al.* 1994 [32]) ont montré leurs limites. D'autre part, le nombre d'objets à indexer dans une vidéo conduit à une base d'index dont la taille énorme semble être difficile à manipuler même avec l'aide de techniques d'indexations avancées (Blott et Weber 1997 [5]).

Partant d'une vidéo au préalable segmentée en plans et on objets mobiles (Bouthemy et Ganansia 1996 [6], Gelgon et Bouthemy 1997 [18]), l'idée de [C4] est de modéliser la variabilité de la signature temporelle d'un objet dans un plan vidéo par une distribution gaussienne multivariée de ce descripteur. L'ensemble des objets est ainsi modélisé par un mélange gaussien et correspond donc à la base d'index de la vidéo considérée. La dimension de cet index est assez réduite et le classement de nouveaux objets de la vidéo se fait simplement par l'utilisation de la discrimination gaussienne associée à ce mélange. Alternativement, les approches concurrentes répertoriées consistent généralement soit à représenter le plan vidéo par la signature d'une seule image du plan comme celle de l'image médiane (O'Connor 1991 [28]), soit à moyenner toutes les signatures de l'objet dans le plan. L'inconvénient de ces deux techniques est ne pas prendre suffisamment en considération la variabilité de l'objet dans le plan, cette variabilité étant par contre naturellement introduite par la matrice de variance dans l'approche par mélange gaussien. Associées à ces techniques assez rudimentaires d'indexation, les méthodes de classement de nouveaux objets se faisait très simplement, par exemple par la distance du χ^2 dans le cas de signatures par histogramme. Comparant ces trois techniques dans le cadre d'une signature par histogramme des couleurs (Swain et Ballard 1991 [36]), il ressort que le nombre d'objects bien classés est supérieur avec la méthode des mélanges gaussiens. Notons qu'un modèle gaussien particulier avec volumes égaux entre composantes a été retenu dans l'experimentation, toutes les autres caractéristiques étant libres par ailleurs entre composantes différentes.

5.3 Applications industrielles

5.3.1 Élicitation de jugements d'experts en fiabilité bayésienne

Ce contrat industriel, mené avec la compagnie EDF dans le cadre d'un post-doctorat avec l'INRIA, est associé à une problématique d'estimation des paramètres d'une loi de fiabilité de Weibull dans un contexte bayésien. Ce travail est donc sans relation avec les mélanges de lois.

Dans le cadre de la fiabilité des structures, l'objectif d'EDF est d'optimiser le programme de maintenance préventive ou les campagnes d'inspection en service sur les composants passifs non réparables des centrales nucléaires françaises. Considérant un modèle de Weibull de vieillissement de tels composants (Clarotti *et al.* 1996 [12]), la fiabilité à l'instant x_1 est donnée par

$$R(x_1|\alpha,\beta) = \exp\left\{-\left(\frac{x_1}{\alpha}\right)^{\beta}\right\}, \tag{5.1}$$

avec α le paramètre d'échelle et β le paramètre de forme. Cependant, les temps de défaillance observés $\mathbf{x} = (x_1, \ldots, x_n)$ et supposés être des réalisations i.i.d. de cette loi sont généralement en nombre insuffisant pour estimer convenablement les paramètres α et β en statistique classique. Afin de compenser ce manque d'information « objective », l'utilisation de données « subjectives » obtenues à partir d'experts est alors envisagée. Il s'agit de définir une loi *a priori* $p(\alpha,\beta)$ sur les paramètres (α,β) pour, grâce au théorème de Bayes, en déduire alors la loi *a posteriori* $p(\alpha,\beta|\mathbf{x})$. Cette loi peut ensuite être employée pour donner des estimations ponctuelles des paramètres, des zones de confiance ou encore une estimation de la fiabilité prédictive $R(t|\mathbf{x})$ au temps de mission t en intégrant le produit $R(t|\alpha,\beta)p(\alpha,\beta|\mathbf{x})$ sur l'espace des paramètres. Néanmoins, afin de faciliter la définition de la loi *a priori* sur (α,β), une reparamétrisation de la fiabilité avec deux paramètres (γ,β) est effectuée, $\gamma = R(t|\alpha,\beta)$ correspondant à la fiabilité au temps de mission. Le logiciel SURVBAYES (Souchois *et al.* 1998 [35]) permet de mettre en œuvre l'estimation bayésienne dans le cas d'une loi *a priori* particulière fondée sur un histogramme de crédibilités. Posant β_0 $(1 < \beta_0 < \infty)$ la limite entre un vieillissement acceptable $(\beta \leq \beta_0)$ et inacceptable $(\beta > \beta_0)$ et γ_0 $(0 < \gamma_0 < 1)$ la limite entre une fiabilité acceptable $(\gamma_0 < \gamma)$ et inacceptable $(\gamma_0 \geq \gamma)$, une loi de probabilité généralisée est définie en fixant des crédibilités $\mathbf{c} = (c_1,c_2,c_3,c_4)$ associées à chacun des quatre quadrants définis par combinaisons des limites de β et de γ.

Ainsi posé, le problème du contrat de recherche est alors de permettre une traduction quantitative des opinions d'experts sous forme des hyperparamètres β_0, γ_0 et $\mathbf{c}$ de la loi *a priori* $p(\beta,\gamma|\beta_0,\gamma_0,\mathbf{c})$ definie ci-dessus. Bien qu'appliqué ici dans une situation particulière, le processus de définition d'une loi *a priori* et de ses hyperparamètres est une difficulté généralement présente dans l'estimation bayésienne. L'élicitation des opinions d'experts s'appuie certes sur une information subjective mais le processus d'élicitation en lui-même doit s'appuyer sur une approche rigoureuse afin d'être crédible et efficace. Comme pour toute recherche scientifique, cette approche doit respecter cinq principes de base (Cooke 1991 [13]) qui sont la reproductibilité, la responsabilité, la vérification empirique, la neutralité et l'équité. Ainsi, dans le cadre d'un système médical, Robert (1991) [30] a montré que la performance du diagnostic médical s'améliorait avec la rigueur de la stratégie d'élicitation. Le choix de l'expert ainsi que la combinaison de différentes opinions d'experts ont aussi été identifiés comme des éléments importants.

Dans la lignée de ces principes fondamentaux, un certain nombre de règles propres aux méthodes d'élicitation des avis d'experts ont été identifiées. Meyer et Booker (1991) [25] ont ainsi mis en évidence deux principaux types de biais qui doivent être pris en considération dans le processus d'élicitation. Il s'agit du biais de motivation, indiquant que la réponse de l'expert a été altérée par le procédé d'élicitation (pression de groupe ou mauvaise compréhension par exemple), ou encore du biais de connaissance, indiquant que le format de réponse attendu n'est pas familier à l'expert (les probabilités par exemple). Afin de fournir un guide rigoureux d'élicitation permettant de réduire ces deux types de biais, Meyer et Booker (1991) [25] ont alors proposé six étapes fondamentales à respecter dans le processus d'élicitation :

1. **Objectif :** Nécessité de bien définir l'objectif du projet.
2. **Questions :** Réduction du biais de motivation par le choix de questions simples, précises et éventuellement commentées.
3. **Experts :** Le choix des experts se fait par leur pairs. Ils doivent être motivés, disponibles et habitués au type du format de la réponse attendue afin de réduire le biais de connaissance. Le concours de plusieurs experts permet aussi de réduire les deux types de biais.
4. **Entretien :** Le choix d'un entretien individuel plutôt que collectif (comme la méthode DELPHI) prend plus de temps mais permet de réduire le biais de motivation. Il devient aussi plus facile de combiner les avis d'experts.
5. **Combinaison des avis :** Si la combinaison de l'avis de différents experts est effectué objectivement, cela contribue à réduire les deux types de biais (Moslesh *et al.* 1988 [26]). Si les opinions se présentent sous forme de distributions de probabilité, Lind et Nowak (1988) [22] présentent quatre caractéristiques importantes, mais impossibles à satisfaire ensemble, qui sont la préservation de la symétrie, la continuité, la préservation de la forme et la convergence bayésienne. Genest et Zidek (1986) [19] recommandent les moyennes arithmétiques ou géométriques pondérées. La combinaison géométrique présente l'avantage de trois des caractéristiques précédentes et évite la multimodalité qui est généralement un frein à la prise de décision. Par ailleurs, les poids doivent être égaux afin de ne pas favoriser un expert.
6. **Documentation :** L'objectif de la phase de documentation est de répondre aux cinq fondements scientifiques de Cooke et de rendre ainsi le processus transparent et crédible.

La stratégie d'élicitation proposée dans [C5] [r9] pour choisir les paramètres $(\beta_0,\gamma_0,\mathbf{c})$ s'est efforcée de mettre en pratique l'ensemble de ces principes généraux. Dans un premier temps, afin de réduire les biais, les paramètres de forme β_0 et de fiabilité au temps de mission γ_0 ont été reformulés afin de devenir plus naturels pour les experts, les crédibilités restant inchangées. Par exemple, le paramètre β_0, très abstrait, a été transformé en un terme 2^{β_0-1} qui représente plus concrètement le vieillissement comme un facteur d'augmentation du nombre de défaillances entre la moitié du temps de mission et le temps de mission complet. Par ailleurs, les experts ont été scindés en deux catégories : les experts de *conception* définissent l'acceptabilité du vieillissement et de la fiabilité (donc β_0 et γ_0) tandis que les experts de *maintenance* indiquent la position de leur équipement dans ces zones d'acceptabilité (donc $\mathbf{c}$). En complément aux valeurs ponctuelles sur les paramètres $(\beta_0,\gamma_0,\mathbf{c})$, une notion attachée à leur incertitude sur la réponse a été prise en compte afin de traduire la connaissance d'un expert en une loi de probabilité $p(\beta_0,\gamma_0,\mathbf{c})$. Cette distribution sur les hyperparamètres s'apparente à une approche bayésienne hiérarchique (Robert 1994 [31]). Dans cette optique, une distribution Bêta, sélectionnée pour sa grande flexibilité, a été retenue pour chaque paramètre scalaire β_0, γ_0, c_1, c_2, c_3 et c_4 (il a été nécessaire de borner supérieurement β_0) et, de plus, il est supposé l'indépendance entre chacun d'entre eux. Les paramètres des lois Bêta sont simplement obtenues en demandant aux experts une valeur médiane sur chaque paramètre, puis un choix entre plusieurs incertitudes prédéfinies. La combinaison des avis sur un paramètre se fait alors par moyenne géométrique avec poids égaux, c'est-à-dire que, notant $p_1(\beta_0),\ldots,p_w(\beta_0)$ les lois lois obtenues par les w experts, on obtient comme loi finale sur β_0 :

$$p(\beta_0) \propto \prod_{i=1}^{w} [p_i(\beta_0)]^{1/w}. \tag{5.2}$$

La distribution résultante est simplement une loi Bêta dont les paramètres correspondent à la moyenne arithmétiques des paramètres des lois Bêta combinées. Ainsi, logiquement, un expert qui a plus de confiance dans son avis aura plus d'influence que d'autres dans la loi résultante.

Bien entendu, le logiciel SURVBAYES utilisé à la suite de ce processus d'élicitation requiert uniquement un triplet $(\beta_0,\gamma_0,\mathbf{c})$. Il est aisé, à partir des lois $p(\beta_0)$, $p(\gamma_0)$, $p(c_1)$, $p(c_2)$, $p(c_3)$ et $p(c_4)$, de donner des valeurs ponctuelles de chacun des paramètres comme une valeur moyenne, une valeur médiane ou encore des valeurs dépendant des fractiles et traduisant des valeurs qualifiées soit de *risquées*, soit de *conservatives*. Enfin, afin de faciliter la mise en pratique de la stratégie d'élicitation proposée, le logiciel ELICITAE écrit en MATLAB à été fourni. Ainsi, des fonctions graphiques permettent de guider l'expert à chacune des étapes du processus. Une application menée par EDF a alors montré l'intérêt de l'approche et du logiciel associé.

5.3.2 Réduction de l'absentéisme dans les collectivités locales

La société de courtage en assurances SOFAXIS, filiale de l'assureur DEXIA, assure le risque d'absence des salariés des établissements hospitaliers et des collectivités locales (mairies par exemple) et assimilées. SOFAXIS se distingue de ses concurrents par le type de prestation d'assurance qui comprend un volet d'accompagnement des établissements pour les aider à prévenir et à abaisser l'absentéisme. Afin de maintenir cet avantage concurrentiel stratégique, il convient pour SOFAXIS de réaliser un progrès significatif sur la prestation d'accompagnement des clients à travers la conception d'une méthode intégrée comportant des actions de diagnostic, de surveillance, de résolution de problème, de formation et de conseil aux établissements. Dans ce contexte, l'IRDQ (Institut de Recherche et de Développement de la Qualité), association loi 1901 implantée à Besançon et à Paris pour le conseil dans le domaine du contrôle de qualité, a été contactée afin de mener une étude de faisabilité pour la conception d'une méthode visant à abaisser l'absentéisme. Pour engager cette recherche, SOFAXIS a mis à la disposition de l'IRDQ trois types d'informations et d'expériences :

1. une macro-typologie *a priori* des processus qui conduisent à une absence ;
2. une base statistique comportant l'historique des absences de 23000 clients ;
3. une équipe de commerciaux connaissant le sujet et un service de prévention qui agit déjà avec un certain nombre de méthodes et d'outils.

C'est ainsi que l'IRDQ a contacté le groupe de statistiques de Besançon pour le traitement associé à la base de données.

L'analyse des données d'absence devait permettre de faire ressortir des types d'absentéisme afin de guider la recherche des mécanismes y conduisant. Comme seule la phase d'analyse de faisabilité était engagée, l'étude s'est restreinte aux collectivités locales pour les arrêts maladies (i.e. hors accidents du travail, arrêt longue maladie et congé maternité). La base de données était constituée de la liste des agents (7082 agents pour le cas présent) ayant été absents l'année en cours. Deux types d'information bien distincts étaient alors présents :

1. Des données qualitatives générales : la région, l'effectif de la collectivité, le sexe de l'agent, la catégorie hiérarchique, l'activité de la collectivité, l'âge de l'agent, son type de travail.
2. Des données numériques d'absence : le nombre de jours indemnisés, le nombre de jours d'arrêt, le montant remboursé et le nombre d'arrêts.

Notons bien que la raison de l'arrêt était absente de la base pour cause de secret médical et que, quoiqu'il en soit, la véritable cause profonde serait peut être même inconnue du médecin (par exemple, problèmes avec la hiérarchie ou familiaux conduisant à une dépression et donc une absence). Le contenu de la base pouvait donc être perçu comme d'une part une information de « qui » est absent (les données qualitatives générales), et d'autre part une information de « comment » se concrétise l'absence (les données numériques d'absence). Au delà d'une étude descriptive de base, l'idée de la démarche retenue était de constituer des groupes homogènes d'absence à partir de ce « comment ». L'hypothèse sous-jacente était alors que des agents absents de la même façon (donc avec le même « comment ») avaient des causes similaires d'absence. Une fois ces groupes identifiés, les autres types d'informations plus « subjectives » fournies par SOFAXIS ou collectées par l'IRDQ permettraient peut être d'identifier la cause essentielle associée à chaque groupe afin de proposer des solutions visant à réduire l'absentéisme.

Une classification par modèle de mélange gaussien a été menée sur les individus « données qualitatives » et décrits dans l'espace des quatre « données quantitatives d'absence » (pour plus de pertinence, certaines de ces données ont été normalisées et transformées en accord avec SOFAXIS auparavant). L'algorithme CEM (Celeux et Govaert 1992 [9]) a été utilisé et le nombre de groupes retenus à été sélectionné par le critère ICL (Biernacki *et al.* 2000 [3]). Ensuite, il a été facile de donner un profil d'absence lié à chaque groupe (groupe des arrêts prolongés mais peu chers, etc.), ainsi que les caractéristiques des agents (taille de collectivité, niveau hiérarchique, etc.) présents dans ces groupes.

Suite aux difficultés financières des assureurs engendrées par l'attentat du 11 septembre 2001 aux États-Unis, SOFAXIS a reporté le contrat qui devait logiquement faire suite à cette étude de faisabilité qui avait donné satisfaction.

5.4 Bilan et perspectives

Le logiciel MIXMOD est, comme nous l'avons vu, déjà opérationnel et diffusé dans les communautés scientifiques et industrielles. Son développement est assuré par les supports humains et financiers qui lui sont associés. En effet, MIXMOD bénéficie d'une part du soutien de l'INRIA par l'opération ODL (Opération de Développement Logiciel) qui finance un ingénieur expert à Grenoble pour deux années (la seconde vient de se terminer), et d'autre part du soutien du CNRS grâce à Florent Langrognet, ingénieur de recherche du CNRS à Besançon et coordinateur informatique et permanent du projet. La validation et les évolutions de MIXMOD se font sous la houlette d'un comité scientifique composé de Gilles Celeux, Directeur de Recherches à l'INRIA, de Gérard Govaert, Professeur à l'Université de Technologie de Compiègne et de moi-même.

Les évolutions prévues à court ou moyen terme pour MIXMOD sont aussi bien d'ordre scientifique qu'informatique. Du point de vue scientifique, la prise en charge d'autres types de mélange, en particulier les mélanges de classes latentes pour les données qualitatives, est en fin de gestation. Il n'est pas exclu aussi, d'implanter de nouvelles fonctionnalités, comme la méthode d'initialisation sur les trajectoires ou encore d'adjonction d'une composante atypique dite « de bruit » qui permettrait de modéliser assez proprement la notion de rejet en distance décrite par exemple par Dubuisson (1990) chap. 3 [15]. En informatique, il est souhaitable de faire évoluer MIXMOD sous forme d'une version disponible en tant que fonction R (ou SPLUS). Par ailleurs, il est envisagé de prendre un stagiaire ingénieur pour optimiser le code de MIXMOD, en particulier concernant tout ce qui est relatif à l'utilisation des bibliothèques mathématiques. Par la même occasion, une comparaison des performances (vitesse d'exécution et capacité de traitement en volume en particulier) avec les logiciels concurrents serait la bienvenue.

En indexation d'images, les outils statistiques sont de plus en plus utilisés afin de prendre en considération les différentes variabilités présentes. En particulier, les mélanges gaussiens, grâce à leur flexibilité et leur facilité d'usage, se répandent largement auprès de ce nouveau public. Dans ce contexte, nous avons décrit certaines propositions comme la modélisation de la distribution des couleurs par ce type de mélange. L'appariement des images de la base avec une image requête se faisait ensuite avec la fonction de vraisemblance des paramètres de mélange, cette technique conduisant ainsi à effectuer un classement des images de la base. Une perspective de recherche concerne alors la détermination d'un seuil de rupture automatique dans les valeurs de vraisemblance afin de sélectionner de façon automatique le nombre d'images de la base qui sont associées à l'image requête.

Enfin, il n'est pas exclut que la compagnie SOFAXIS relance le contrat de réduction de l'absentéisme dans les collectivités locale pour faire suite à l'étude de faisabilité déjà menée à son terme. En effet, cette problématique était un enjeu stratégique dans son métier et la problématique de classification identifiée semblait être un point de départ fort prometteur pour aborder ce sujet de façon objective.

Bibliographie

[1] D. Alleysson and J. Hérault. Variability in color discrimination data explained by a generic model of retina with non linear and adaptative processing. In Society Annual Meeting, editor, *International Color Vision*, Göttingen, August 1999.

[2] S. Benayoun, H. Bernard, P. Bertolino, M. Gelgon, C. Schmid, and F. Spindler. Structuration de vidéos pour des interfaces de consultations avancées. In *CORESA98 – Journée détudes et d'échanges COmpression et REpresentation des Signaux Audiovisuels*, June 1998.

[3] 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.

[4] 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 & Data Analysis*, 41:561–575, 2003.

[5] S. Blott and R. Weber. A simple vector-approximation file for similarity in high-dimensional vector spaces. Technical Report 19, ESPRIT project HERMES (no. 9141), March 1997.

[6] P. Bouthemy and F. Ganansia. Video partitioning and camera motion characterisation for content-based video indexing. In *3rd IEEE International Conference on Image Processing*, September 1996.

[7] C. Carson, S. Belongie, H. Greenspan, and Malik J. Blobword: image segmentation using expectation-maximisation and its application to image querying. *IEEE to Pattern Analysis and Machine Intelligence*, 1999. Submitted.

[8] 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(1):73–92, 1985.

[9] G. Celeux and G. Govaert. A classification EM algorithm and two stochastic versions. *Computational Statistics & Data Analysis*, 14:315–332, 1992.

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

[11] P. Cheeseman and J. Stutz. Bayesian classification (autoclass): Theory and results. In *Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA.*, pages 61–83, 1996.

[12] C. A. Clarotti, B. Villain, and H. Procaccia. Bayesian analysis of general failure data from an aging distribution: Advances in numerical methods. In *ESREL'96-PSAM III Conference*, pages 1289–1294, Crete, 1996.

[13] R. Cooke. *Experts in Uncertainty – Opinion and Subjective Probability in Science.* Oxford University Press, New-York, 1991.

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

[15] B. Dubuisson. *Diagnostic et reconnaissance des formes.* Hermès, Paris, 1990.

[16] C. Fraley and A. E. Raftery. Mclust: Software for model-based cluster analysis. *Journal of Classification*, 16:297–306, 1999.

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

[18] M. Gelgon and P. Bouthemy. A region-level graph labeling approach to motion-based segmentation. In *CVPR*, pages 514–519, Puerto Rico, June 1997.

[19] C. Genest and V. Zidek. Comining probability distributions: a critique and an annoted bibliography. *Statistical Science*, 1(1):114–148, 1986.

[20] H. Greenspan, J. Goldberger, and L. Riddel. A continuous probabilistic framework for image matching. Technical report, University of Tel-Aviv, 2000.

[21] G. Healey and Q. T. Luong. *Color in computer vision: Recent progress*, chapter Chapter 2.3, pages 283–312. Handbook of Pattern Recognition and Computer Vision. 1998.

[22] C. Lind and A. Nowak. Pooling expert opinions on probability distributions. *Journal of Engineering Mechanics*, 114(2):328–341, 1988.

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

[24] G.J. Mclachlan, D. Peel, K. E. Basford, and P. Adams. Fitting of mixtures of normals and t-components. *Journal of Statistical Software*, 4(2), 1999.

[25] M. Meyer and J. Booker. *Eliciting and Analysing Expert Judgment – A Practical Guide.* Academic Press, 1991.

[26] A. Moslesh, V. Bier, and G. Apostolakis. A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering and System Safety*, 20:63–85, 1988.

[27] H. Murase and S. K. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[28] B.C. O'Connor. Selecting key frames of moving image documents: A digital environment for analysis and navigation. *Microcomputers for Information Management*, 8(2):119–133, 1991.

[29] J. Puzicha, J.M Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *International Conference on Computer Vision*, pages 1165–1172, Kerkyra, Greece, September 1999.

[30] C. Robert. *Modèles statistiques pour l'I.A. – L'exemple du diagnostic médical (Statistical models for AI – The example of medical diagnosis).* Masson, 1991.

[31] C. P. Robert. *The Bayesian Choice : a Decision-Theoretic Motivation.* New York: Springer-Verlag, 1994.

[32] L. A. Rowe, J. S. Boreczky, and C. A. Eads. Indexes for user access to large video databases. In *IS. T SPIE Conference on Storage and Retrieval for Image and Video Databases II*, pages 150–161, 1994.

[33] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[34] S. W. Smoliar, H. J. Zhang, C. Y. Low, and Wu J. H. Video parsing, retrieval and browsing: An integrated and content-based solution. In *ACM Multimedia*, pages 15–24, 1995.

[35] T. Souchois and B. Villain. Survbayes: A software for bayesian reliability analysis of nuclear power plant components subjected to aging. In *ESREL'98 Conference*, pages 937–944, Balkema, Rotterdam, 1998.

[36] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[37] C. S. Wallace and D. L. Dowe. Intrinsic classification by mml – the snob program. In *Proceedings 7th Australian Joint Conference on Artificial Intelligence*, pages 37–44, Singapore, 1994. World Scientific Publishing.

[38] J. H. Wolfe. Normix: Computational methods for estimating the parameters of multivariate normal mixtures of distributions. Research Memo SRM 68–2, US Naval Personnel Research Activity, San Diego, 1967. U.S. Naval Personnel Research Activity.

[39] G. Wyszecki and W. S. Stiles. *Color Science : Concepts and Methods, Quantitative Data and Formulae.* Wiley, second edition edition, 1982.

Conclusion générale

On aura certainement compris que les modèles de mélanges finis de lois de probabilités sont à l'intersection de nombreuses applications dans des domaines variés et de non moins nombreuses questions théoriques ouvertes. Cette synergie application/théorie se révèle être une source prolifique de propositions pour le chercheur. C'est donc naturellement que nous envisageons de garder cet axe de recherche parmi les futures thématiques qui seront développées.

Chacun des chapitres de ce mémoire se termine par une partie où les perspectives et extensions liées aux travaux décrits sont présentées. Il ne s'agit donc pas ici de reprendre ces différents développements mais, au contraire, de proposer des perspectives moins corrélées avec ces travaux antérieurs. Un axe majeur de recherche futur concerne alors l'utilisation des mélanges de lois avec des distributions ou des données particulières, ces deux concepts étant parfois étroitement liés. Cette idée avait déjà été amorcée par l'étude des données groupées dans le cas gaussien ou encore par l'étude des données binaires dans le cas de lois de Bernoulli. Il s'agit en particulier de s'intéresser aux données fréquentielles pour lesquelles un mélange de lois de Dirichlet pourrait être utilisé. Mentionnons aussi notre intérêt pour les données manquantes, les données de rangs et les données répétées. Enfin, le couplage de données de nature différente est semble-t-il de grande importance en pratique et il peut s'avérer utile d'explorer des modèles qui passent outre l'indépendance des variables de catégories différentes en prenant en considération d'éventuelles corrélations conditionnellement aux composantes du mélange.