ANR ClasSel
Sélection de modèles en classification croisée
Livrable 2.2

Août 2010

Table des matières

Introduction

Deux pistes ont été lancés pour proposer des solutions au problème du choix du nombre de classes, et plus généralement du choix du du modèle, en classification croisée.

- La première consiste à étendre le critère asymptotique BIC au modèle des blocs latents. Ce travail, encore en cours, fait l'objet du premier chapitre et a été mené principalement par Aurore Lomet, doctorante à Heudiasyc ;
- La seconde approche consiste à s'appuyer sur le principe du *bootstrap* ; pour l'instant, nous avons commencé à étudier cette approche sur les modèles de mélanges classiques avant de l'étendre, si les résultats sont satisfaisant, au modèle des blocs latents. Ce travail fait l'objet du second chapitre et a été mené par Fida El Baf, étudiante en Post-doc et par Boubacar Diawara stagiaire du Master M2-MIGS du département de Mathématiques de l'Université de Bourgogne, tous deux rémunérés grâce aux crédits de l'ANR ClasSel.

Chapitre 1

BIC

1.1 Introduction

La sélection de modèle est un sujet récurrent en statistique et bien étudié dans la littérature. L'objectif est de choisir un modèle considéré comme le « meilleur »parmi un ensemble . La validation d'un modèle peut se faire par différentes méthodes : le pourcentage explicatif (pourcentage de bien classés), des tests de significativité, l'étude des résidus, un critère de sélection...

En classification croisée (modèle de blocs latents), peu de ces méthodes sont applicables. L'utilisation de pourcentage explicatif entraîne une surestimations du nombre de classes. En effet, plus le nombre de classes est grand, plus le pourcentage de bien classés sera grand. Ainsi, on tend vers une donnée par classe obtenant une classification avec aucune donnée mal classée, mais qui n'a pas de sens. Par ailleurs, l'utilisation des tests de significativité des paramètres permet de déterminer le nombre de paramètres d'un modèle mais pas son type. Enfin, l'étude des résidus n'est réalisable que si la signature est définie, ce qui n'est pas le cas en classification.

En revanche, les critères de sélection offrent de nombreux avantages et sont adaptés à la classification croisée. Développée dans les années 1970, cette théorie permet la comparaison de différents modèles, mais, surtout la sélection du « vrai »ou du « meilleur »modèle au sens d'un critère. Les plus généralement utilisés sont les critères AIC (*Akaike Information Criterion*) développé en 1973 par Akaike et BIC (*Bayesian Information Criterion*), introduit par (Schwarz, 1978), ainsi que leurs dérivés.

Le critère BIC, aussi connu sous le nom de *Schwarz Criterion*, est utilisé pour la sélection de modèles paramétriques ayant ou non un nombre différent de paramètres. Il s'applique dans un contexte bayésien de sélection de modèle : les paramètres et les modèles sont vus comme des variables aléatoires munies d'une distribution. Il s'agit d'une approximation de la probabilité du modèle conditionnellement aux données qui se traduit par une vraisemblance pénalisée. Le « meilleur »modèle est alors celui qui minimise ce critère.

Ce critère est bien adapté au modèle « blocs latents »Govaert et Nadif (2010) qui est une extension du modèle de mélange pour lequel les partitions lignes et colonnes sont des variables latentes. Il s'agit d'un modèle paramétrique s'inscrivant dans un cadre bayésien : les paramètres du modèle de blocs latents sont

vues comme des variables aléatoires. Ainsi, une extension du critère BIC prend son sens avec l'estimation de différents modèles ayant un nombre variable de paramètres. En effet, le nombre de paramètres dépendant du nombre de classes et le critère prenant en compte la complexité du modèle, BIC pourrait répondre au problème du choix du type de modèle et du nombre de classes.

Dans ce document, une étude préliminaire sur la notion de « vrai »et de « meilleur »modèle est menée pour caractériser le type de modèle choisi par le critère BIC. Ce dernier est exposé dans son développement général. Puis, une extension à 2 dimensions pour la classification croisée est proposée et discutée.

1.2 Étude préliminaire : les notions de « vrai »et « meilleur »modèle

La sélection de modèle amène la question de « vrai »et « meilleur »modèles. Il existe une variété de critères de sélection qui ne mènent pas toujours au même modèle choisi. N'étant pas tous l'estimation d'une même quantité et ne considérant pas forcément le même ensemble initial de modèles, les critères peuvent diverger quant au choix d'un modèle. Il est alors nécessaire de définir avant leurs utilisations les ensembles considérés. Quel type de modèle est choisi par le critère, et comment le caractériser ? Ainsi, la notion de « vrai »modèle, couramment utilisées dans la littérature, est à définir car elle permet l'identification du type de modèle sélectionné. Cette définition implique celle du « meilleur »modèle qui est associé à un critère.

1.2.1 Le « vrai »modèle

Sur des données simulées, le modèle génératif réel est défini dès le début de l'expérience. La famille de modèles et les paramètres sont fixes et connus. Dans le cas de données réelles, des connaissances a priori sur le modèle peuvent intervenir, mais, le plus souvent, l'identification du modèle fait suite à une expertise de la part de l'utilisateur. Dans ces deux situations, un des objectifs de l'analyse de données est de trouver un modèle permettant de les expliquer. Ainsi, un ensemble de modèles différents peut être défini. En effet, le choix entre plusieurs méthodes statistiques, les algorithmes, les types de modèles amène souvent à une variété de modèles possibles. Obtenant un ensemble pour un même jeu de données, se pose alors la question de savoir quel est le « vrai »modèle. Est-ce celui qui génère les données ou celui vers lequel elles tendent ? Peut-il être choisi par un critère ? Son existence est-elle certaine ?

La notion de « vrai »modèle apparaît dans la littérature sous différentes définitions. Ainsi, (Burnham et Anderson, 2004) définissent le « vrai »modèle comme « le modèle mathématique qui exprime exactement l'entière réalité ». Or, les modèles estimables ne permettent pas d'expliquer exactement toutes les données. Dans le cas de simulations, les données sont simples et le modèle théorique qui les a générées peut être considéré comme le vrai modèle. En revanche, dans le cas de données réelles qui sont généralement plus difficiles à expliquer par un modèle, le modèle estimé ne permet pas de prendre en compte la totalité des informations. Ainsi, selon cette définition, le « vrai »modèle ne peut être estimé. (Box et Draper, 1987) arrive a une conclusion similaire : « Essentiellement , tous les modèles sont faux, mais certains sont utiles".

Par ailleurs, Ye *et al.* (2008) décrivent le vrai modèle comme celui ayant « générée les données observées". Dans le cas des données simulées, le vrai modèle est alors le modèle théorique initial. Selon cette définition, l'estimation du vrai modèle de données réelles est alors possible dès lors qu'il permet de les générer.

Dans cette étude, la définition du « vrai »modèle résulte d'un choix pratique. Les expériences du critère BIC sont réalisées en 2 temps : en premier lieu, sur des données simulées, puis avec des données réelles déjà utilisées dans la littérature. Les données simulées offrent l'avantage de connaître le modèle qui les génère et, par conséquent, d'avoir un point de comparaison. Puisqu'il est pris en référence, le modèle génératif est le « vrai »modèle. Les performances du critère BIC sont alors évaluées selon sa capacité à choisir le vrai modèle ou celui se rapprochant le plus de celui ayant généré les données au sens de la divergence de Kullback-Leiber définie pour f et g, deux fonctions de densités quelconques par :

$$d(f,g) = \int \log\left(\frac{f(x)}{g(x)}\right) f(x)dx \tag{1.1}$$

Dans le cas de données réelles, le « vrai »modèle est inconnu. Il peut être estimé. Le modèle choisi n'est donc pas forcément le « vrai »modèle mais le « meilleur" »au sens du critère BIC si le vrai modèle n'appartient pas à l'ensemble de départ.

1.2.2 Le « meilleur »et « quasi-vrai »modèle

Le critère BIC est une approximation de la probabilité a posteriori du modèle. Il permet de sélectionner un modèle parmi un ensemble fini. Si le vrai modèle fait partie de la collection, le critère BIC permet de choisir le vrai modèle quand la taille de l'échantillon est grande. En revanche, s'il n'appartient pas à cette collection, le modèle choisi par le critère BIC est un modèle considéré comme le « meilleur".

Pour (Lebarbier et Mary-Huard, 2004), le « meilleur »modèle, au sens du critère BIC, tend (pour une taille d'échantillon tendant vers l'infini) vers le « quasi-vrai »modèle quand le vrai modèle ne fait pas partie de la collection initiale. Supposant que les modèles sont emboîtés, le « quasi-vrai »modèle est le modèle le plus proche du vrai selon la divergence de Kullback-Leibler.

Ainsi, la probabilité a posteriori du modèle tend vers 1 pour le quasi-vrai modèle quand la taille de l'échantillon tend vers l'infini et 0 pour les autres.

$$P(M_i|X) \approx \frac{\exp(-\frac{1}{2}BIC_i - BIC_{min})}{\sum_a(-\frac{1}{2}BIC_i - BIC_a)} \tag{1.2}$$

avec X les données, M_i le modèle appartenant à la collection et a l'indice du modèle appartenant à la collection, BIC_i les critère BIC calculé pour le modèle M_i.

Le meilleur modèle choisi par le critère BIC n'est donc pas forcément le vrai mais le quasi-vrai. Ce dernier pouvant être éloigné du vrai, il est nécessaire que le vrai modèle fasse partie de la collection initiale pour garantir la justesse du modèle choisi.

1.3 Généralité : le critère BIC

Cette section aborde la construction du critère BIC dans sa forme classique. Le développement suivant s'appuie principalement sur les papiers de (Lebarbier et Mary-Huard, 2004) et de (Raftery, 1995) qui expliquent l'ensemble des étapes nécessaires à l'écriture du critère BIC. Une démonstration analogue sera proposée pour la classification croisée.

Soient :

- un *n-échantillon* $\mathbf{X} = (X_1, ...X_n)$ de variables aléatoires indépendantes de densité inconnue f que l'on cherche à estimer
- une ensemble fini de modèles $(M_1, ..., M_A)$
- g_{M_i} la densité associé au modèle M_i de paramètre θ_i
- Θ_i l'espace de dimension K_i auquel appartient θ_i

L'objectif est de choisir un M_i parmi la collection de modèles. Se plaçant dans un contexte bayésien, $(M_1, ..., M_n)$ et θ_i sont des variables aléatoires munies d'une distribution *a priori*.

L'écriture du critère BIC est réalisée en 3 étapes :

1. On pose le problème : maximisation de $P(M_i|\mathbf{X})$ sous l'hypothèse d'équiprobabilité des modèles.
2. Pour maximiser cette probabilité, on maximise $P(\mathbf{X}|M_i)$.
3. On utilise l'approximation de Laplace de cette probabilité et on néglige les termes constants et d'erreur.

1.3.1 Définition du problème

Choisir le modèle le plus probable revient à maximiser la probabilité a posteriori du modèle $P(M_i|\mathbf{X})$. Cette dernière étant fonction du critère BIC, le modèle choisi est alors défini par :

$$M_{BIC} = \arg\max_{M_i} P(M_i|\mathbf{X}) \tag{1.3}$$

1.3.2 Maximisation de $P(\mathbf{X}|M_i)$

Ne connaissant pas $P(M_i|\mathbf{X})$, une réécriture de cette probabilité est réalisée, obtenant :

$$P(M_i|\mathbf{X}) = \frac{P(\mathbf{X}|M_i)P(M_i)}{P(\mathbf{X})} \tag{1.4}$$

$P(\mathbf{X})$ ne dépend pas du modèle M_i et si $P(M_i)$ est constant alors $P(M_i|\mathbf{X}) \equiv P(\mathbf{X}|M_i)$. On travaille alors avec $P(\mathbf{X}|M_i)$ à laquelle la formule des probabilités totales.

$$\begin{aligned} P(\mathbf{X}|M_i) &= \int_{\Theta_i} P(\mathbf{X}, \theta_i|M_i)d\theta_i \\ &= \int_{\Theta_i} P(\mathbf{X}|\theta_i, M_i)P(\theta_i|M_i)d\theta_i \end{aligned}$$

1.3.3 L'approximation de Laplace

Ne sachant pas calculer cette intégrale, on applique l'approximation de Laplace définie par : Soit une fonction $h : \mathbb{R}^d \to \mathbb{R}$ telle que h est deux fois différentiable sur $\mathbb{R}^d$ et atteint un unique maximum sur $\mathbb{R}^d$ en u^*. On a alors :

$$\int_{\mathbb{R}^d} \exp(nh(u))du = \exp(nh(u^*))\Big(\frac{2\pi}{n}\Big)^{d/2}|-h''(u^*)|^{-1/2} + O(n^{-1}) \qquad (1.5)$$

Pour se faire, on pose :

$$h(\theta_i) = \frac{1}{n}\sum_{k=1}^{n} \log P(X_k|\theta_i^*, M_i) + \frac{\log(P(\theta_i|M_i))}{n} \qquad (1.6)$$

On obtient alors :

$$\begin{aligned} \log(P(\mathbf{X}|M_i)) &= \log(P(\mathbf{X}|\theta_i^*, M_i)) + \log(P(\theta_i^*|M_i)) \\ &+ \frac{K_i}{2}\log(2\pi) - \frac{K_i}{2}\log(n) - \frac{1}{2}\ln(|A_i^*|) + O(n^{-1}) \end{aligned}$$

avec $\theta_i^* = \arg\max_{\theta_i \in \Theta_i} Ln(\theta_i)$ et A_i^* la hessienne de h en θ_i^*

A l'asymptotique, θ_i^* peut être remplacé par l'estimateur du maximum de vraisemblance $\hat{\theta}_i$ avec $\hat{\theta}_i = \arg\min_{\theta_i} \frac{1}{n}P(\mathbf{X}|\theta_i, M_i)$. De même, A_i^* est remplacée par l'information de Fisher $I_{\hat{\theta}_i} = -E\Big(\Big[\frac{\partial^2 \log(P(X|\theta_i, M_i))}{\partial\theta_i^j \partial_i^l}\Big]_{j,l}|_{\theta_i=\hat{\theta}_i}\Big)$.

$$\begin{aligned} \log(P(\mathbf{X}|M_i)) &= \overbrace{\log(P(\mathbf{X}|\hat{\theta}_i, M_i)) - \frac{K_i}{2}\log(n)}^{\text{tend vers } -\infty \text{ avec n}} \\ &+ \underbrace{\log(P(\hat{\theta}_i|M_i)) + \frac{K_i}{2}\log(2\pi) - \frac{1}{2}\ln(|I_{\hat{\theta}_i}|)}_{O(1)} + O(n^{-1/2}) \end{aligned}$$

En effet, $\log(P(\hat{\theta}_i|M_i))$ ne dépend pas de n, de même pour $\frac{K_i}{2}\log(2\pi)$. L'information de Fisher s'écrit pour une une variable ; elle ne dépend pas de n, elle est donc constante pour la taille de l'échantillon (cf Annexe B). Par ailleurs,

$$BIC \equiv -2\log\widehat{P(M_i|\mathbf{X})} \qquad (1.7)$$

Donc, en négligeant les termes d'erreur et les termes constants à tous les modèles, BIC s'écrit :

$$BIC_i = -2\log P(\mathbf{X}|M_i) \approx -2\log(P(\mathbf{X}|\hat{\theta}_i, M_i)) + K_i\log(n) \qquad (1.8)$$

$$BIC = -2L(\hat{\theta}) + K\log(n) \qquad (1.9)$$

Le meilleur modèle choisi par ce critère est celui qui minimise BIC.

Remarques :

– L'erreur en $O(1)$ peut perturber le choix du modèle.

– h est une fonction concave.

1.4 Application du critère BIC à la classification croisée

Le critère BIC n'est pas directement applicable au modèle de blocs latents. La pénalité de BIC dépend de la taille de l'échantillon. Or, en classification croisée, un tableau est de taille $n \times d$. Plusieurs tailles d'échantillon peuvent alors être proposée. On peut considérer un tableau comme une variable aléatoire ou alors considérer chaque valeur du tableau comme des variables. Dans son modèle de bloc latent, (Govaert et Nadif, 2009) écrit la densité du modèle en considérant le tableau comme variable aléatoire. Chaque tableau correspond donc à une variable aléatoire. Ne travaillant généralement, que sur un seul tableau par classification, l'échantillon est souvent de taille 1. Or le critère BIC n'est écrit pour une grande taille d'échantillon. Il est alors nécessaire de proposer une extension au critère BIC appliquée à la classification croisée.

Un raisonnement analogue à celui exposé dans la section précédente et présenté par Lebarbier et Raftery est ainsi proposé. Le développement du critère BIC est réalisé alors en 4 étapes (les 3 précédentes et une étape propre au modèle de blocs latents) :

1. On pose l'objectif : maximisation de $P(M|X)$.
2. Pour maximiser cette probabilité, maximisation de $P(\mathbf{X}|M)$.
3. Utilisation de l'approximation variationnelle étendue pour calculer $P(X|M)$,.
4. Utilisation de l'approximation de Laplace de l'approximation variationnelle étendue ; puis, on néglige les termes constants et d'erreur.

1.4.1 L'objectif

Pour alléger l'écriture, on sous-entend les indices liés au modèle i. Comme précédemment, on cherche à maximiser la probabilité à posteriori du modèle $P(M|X)$ afin d'obtenir le plus probable.

$$M_{BIC} = \arg\max_{M} P(M|X) \tag{1.10}$$

1.4.2 Maximisation de $P(X|M)$

On écrit : $P(M|X) = P(X|M)P(M)/P(X)$.

$$\begin{aligned} P(X|M) &= \int_{\Theta} P(X,\theta|M)d\theta \\ &= \int_{\Theta} P(X|\theta,M)P(\theta|M)d\theta \end{aligned}$$

De plus, le modèle bloc latent est défini par :

$$P(X|\theta,M) = \sum_{(z,w)\times(Z,W)} \prod_{i} \pi_{z_i} \prod_{j} \rho_{w_j} \prod_{i,j,k,l} f_{z_i w_j}(x_{ij},\theta) \tag{1.11}$$

On écrit $g(\theta) = \log(P(X|\theta,M)P(\theta|M))$. On a donc $P(X|M) = \int_{\Theta} e^{g(\theta)}d\theta$.

Comme vu dans le cas général, on souhaite utiliser l'approximation de Laplace. Si on suit une démonstration analogue à la section 3, on devrait diviser

g par 1 (taille de l'échantillon) pour obtenir h ($g(\theta) = h(\theta)$). Ce raisonnement pose plusieurs problèmes.

Le premier des problèmes est la taille de l'échantillon. Dans le cas du modèle de blocs latents, l'échantillon est de taille 1 (le tableau à classifier). On ne travaille donc pas à l'asymptotique. Or, une des condition pour l'application de l'approximation de Laplace est l'existence d'un unique maximum. Il faut donc prouver que la fonction h, fonction de vraisemblance, admet un unique maximum. Une des propriété des fonctions de vraisemblance est que la probabilité de l'unicité du maximum ne tend vers 1 que pour des grandes tailles d'échantillon. Dans le cas présent, on ne peut donc pas prouver son unicité et donc appliquer l'approximation de Laplace. Par ailleurs, même si on admettait l'unicité du maximum, on ne pourrait, quand même, pas faire l'approximation du maximum par celui du maximum de vraisemblance qui ne se justifie que pour une grande taille d'échantillon (cf annexe A).

Le second problème s'explique par l'écriture de la densité du modèle Bloc latent. Travaillant sur l'ensemble des partitions possibles, une optimisation de la probabilité $P(\mathbf{X}|\theta, M)$ (et donc de h) est difficile à réaliser. On ne pourrait donc pas proposer des estimations du maximum de vraisemblance pour θ.

1.4.3 L'approximation variationnelle étendue

Une solution envisagée à ces problèmes est l'approche variationnelle étendue utilisée dans l'algorithme EM. Il s'agit d'une méthode qui vise à factoriser une probabilité conditionnelle. Le principe est d'admettre l'égalité entre $P(z, w|x)$ et $P(z|x)P(w|x)$. On peut ainsi travailler sur l'ensemble des x_{ij} et non plus sur l'ensemble des partitions possibles. L'échantillon se définit alors par $\mathbf{X} = (X_{11}, ..., X_{nd})$ de taille $n \times d$. On peut donc faire tendre nd vers ∞ pour obtenir l'asymptotique nécessaire à l'écriture du critère BIC.

Avec cette approche, la log-vraisemblance classifiante du modèle bloc latent est définie par :

$$L_c(\mathbf{z}, \mathbf{w}, \theta) = \sum_{i,k} z_{ik} \log \pi_k + \sum_{j,l} w_{jl} \log \rho_l + \sum_{i,j,k,l} z_{ik} w_{jl} \log f_{kl}(x_{ij}, \alpha) \quad (1.12)$$

En pratique, on ne connaît pas $\mathbf{z}$ et $\mathbf{w}$. Les z_{ik} et w_{jl} sont alors estimés par maximisation alternée du critère flou de classification Govaert et Nadif (2009) :

$$F_c(\mathbf{s}; \theta) = L_c(\mathbf{s}, \theta) - \sum_{i,k} s_{ik} \log s_{ik} \quad (1.13)$$

avec $s_{ik} = P(z_{ik} = 1|\mathbf{x}, \theta)$.

De même, on a $t_{jl} = P(w_{jl} = 1|\mathbf{x}, \theta)$. La log-vraisemblance classifiante estimée est alors :

$$\tilde{L}(\theta) = \arg\max_{st} F_c(\mathbf{s}, \mathbf{t}, \theta) \quad (1.14)$$

avec :

$$F_c(\mathbf{s},\mathbf{t},\theta) = L_{CR}(\mathbf{s},\mathbf{t},\theta) + H(\mathbf{s}) + H(\mathbf{t}) + \sum_{i,k} s_{ik}\log\pi_k + \sum_{j,l} t_{jl}\log\rho_l$$

$$L_{CR}(\mathbf{s},\mathbf{t},\theta) = \sum_{i,j,k,l} s_{ik}t_{jl}\log f_{kl}(x_{ij},\alpha)$$

$$H(\mathbf{s}) = -\sum_{i,k} s_{ik}\log s_{ik}$$

$$H(\mathbf{t}) = -\sum_{j,l} t_{jl}\log t_{jl}$$

Remarque 1 : Dans *A view of the EM algorithm that justifies incremental, sparse, and other variants* Neal et Hinton (1998) (théorème 2) montrent que le maximum atteint pour $F_c(\mathbf{s},\mathbf{t},\theta)$ est égal à celui de $L_c(\mathbf{z},\mathbf{w},\theta)$.

Remarque 2 : Pour (Jaakkola, 2000), si les distributions a posteriori des variables latentes sont presque indépendantes, alors l'approximation variationnelle est presque parfaite. Plus elles sont liées, plus les résultats se dégradent. Ainsi, on peut considérer $\tilde{M} \approx M$ quand les variables latentes ont des distributions presque indépendantes et cette égalité pourra être discutée en cas cas de forte dépendance.

Suivant cette approximation, la probabilité des données $\mathbf{X}$ connaissant les paramètres et le modèle devient :

$$\log\tilde{P}(\mathbf{X}|\theta,M) = \tilde{L}(\theta) \tag{1.15}$$

1.4.4 Application de l'approximation de Laplace

Pour appliquer l'approximation de Laplace, on pose :

$$h_{nd}(\theta) = \frac{1}{nd}\log\tilde{P}(\mathbf{X}|\theta,M) + \frac{1}{nd}P(\theta|M) \tag{1.16}$$

$$= \frac{1}{nd}\tilde{L}(\theta) + \frac{1}{nd}P(\theta|M) \tag{1.17}$$

On a alors :

$$\int_\Theta e^{ndh_{nd}(\theta)}d\theta = e^{ndh_{nd}(\theta^*)}(\frac{2\pi}{nd})^{K/2}|-h''_{nd}(\theta^*)|^{1/2} + O((nd)^{-1})$$

$$\log(P(X|M)) = \tilde{L}(\theta^*) - \frac{K}{2}\log(nd) + \frac{K}{2}\log(2\pi) + \log(P(\theta^*|M)) - \frac{1}{2}\log(|-h''_{nd}(\theta^*)|) + O((nd)^{-1})$$

A l'asymptotique, on remplace alors θ^* par $\theta_{MV} = \hat{\theta}$ et $-h''_{nd}(\theta^*)$ par $I_{\hat{\theta}}$ (Annexe). En négligeant les termes d'erreur et les termes en $O(1)$, on obtient :

$$\log(P(X|M)) \approx \tilde{L}(\hat{\theta}) - \frac{K}{2}\log(nd) \tag{1.18}$$

BIC s'écrit alors :

$$BIC = -2\tilde{L}(\hat{\theta}) + K\log(nd) \tag{1.19}$$

avec θ les paramètres du modèle, K la dimension de θ , $n \times d$ la taille de l'échantillon.

1.4.5 Remarques, discussion, perspective

L'application du critère BIC au modèle de blocs latents nécessite une étape supplémentaire par rapport à son développement classique. L'approximation variationnelle étendue permet d'obtenir une forme explicite et calculable du critère. BIC s'écrit donc suite à deux approximations (variationnelle et Laplace) qui peut perturber les résultats. De plus, il est préférable d'utiliser des tableaux de grandes tailles puisque des points clefs du développement ne sont vrais qu'à l'asymptotique.

Enfin, des expériences sur différents jeux de données seront réalisées pour vérifier la précision du critère BIC appliqué au modèle de blocs latents.

1.5 Essais numériques

1.5.1 Choix du nombre de classes

Conditions expérimentales

Dans les premiers essais, on considère un seul modèle dont on veut retrouver le nombre de classes. Il s'agit d'un modèle de blocs latents dont les proportions associées à chaque classe sont égales, mais dont les moyennes et les variances sont différentes. Un tableau est généré aléatoirement suivant des paramètres fixés (paramètres du modèle et taille du tableau). Chaque essai correspond à un pourcentage de mal classés MAP 5%, 12% et 20% correspondant respectivement à des classes séparées, moyennement séparées et peu séparées. Les proportions, les moyennes sont fixes ; seules les variances changent pour obtenir ces taux de mal-classés MAP. Pour le cas symétrique, on définit ainsi les paramètres suivant :

$$\boldsymbol{\pi} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix}$$

$$\boldsymbol{\rho} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{pmatrix} 1 & -1 & -1 \\ -1 & -1 & 1 \\ -1 & 1 & -1 \end{pmatrix}$$

σ^2	5%	12%	20%
50×50	$\begin{pmatrix} 10.76 & 12.105 & 5.38 \\ 8.07 & 8.74 & 10.54 \\ 5.38 & 12.11 & 5.38 \end{pmatrix}$	$\begin{pmatrix} 18.74 & 21.08 & 9.37 \\ 14.06 & 15.23 & 18.37 \\ 9.37 & 21.08 & 9.37 \end{pmatrix}$	$\begin{pmatrix} 31.52 & 35.46 & 15.76 \\ 23.64 & 25.61 & 30.89 \\ 15.76 & 35.46 & 15.76 \end{pmatrix}$
200×200	$\begin{pmatrix} 30.12 & 31.38 & 25.10 \\ 27.61 & 28.24 & 29.92 \\ 25.10 & 31.38 & 25.10 \end{pmatrix}$	$\begin{pmatrix} 43.68 & 45.50 & 36.40 \\ 40.04 & 40.95 & 43.39 \\ 36.40 & 45.50 & 36.40 \end{pmatrix}$	$\begin{pmatrix} 58.68 & 61.13 & 48.90 \\ 53.79 & 55.01 & 58.29 \\ 48.90 & 61.13 & 48.90 \end{pmatrix}$
500×500	$\begin{pmatrix} 85.56 & 89.13 & 71.30 \\ 78.43 & 80.21 & 84.99 \\ 71.30 & 89.13 & 71.30 \end{pmatrix}$	$\begin{pmatrix} 135.96 & 141.63 & 113.30 \\ 124.63 & 127.46 & 135.05 \\ 113.30 & 141.63 & 113.30 \end{pmatrix}$	$\begin{pmatrix} 196.32 & 204.50 & 163.60 \\ 179.96 & 184.05 & 195.01 \\ 163.60 & 204.50 & 163.60 \end{pmatrix}$

Il s'agit donc d'un modèle ayant 3 classes en ligne et 3 classes en colonne. Le tableau correspondant est classifié par l'algorithme EM.

Évolution du critère BIC

Pour chaque combinaison de classes possibles (de 1 à 4), on calcule le critère BIC. Le modèle préféré est celui obtenant la plus petite valeur. Cette opération est réalisée sur 20 tableaux différents issues de simulations ayant des paramètres identiques. Ainsi, pour chaque taille de tableau et pour chaque taux de mal-classés, on réalise 20 fois la même opération.

1.6 Annexe

1.6.1 Annexe A

On veut montrer que :

$$h(\theta) \underset{nd\to\infty}{\longrightarrow} \frac{1}{nd}\tilde{L}(\theta)$$

avec $h(\theta) = \frac{1}{nd}\tilde{L}(\theta) + \frac{1}{nd}P(\theta|M)$.

Quand nd tend vers l'infini alors $\frac{1}{nd}P(\theta|M)$ tend vers 0 puisque $P(\theta|M)$ ne dépend pas de nd. Il faut donc monter que $\frac{1}{nd}\tilde{L}(\theta)$ ne tend pas vers 0.

A l'asymptotique, $\frac{1}{nd}\tilde{L}(\theta)$ est équivalent à $\frac{1}{nd}L_{CR}(\mathbf{s},\mathbf{t},\theta)$. En effet, on a :

$\frac{H(\mathbf{s})}{nd} \to 0$ car $H(s)$ est de l'ordre de n (somme sur n)
$\frac{H(\mathbf{t})}{nd} \to 0$ car $H(t)$ est de l'ordre de d (somme sur d)
$\frac{\sum_{i,k} s_{ik}\log(\pi_k)}{nd} \to 0$ car la somme est de l'ordre de n (somme sur n)
$\frac{\sum_{j,l} t_{jl}\log(\rho_l)}{nd} \to 0$ car la somme est de l'ordre de d (somme sur d)

Or,

$$\begin{aligned}
\frac{1}{nd}\tilde{L}(\theta) &\geq \frac{1}{nd}\sum_{i,j,k,l}\frac{1}{g}\frac{1}{m}\log f_{kl}(x_{ij},\alpha) \\
&\geq \frac{1}{nd}\sum_{i,j}\log\underbrace{(\prod_{k,l} f_{kl}(x_{ij},\alpha))^{1/gm}}_{=g(x_{ij})} \\
&\geq \underbrace{\frac{1}{nd}\sum_{i,j}\log g(x_{ij})}_{O(1)\text{ car } g(x_{ij})\text{ fonction de densité}}
\end{aligned}$$

Donc $\frac{1}{nd}\tilde{L}(\theta)$ ne tend pas vers 0. On a donc prouver que $h(\theta) \underset{nd\to\infty}{\longrightarrow} \frac{1}{nd}\tilde{L}(\theta)$. On peut donc remplacer θ^* par $\widehat{\theta}$ et $-h''_{nd}(\theta^*)$ par $I_{\hat{\theta}}$ (cf Annexe B).

1.6.2 Annexe B

On veut montrer que l'information de Fisher $I_{\hat{\theta}}$ est constante quand n tend vers l'infini.

$$I_{\hat{\theta_i},n} = -E\Big(\Big[\frac{\partial^2 \log(P(\mathbf{X}|\theta_i, M_i))}{\partial\theta_i^j \partial_i^l}\Big]_{j,l}{}_{|\theta_i=\hat{\theta}_i}\Big)$$

Si $\mathbf{X}$ est un vecteur de variables indépendantes, alors on peut utiliser la propriété d'additivité de l'information de Fisher. De plus, si $\mathbf{X}$ ne dépend pas de θ_i et que chaque variable soit identiquement distribuée, alors chaque observation apporte la même information au paramètre θ. On a donc :

$$I_{\hat{\theta_i},n} = nI_{\hat{\theta_i}}$$

où $I_{\hat{\theta_i}}$ est l'information de Fisher pour une observation.

De plus, $h \approx \frac{1}{n}\sum_{k=1}^{n} \log P(X_k|\theta_i^*, M_i)$, on peut donc remplacer A_i^* par $I_{\hat{\theta}}$ qui est une quantité qui ne dépend pas n. L'information de Fisher est donc constante pour n qui tend vers l'infini.

Chapitre 2

Utilisation du Bootstrap

2.1 Modélisation du problème de classification

2.1.1 Présentation

Pour introduire un modèle théorique au problème de classification, on fait une hypothèse de statistique, en considérant les données comme une réalisation de vecteurs aléatoires indépendants et identiquement distribués. Dans le cadre des approches paramétriques, puisque c'est ce qui nous intéresse, on fait des hypothèses sur les distributions de probabilités de ces vecteurs aléatoires. En effet, on s'appuie sur des modèles probabilistes pour caractériser ces distributions de probabilités. Comme précédemment annoncé dans l'introduction, dans la suite de ce chapitre nous développons les modèles de mélange fini.

2.1.2 Modèles de mélange fini

La première utilisation de ces modèles remonte en 1886 par Newcomb, pour la détection de points aberrants, puis par Pearson en 1894 pour l'identification de deux populations de crabes. Elle est parfaitement adaptée à la détection des classes dans une population.

Pour répartir les observations[1] en g classes, on suppose qu'elles ont été générées à partir de g distributions homogènes. Chaque classe k est ainsi caractérisée par une distribution de probabilité de densité φ_k de paramètre α_k pour $1 \leq k \leq g$. La distribution de probabilité de la population entière est le mélange de ces distributions avec une certaine proportion π_k, $1 \leq k \leq g$, pour chacune. Elle est donnée en un point x par la densité :

$$f(x, \theta) = \sum_{k=1}^{g} \pi_k \varphi_k(x; \alpha_k) \tag{2.1}$$

avec

- $\theta = (\pi_1, \cdots, \pi_g, \alpha_1, \cdots, \alpha_g)$: les paramètres du modèle ;
- $\pi = (\pi_1, \cdots, \pi_g)$: les proportions du mélange, $\sum_{k=1}^{g} \pi_k = 1$;

[1] Nous ne considérons ici que le cas où les observations ne prennent que des valeurs continues dans R^p

– $\alpha = (\alpha_1, \cdots, \alpha_g)$: les paramètres des densités φ_k

Nous avons considéré que les densités φ_k des composantes appartiennent à une famille paramétrée $\varphi(., \alpha)$. Dans la suite de ce document, cette famille sera considérée gaussienne. Pour simplifier les notations, les φ_k seront tout simplement notés φ. Les paramètres α_k suffisent à les identifier.

Les observations représentent un échantillon de taille n de vecteurs aléatoires $x = (x_1, \ldots, x_n)$ indépendamment et identiquement distribués issu de la distribution f où chaque individu i est mesuré par un vecteur $x_i = (x_{i1}, \ldots, x_{ip})$. Les données sont représentées par une matrice de dimension $X = (n, p)$.

Classifier ces individus revient donc à rechercher une partition $z = (z_1, \ldots, z_n)$ (notation vectorielle) en g classes où $z_i \in \{1, \cdots, g\}$ indique la classe de l'individu i. Par la suite les z_i seront décomposés de la manière suivante : $z_i = (z_{i1}, \cdots, z_{ig})$(notation matricielle) avec $z_{ik} = 1$ si l'individu i appartient à la classe k et $z_{ik} = 0$ sinon.

2.1.3 Modèles de mélange Gaussien multidimensionnel

Dans ces modèles, chaque φ est caractérisé par une densité normale multidimensionnelle .

$$\varphi(x, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{p}{2}} |\sum_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu_k)' \Sigma_k^{-1}(x - \mu_k)\right) \tag{2.2}$$

où

– μ_k est le vecteur moyenne de la classe k

– Σ_k : matrice de variance-covariance de la classe k

La densité du mélange pour un vecteur $x \in R^p$ s'écrit alors

$$f(x, \theta) = \sum_{k=1}^{g} \frac{\pi_k}{(2\pi)^{\frac{p}{2}} |\sum_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu_k)' \Sigma_k^{-1}(x - \mu_k)\right) \tag{2.3}$$

avec $\theta = (\pi_1, \cdots, \pi_g, \mu_g, \cdots, \mu_1, \Sigma_1, \cdots, \Sigma_g)$

Les classes associées aux composantes du mélange sont caractérisées par des ellipsoïdes.

La fréquente utilisation des mélanges de distributions gaussiennes en classification automatique est expliquée par par DANG van Mô selon ces termes : «Ceci tient d'une part au fait que la distribution normale modélise de façon adéquate un grand nombre de phénomènes aléatoires. D'autre part, les mélanges gaussiens permettent de retrouver un certain nombre de critères de classification traditionnels»[2]

2.1.4 Mélanges gaussiens parcimonieux

(Banfield et Raftery, 1993) et (Celeux et Govaert, 1995) proposent une paramétrisation des matrices Σ_k de la façon suivante :

$$\Sigma_k = \lambda_k D_k A_k D_k' \tag{2.4}$$

où

[2]Le lien entre le certains modèles de mélange et les critères classiques de classifications est bien détaillé dans la thèse de (Dang, 1998)

- $\lambda_k = |\Sigma_k|^{\frac{1}{p}}$: désigne le volume de l'ellipsoïde.
- D_k : est la matrice des vecteurs propres de Σ_k. Elle détermine les directions de l'ellipsoïde.
- A_k : est la matrice diagonale des valeurs propres normalisés de Σ_k telle que $|A_k| = 1$. Elle donne la forme de l'ellipsoïde.

Le paramètre θ du mélange est finalement égal à :

$$\theta = (\pi_1, \cdots, \pi_g, \mu_1, \cdots, \mu_g, \lambda_1, \cdots, \lambda_g, A_1, \cdots, A_g, D_1, \cdots, D_g) \qquad (2.5)$$

Avec cette décomposition des matrices Σ_k de variance-covariance, différentes contraintes peuvent être imposées sur les paramètres λ_k, D_k, et A_k.Nous obtenons ainsi les mélanges parcimonieux plus simples à interpréter. Ces modèles peuvent être regroupés dans quatre grandes familles[3].

La famille sphérique. Dans cette famille on impose à la matrice A_k d'être la matrice identité I. Deux modèles parcimonieux existent dans cette famille : $\Sigma_k = \lambda I$ (volume identique) et $\Sigma_k = \lambda_k I$ (volume différent) pour $1 \leq k \leq g$. De plus on peut utiliser les paramètres π_k pour identifier les cas ou les classes ont les même proportions ou pas. En faisant varier le nombre g de composantes dans le mélange, cette famille s'agrandit avec des sous modèles dont un résumé est donné dans la table ci-dessous. Ici, on se limite à 4 classes maximum.

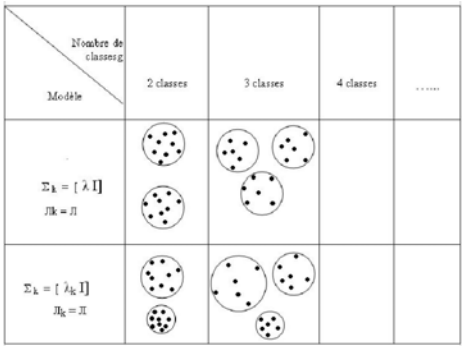


FIG. 2.1 – 4 modèles sphériques

	1	2	3	4
$\Sigma_k = \lambda I; \pi_k = \pi = \frac{1}{g}$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\Sigma_k = \lambda_k I; \pi_k = \pi = \frac{1}{g}$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\Sigma_k = \lambda I; \pi_k \neq$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\Sigma_k = \lambda_k I; \pi_k \neq$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

TAB. 2.1 – Les modèles sphériques

Dans la suite, pour tester les critères de sélection, nous nous limiterons uniquement à cette famille. Chaque ligne de la table : 2.1 correspond à une sous famille sur laquelle des recherches de modèle peuvent être lancées.

[3]ce sont les familles quadratique, générale, diagonale et sphérique dont une description est donnée dans Govaert (2003) selon les contraintes dont la famille sphérique

2.1.5 Estimation des paramètres du mélange par l'algorithme EM

Pour estimer les paramètres du modèle nous faisons appel à la méthode du maximum de vraisemblance qui est de loin la méthode la plus utiliser dans le cadre des modèles de mélange. Elle consiste à maximiser la log-vraisemblance dont une définition est donnée comme suit :

Definition 1. ***log-vraisemblance, log-vraisemblance complétée***
Soit $x = (x_1, \cdots, x_n)$ un échantillon de variables aléatoires indépendantes et identiquement distribuées de densité $f(x_i, \theta)$. La log-vraisemblance de x est définie par :

$$L(\theta, x) = \log\left(\prod_{i=1}^{n} f(x_i, \theta)\right) = \sum_{i=1}^{n} \log f(x_i, \theta) \tag{2.6}$$

ou encore, en remplaçant $f(x_i, \theta)$ par sa valeur on a :

$$L(\theta, x) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{g} \pi_k \varphi_k(x_i; \alpha_k)\right) \tag{2.7}$$

Pour définir la log-vraisemblance complétée, on suppose que l'échantillon x provient d'une population P dont une partie z nous est cachée. z est appelé *information manquante*. Dans le cadre de notre étude z correspond à la partition $(z_1, \cdots, z_n)$ que l'on cherche. Elle est liée à x sous la forme $y = (x, z)$ avec $(x, z) = ((x_1, z_1), \cdots, (x_n, z_n))$. y est appelé *données complétées*. La log-vraisemblance complétée est définie à partir des données complétées par :

$$L(\theta, y) = \sum_{i=1}^{n} \log f(y_i, \theta)$$

En remarquant que $f(y; \theta) = f(y, x; \theta) = f(y|x; \theta) f(x; \theta)$ où $x = h(y)$ cette dernière relation peut encore s'écrire :

$$\begin{aligned} L(\theta, y) &= \sum_{i=1}^{n} \log\left(f(y_i|x_i, \theta) f(x_i, \theta)\right) \\ &= L(\theta, x) + \sum_{i=1}^{n} \log f(y_i|x_i, \theta) \\ &= L(\theta, x) + log f(y|x, \theta) \qquad \forall y \in h^{-1}(x) \end{aligned}$$

Où $f(y_i|x_i, \theta)$ désigne densité de probabilité de y_i sachant x_i et θ.

Par ailleurs nous avons :

$$f(y_i, \theta) = f(x_i, z_i, \theta) = \pi_{z_i} \varphi(x_i, \alpha_{z_i}) \tag{2.8}$$

où π_{z_i}, α_{z_i} correspondent respectivement à la probabilité de la classe z_i à laquelle appartient l'individu i ; et les paramètres de la densité de probabilité de cette

classe.
En remplaçant, dans $L(\theta, y)$, $f(y_i, \theta)$ par sa valeur (2.9) nous obtenons.

$$\begin{aligned} L(\theta, y) &= \sum_{i=1}^{n} \log f(y_i, \theta) \\ &= \sum_{i=1}^{n} \log\left(\pi_{z_i} \varphi(x_i, \alpha_{z_i})\right) \\ &= \sum_{i,k} z_{ik} \log\left(\pi_k \varphi(x_i, \alpha_k)\right) \text{(Scott et Symons)} \end{aligned}$$

Pour retrouver les paramètres estimés du modèle par *maximum de vraisemblance*[4] on cherche les paramètres θ qui maximisent la probabilité à posteriori $f(x|\theta)$. Cela revient à chercher θ qui maximise la vraisemblance.

$$\hat{\theta}_{MV} = arg\ max_\theta\ L(\theta, x) \tag{2.9}$$

Étant donné que la résolution directe des *équations de vraisemblance*[5] dans le cadre des modèles de mélange ne conduit généralement pas à une solution analytique, on utilise l'algorithme EM Dempster *et al.* (1977) pour estimer le maximum de vraisemblance.

EM est un algorithme itératif qu'on initiale avec une valeur initiale choisie arbitrairement des paramètres $\theta^{(0)}$. Connaissant les paramètres à l'étape p on estime les paramètres à l'étape $p+1$ en cherchant $\theta^{(p+1)}$ de façon à maximiser l'espérance de la log-vraisemblance complétée conditionnellement à x et $\theta^{(p)}$.

$$\theta^{(p+1)} = arg\ max_\theta\ Q\left(\theta, \theta^{(p)}\right) \tag{2.10}$$

avec

$$Q\left(\theta, \theta^{(p)}\right) = E\left(L(\theta; y)|x, \theta^{(q)}\right) \tag{2.11}$$

Un lancer de l'algorithme pour le modèle de mélange gaussien est donné par

---

Algorithme

$q \leftarrow 0$

`Initialiser les proportions, centres et variances à une valeur initiale` $\theta^{(0)}$

`répéter`

`* Etape E : calculer les probabilités d'appartenance` $t_{ik}^{(q+1)}$

`pour i allant de 1 à n ; k de 1 à` g `faire`

---

[4] L'estimateur du maximum de vraisemblance est asymptotiquement sans biais, efficace et a une distribution normale

[5] ces équations s'obtiennent en posant que les dérivées partielles de la log-vraisemblance par rapport au vecteur de paramètres θ sont égales à zéro

$t_{ik}^{(q+1)} \leftarrow \frac{\pi_k^{(q)} \varphi(x_i|\alpha_k^{(q)})}{\sum_{j=1}^{g} \pi_k^{(q)} \varphi(x_i|\alpha_h^{(q)})}$
`* Etape M : recalculer les paramètres` $\theta^{(q+1)}$

`pour k allant de 1 à` g `faire`
$n_{ik}^{(q+1)} \leftarrow \sum_{i=1}^{n} t_{ik}^{(q+1)} \qquad \pi_k^{(q+1)} \leftarrow \frac{n_k^{(q+1)}}{n}$
$\mu_k^{(q+1)} \leftarrow \frac{1}{n_k^{(q+1)}} \sum_{i=1}^{n} t_{ik}^{(q+1)} x_i$;
$\Sigma_k^{q+1} \leftarrow \frac{1}{n_k^{(q+1)}} \sum_{i=1}^{n} t_{ik}^{(q+1)} \left(x_i - \mu_k^{(q+1)}\right) \left(x_i - \mu_k^{(q+1)}\right)'$
$q \leftarrow q + 1$
`jusqu'à` $t^{(q)} \approx t^{(q+1)}$ $[ou \quad q = QMAX]$
$\hat{\theta} \leftarrow \hat{\theta}^{(q)}$

2.1.6 Critères de sélection de Modèle

La définition des critères de sélection de modèle dans les approches paramétriques (probabilistes) dans un problème de modélisation peut être résumé comme suit :

On dispose d'un échantillon $x = (x_1, \cdots, x_n)$ de taille n de variables indépendantes identiquement distribuées de densité f inconnue. Pour estimer f, on se donne une famille de modèles $\mathcal{M} = \{M_1, \cdots, M_p\}$. Un modèle M_i est défini par le couple (m_i, g_i) qui correspondent respectivement à une des deux variantes $[\pi, \lambda I]$ ou $[\pi_k, \lambda I]$ du modèle de mélange gaussien et le nombre de classes dans le mélange. Il s'agit de choisir un modèle dans cette famille de modèles. Dans la suite, on distinguera les deux sous familles où $m_i = [\pi, \lambda I]$ ou $m_i = [\pi_k, \lambda I]$.

Les critères d'information de sélection de modèle BIC, AIC, ICL sont basés sur la vraisemblance ou la vraisemblance complétée du modèle à k classes. Étant donné que la vraisemblance ou la vraisemblance complétée augmentent avec le nombre de classe, on ne peut pas se baser uniquement sur elles. C'est pourquoi dans ces trois critères on les pénalise avec le nombre paramètres à estimer. Pour un modèle M_i, leur formulation générale est donnée par :

$$C(M_i) = -2\left(max_{f_i \in M_i}\ L\left(f_i\right)\right) + \gamma_C * \nu\left(M_i\right) \tag{2.12}$$

où
- $L\left(f_i\right)$ est la vraisemblance pour une densité $f_i \in M_i$ caractérisée par ses paramètres θ ;
- γ_C est un coefficient de pénalisation de la complexité spécifique à chaque critère ;
- $\nu\left(M_i\right)$ est le nombre de paramètres à estimer dans le modèle.

Avec ces critères, on cherche à s'approcher le plus possible du «vrai modèle» dont la définition diffère selon les critères. Pour BIC et ICL, «le vrai modèle» est le modèle M_i qui maximise la probabilisé à posteriori $P\left(M_i|x\right)$. Alors que pour AIC c'est le modèle qui a généré les données. Chaque critère retient le modèle qui sa valeur minimale.

Minimiser l'expression (2.12) revient à réaliser un compromis entre maximiser la vraisemblance et minimiser la complexité du modèle. On peut utiliser ces

critères dans le cadre particulier de notre étude, puisque choisir un nombre de classe revient à choisir un modèle. Dans la suite nous identifierons les densités $f_i \in M_i$ par leurs paramètres θ.

Les deux critères BIC et ICL se placent dans un cadre bayésien : θ_i et M_i sont considérés comme des variables aléatoires munies d'une distribution à priori. La distribution à priori de M_i est notée $P(M_i)$. Pour un modèle M_i fixé, la distribution à priori de θ_i est notée $P(\theta_i|M_i)$. On peut ainsi donner des poids plus importants à certains modèles si on possède des informations à priori, sinon, la distribution des modèles M_i est généralement considérée uniforme. Avec des considérations asymptotiques la distribution à priori des θ_i n'intervient pas dans la la forme de nos critères. Ces critères cherchent à sélectionner respectivement un modèle M_i maximisant une certaine quantité. Nous définissons leur expressions dans les paragraphes suivants.

Avant d'aborder la définition des critères de sélection, rappelons quelques définitions dont on aura besoin.

Definition 2. ***biais d'un estimateur***
Soit $f(.,\theta)$ une fonction de densité de paramètre inconnu. Soit $\hat{\theta}$ un estimateur de ce paramètre. Le biais de $\hat{\theta}$ est défini par la quantité

$$biais\left(\hat{\theta}\right) = E\left[\hat{\theta}\right] - \theta$$

Il mesure l'écart systématique (non aléatoire) qu'existe entre l'estimateur $\hat{\theta}$ et la vraie valeur du paramètre θ. Un estimateur est dit non biaisé (sans biais) si son biais est nul.

Definition 3. ***Pseudo-distance de Kullback-Leibler*** *Soient f_i, f_j respectivement les densités des modèles M_i et M_j. La pseudo-distance de Kullback-Leibler entre f_i et f_j est définie par :*

$$d_{KL}(f_i, f_j) = \int_{\Omega} \log\left(\frac{f_i(x)}{f_j(x)}\right) f_i(x)\, dx$$

Definition 4. ***"Quasi-vrai" modèle***
Soit M_i un modèle de la famille $\mathcal{M}$. M_i est dit "quasi-vrai" modèle Lebarbier et Mary-Huard (2004) au vu de X si la limite quand n tend vers ∞ de $P(M_i|X)$ tend vers 1.

Bayesian Information Criterion (BIC)

Le critère BIC Schwarz (1978) cherche le modèle M_{BIC} qui maximise la probabilité à posteriori $P(M_i|x)$.

$$M_{BIC} = argmax_{M_i} P(M_i|x)$$

Nous donnons ici son expression pour un modèle M_i, sans rentrer dans sa construction Lebarbier et Mary-Huard (2004).

$$BIC(M_i) = -2L(\widehat{\theta_i}; x) + \nu_i \log(n)$$

où $\widehat{\theta_i}$ et ν_i sont respectivement l'estimateur du maximum de vraisemblance de θ_i (les paramètres du modèles M_i) et le nombre de paramètres libres.

Le critère BIC cherche à choisir le modèle plus probable au vu des données. Il cherche à s'approcher le plus possible au "quasi-vrai" modèle. Toute fois le "quasi-vrai" modèle peut être très éloigné du vrai modèle au sens de Kullback-Leibler si ce dernier n'est pas dans la famille des modèles. Dans ce cas une probabilité à posteriori aussi élevée soit-elle ne veut rien dire.

Akaike Information Criterion (AIC)

Soient M_i un modèle de la famille $\mathcal{M}$ et x un échantillon de vecteurs aléatoires indépendantes et identiquement distribués de la fonction densité f inconnue. Soit $\hat{\theta}_i$ l'estimateur du maximum de vraisemblance de M_i. Notons par f_{M_i} et ν_i respectivement la densité de l'estimateur du maximum de vraisemblance de M_i et le nombre de paramètres de M_i .

Le critère AIC Akaike (1973) cherche le modèle M_{AIC} pour lequel la densité du maximum de vraisemblance est la plus proche à la fonction densité f qui a généré les données, au sens de la distance de Kullback-Leibler.

$$M_{AIC} = argmin_{M_i} E(d_{KL}(f, f_{M_i}))$$

Pour un modèle M_i son expression est donnée par :

$$AIC(M_i) = -2L(\widehat{\theta_i}; x) + 2\nu_i$$

Integrated Completed Likelihood (ICL)

Le critère ICL Biernacki *et al.* (2000) est calculé à partir de la vraisemblance complétée. Avec ce critère on cherche le modèle M_{ICL} qui maximise la probabilité à posteriori $P(M_j|y)$ avec $y = (x, z)$. Il pénalise la vraisemblance complétée. La vraisemblance complétée d'un modèle M_j est définie par :

$$L_C(\widehat{\theta_j}; x, \widehat{z}) = L(\widehat{\theta_j}; x) + \sum_{i,k} \hat{z}_{ik} \log t_{ik}$$

où

- $\widehat{z} = MAP\left(\widehat{\theta_j}\right)$ le maximum à posteriori de z. Pour un individu i, le MAP consiste à trouver la classe k qui maximise la probabilité à posteriori $p\left(z_i = k|x_i, \hat{\theta}_j\right)$

$$\hat{z}_i(x_i) = arg\ max_k\ p(z_i = k|x_i, \hat{\theta}_j)$$

- $\hat{z}_{ik}$ égal à 1 si l'individu i appartient à la classe k ; et zéro sinon.
- $t_{ik} = p\left(z_i = k|x_i, \hat{\theta}_j\right) = \frac{\pi_k \varphi(x_i|\alpha_k)}{\sum_{l=1}^{g} \pi_l \varphi(x_i|\alpha_l)}$ est la probabilité que l'individu i appartienne à la classe k connaissant θ_j

Le critère ICL est donnée par :

$$ICL(M_j) = -2L_C(\widehat{\theta_j}; x, \widehat{z}) + \nu_j \log(n)$$

ICL est lié au critère BIC par :

$$ICL\left(M_j\right) = BIC\left(M_j\right) + \sum_{ik} \hat{z}_{ik} \log t_{ik}$$

2.2 Le bootstrap dans la classification

Pour le calcul des critères bootstrap, revenons sur le contexte des modèles de mélange. Dans le chapitre précédent, nous avons fait l'hypothèse que les données observées constituent un échantillon aléatoire de taille n de vecteurs aléatoires identiques et indépendamment distribués. Cet échantillon est tiré d'une population P de distribution $f\left(., \theta\right)$. Ensuite, nous avons supposé que f faisait partie de la famille de modèles de mélange. Pour estimer les paramètres de ces modèle nous avons utilisé la méthode du maximum de vraisemblance. On calculait ensuite la vraisemblance et/ ou la vraisemblance complétée de ces paramètres estimés.

Comme les critères de sélection de modèle précédents nous définissons ici des critères basés sur les estimateurs de la vraisemblance à partir de la méthode statistique *bootstrap*.

Ici le bootstrap est utilisé pour corriger l'erreur de l'estimateur de la vraisemblance ou de la vraisemblance complétée. Avant cela, expliquons en quoi consiste cette technique.

Le mot bootstrap provient de l'expression anglaise « to pull oneself up by one's bootstrap »Efron et Tibshirani (1993), qui signifie littéralement « se soulever en tirant sur les languettes de ses bottes ». Le bootstrap fait partie des techniques d'inférence statistique basées sur l'utilisation des ordinateurs.

2.2.1 Principe du bootstrap

Soit $x = (x_1, \cdots, x_n)$ un échantillon aléatoire de vecteurs identiques et indépendamment distribués de f. Le principe de la méthode du bootstrap est de considérer l'échantillon x comme la population tout entière et d'y prélever B échantillons de taille n chacun par des tirages uniformes avec remise parmi les n observations. Les échantillons ainsi obtenus sont appelés des *échantillons bootstrap*.

Soit M_i un modèle de famille $\mathcal{M}$ et θ_i les paramètres de M_i. Dans la suite de ce chapitre nous utiliserons les notations suivantes :

- $x = (x_1, \cdots, x_n)$: l'échantillon de départ
- $x_1^*, x_2^*, \cdots, x_B^*$: les B échantillons bootstrap, avec $x_b^* = (x_{b1}^*, x_{b2}^*, \ldots, x_{bn}^*)$
- $\hat{\theta}_i$ estimateur du maximum de vraisemblance de θ_i
- $\hat{\theta}_i^b$: les paramètres estimés sur l'échantillon bootstrap x_b^*
- $L(\hat{\theta}_i, x)$: la vraisemblance de $\hat{\theta}_i$, calculée à partir des données de départ
- $L\hat{\theta}_i^b, x)$: la vraisemblance de $\hat{\theta}_i^b$, calculée à partir des données de départ
- $L_C(\hat{\theta}_i, x, \hat{z})$: la vraisemblance classifiante de $\hat{\theta}_i$, calculée à partir des données de départ
- $L_C(\hat{\theta}_i^b, x, \hat{z})$: la vraisemblance classifiante de $\hat{\theta}_i^b$, calculée à partir des données de départ
- $L(\hat{\theta}_i^b, x_b^*)$: la vraisemblance de $\hat{\theta}_i^b$, calculée à partir de x_b^*

– $L_C(\hat{\theta}_i^b, x_b^*, \hat{z}_b)$: la vraisemblance classifiante de $\hat{\theta}_i^b$, calculée à partir de x_b^*

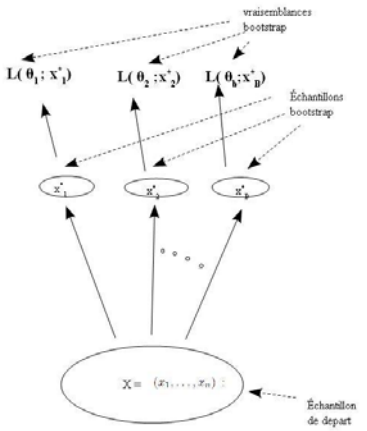


FIG. 2.2 – Schéma d'une procédure bootstrap

2.2.2 Critères de sélection bootstrap

Les critères bootstrap sont basés sur la correction de l'erreur de l'estimateur de la vraisemblance ou de la vraisemblance complétée de θ, paramètres du modèle M, à partir de l'échantillon x.

Critères bootstrap naïfs

Ce premier critère correspond à la correction bootstrap « naïve »de la vraisemblance. Considérons un modèle $M_i \in \mathcal{M}$ de paramètres θ_i. On estime les $\hat{\theta}_i^b$, estimateurs de θ_i sur les échantillons bootstrap x_b^*. On calcule ensuite les vraisemblances $L(\hat{\theta}_i^b; x)$ de $\hat{\theta}_i^b$ sur les données de départ x. Le critère bootstrap naïf de M_i noté par $C_{\text{naïf}}(M_i)$ est donné par :

$$C_{\text{naïf}}(M_i) = \frac{1}{B}\sum_{b=1}^{B} L(\hat{\theta}_i^b; x)$$

De la même manière on calcule le critère $CC_{\text{naïf}}(M_i)$ en remplaçant la vraisemblance par la vraisemblance classifiante. Il est donné par :

$$CC_{\text{naïf}}(M_i) = \frac{1}{B}\sum_{b=1}^{B} L_C(\hat{\theta}_i^b; x, \hat{z})$$

On retient avec ces critères le modèle M_i qui maximise leurs quantités.

Critères bootstrap basés sur l'optimisme

Le critère suivant qu'on utilise ici est le critère bootstrap basé sur «l'optimisme» que l'on note C_{opt}. Nous avons utilisé l'échantillon x dans le calcul du maximum de vraisemblance $\hat{\theta}$ et ce même échantillon nous a servit à calculer la vraisemblance de ces paramètres obtenus. (Efron et Tibshirani, 1993)

définissent «l'optimisme» par la quantité par laquelle on sous estime la vraisemblance de θ. Une première estimation de l'optimisme consiste à faire la moyenne des différences entre les vraisemblances des $\hat{\theta}_i^b$ sur x et celles des $\hat{\theta}_i^b$ sur les x_b^*.

$$C_{opt}(M_i) = L(\widehat{\theta}_i; x) + \frac{1}{B}\sum_{b=1}^{B}(L(\hat{\theta}_i^b; x) - L(\hat{\theta}_i^b; x_b^*))$$

On utilise la même procédure de correction sur les vraisemblances classifiantes dans la variante CC_{opt} du bootstrap basé sur l'optimisme. Le CC_{opt} est défini par

$$CC_{opt}(M_i) = L_c(\widehat{\theta}_i; x, \hat{z}) + \frac{1}{B}\sum_{b=1}^{B}(L_c(\hat{\theta}_i^b; x, \hat{z}) - L_c(\hat{\theta}_i^b; x_b^*, \hat{z}_b))$$

Critères bootstrap 632

Les deux derniers critères bootstrap C_{632} et CC_{632} s'inscrivent aussi dans le cadre de la correction de l'estimateur de la vraisemblance. Une solution correction est de ne prendre en compte dans le calcul des vraisemblance et vraisemblances classifiantes des $\hat{\theta}_b$ que les seuls individus qui n'ont pas servi dans leur estimation. Cette dernière solution est biaisée par pessimisme. La probabilité pour qu'un individu x_i soit tiré dans un échantillon de taille n est égale à : $1 - \left(1 - \frac{1}{n}\right)^n$ qui tend vers 0.632 quand n tend vers ∞. (Efron et Tibshirani, 1993) proposent une solution de compromis entre ces deux situations en faisant intervenir la valeur .632. Ces critères donnent pour un modèle M_i les formules suivantes :

$$C_{632}\,(M_i) = L(\widehat{\theta}_i; x) + 0.632\sum_{j=1}^{n}\frac{1}{|B_j|}\sum_{b\in B_j}\left(L(\hat{\theta}_i^b; x_j) - L(\widehat{\theta}_i; x_j)\right)$$

$$CC_{632}\,(M_i) = L_C(\widehat{\theta}_i; x, \hat{z}) + 0.632\sum_{j=1}^{n}\frac{1}{|B_j|}\sum_{b\in B_j}\left(L_C(\hat{\theta}_i^b; x_j, \hat{z}_i) - L_C(\widehat{\theta}_i; x_j, \hat{z}_j)\right)$$

avec

$$B_j = \{b \in \{1, \ldots, B\}\,/x_j \notin x_b^*\}$$

2.2.3 Critères basés sur les partitions

Reprenons l'ensemble $E = \{1, \cdots, i, \cdots, n\}$, où chaque individu i est caractérisé par un vecteur de variables $x_i = (x_{i1}, \cdots, x_{ip})$. Rappelons que le but principal de la classification est de former des classes avec les individus de E. On suppose avec le modèle de mélange qu'il existe des partitions $P = \{C_1, \cdots, C_g\}$ en g classes. Chaque classe est alors caractérisée par une distribution gaussienne. On se pose la question de savoir quelle partition reflète la réalité. On aimerait bien avoir "la vrai partition v" des individus, s'il y en a, pour ensuite le comparer aux partitions engendrées par les modèles de mélange. Cette partition, si elle existe, nous est inconnu. Il faut donc trouver un autre moyen pour trouver le modèle qui s'approche le plus à cette partition.

Puisqu'on ne peut pas comparer les partitions générées par les modèles avec la vrai partition, on va les étudier en terme de stabilité. Après avoir estimer les paramètres $\hat{\theta}_b$ du modèle sur les échantillons bootstrap, on peut créer des partions P_b de $E = \{x_1, \cdots, x_i, \cdots, x_n\}$ avec ces nouveaux paramètres. On compare ces partitions avec celle obtenue avec $\hat{\theta}$ sur les données de départ. Ces comparaisons nous donnent une idée sur la stabilité de chaque modèle.

Nous présentons dans cette section deux critères de comparaison de partitions, *l'indice de Rand* et *l'information mutuelle.*

Indice de Rand

Soient $P = \{C_1, \cdots, C_g\}$ et $P_b = \{C_{b1}, \cdots, C_{bg}\}$ deux partitions de E en g classes. Nous définissons par $TabCont_{\{P,P_b\}}$, la table de contingence ou encore table des accords-désaccords entre P et P_b.

$P \setminus P_b$	C_{b1}	$\cdots$	C_{bj}	$\cdots$	C_{bg}	
C_1			$\vdots$			
$\vdots$						
C_i	$\cdots$		n_{ij}		$\cdots$	$n_{i.}$
$\vdots$						
C_g			$\vdots$			
			$n_{.j}$			$n_{..} = n$

TAB. 2.2 – table de contingence entre P et P_b

Pour définir l'indice de Rand, on considère a, d, c, et b les nombres qui représentent respectivement les paires d'individus qui sont dans la même classe pour P et P_b, les paires d'individus qui sont dans deux classes différentes dans P et P_b, les paires d'individus qui sont classés dans une même classe dans P mais dans deux classes différentes selon P_b et les paires d'individus qui sont dans deux classes différentes selon P et dans une même pour P_b.

$$a = \sum_{ij} \binom{n_{ij}}{2}$$

$$b = \sum_{i} \binom{n_{i.}}{2} - \sum_{ij} \binom{n_{ij}}{2}$$

$$c = \sum_{j} \binom{n_{.j}}{2} - \sum_{ij} \binom{n_{ij}}{2}$$

$$d = \binom{n}{2} + \sum_{ij} \binom{n_{ij}}{2} - \sum_{i} \binom{n_{i.}}{2} - \sum_{j} \binom{n_{.j}}{2}$$

	$\hat{m}$ classe dans P_b	$\neq$ classes dans P_b
$\hat{m}$ classe dans P	a	b
$\neq$ classes dans P	c	d

TAB. 2.3 – table des accords-désaccords

L'indice de Rand mesure le pourcentage des accords entre les deux partitions P et P_b.

$$R(P, P_b) = \frac{a+d}{a+b+c+d}$$

Il prend ses valeurs dans $[0, 1]$. Si $R(P, P_b) = 1$ alors P, P_b désignent la même partition et si $R(P, P_b) = 0$ alors P, P_b sont en désaccord total.

Dans certaines configurations de partitions ce critère n'est jamais nul. Dans ce travail nous utilisons l'indice de Rand corrigé Rc pour pallier à ce problème d'échelle.

$$Rc(P, P_b) = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{n}{2}}}$$

Information mutuelle

Le deuxième critère de cette partie est *l'information mutuelle* dans sa version ajustée. Elle mesure la dépendance entre deux partitions P et Q de l'ensemble E. Pour pouvoir l'utiliser comme critère de sélection, on va mesurer la stabilité des modèles en calculant la dépendance entre une partition P, des données de départ, engendrée par l'estimateur $\hat{\theta}$ des paramètres θ du modèle et les partitions bootstrap P_b, des données de départ, engendrées par les estimateurs bootstrap $\hat{\theta}_b$ de θ.

Considérons la table de contingence $TabCont_{\{P,P_b\}}$ 2.2 définie dans le paragraphe précédent. On définit les probabilités des classes de la partition P par : $P(C_i) = \frac{n_{i.}}{n}$ et celles des classes de P_b par : $P(C_j) = \frac{n_{.j}}{n}$. La probabilité pour qu'un individu fasse partie du couple (C_i, C_j) de classes de P et P_b est donnée par :$P(C_i, C_j) = \frac{n_{ij}}{n}$.

L'information mutuelle MI et sa version ajustée AMI entre P et P_b sont données par les formules suivantes :

$$MI(P, P_b) = \sum_{i=1}^{g} \sum_{j=1}^{g} P(C_i, C_j) \log \frac{P(C_i, C_j)}{P(C_i) P(C_j)}$$

$$AMI(P, P_b) = \frac{MI(P, P_b) - E\{MI(P, P_b)\}}{max\{H(P), H(P_b)\} - MI(P, P_b)}$$

où

$H(P) = -\sum_{i=1}^{g} P(C_i) \log P(C_i)$, mesure l'entropie de la partition P

$H(P_b) = -\sum_{j=1}^{g} P(C_j) \log P(C_j)$, mesure l'entropie de la partition P_b

et

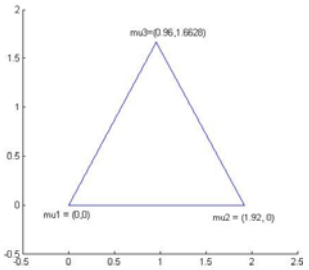
$$E\left\{MI\left(P,P_b\right)\right\}=\sum_{i=1}^{g}\sum_{j=1}^{g}\sum_{n_{ij}=(a_i+b_j-n)^+}^{\min(a_i,b_j)}\frac{n_{ij}}{n}\log\left(\frac{n*n_{ij}}{a_ib_j}\right)\frac{a_i!b_j!\left(n-a_i\right)!\left(n-b_i\right)!}{n!n_{ij}!\left(a_i-n_{ij}\right)!\left(b_j-n_{ij}\right)!\left(n-a_i-b_j+n_i\right.}$$

avec $\left(a_i+b_j-n\right)^+=\max\left(0,a_i+b_j-n\right)$.

2.3 Application des critères de sélection

2.3.1 Présentation des données

Afin d'appliquer les critères de sélection de modèles, nous allons générer des données selon des modèles de mélange de trois classes. Considérons pour cela le triangle équilatéral Δ de sommets $\mu_1=(0,0),\mu_2=(1.92,0),\mu_3=(0.96,1.6628)$.



Nous procédons par expérience[6]. Dans ces expériences j'utilise les modèles de mélange parcimonieux à volume identique $\Sigma_k=\lambda I$ de la famille sphérique , pour générer des données autour des trois sommets du triangle Δ. Les paramètres d'un tel modèle peuvent être décrits par $\theta=(\pi_1,\pi_2,\pi_3,\mu_1,\mu_2,\mu_3,\lambda)$. Selon les expériences, ces classes seront considérées comme étant de proportions égales (équiprobables) ou pas.

Nous définissons les mélanges à classes bien séparées , modérément et mal séparées respectivement par les taux d'erreur à 7.5%, 18% et 30%. Les volumes λ correspondant à ces taux d'erreu sont $\lambda=0.313$ pour les classes bien séparées, $\lambda=0.5986$ pour les classes modérément séparées et enfin $\lambda=1.21$ pour les classes mal séparées.

Les conditions de simulation pour chaque expérience sont définies dans un fichier *exp*. Ces conditions correspondent au modèle de mélange utilisé, aux paramètres $\theta=(\pi_1,\pi_2,\pi_3,\mu_1,\mu_2,\mu_3,\lambda)$ de ce modèle, le nombre d'individus n dans le mélange, le nombre d'échantillons bootstrap B à générer, la liste $liste\_g$ du nombre de classes pour chaque estimation, le nombre de simulations $nsimul$ et enfin le modèle $model\_algo$ qui servira pour l'algorithme EM. Le contenu d'un fichier *exp* ressemble à ceci :

– theta.model = $[pi, lambda, i]$;

[6]Une expérience correspond à la définition d'un fichier *exp*, à l'estimation des paramètres des modèles et des partitions, aux calculs et comparaisons des critères de sélection

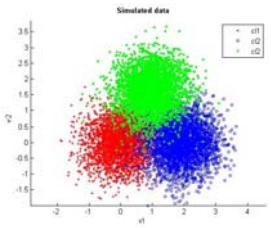


FIG. 2.3 – Nuage de points (classes bien séparées à proportions identiques)

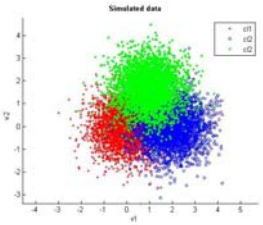


FIG. 2.4 – Nuage de points (classes modérément séparées à proportions identiques)

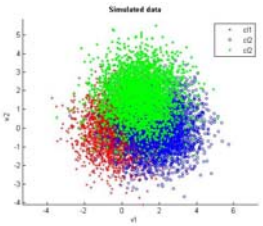


FIG. 2.5 – Nuage de points(classes mal séparées à proportions identiques)

- theta.mu = [0 0; 1.92 0; 0.96 1.6628] ;
- theta.lambda = 0.313 ;
- n = 200 ;
- B = 30 ;
- liste_g = [1 :4] ;
- model_algo = $[pi\_k, lambda, i]$;
- nsimul = 200 ;

Une fois les conditions expérimentales établies, nous allons générer avec la fonction *GaussMixt_ experiment* les données de chaque simulation en deux étapes.

échantillon de départ : En un premier nous faisons appel à la fonction *GaussMixtRnd* pour simuler les données *data* où chaque individu est répresententé par ses coordonnées (x1, x2) dans R^2 et son numéro de classe z.

..	x1	x2	z
1	0.7180	2.0182	3.0000
2	-0.9318	-0.3365	1.0000
3	1.9901	0.3084	2.0000
4	2.0809	-0.6153	2.0000
5	0.3186	1.7109	3.0000
10	2.0177	0.6918	2.0000
..	...	....	...
196	-0.1901	0.3836	1.0000
197	0.3223	1.3065	3.0000
198	-0.1181	-0.5609	1.0000
199	0.6659	-0.1038	1.0000
200	0.3355	1.0731	3.0000

TAB. 2.4 – Tableau de données

A partir de cet échantillon, nous tirons 30 échantillons bootstrap dont les indices des individus retenus sont stockés dans la matrice Ind_b.

Estimation des paramètres : Dans cette étape nous faisons successivement l'hypothèse que les données de l'échantillon de départ ainsi celles des échantillons bootstrap ont été générées par un modèle de mélange $\Sigma_k = \lambda I$ à 1 , 2, 3 ou 4 classes à proportions différentes. Nous estimons ensuite les paramètres de ces modèles et les partitions à posteriori associées $\hat{z}$ avec l'algorithme EM.

Comme il a été souligné dans le deuxième chapitre, cet algorithme cherche le maximum de la vraisemblance en partant d'une position initiale θ_0 choisie au hasard. La solution trouvée est un maximum local. La solution du EM dépend du point de l'algorithme départ. Pour diminuer cette dépendance, nous lançons EM à 20 reprises, avec une nouvelle position initiale à chaque fois. On retient la solution $\hat{\theta}$ qui donne maximise la vraisemblance.

Pour l'estimation des paramètres correspondant aux échantillons bootstrap, nous initialisons l'algorithme EM avec la solution de l'estimation des paramètres du modèle de l'échantillon de départ, et on itère l'algorithme jusqu'à convergence vers une solution.

Pour la convergence de l'algorithme, nous utilisons une technique de test de stabilisation de la vraisemblance. Entre deux itérations q et $q+1$ on vérifie la relation

$$|\frac{L(\theta^{q+1}) - L(\theta^q)}{L(\theta^q)} < \epsilon \text{ avec} \epsilon = 10^{-16}.$$

Si cette relation est vérifiée on sort de l'algorithme sinon on continue jusqu'au nombre d'itération maximum $QMAX = 1000$.

Nous utilisons ensuite ces paramètres estimés pour classifier les individus par MAP (Maximum à posteriori).

Ces estimations sont effectuées avec la fonction *GaussMixtXem*. À la suite de cette étape nous créons 200 fichiers *result.sav*, dans lesquels on sauvegarde les données des 200 simulations. En résumé un fichier *result.sav* contient :

- *data* données et partition de l'échantillon de départ
- Ind_b matrices des indices des individus retenus dans les échantillons boostrap
- *res*0 estimation des paramètres des modèles à 1, 2, 3 et 4 classes ; les partitions MAP $\hat{z}$ correspondantes à l'échantillon de départ
- *resb* estimation des paramètres des modèles à 1, 2, 3 et 4 classes ; les partitions MAP $\hat{z}$ correspondantes aux 30 échantillons bootstrap

A partir de là, nous possédons la quasi-totalité des informations requises pour la calcul des critères. Ces calculs constituent la deuxième phase de chaque expérience. La fonction *GaussMixt_critrs* nous donne en sortie un fichier "*criters.sav*" dans lequel on sauvegarde les matrices suivantes :

- tab_err (200×6) matrice des taux d'erreurs pour les modèles à 1, 2, 3 , 4, 5 et 6 classes.
- tab_Loglik (200×6) matrice des log-vraisemblances pour les modèles à 1, 2, 3 , 4, 5 et 6 classes.
- $tab\_LoglikC$ (200×6) matrice des log-vraisemblances complétées pour les modèles à 1, 2, 3 , 4, 5 et 6 classes.
- tab_BIC (200×6) matrice des critères BIC pour les modèles à 1, 2, 3 , 4, 5 et 6 classes.
- tab_AIC (200×6) matrice des critères AIC pour les modèles à 1, 2, 3 , 4, 5 et 6 classes.
- tab_ICL (200×6) matrice des critères ICL pour les modèles à 1, 2, 3 , 4, 5 et 6 classes.
- tab_Cnaive (200×6) matrice des critères bootstrap C_{naive} pour les modèles à 1, 2, 3 , 4, 5 et 6 classes.
- tab_Copt (200×6) matrice des critères bootstrap optimistes C_{opt} pour les modèles à 1, 2, 3 , 4, 5 et 6 classes.
- tab_C632 (200×6) matrice des critères bootstrap 632 C_{632} pour les modèles à 1, 2, 3 , 4, 5 et 6 classes.

Les fonctions *GaussMixt_MutInfo* et *GaussMixt_IndRand* nous permettent de calculer les moyenne des informations mutuelles et indice de Rand entre les partitions P et 30 partitions bootstrap P_b pour chaque simulation d'une expérience. Ces valeurs sont sauvegardées dans les fichiers *MutInfo.sav* et *IndRand.sav* dans lesquelles on retrouve les matrices $MEAN\_MAP$. Ces ma-

trices comportent 200 lignes et 6 colonnes. En plus de ces matrices, les deux fichiers *MutInfo.sav* et *IndRand.sav* contiennent respectivement les matrices $MINFO$ et $IRAND$, où l'élément $MINFO\,(i,j)$ respectivement $IRAND\,(i,j)$ sont les calculs de l'information mutuelle et l'indice de Rand entre les partitions MAP, des estimateurs $\hat{\theta}$ pour chaque modèle et la vraie partition z. Dans la réalité, on ne peut pas effectuer ces derniers calculs puisqu'on ne connaît pas la vrai partition z. Ils nous donnent une idée sur la qualité de l'estimation.

	1	2	3	4
1	0.6450	0.5200	0.0600	0.3500
2	0.5900	0.4850	0.1050	0.3950
3	0.6250	0.4250	0.0500	0.2750
4	0.6100	0.4100	0.0950	0.1100
5	0.6400	0.5100	0.0800	0.3550
...	...	...	...	...
196	0.6450	0.4100	0.0700	0.5050
197	0.6450	0.4450	0.0950	0.4250
198	0.6100	0.4100	0.1100	0.4550
199	0.6450	0.3750	0.0350	0.4300
200	0.6450	0.5450	0.0850	0.3850

TAB. 2.5 – Tableau des taux d'erreurs

	1	2	3	4
1	-186.6269	-192.3024	-159.6856	-165.5744
2	-177.5568	-182.4355	-165.0238	-172.2762
3	-196.8803	-203.3319	-183.1599	-184.9732
4	-190.5991	-194.3462	-169.4902	-172.9438
5	-194.0298	-200.8216	-165.7645	-171.6818
...	...	...	...	...
196	-210.4457	-214.2946	-178.8555	-184.7978
197	-183.2272	-190.0925	-161.3387	-166.2061
198	-187.8983	-193.4768	-175.2798	-181.7450
199	-193.0709	-199.8442	-161.2007	-166.5749
200	-189.7963	-194.6474	-173.2831	-178.1051

TAB. 2.6 – Tableau des critères BIC

Pour chaque expérience on crée un dossier *exp(i)* qui porte son numéro et contient les fichiers crées ci-dessus. Nous étudions les performances des critères de sélection selon la complexité du modèle de base.

Dans les expériences qui vont suivre, nous gardons certaines valeurs des conditions expérimentales fixes. Il s'agit de :
theta.mu = $[0\ 0; 1.92\ 0; 0.96\ 1.6628]$;
n = 200 ; B = 30 ; nsimul = 200 ; liste_g = $[1:6]$;
model_algo = $[\pi_k, \lambda, i]$;

2.3.2 Expérience 1 : 6 classes à proportion identiques et bien séparées

Nous utilisons dans cette première expérience le modèle theta.model = $[\pi, \lambda, i]$; et $\lambda = 0.313$. On trouve les résultats suivants :

critères \ g	1	2	3	4	5	6
BIC	0	0	199	1	0	0
AIC	0	0	138	37	18	7
ICL	53	0	146	1	0	0
C_{naive}	0	0	44	53	47	56
C_{opt}	0	0	147	38	11	4
C_{632}	0	0	155	36	6	3
CC_{naive}	7	0	184	7	0	0
CC_{opt}	15	0	174	10	1	0
CC_{632}	15	0	180	5	0	0

TAB. 2.7 – table des modèles retenus

critères \ g	1	2	3	4	5	6
$IRAND$	0	0	189	11	0	0
$MEAN\_MAP$	0	0	200	0	0	0

TAB. 2.8 – table des indices de Rand

critères \ g	1	2	3	4	5	6
$MINFO$	0	0	198	2	0	0
$MEAN\_MAP$	0	0	200	0	0	0

TAB. 2.9 – table des informations mutuelles

Sur les 200 simulations, les moyennes des indice de Rand et des informations mutuelles donnent les meilleurs résultats avec 200 bon modèles. Le BIC donne pratiquement le même résultat avec un bon modèle en moins. Les critères bootstrap classifiants fournissent des résultats similaires dans cette expérience. A noter aussi la proximité des résultats du C_{opt} et C_{632}.

2.3.3 Expérience 2 : classes à proportion identiques et modérément séparées

Le modèle de mélange utilisé pour générer les données de cette expérience est défini par theta.model = $[\pi, \lambda, i]$. Avec $\lambda = 0.5986$, le nuage de points est d'avantage mélangé. Les trois tableaux suivants résume les résultats obtenus.

critéres \ g	1	2	3	4	5	6
BIC	170	0	30	0	0	0
AIC	117	1	57	17	4	4
ICL	200	0	0	0	0	
C_{naive}	116	0	22	18	15	29
C_{opt}	119	1	64	12	3	1
C_{632}	119	1	69	9	1	1
CC_{naive}	200	0	0	0	0	0
CC_{opt}	200	0	0	0	0	0
CC_{632}	200	0	0	0	0	0

TAB. 2.10 – table des modèles retenus

critères\ g	1	2	3	4	5	6
IRAND	0	0	172	24	3	1
MEAN_MAP	0	6	159	23	7	5

TAB. 2.11 – table des indices de Rand

critères\ g	1	2	3	4	5	6
MINFO	0	0	184	14	2	0
MEAN_MAP	0	1	152	19	10	18

TAB. 2.12 – table des informations mutuelles

Avec un taux d'erreur égal à 18 la performance globale des critères diminue. Toute fois les deux moyennes des indices de Rand et de informations mutuelles donne les meilleurs résultats. A remarquer la forte baisse de performance du BIC et les critères bootstrap classifiants. Une fois de plus les résultats du C_{opt} et C_{632} sont très proches. x

2.3.4 Expérience 3 : classes à proportion identiques et mal séparées

On utilise pour une dernière fois le modèle theta.model = $[\pi, \lambda, i]$. Le volume λ des classes est plus grand ici, il vaut 1.21. le nuage de points est plus mélangé .

critères\ g	1	2	3	4	5	6
BIC	200	0	0	0	0	
AIC	110	20	37	20	5	8
ICL	200	0	0	0	0	
C_{naive}	24	17	33	35	36	55
C_{opt}	122	22	32	17	3	4
C_{632}	123	23	35	13	2	4
CC_{naive}	200	0	0	0	0	0
CC_{opt}	199	1	0	0	0	0
CC_{632}	200	0	0	0	0	0

TAB. 2.13 – table des modèles retenus

critères\ g	1	2	3	4	5	6
IRAND	0	4	121	48	18	9
MEAN_MAP	0	20	38	47	36	59

TAB. 2.14 – table des indices de Rand

critères\ g	1	2	3	4	5	6
MINFO	0	0	96	59	34	11
MEAN_MAP	0	3	17	31	30	119

TAB. 2.15 – table des informations mutuelles

Cette expérience confirme, la contre performance du BIC et les bootstrap classifiants dans le cas où les classes ne sont pas bien séparées. Le résultat du AIC diminue aussi mais est meilleur plus que BIC. Les deux critères C_{632} et C_{opt} donnent des résultats du même ordre que AIC

2.3.5 Expérience 4 : classes à proportions différentes et bien séparées

A partir de cette expériences les données seront générées avec un modèle à proportions différentes theta.model = $[\pi_k, \lambda, i]$. Le vecteur des proportions theta.π_k est égal à $[\pi_1, \pi_2, \pi_3] = [0.5, 0.15, 0.35]$. Les volumes des classes sont égales et valent $\lambda = 0.313$.

critères\ g	1	2	3	4	5	6
BIC	0	0	200	0	0	
AIC	0	0	142	40	12	6
ICL	22	0	178	0	0	0
C_{naive}	0	0	45	49	53	53
C_{opt}	0	0	142	49	7	2
C_{632}	0	0	147	48	4	1
CC_{naive}	4	0	188	8	0	0
CC_{opt}	6	0	184	10	0	0
CC_{632}	6	0	189	5	0	0

TAB. 2.16 – table des modèles retenus

critères\ g	1	2	3	4	5	6
IRAND	0	0	183	14	3	0
MEAN_MAP	0	6	194	0	0	0

TAB. 2.17 – table des indices de Rand

critères\ g	1	2	3	4	5	6
MINFO	0	0	192	8	0	0
MEAN_MAP	0	11	188	1	0	0

TAB. 2.18 – table des informations mutuelles

2.3.6 Expérience 5 : classes à proportions différentes et modérément séparées

Nous restons dans les mêmes conditions que l'expérience 5 ; on donne la valeur 0.5986 au volume λ des classes pour avoir des classes modérément séparées.

critères\ g	1	2	3	4	5	6
BIC	87	38	75	0	0	0
AIC	1	6	136	32	17	8
ICL	200	0	0	0	0	0
C_{naive}	0	1	64	44	33	58
C_{opt}	1	8	143	33	11	4
C_{632}	2	8	152	27	8	3
CC_{naive}	198	0	2	0	0	0
CC_{opt}	200	0	0	0	0	0
CC_{632}	200	0	0	0	0	0

TAB. 2.19 – table des modèles retenus

critères\ g	1	2	3	4	5	6
IRAND	0	2	176	19	2	1
MEAN_MAP	0	27	155	11	2	5

TAB. 2.20 – table des indices de Rand

critères\ g	1	2	3	4	5	6
MINFO	0	0	171	26	2	1
MEAN_MAP	0	21	140	16	6	17

TAB. 2.21 – table des informations mutuelles

2.3.7 Expérience 6 : classes à proportions différentes et mal séparées

Avec $\lambda = 1.21$, le nuage de points est plus mélangé.

critères\ g	1	2	3	4	5	6
BIC	190	10	0	0	0	0
AIC	70	39	66	13	7	5
ICL	200	0	0	0	0	0
C_{naive}	15	19	43	47	26	50
C_{opt}	84	46	55	9	3	3
C_{632}	88	47	54	8	3	0
CC_{naive}	200	0	0	0	0	0
CC_{opt}	200	0	0	2	0	0
CC_{632}	200	0	0	0	0	0

TAB. 2.22 – table des modèles retenus

critères\ g	1	2	3	4	5	6
IRAND	0	16	124	36	16	8
MEAN_MAP	0	24	62	33	22	59

TAB. 2.23 – table des indices de Rand

critères\ g	1	2	3	4	5	6
MINFO	0	3	110	54	27	6
MEAN_MAP	0	7	33	23	28	109

TAB. 2.24 – table des informations mutuelles

Commentaires des 4, 5 et 6 : Nous avons regroupé les commentaires de ces expériences puisque elles donnent quasiment les mêmes résultats et révèlent les même tendances que dans les expériences. Le BIC,$CC_{\text{naïf}}$, CC_{opt} CC_{632} et les $MEAN_MAP$ donnent de très bon résultats pour des classes bien séparées tandis que AIC, $C_{\text{naïf}}$, C_{opt} et C_{632} donnent des résultats en dessous de ceux cités dessus, mais ne s'effondre pas quand les données sont plus mélangés.

Bibliographie

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In* Petrov, B. et Csaki, F., éditeurs : *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademiai Kiado.

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

Biernacki, C., Celleux, G. et Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on patern analysis and machine intelligence*, 22(7):719–725.

Box et Draper (1987). Empirical Model-Building and Response Surfaces. page 424.

Burnham, K. et Anderson, D. (1998). *Model selection and inference : a pratical information - theoretic approach*. Springer-Verlag New York Inc, New York.

Burnham, K. et Anderson, D. (2004). Multimodel inference : understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261.

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

Colletaz, G. (2007). Les critères de sélection MASTER 1 ESA.

Dang, V. M. (1998). *Classification de données spatiales : modèles probabilistes et critères de partitionnement*. Thèse de doctorat, Université de technologie de Compiègne.

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

Efron, B. et Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

El-Baf, F. (2010). Etude comparative sur les critères de sélection. Rapport technique interne, Université de technologie de Compiègne.

Fraley, C. et Raftery, A. (1998). How many clusters ? which clustering method ? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578.

Govaert, G. (2003). *Analyse des données*, chapitre Classifcation et modèle de mélange. Lavoisier.

Govaert, G. et Nadif, M. (2009). Un modèle de mélange pour la classification croisée d'un tableau de données continues. *In CAp 09, 11e conférence sur l'apprentissage artificiel*, pages 287–302, Hammamet, Tunisie.

Govaert, G. et Nadif, M. (2010). Latent block model for contingency table. *Communications in Statistics - Theory and Methods*, 39(3):416–425.

Jaakkola, T. (2000). Tutorial on variational approximation methods. *Advanced mean field methods : theory and practice*, pages 129–159.

Jaakkola, T. et Jordan, M. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Lebarbier, E. et Mary-Huard, T. (2004). Le critère bic : fondements théoriques et interprétation.

Neal, R. et Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in graphical models*, 89:355–368.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological methodology*, 25:111–163.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared ? A conflict between model indentification and regression estimation. *Biometrika*, 92(4): 937.

Ye, M., Meyer, P. et Neuman, S. (2008). On model selection criteria in multimodel analysis. *Water Resources Research*, 44(3):3428.