Classification simultanée à base de mélanges gaussiens pour des échantillons d'origines multiples

Alexandre Lourme, Christophe Biernacki

CLASSIFICATION SIMULTANÉE À BASE DE MÉLANGES GAUSSIENS POUR DES ÉCHANTILLONS D'ORIGINES MULTIPLES

Alexandre Lourme [†,1] & Christophe Biernacki [‡,2]

[†] *IUT département GB, 361 rue du Ruisseau, 40000 Mont de Marsan, France.*

[‡] *Université Lille 1 & CNRS, 59655 Villeneuve d'Ascq, France.*

RÉSUMÉ : Habituellement, quand on cherche à partitionner des échantillons provenant de populations différentes, on met en oeuvre plusieurs procédures de classification indépendantes. Mais lorsque les échantillons sont décrits par les mêmes variables et que la partition a la même signification d'un échantillon à l'autre (mâles/femelles par exemple), nous proposons d'appliquer une procédure dite de classification simultanée. Il s'agit, dans un contexte de mélanges gaussiens, d'établir, sous des hypothèses simples et réalistes, un lien stochastique linéaire entre les composantes normales des populations, ce qui permet d'estimer simultanément le paramètre de tous les mélanges afin de classer les individus des différents échantillons. Plusieurs modèles de contraintes sont proposés, portant à la fois sur le paramètre des mélanges (*modèles intra-populations*) et sur le lien linéaire entre les populations (*modèles inter-populations*). Le critère BIC, par exemple, permet de choisir un de ces modèles, mais aussi de comparer les deux méthodes de classification, indépendante et simultanée. Grâce à un exemple biologique, où des échantillons d'oiseaux provenant de trois espèces différentes doivent être classés selon leur sexe, nous observons que la classification simultanée donne des résultats prometteurs.

MOTS-CLÉS : classification non supervisée, mélanges gaussiens, relation stochastique, choix de modèle, variables biologiques.

ABSTRACT: When several samples, described by the same variables and possibly arising from different populations, have to be split into a same meaning partition, it is common to perform several independant clustering processes. But, in a Gaussian mixture model-based clustering context, one may establish, under few assumptions, some realistic parametric link between the Gaussian components of the mixtures, which allows to estimate simultaneously all mixture parameters in order to classify all individuals. Several parsimonious models are proposed, combining classical constraints on the Gaussian component parameters (*intrapopulation models*) and models of constraints set on the parametric link (*interpopulation models*). BIC criterion for instance, allows to choose a model among those ones and, on the other hand, to compare both clustering methods, the independent and the simultaneous one.

[1] alexandre.lourme@univ-pau.fr

[2] biernack@univ-lille1.fr

Some experiences on three samples of different seabird species, which have to be split according to their sex, are provided, which show that simultaneous clustering improves the estimated partitions obtained by independent clustering.

KEY WORDS: Model-based clustering, Gaussian mixtures, stochastic relationship, model choice, biological features.

1 Problématique

On dispose de H échantillons qui doivent être partitionnés en K groupes de même signification. On suppose de plus que les variables descriptives sont identiques d'un échantillon à l'autre. Pour fixer les idées, chaque échantillon x^h $(h \in \{1 \dots H\})$ est constitué de n^h individus x_i^h $\left(i \in \{1 \dots n^h\}\right)$ de $\mathbb{R}^d$. Sa partition est une matrice $z^h \in \{0,1\}^{n^h \times K}$ dont le coeficient $z_{i,k}^h$ vaut 1 si et seulement si x_i^h est affecté au groupe k. On se place en outre exclusivement dans le cas gaussien, c'est à dire que les couples $\left(x_i^h, z_i^h\right)_{i=1\dots n^h}$ sont supposés être des réalisations de vecteurs indépendants et identiquement distribués à un vecteur aléatoire $\left(X^h, Z^h\right)$ tel que :

$$Z^h \sim \mathscr{M}_K(1, \pi_1^h, \dots, \pi_K^h) \quad \text{et} \quad (X^h | Z_k^h = 1) \sim \mathscr{N}_d(\mu_k^h, \Sigma_k^h).$$

$\mathscr{M}_K(1, \pi_1^h, \dots, \pi_K^h)$ désigne la loi multinomiale K-dimensionnelle d'ordre 1, de paramètre $(\pi_1^h, \dots, \pi_K^h)$ $(\pi_k^k \geq 0$ et $\sum_{k=1}^{K} \pi_k^h = 1)$ et $\mathscr{N}_d(\mu_k^h, \Sigma_k^h)$ désigne la loi normale d-dimensionnelle centrée en μ_k^h et dont Σ_k^h est la matrice des covariances. Ainsi, on identifie la population P^h dont provient l'échantillon x^h, à un mélange de K composantes normales paramétrées par $\psi_k^h = \left(\pi_k^h, \mu_k^h, \Sigma_k^h\right)$ $(k \in \{1, \dots, K\})$. Dans ce contexte, on peut entreprendre H procédures de classification indépendantes, en estimant séparément le paramètre $\psi^h = \left\{\psi_k^h\right\}$ de chaque population P^h (McLachlan et Peel (2000)). Mais ces populations ne sont pas sans lien. En effet, les variables qui les décrivent sont identiques, les groupes recherchés sont en nombre égal et ces groupes ont la même signification d'une population à l'autre. Nous cherchons, dans ce travail, à formaliser cette information dans le but d'améliorer globalement les partitions estimées.

2 Principe de la classification simultanée

L'idée de base de ce travail est la suivante. Pour tout $h \in \{2 \dots H\}$ et tout $k \in \{1 \dots K\}$, on suppose qu'il existe une application $\xi_k^h : \mathbb{R}^d \to \mathbb{R}^d$ telle que :

$$(X^h | Z_k^h = 1) \sim \xi_k^h (X^1 | Z_k^1 = 1). \tag{1}$$

Cette hypothèse signifie que, conditionnellement aux groupes recherchés, les populations P^h $(h = 2 \dots H)$ sont obtenues par transformation stochastique de P^1. Par ailleurs, puisque les échantillons sont décrits par les mêmes variables, il est naturel de supposer qu'une variable

dans une population dépend fortement de la même variable dans une autre population. On suppose donc que la $j^{\text{e}}(j \in \{1 \dots d\})$ composante $(\xi_k^h)^{(j)}(x)$ du vecteur $\xi_k^h(x)$ ne dépend que de la j^{e} composante $x^{(j)}$ de x, ce qui se traduit par : $\forall (x,y) \in \mathbb{R}^d \times \mathbb{R}^d, x^{(j)} = y^{(j)} \Rightarrow (\xi_k^h)^{(j)}(x) = (\xi_k^h)^{(j)}(y)$. En d'autres termes, $(\xi_k^h)^{(j)}$ correspond à une application de $\mathbb{R}$ dans $\mathbb{R}$ qui transforme la variable normale $(X^1|Z_k^1 = 1)^{(j)}$ en une autre variable normale $(X^h|Z_k^h = 1)^{(j)}$. Si on suppose de plus que $(\xi_k^h)^{(j)}$ est continûment dérivable, alors cette application est forcément affine (De Meyer, Roynette, Vallois et Yor (2000)). En conséquence, pour tout $h \in \{1 \dots H\}$ et tout $k \in \{1 \dots K\}$, il existe une matrice $D_k^h \in \mathbb{R}^{d \times d}$ diagonale et un vecteur $b_k^h \in \mathbb{R}^d$ tels que :

$$(X^h|Z_k^h = 1) \sim D_k^h(X^1|Z_k^1 = 1) + b_k^h. \tag{2}$$

La relation (1) est la clef de la classification simultanée et (2) est la forme affine qu'elle prend sous les deux hypothèses précédentes. La relation (2) établit un lien paramétrique entre la population P^1 et les autres populations. Estimer conjointement le paramètre de ce lien et le paramètre de P^1 permet d'estimer les paramètres gaussiens des autres populations P^h $(h \in \{2, \dots, H\})$. On établit alors une partition de tous les échantillons en utilisant le principe du *Maximum A Posteriori* (MAP).

3 Modèles de contrainte

La relation (2) établit un lien stochastique linéaire entre chaque composante ψ_k^1 de P^1 et la composante ψ_k^h de P^h $(h \geq 2)$. On peut donc reparamétrer le modèle considéré par $\theta = (\theta_r, \theta_l)$ où $\theta_r = \{(\pi_k^1, \mu_k^1, \Sigma_k^1)\}$ fait office de paramètre de référence, $\theta_l = \{(\pi_k^h, D_k^h, b_k^h); k = 1, \dots, K; h = 2, \dots, H\}$ contenant, lui, le paramètre du lien conditionnel linéaire entre P^1 et les autres populations.

Plusieurs modèles parcimonieux sont envisagés pour le mélange P^1. Ses composantes peuvent être homoscédastiques $(\Sigma_k^1 = \Sigma)$ ou hétéroscédastiques $(\Sigma_k^1 = \Sigma_k)$, ses proportions mélange peuvent être égales $(\pi_k^1 = \pi)$ ou différentes $(\pi_k^1 = \pi_k)$. Ces modèles, qui concernent le paramètre de référence θ_r, sont appelés des *modèles intra-populations*. On peut les combiner avec d'autres modèles parcimonieux portant, eux, sur le paramètre du lien θ_l, et appelés pour cette raison *modèles inter-populations*. Dans le cas général, les matrices D_k^h sont sans contrainte. Elles peuvent aussi être indépendantes du groupe $(D_k^h = D^h)$, de la population $(D_k^h = D_k)$, des variables $(D_k^h = \alpha_k^h \mathrm{I})$, être indépendantes à la fois du groupe et des variables $(D_k^h = \alpha^h \mathrm{I})$, ou toutes égales à l'identité $(D_k^h = \mathrm{I})$. De la même façon, les vecteurs b_k^h peuvent être sans contrainte, indépendants du groupe $(b_k^h = b^h)$, ou tous nuls $(b_k^h = 0)$. Enfin, les proportions des mélanges peuvent être libres $(\pi_k^h = \pi^h)$ ou égales d'une population à l'autre $(\pi_k^h = \pi)$.

La relation (2) qui permet de caractériser un modèle de classification simultanée, semble faire de P^1 la population de référence dont on déduit le paramètre des autres mélanges. Or le choix des labels des populations est arbitraire et on souhaite que le modèle estimé puisse exister sous n'importe quel choix initial du label des populations. Un modèle possèdant cette propriété

est dit *Invariant à la Numérotation des Populations* (INP).

Enfin, le label des groupes dans chaque mélange est lui aussi arbitraire et pour ne pouvoir apparier les groupes des différentes populations que de façon unique, le paramètre θ doit être identifiable.

Le tableau 1 indique parmi les combinaisons de modèles intra et inter-populations envisagés précédemment ceux qui sont INP et identifiables.

Table 1: *Identifiabilité et propriété INP pour les combinaisons de modèles intra et inter-populations.*
On note "." un modèle non INP, "$\circ$" désigne un modèle INP mais non identifiable, et "$\bullet$", un modèle INP et identifiable. Certains modèles INP sont identifiables quand $H = 2$ et non identifiables quand $H \geq 3$. Ces modèles sont indiqués par "$\odot$".
Les modèles où les matrices D_k^h sont indépendantes de la population ($D_k^h = D_k$), ne sont pas INP si $h \geq 3$, et lorsque $h = 2$, ils se ramènent au modèle D_k^h. Pour cette raison, nous ne les mentionnons pas dans le tableau ci-dessous.

			Modèles intra-populations			
			π		π_k	
Modèles inter-populations			Σ	Σ_k	Σ	Σ_k
π (π^h)	I, α^hI, D^h	0	$\bullet$ (.)	$\bullet$ (.)	$\bullet$ ($\bullet$)	$\bullet$ ($\bullet$)
		b^h	$\bullet$ (.)	$\bullet$ (.)	$\bullet$ ($\bullet$)	$\bullet$ ($\bullet$)
		b_k^h	$\circ$ (.)	$\bullet$ (.)	$\odot$ ($\circ$)	$\bullet$ ($\bullet$)
	α_k^hI, D_k^h	0	. (.)	$\bullet$ (.)	. (.)	$\bullet$ ($\bullet$)
		b_k^h	. (.)	$\bullet$ (.)	. (.)	$\bullet$ ($\bullet$)

4 Estimation du paramètre

Le paramètre θ est estimé en maximisant sa logvraisemblance (On pose $D_k^1 = \mathrm{I}$ et $b_k^h = 0$.) :

$$\ell(\theta) = \sum_{h=1}^{H} \sum_{i=1}^{n^h} \ln \sum_{k=1}^{K} \pi_k^h \phi_d \left(x_i^h ; D_k^h \mu_k^1 + b_k^h, D_k^h \Sigma_k^1 D_k^h \right), \tag{3}$$

grâce à un algorithme GEM (Dempster, Laird et Rubin (1977)). L'étape E est un calcul classique des probabilités conditionnelles $t_{i,k}^h$ que chaque individu x_i^h appartienne aux différents groupes. L'étape GM consiste, sous les contraintes imposées par le modèle considéré, à accroitre l'espérance conditionnelle de la logvraisemblance complétée de θ, en la maximisant alternativement par rapport à chacune des composantes $\{\pi_k^h\}$, $\{\mu_k^1\}$, $\{\Sigma_k^1\}$, $\{D_k^h\}$, $\{b_k^h\}$ et en utilisant ses propriétés de concavité partielle. A chaque itération les composantes de P^1 ont des estimateurs explicites pour les centres :

$$\mu_k^1 = \frac{1}{n_k} \sum_{h=1}^{H} \sum_{i=1}^{n^h} t_{i,k}^h \left(D_k^h \right)^{-1} \left(x_i^h - b_k^h \right), \tag{4}$$

où $n_k = \sum_{h=1}^{H} n_k^h$ avec $n_k^h = \sum_{i=1}^{n^h} t_{i,k}^h$, et pour les matrices de covariances :

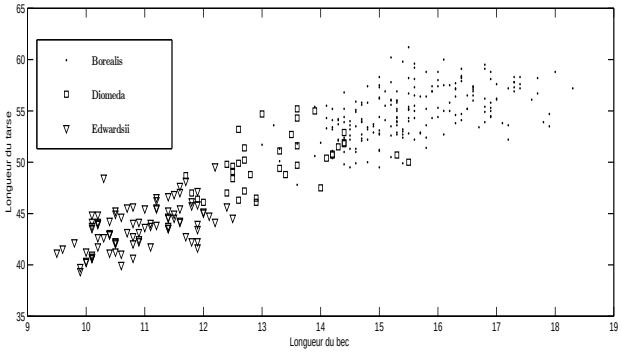
$$\Sigma_k^1 = \frac{1}{n_k} \sum_{h=1}^{H} \sum_{i=1}^{n^h} t_{i,k}^h \left[\left(D_k^h \right)^{-1} \left(x_i^h - b_k^h \right) - \mu_k^1 \right] \left[\left(D_k^h \right)^{-1} \left(x_i^h - b_k^h \right) - \mu_k^1 \right]', \tag{5}$$

par exemple, si l'on suppose les mélanges hétéroscédastiques. Quand les proportions des mélanges sont supposées libres, on les estime grâce à : $\pi_k^h = \frac{n_k^h}{n^h}$. En outre, l'estimateur des vecteurs b_k^h est explicite, ainsi que celui des matrices D_k^h dans les modèles $D_k^h = \alpha^h \mathrm{I}$ et $D_k^h = \alpha_k^h \mathrm{I}$. Lorsque les matrices D_k^h ne sont pas des matrices d'homothétie, on les estime de façon approchée.

5 Application biologique

Thibault, Bretagnolle et Rabouam (1997) décrivent trois échantillons d'oiseaux de mer ($H = 3$) vivant dans différentes zones géographiques. *Borealis* (échantillon S^1, taille $n^1 = 206$ individus, 45% de femelles) vivent dans les îles de l'Atlantiques (Açores, Canaries, etc.), *Diomeda* (échantillon S^2, taille $n^2 = 38$ individus, 58% de femelles), dans les îles méditerranéennes (Baléares, Corse, etc.), et *Edwardsii* (échantillon S^3, taille $n^3 = 92$ individus, 52% de femelles), dans les îles du Cap Vert. Les oiseaux de chaque échantillon sont décrits par les mêmes cinq variables morphologiques ($d = 5$) : longueur et hauteur du bec, longueur du tarse, des ailes et de la queue. D'après la figure 1, qui représente les oiseaux par la longueur de leur bec et la longueur de leur tarse, les trois échantillons semblent provenir de trois populations bien distinctes.

Figure 1: *Trois populations d'oiseaux décrites par les mêmes variables.*



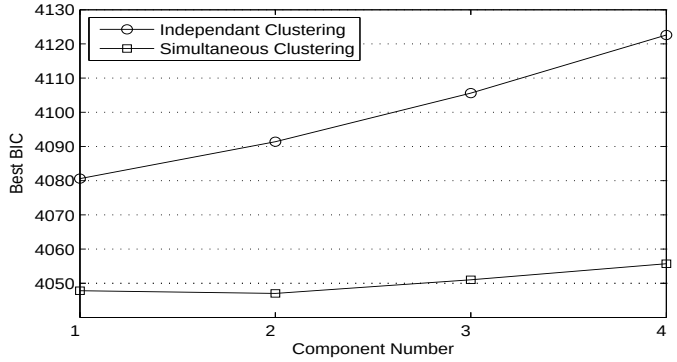
On souhaite retrouver le sexe des oiseaux ($K = 2$). Les valeurs de BIC (Lebarbier, Mary-Huard (2004)) et le taux d'erreur correspondant on été calculés pour chacun des 66 modèles du tableau 1 qui sont INP. Le tableau 2 présente quelques unes de ces valeurs ainsi que les valeurs de BIC et au taux d'erreur que l'on obtient en classant indépendamment les échantillons. Sur cet exemple, la classification simultanée semble marquer un avantage sur la classification indépendante : en effet, non seulement la classification simultanée est retenue par le critère BIC, mais elle conduit également à l'erreur optimale.

Table 2: *Valeur de BIC et (taux d'erreur apparent) obtenus par classification simultanée (2 groupes) des oiseaux.*

			π		π_k	
			Σ	Σ_k	Σ	Σ_k
π	α_k^h	0	.	4282.9 (32.14)	.	4279.4 (38.69)
		b_k^h	.	4110.4 (12.50)	.	4110.4 (16.07)
	D^h	0	**4047.0** (**10.42**)	4071.9 (11.61)	4049.7 (11.31)	4073.9 (11.88)
		b^h	4071.8 (10.71)	4096.9 (12.20)	4074.7 (10.71)	4099.3 (14.58)
		b_k^h	4094.9 (33.33)	4122.2 (11.31)	4101.9 (41.96)	4122.7 (15.77)
Indépendante			4091.9 (55.65)	4152.4 (48.21)	4091.4 (54.76)	4147.4 (44.35)

Par ailleurs, nous pourrions ignorer le nombre de groupes, et supposer la valeur de K inconnue. On peut donc étendre l'ensemble des modèles en compétition en supposant successivement $K = 1, 2, 3, 4$, par exemple. La figure 2 représente, pour chacun des cas précédents, la meilleure valeur de BIC obtenue en classification simultanée et en classification indépendante. On observe que la classification simultanée retient 2 classes contrairement à la classification indépendante.

Figure 2: *Meilleur BIC pour différents nombres de groupes, en classification indépendante et simultanée.*



Bibliographie

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

[2] Dempster, A.P., Laird, N.M. and 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.

[3] Lebarbier, E. et Mary-Huard T. (2006). Le critère BIC, fondements théoriques et interprétation, *Journal de la Société Française de Statistique*, **1**, 39-57.

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

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