Soumission

Journal de la Société Française de Statistique

Classification simultanée de plusieurs échantillons sous contrainte d'égalité des entropies de partition

Title: Simultaneous Clustering of Several Samples by Equalizing their Partition Entropy

Alexandre Lourme[1] et Christophe Biernacki[2]

Résumé : La classification non supervisée basée sur des modèles de mélange est devenue usuelle pour identifier des sous-populations dans un jeu de données. Ici, nous nous intéressons au cas fréquent où plusieurs échantillons provenant de populations distinctes doivent être classifiés en des partitions *a priori* de même signification. Nous supposons que le pouvoir discriminant des variables décrivant spécifiquement les différents échantillons est globalement égal. Nous traduisons cette hypothèse en imposant que l'entropie de la partition soit homogène d'une population à l'autre, ce qui nous conduit à une méthode de classification dite simultanée. Un algorithme inspiré d'EM, et baptisé ẼM, permet de réaliser cette estimation par maximum de vraisemblance sous contrainte, pour n'importe quel modèle de mélange puisque la contrainte entropique est active uniquement à l'étape E. Des résultats expérimentaux obtenus sur des données simulées d'une part et sur des données réelles issues de la biologie d'autre part, illustrent le potentiel de notre approche.

Abstract: Model-based clustering is now a standard tool for revealing some subpopulations in a data set. We focus here on the frequent situation where several data sets have to be classified into partitions with *a priori* identical meanings. We assume that the discriminant ability of the variables involved specifically in the different data sets, is globally invariant. This assumption is formalized by setting that the entropy of the partition is homogeneous through the populations, which leads us to a so-called simultaneous clustering method. A pseudo EM algorithm, called ẼM, allows to perform this maximum likelihood estimation under constraint for any mixture model since the entropic constraint is only involved in the E step. A real numerical example on biological data leads to encouraging results.

Mots-clés : modèles de mélange, algorithme EM, lien entre populations, données biologiques
Keywords: mixture models, EM algorithm, link between populations, biological data
Classification AMS 2000 : , ,

1. Introduction

La classification non supervisée vise à partitionner un échantillon afin de révéler une structure cachée d'intérêt. Dans un contexte probabiliste, il est maintenant classique de supposer que les données proviennent d'un mélange fini de lois de probabilités paramétriques puis d'affecter chaque individu à la classe de plus forte probabilité conditionnelle (voir [16] pour une vue d'ensemble). La distribution des composantes du mélange varie en fonction de la nature des données à classifier : dans le cas continu, il s'agit le plus souvent de gaussiennes ou de Student multivariées ([16], chap. 3 et 7 respectivement) ; dans le cas discret, les lois de Poisson multivariées peuvent être choisies [11] ; dans le cas catégoriel, généralement des lois multinomiales multivariées avec

[1] Université de Pau & IUT Génie Biologique, Mont de Marsan, France.
E-mail : Alexandre.Lourme@univ-pau.fr
[2] Université Lille 1, CNRS & INRIA, Villeneuve d'Ascq, France.
E-mail : Christophe.Biernacki@math.univ-lille1.fr

Soumis au Journal de la Société Française de Statistique
File: lourme_biernacki.tex, compiled with jsfds, version : 2009/12/09
date: 19 novembre 2011

indépendance des variables sont implémentées conduisant au modèle dit des classes latentes [9] ; dans le cas de données de rangs, le modèle Φ de Mallows sert de référence [17] ; dans le cas de données directionnelles, les lois de von Mises ou encore de von Mises-Fisher sont incontournables [15]. Ces exemples ne sont pas exhaustifs, car la flexibilité des mélanges permet à chaque praticien d'inclure la loi de son choix dans son étude. Dans chacune de ces situations, la classification à base de modèles a permis de traiter avec succès de nombreuses applications relevant de multiples champs d'activité : génétique [18], médecine [16], imagerie par résonnance magnétique [1], astronomie [6], finance [14], biologie [3]... En conséquence, de nos jours l'usage de ces modèles pour la classification peut être considérée comme familière à tous les statisticiens et également à un nombre croissant de praticiens.

Il est en outre fréquent d'avoir à classifier non pas un mais plusieurs échantillons, provenant éventuellement de populations distinctes, dans le contexte spécifique suivant : la partition attendue possède la même interprétation d'un échantillon à l'autre.

En biologie par exemple, [20] cherchent à déterminer le sexe de trois échantillons d'oiseaux marins correspondant chacun à une sous-espèce de puffins cendrés. Les échantillons sont décrits par un même ensemble de cinq variables morphométriques (dimension du tarse, des ailes, *etc.*), mais la distribution de chacune de ces variables change d'une sous-espèce à l'autre. Dans ce cas, une procédure usuelle de classification peut être appliquée indépendamment à chaque échantillon. Or les échantillons ne sont pas sans lien : ils sont constitués d'unités statistiques de même nature (des puffins cendrés) et ils sont décrits par des caractéristiques biométriques de même signification. On peut donc raisonnablement s'attendre à une certaine connexion entre les mâles (resp. les femelles) de chaque échantillon.

De la même façon en finance, [8] cherchent à distinguer parmi deux échantillons d'entreprises considérées à un an d'intervalle (2002 et 2003), les sociétés en bonne santé ou en faillite. Les firmes des deux échantillons sont décrites par des ratios financiers de même signification relatant leur performance, leur efficacité, leur liquidité, *etc*. Ces ratios évoluent de façon significative d'une année à l'autre mais sans changer, d'après les financiers, la typologie attendue (bonne santé/faillite) des entreprises. Là encore, il est raisonnable de supposer qu'il existe un lien entre les partitions de 2002 et 2003.

Dans le cadre des mélanges gaussiens et de Student multivariés, les auteurs ont établi sous des hypothèses assez faibles un lien stochastique, conditionnel et affine entre populations [13, 14]. Ces classifications simultanées particulières ont été appliquées avec succès respectivement aux cas biologique et financier précédents. Dans ce travail, nous souhaitons étendre le principe de la classification simultanée sans se restreindre aux modèles gaussiens ou de Student d'une part, et sans se limiter à une exacte concordance entre les variables des différentes populations d'autre part. Notre hypothèse de travail fondamentale est de supposer que les variables décrivant les différents échantillons ont globalement le même pouvoir discriminant, ce que nous traduisons par une hypothèse d'égalité d'entropie des partitions de chaque échantillon.

Le papier s'organise de la façon suivante. Dans la partie 2 nous présentons le nouveau modèle de classification simultanée sous contrainte entropique. Ensuite, dans la partie 3, un algorithme d'estimation spécifique, baptisé ẼM de part son fort lien à EM, est décrit et évalué. Il a pour rôle

Soumis au Journal de la Société Française de Statistique
File: lourme_biernacki.tex, compiled with jsfds, version : 2009/12/09
date: 19 novembre 2011

de maximiser un critère différent, mais asymptotiquement égal, à la vraisemblance. Son avantage est d'être applicable à tous les types de mélanges, la contrainte entropique étant uniquement active à l'étape E, renommée pour l'occasion Ẽ car il ne s'agit donc plus d'une étape E à proprement parler. La partie 4 illustre les performances de la nouvelle méthode sur des données biologiques et la dernière partie (partie 5) conclut le papier.

2. Classification indépendante et simultanée avec lien entropique

Notre objectif est de classifier H échantillons en K groupes chacun. Nous décrivons dans un premier temps la classification à base de modèles de mélange (sous-partie 2.1) dans cette situation apparemment plus complexe (H échantillons au lieu d'un seul) car ce sera ensuite pratique pour faire le lien avec la classification simultanée sous contrainte d'égalité entropique (sous-partie 2.2). Chaque échantillon $\mathbf{x}^h$ ($h \in \{1,\dots,H\}$) comporte n^h individus $\mathbf{x}_i^h$ ($i = 1,\dots,n^h$) d'un espace d dimensionnel $\mathscr{X}$ dépendant de leur nature ($\mathscr{X}$ peut être $\mathbb{R}^d$ dans le cas continu, $\{0,1\}^d$ dans le cas binaire, *etc.*), et provient d'une population P^h. Chaque population est composée non seulement du même nombre de classes K mais aussi les partitions sont supposées toutes avoir la même signification. Enfin, chaque population n'est pas nécessairement décrite par les mêmes d variables mais, globalement, on fait l'hypothèse qu'elles ont le même pouvoir discriminant pour décrite les partitions recherchées.

On notera dans la suite les indices h,i,k avec leur domaine de variation respectif : $h = 1,\dots,H$, $i = 1,\dots,n^h$, $k = 1,\dots,K$. Par souci de clarté et de concision, on se permettra alors d'écrire des sommes sur ces indices sans indiquer les domaines de variation.

2.1. Solution habituelle : plusieurs classifications indépendantes

La classification habituelle à base de modèles suppose que les individus $\mathbf{x}_i^h$ de chaque échantillon $\mathbf{x}^h$ proviennent de façon indépendante d'un vecteur aléatoire $\mathbf{X}^h$ suivant un mélange de K lois de probabilités :

$$\mathrm{p}(\mathbf{x}_i^h;\boldsymbol{\theta}^h) = \sum_k \alpha_k^h \mathrm{p}(\mathbf{x}_i^h;\boldsymbol{\beta}_k^h), \quad \mathbf{x}_i^h \in \mathscr{X}.$$

Les coefficients α_k^h correspondent aux proportions du mélange ($\alpha_k^h > 0$, et $\sum_k \alpha_k^h = 1$) ; les $\boldsymbol{\beta}_k^h$ correspondent aux paramètres de la loi de la k^{e} composante de la population P^h. Les paramètres $\boldsymbol{\beta}_k^h$ peuvent être continus, discrets ou encore une association des deux. Le paramètre global de la loi de P^h est alors noté $\boldsymbol{\theta}^h = (\boldsymbol{\theta}_k^h)_{\{k\}}$ où $\boldsymbol{\theta}_k^h = (\alpha_k^h, \boldsymbol{\beta}_k^h)$. On suppose en outre que les hypothèses standards (voir [16], chap. 1) sur la loi mélange sont vérifiées comme l'identifiabilité à une permutation près des composantes.

La composante qui a généré l'individu $\mathbf{x}_i^h$ est une donnée manquante. Nous la représentons par un vecteur binaire $\mathbf{z}_i^h \in \{0,1\}^K$ dont le k^{e} terme z_{ik}^h est égal à 1 si et seulement si $\mathbf{x}_i^h$ provient de la k^{e} composante du mélange. Le vecteur $\mathbf{z}_i^h$ provient alors d'un vecteur aléatoire $\mathbf{Z}^h$ de distribution multinomiale K-variée d'ordre 1 et de paramètre $(\alpha_1^h,\dots,\alpha_K^h)$. Cette partition inconnue est notée $\mathbf{z}^h = \{\mathbf{z}_1^h,\dots,\mathbf{z}_{n^h}^h\}$. Le modèle suppose que les données complètes $(\mathbf{x}_i^h,\mathbf{z}_i^h)_{\{i\}}$ sont des réalisations i.i.d. du couple de vecteurs aléatoires $\left(\mathbf{X}^h,\mathbf{Z}^h\right)$ de $\mathscr{X} \times \{0,1\}^K$.

Soumis au Journal de la Société Française de Statistique
File: lourme_biernacki.tex, compiled with jsfds, version : 2009/12/09
date: 19 novembre 2011

L'estimation du paramètre $\boldsymbol{\theta} = (\boldsymbol{\theta}^h)_{\{h\}}$ regroupant toutes les populations est classiquement réalisée par maximisation de la log-vraisemblance

$$\ell(\boldsymbol{\theta};\mathbf{x}) = \sum_{h,i} \ln\left[\mathrm{p}(\mathbf{x}_i^h;\boldsymbol{\theta}^h)\right] = \sum_h \ell^h(\boldsymbol{\theta}^h;\mathbf{x}^h), \tag{1}$$

évaluée en les données observées $\mathbf{x} = \bigcup_h \mathbf{x}^h$. De façon équivalente, cela revient à maximiser indépendamment chaque log-vraisemblance $\ell^h(\boldsymbol{\theta}^h;\mathbf{x}^h)$ du paramètre $\boldsymbol{\theta}^h$ calculée sur chaque échantillon $\mathbf{x}^h$. L'utilisation d'un algorithme EM [7] pour cette optimisation est alors une méthode standard et on pourra se référer de nouveau à [16] pour les détails.

Chaque partition $\mathbf{z}^h$ est alors estimée par le principe du Maximum A Posteriori (MAP) qui consiste à retenir la composante de plus grande probabilité conditionnelle pour chaque individu de $\mathbf{x}^h$. Calculée en l'estimateur par maximum de vraisemblance $\hat{\boldsymbol{\theta}}$, cette probabilité pour l'individu $\mathbf{x}_i^h$ est donnée par

$$t_{ik}^h(\hat{\boldsymbol{\theta}}^h) = \mathrm{E}\left(Z_k^h \middle| \mathbf{X}^h = \mathbf{x}_i^h; \hat{\boldsymbol{\theta}}^h\right), \tag{2}$$

et ainsi $\hat{\mathbf{z}}_{ik}^h = 1$ si $t_{ik}^h(\hat{\boldsymbol{\theta}}^h) \geq t_{ik'}^h(\hat{\boldsymbol{\theta}}^h)$ pour tout $k' \neq k$. On note dans la suite $\mathbf{t}^h(\boldsymbol{\theta}^h) = \left(t_{ik}^h(\boldsymbol{\theta}^h)\right)_{\{i,k\}}$ et $\mathbf{t}(\boldsymbol{\theta}) = \left(\mathbf{t}^1(\boldsymbol{\theta}^1), \ldots, \mathbf{t}^H(\boldsymbol{\theta}^H)\right)$.

2.2. Solution proposée : homogénéité entre populations de l'entropie de partition

Rappelons que chaque population n'est pas nécessairement décrite par les mêmes d variables mais on a fait l'hypothèse forte qu'elles ont globalement le même pouvoir discriminant pour décrire les partitions recherchées. Nous proposons de quantifier cette information par l'égalité des entropies de partition $\mathscr{E}(\boldsymbol{\theta}^h)$ de chaque population P^h, donc

$$\mathscr{E}(\boldsymbol{\theta}^1) = \ldots = \mathscr{E}(\boldsymbol{\theta}^H), \tag{3}$$

où

$$\mathscr{E}(\boldsymbol{\theta}^h) = \mathrm{E}\left[-\sum_k \mathrm{E}\left(Z_k^h \middle| \mathbf{X}^h;\boldsymbol{\theta}^h\right) \ln \mathrm{E}\left(Z_k^h \middle| \mathbf{X}^h;\boldsymbol{\theta}^h\right); \boldsymbol{\theta}^h\right].$$

La quantité $\mathscr{E}(\boldsymbol{\theta}^h)$ mesure le pouvoir séparateur des variables dans la population P^h : elle est positive et croît quand les composantes se chevauchent, c'est-à-dire quand les probabilités conditionnelles $\mathrm{E}\left(Z_k^h \middle| \mathbf{X}^h;\boldsymbol{\theta}^h\right)$ s'éloignent des valeurs 0 ou 1, synonymes de très bonne séparation.

En pratique, l'entropie $\mathscr{E}(\boldsymbol{\theta}^h)$ peut être estimée par sa version empirique

$$e(\mathbf{t}^h(\boldsymbol{\theta}^h)) = -\frac{1}{n^h}\sum_{i,k} t_{ik}^h(\boldsymbol{\theta}^h) \ln t_{ik}^h(\boldsymbol{\theta}^h) \tag{4}$$

et la contrainte (3) devient dans ce cas

$$e(\mathbf{t}^1(\boldsymbol{\theta}^1)) = \ldots = e(\mathbf{t}^H(\boldsymbol{\theta}^H)). \tag{5}$$

Soumis au Journal de la Société Française de Statistique
File: lourme_biernacki.tex, compiled with jsfds, version : 2009/12/09
date: 19 novembre 2011

La maximisation de la log-vraisemblance (1) doit maintenant être réalisée sous la contrainte (5). Dans un algorithme EM, l'étape E n'est pas affectée par cette contrainte qui porte uniquement sur le paramètre $\boldsymbol{\theta}$ de toutes les populations réunies. Par contre l'étape M conduit à une maximisation de l'espérance conditionnelle de la log-vraisemblance très délicate dû au caractère hautement non linéaire de la contrainte entropique (5). Nous présentons dans la partie suivante un algorithme spécifique pour répondre à cette difficulté.

Remarque. En théorie la contrainte (5) ne permet pas d'éviter la dégénérescence que l'on peut observer dans certaines familles de mélanges. En effet, lorsque la vraisemblance d'un mélange n'est pas majorée, comme celle des mélanges gaussiens par exemple (voir [2]), elle reste non majorée si on la soumet à (5). Cependant, d'un point de vue empirique, la contrainte (5) semble diminuer le risque de dégénérescence puisqu'aucune occurrence de ce phénomène n'a été observée dans les expériences relatées dans les parties 3.3 et 4.

3. Estimation par un pseudo algorithme EM

3.1. Difficultés pour utiliser un algorithme EM

[10] propose d'interpréter l'algorithme EM comme une optimisation alternée du critère suivant sur le couple $(\boldsymbol{\theta}, \mathbf{s})$:

$$C(\boldsymbol{\theta}, \mathbf{s}) = \sum_{h,i,k} s_{ik}^h \ln[\alpha_k^h \mathrm{p}(\mathbf{x}_i^h; \boldsymbol{\beta}_k^h)] - \sum_{h,i,k} s_{ik}^h \ln s_{ik}^h, \tag{6}$$

où $\mathbf{s} = (\mathbf{s}^h)_{\{h\}}$ avec $\mathbf{s}^h = (s_{ik}^h)_{\{i,k\}}$ tel que $s_{ik}^h \in [0,1]$ et $\sum_k s_{ik}^h = 1$. Partant d'un paramètre initial $\boldsymbol{\theta}^{[0]}$, l'itération q de EM, dans le cas de la contrainte entropique (5), s'exprime alors par

– **Étape E** : résoudre

$$\mathbf{s}^{[q]} = \arg\max_{\mathbf{s}} C(\boldsymbol{\theta}^{[q-1]}, \mathbf{s}),$$

ce qui conduit à $\mathbf{s}^{[q]} = \mathbf{t}(\boldsymbol{\theta}^{[q-1]})$;

– **Étape M** : résoudre

$$\begin{cases} \boldsymbol{\theta}^{[q]} = \arg\max_{\boldsymbol{\theta}} C(\boldsymbol{\theta}, \mathbf{s}^{[q]}) \\ \text{s.c. } e(\mathbf{t}^1(\boldsymbol{\theta}^1)) = \ldots = e(\mathbf{t}^H(\boldsymbol{\theta}^H)), \end{cases}$$

la solution de ce problème sous contrainte dépendant du modèle considéré.

L'algorithme converge vers un point fixe $(\boldsymbol{\theta}^{[*]}, \mathbf{s}^{[*]})$ qui vérifie $C(\boldsymbol{\theta}^{[*]}, \mathbf{s}^{[*]}) = \ell(\boldsymbol{\theta}^{[*]}; \mathbf{x})$, exhibant ainsi l'étroite relation entre le critère C et la log-vraisemblance. L'étape M est cependant difficile à résoudre étant donné le caractère hautement non linéaire de la contrainte.

Il est alors essentiel de remarquer pour l'approche alternative que nous allons présenter dans la partie suivante que la contrainte entropique est automatiquement vérifiée (donc non active) aussi à l'étape E, donc que l'algorithme EM suivant est strictement équivalent à l'algorithme EM précédent et sans effort supplémentaire :

– **Étape E** : résoudre

$$\begin{cases} \mathbf{s}^{[q]} = \arg\max_{\mathbf{s}} C(\boldsymbol{\theta}^{[q-1]}, \mathbf{s}) \\ \text{s.c. } e(\mathbf{s}^1) = \ldots = e(\mathbf{s}^H) \, ; \end{cases}$$

Soumis au Journal de la Société Française de Statistique
File: lourme_biernacki.tex, compiled with jsfds, version : 2009/12/09
date: 19 novembre 2011

– **Étape M** : résoudre

$$\begin{cases} \boldsymbol{\theta}^{[q]} = \arg\max_{\boldsymbol{\theta}} C(\boldsymbol{\theta}, \mathbf{s}^{[q]}) \\ \text{s.c. } e(\mathbf{t}^1(\boldsymbol{\theta}^1)) = \ldots = e(\mathbf{t}^H(\boldsymbol{\theta}^H)). \end{cases}$$

3.2. Présentation de l'algorithme ẼM

Partant de la remarque précédente, nous proposons alors l'algorithme *ad hoc* suivant dérivé de EM qui relâche la contrainte entropique imposée à l'étape M tout en la maintenant par contre à l'étape E. Cette étape est renommée pour l'occasion Ẽ puisque il ne s'agit plus à strictement parler d'une étape E classique à cause de la contrainte réellement active cette fois.

– **Étape Ẽ** : résoudre

$$\begin{cases} \mathbf{s}^{[q]} = \arg\max_{\mathbf{s}} C(\boldsymbol{\theta}^{[q-1]}, \mathbf{s}) \\ \text{s.c. } e(\mathbf{s}^1) = \ldots = e(\mathbf{s}^H) \ ; \end{cases}$$

– **Étape M** : résoudre

$$\boldsymbol{\theta}^{[q]} = \arg\max_{\boldsymbol{\theta}} C(\boldsymbol{\theta}, \mathbf{s}^{[q]}).$$

Cet algorithme, noté ẼM, vise donc à maximiser $C(\boldsymbol{\theta}, \mathbf{s})$ non plus sous la contrainte initiale (5) portant sur $\boldsymbol{\theta}$ mais sous la contrainte nouvelle $e(\mathbf{s}^1) = \ldots = e(\mathbf{s}^H)$ portant cette fois sur $\mathbf{s}$ mais de même nature. On ne cherche plus alors à maximiser la vraisemblance, c'est pourquoi il sera nécessaire de discuter des propriétés des estimateurs dans la partie suivante. Cependant l'avantage de cette approche est double :

– D'une part ẼM permet de grandement simplifier le problème d'optimisation en reportant à l'étape E, la contrainte (5) portant initialement sur la log-vraisemblance attendue à l'étape M. En effet la partition $\mathbf{s}$ optimale à l'étape E s'exprime par

$$s_{ik}^{h[q]} = \left\{ 1 + \sum_{k' \in \{1,\ldots,K\} \setminus \{k\}} \left[t_{ik'}^h \left(\boldsymbol{\theta}^{h[q-1]} \right) \Big/ t_{ik}^h \left(\boldsymbol{\theta}^{h[q-1]} \right) \right]^{1/(1-\lambda^h/n^h)} \right\}^{-1}$$

où les $(\lambda^h)_{\{h\}}$ sont des multiplicateurs de Lagrange tels que $\sum_h \lambda^h = 0$ et dont la valeur est approchée numériquement grâce à la contrainte entropique (5) portant sur $\mathbf{s}^{[q]}$. Ainsi le coût d'ẼM réside uniquement dans la détermination à chaque étape E, du vecteur $(\lambda^1, \ldots, \lambda^H)'$ comme zéro d'une fonction numérique de H variables, soumise à $H-1$ contraintes.

– D'autre part ẼM est utilisable sans effort supplémentaire pour n'importe quel type de modèle de mélange. En effet, l'étape Ẽ est indépendante du modèle et l'étape M correspond à une étape M classique du modèle considéré.

3.3. Evaluation des performances de ẼM

Considérations asymptotiques On a vu que le nouvel algorithme proposé ẼM ne vise pas à maximiser une vraisemblance à proprement parler, ce qui pose naturellement la question de la convergence des estimateurs de $\boldsymbol{\theta}$ obtenus. Nous allons répondre en deux temps. Nous notons $\hat{\boldsymbol{\theta}}$ et $\tilde{\boldsymbol{\theta}}$ respectivement l'estimateur du maximum de ℓ *sans* contrainte entropique obtenu par EM

(modèle noté $\hat{m}$) et l'estimateur du maximum de *C avec* contrainte entropique obtenu par ẼM (modèle noté $\tilde{m}$).

Supposons tout d'abord que le modèle $\tilde{m}$ soit exact. Dans ce cas, l'algorithme EM *sans* contrainte entropique est asymptotiquement (l'asymptotique portant sur les tailles d'échantillons $(n^h)_{\{h\}}$) équivalent à ẼM car la contrainte $e(\mathbf{t}^1(\hat{\boldsymbol{\theta}}^1)) = \ldots = e(\mathbf{t}^H(\hat{\boldsymbol{\theta}}^H))$ est asymptotiquement vérifiée. Comme $\hat{\boldsymbol{\theta}}$ est un estimateur convergent (propriétés habituelles du maximum de vraisemblance), on déduit que $\tilde{\boldsymbol{\theta}}$ est aussi convergent.

En général, on ne connaît pas bien sûr lequel des deux modèles $\tilde{m}$ ou $\hat{m}$ doit être retenu. Dans ce cas, on peut utiliser le critère BIC [19] défini comme suit pour chaque modèle $\hat{m}$ et $\tilde{m}$ respectivement :

$$\widehat{\mathrm{BIC}} = -2\ell(\hat{\boldsymbol{\theta}}) + |\boldsymbol{\theta}| \ln(n)$$

et

$$\widetilde{\mathrm{BIC}} = -2C(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{s}}) + (|\boldsymbol{\theta}| - H + 1) \ln(n),$$

où $|\boldsymbol{\theta}|$ dénote le nombre de paramètres *continus* dans $\boldsymbol{\theta}$, $n = \sum_h n^h$ correspond à la taille totale d'échantillon et $\tilde{\mathbf{s}}$ est l'optimum de $\mathbf{s}$ obtenu dans ẼM. En pratique, on retient le modèle qui conduit à la valeur la plus faible de BIC. Bien entendu, la définition du critère BIC pour le modèle $\tilde{m}$ est *ad hoc* car on ne pénalise pas une log-vraisemblance à strictement parler mais cette définition sera efficace pour choisir entre $\hat{m}$ et $\tilde{m}$ comme on le voit maintenant. En effet, deux situations se présentent :

- si $\tilde{m}$ est le vrai modèle, BIC le retiendra asymptotiquement car (i) on a vu précédemment que EM et ẼM sont dans ce cas asymptotiquement équivalents et (ii) le critère BIC est convergent (voir [12] pour les propriétés de BIC) ;
- si $\hat{m}$ est le vrai modèle, BIC choisira $\hat{m}$ asymptotiquement car cette fois-ci $n^{-1}\ell(\hat{\boldsymbol{\theta}}) - n^{-1}C(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{s}}) \xrightarrow{p} a > 0$ asymptotiquement.

Considérations empiriques à taille finie Pour compléter ces propriétés de convergence et de choix de modèle à taille finie d'échantillons, nous menons maintenant quelques expériences sur données simulées.

Le vrai modèle consiste en deux mélanges gaussiens ($H = 2$) définis sur $\mathbb{R}^2$. Chaque mélange est homoscédastique d'ordre deux ($K = 2$) et ses composantes ont le même poids. Dans le vrai modèle, l'entropie globale des classes est homogène d'un mélange à l'autre.

Le modèle inféré consiste en deux mélanges de même nature que le vrai modèle. Chacun d'eux est homoscédastique (homoscédasticité intrapopulation) d'ordre deux, aux composantes de même poids. On suppose de l'entropie des classes qu'elle est soit homogène soit libre d'un mélange à l'autre. Dans le premier cas l'estimateur $\tilde{\boldsymbol{\theta}}$ du paramètre est obtenu par l'algorithme ẼM. Dans le second cas, l'estimateur $\hat{\boldsymbol{\theta}}$ du paramètre est obtenu par EM.

La taille $n_k^h \in \{10, 20, 50, 100, 200\}$ des données conditionnelles (à la classe k et à la population h) étant fixée, 200 échantillons sont générés, sur lesquels les paramètres $\tilde{\boldsymbol{\theta}}$ et $\hat{\boldsymbol{\theta}}$ sont inférés. Le tableau 1 donne le pourcentage d'occurrences où BIC choisit le modèle $\tilde{m}$ plutôt que $\hat{m}$ et l'erreur de classement empirique moyennée $\tilde{\tau}$ et $\hat{\tau}$ obtenue par $\tilde{m}$ et $\hat{m}$ respectivement (l'écart-type relatif est donné entre parenthèses). En outre, deux situations distinctes sont considérées pour le vrai modèle : les composantes des mélanges se chevauchent fortement (30% d'erreur de Bayes) ou

faiblement (5% d'erreur de Bayes).

	Fort chevauchement			*Faible* chevauchement		
n_k^h	$\%[\widetilde{\text{BIC}} > \widehat{\text{BIC}}]$	$\hat{\tau}$	$\tilde{\tau}$	$\%[\widetilde{\text{BIC}} > \widehat{\text{BIC}}]$	$\hat{\tau}$	$\tilde{\tau}$
10	96.0	34.02 (7.96)	34.65 (7.78)	97.0	13.80 (9.40)	14.50 (9.90)
20	96.0	34.97 (5.89)	35.21 (5.74)	98.0	9.10 (6.60)	9.10 (6.50)
50	98.5	36.18 (4.94)	36.37 (5.11)	96.5	7.20 (4.38)	7.04 (3.80)
100	99.5	36.47 (4.71)	36.32 (4.73)	98.0	6.26 (2.33)	6.14 (1.82)
200	100.0	36.30 (4.80)	36.30 (4.30)	99.5	6.32 (2.71)	6.00 (1.68)

TABLEAU 1. *Considérations empiriques pour ẼM dans le cas d'un chevauchement fort puis faible des composantes (erreur de Bayes de respectivement* 30% *et* 5%*).*

Quelle que soit la taille d'échantillon et quel que soit le degré de chevauchement des composantes dans le vrai modèle, on observe que BIC préfère très largement $\tilde{m}$ à $\hat{m}$. Ainsi BIC détecte bien l'homogénéité de l'entropie comme une contrainte du vrai modèle, même pour une taille d'échantillon modeste.

Le comportement du taux d'erreur est quant à lui plus sensible que celui de BIC au chevauchement des classes et dans une moindre mesure à la taille d'échantillon. Lorsque les composantes sont fortement séparées, ce taux détecte bien l'homogénéité de l'entropie comme une contrainte vraie car $\tilde{\tau}$ est meilleur que $\hat{\tau}$ même pour une assez faible taille d'échantillon. Par contre, lorsque les composantes des mélanges se chevauchent fortement, la différence entre les taux d'erreurs liés à $\tilde{m}$ et $\hat{m}$ est moins significative que dans le cas précédent. De plus la convergence de $\tilde{\tau}$ (comme celle de $\hat{\tau}$) est à la fois plus lente et plus erratique que lorsque les composantes sont plus séparées.

Ainsi, ces expériences sont concluantes pour nous rassurer sur le fait que ẼM permet bien d'une part de produire une valeur de BIC « exploitable » et d'autre part d'accélérer la vitesse de convergence des estimateurs du maximum de vraisemblance lorsque le modèle $\tilde{m}$ est réaliste. En outre, il apparaît que cette accélération est d'autant plus importante que les classes sont bien séparées dans le vrai modèle.

Considérations de vitesse d'estimation ẼM se distingue d'EM par l'étape Ẽ où les probabilités conditionnelles sont estimées sous la contrainte (5). L'estimation de $\boldsymbol{\theta}$ par ẼM est donc plus longue que par EM. L'écart entre la durée d'exécution des deux algorithmes dépend du nombre et de la taille des échantillons ainsi que du nombre de classes par échantillon ; mais cet écart ne dépend pas, en revanche, de la dimension de l'espace, ni du chevauchement des classes.

Le tableau 2 indique l'évolution du rapport entre la durée d'ẼM et celle d'EM, lorsque n_k^h (la taille des échantillons conditionnels) croît. Pour chaque valeur de $n_k^h \in \{50, 100, \ldots, 500\}$, deux échantillons de $\mathbb{R}^2$ sont générés, chacun selon un mélange gaussien homoscédastique aux composantes de même poids. Le modèle inféré par EM et par ẼM consiste en deux mélanges gaussiens de même nature que ceux du vrai modèle. Les deux algorithmes ẼM et EM comportent le même nombre d'itérations réalisées sur la même machine, à partir de la même valeur initiale du paramètre. On remarque que le rapport entre la durée d'ẼM et celle d'EM, diminue quand n_k^h

augmente, ce qui signifie que le coût additionnel de l'étape Ẽ devient négligeable quand la taille d'échantillon augmente. Ce comportement a aussi été observé lorsque l'homogénéité de l'entropie des classes n'est pas une hypothèse du vrai modèle (expériences non reportées dans cet article).

n_k^h	50	100	150	200	250	300	350	400	450	500
ẼM/EM	11.9	9.2	7.2	5.9	4.8	4.2	3.4	3.2	2.9	2.6

TABLEAU 2. *Rapport entre la durée de l'estimation par ẼM et par EM lorsque la taille des échantillons conditionnels (n_k^h) croît.*

4. Expériences numériques sur données biologiques

4.1. Données continues avec modèle gaussien

Nous considérons ici deux échantillons d'oiseaux de mer ($H = 2$) vivant dans des zones géographiques distinctes, correspondant chacun à une sous-espèce de Puffins cendrés [20] : *Calonectris diomedea borealis* ($n^1 = 206$ oiseaux dont 45% de femelles) et *Calonectris diomedea diomedea* ($n^2 = 38$ oiseaux dont 58% de femelles). Les échantillons sont décrits par le même ensemble de cinq variables morphométriques ($d = 5$) : la longueur et la hauteur du bec, le tarse, la longueur de l'aile et de la queue. La figure 1 représente les deux échantillons dans le plan canonique : hauteur $\times$ longueur du bec. Il est plausible d'après les biologistes ayant mesuré les variables, que les mâles et les femelles soient issus de distributions presque gaussiennes et mélangés de façon similaire dans les deux sous-espèces. Ainsi les descripteurs biométriques disponibles auraient la même aptitude chez *borealis* et chez *diomedea*, à permettre de déterminer le genre des oiseaux. Nous allons voir que l'inférence d'un modèle par l'algorithme ẼM corrobore cette hypothèse.

Nous proposons deux situations classiques quant aux mélanges gaussiens modélisant les oiseaux de chaque espèce : les composantes sont homoscédastiques ($\mathbf{\Sigma}^h$) ou hétéroscédastiques ($\mathbf{\Sigma}_k^h$). Dans les deux cas les proportions du mélange sont supposées égales, ce qui est naturel dans le cas de l'équilibre de l'espèce. Chacun de ces deux modèles *intra-populations* ainsi obtenus peut être combiné avec les deux modèles *inter-populations* précédents $\tilde{m}$ et $\hat{m}$. Rappelons que $\tilde{m}$ suppose un chevauchement homogène des composantes et l'estimation se fait par l'algorithme ẼM tandis que $\hat{m}$ n'impose aucun lien et l'estimation se fait par un algorithme EM classique (voir [16], chap. 3, par exemple pour une description détaillée du modèle et de l'algorithme). Le tableau 3 indique la valeur de BIC et le taux d'erreur de classement des puffins, obtenus par les différentes combinaisons de modèles intra- et inter-populations.

	$\mathbf{\Sigma}^h$	$\mathbf{\Sigma}_k^h$
$\tilde{m}$	**6092.8** (**11.48**)	6201.6 (16.80)
$\hat{m}$	6095.0 (12.30)	6201.8 (45.08)

TABLEAU 3. *BIC (et % d'erreur de classement) obtenus pour la classification des borealis et des diomedea.*

Soumis au Journal de la Société Française de Statistique
File: lourme_biernacki.tex, compiled with jsfds, version : 2009/12/09
date: 19 novembre 2011

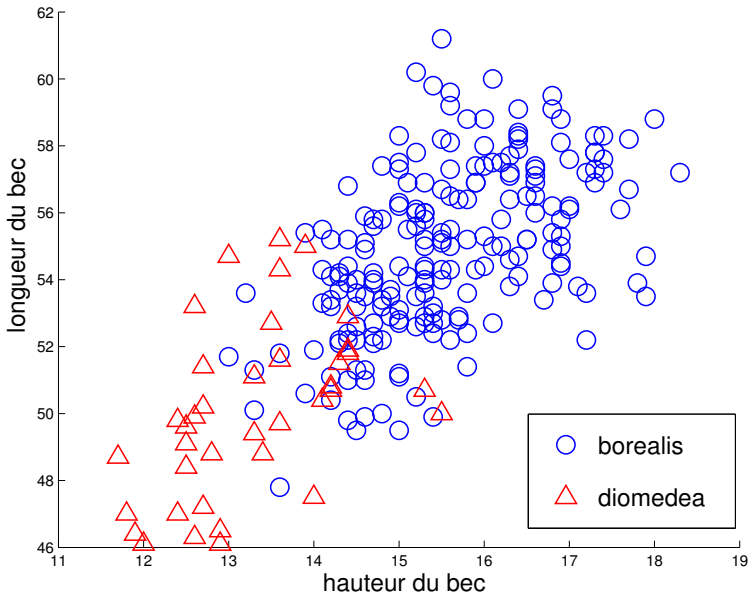
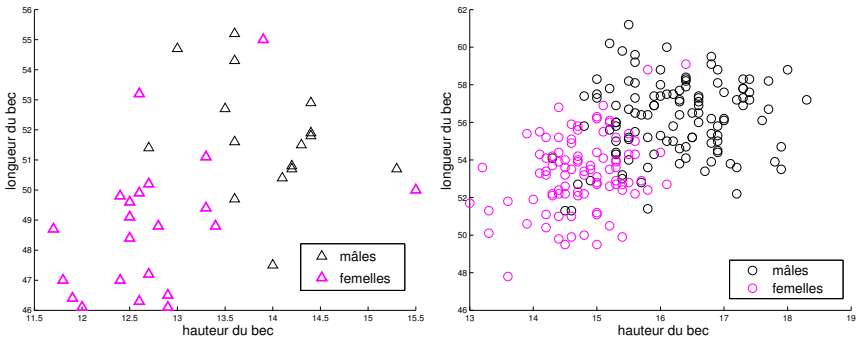


FIGURE 1. *Deux sous-espèces de Calonectris diomedea (Puffins cendrés).*



(a) *La partition recherchée des diomedea* (b) *La partition recherchée des borealis*

FIGURE 2. *Un chevauchement similaire des mâles et des femelles dans les deux populations d'oiseaux ?*

On observe que la meilleure combinaison de modèles inter- et intra-populations est obtenue dans le cas homoscédastique avec $\tilde{m}$, tant du point de vue du critère BIC que du critère d'erreur de classement. Ceci corrobore l'hypothèse des biologistes selon laquelle le chevauchement des mâles et des femelles est homogène parmi *borealis* et *diomedea* : les descripteurs biométriques ont donc globalement la même aptitude, chez ces deux sous-espèces, à séparer les mâles des femelles. Deux éléments supplémentaires plaident en faveur de cette hypothèse. D'une part les figures 2(a) et 2(b) indiquant le genre des oiseaux, montrent un chevauchement similaire des mâles et des femelles dans les deux sous-espèces. D'autre part l'inférence d'un mélange gaussien homoscédastique sur chaque échantillon (en supposant le genre connu cette fois) conduit à un recouvrement similaire des mâles et des femelles puisque l'entropie normalisée (4) est de 0.14 chez les *borealis* et de

0.08 chez les *diomedea*, ces deux valeurs étant non significativement différentes comme des tests statistiques standards pourraient aisément le vérifier.

4.2. Données binaires avec modèle des classes latentes

Nous considérons maintenant un échantillon de 69 oiseaux provenant de deux sous-espèces *lherminieri* (34 individus) et *subalaris* (35 individus) tous deux de la famille des *Procellaridae* et décrits par les quatre mêmes variables binaires [4] : sourcil (présence/absence), collier (continu ou presque/discontinu), sous-caudale (couleur unie/couleur panachée) et liseret (présence/absence). Nous considérons les deux populations ($H = 2$) des mâles (47.83% de l'échantillon) et les femelles et nous cherchons à retrouver l'espèce de chaque oiseau ($K = 2$). Nous mettons de nouveau en compétition les modèles inter-populations $\tilde{m}$ et $\hat{m}$ mais cette fois dans le cas du modèle des classes latentes [9] pour lequel on impose les proportions égales (les deux espèces sont en effectif semblable dans l'exemple). On pourra se référer à [5] pour une description de l'algorithme EM associé au modèle des classes latentes. Le tableau 4 indique la valeur de BIC et le taux d'erreur de classement des oiseaux pour les deux modèles inter-population $\tilde{m}$ et $\hat{m}$ et l'unique modèle intra-population.

$\tilde{m}$	294.2 (**15.94**)
$\hat{m}$	295.1 (18.84)

TABLEAU 4. *BIC (et % d'erreur de classement) obtenus pour la classification des lherminieri et des subalaris.*

Il ressort que le modèle $\tilde{m}$ est de nouveau retenu par BIC et conduit également au taux d'erreur le plus faible des deux modèles en compétition. Cela confirme l'hypothèse des biologistes qui considèrent qu'il y a peu de différences entre mâles et femelles pour les variables binaires de cette étude et que, par conséquence immédiate, les entropies sont quasiment égales entre sous-espèces d'un sexe à l'autre.

5. Conclusion et perspectives

Ce travail propose un modèle de liaison entre plusieurs échantillons à classifier en établissant un lien d'égalité entropique de partition entre eux. Cette hypothèse permet de s'affranchir de l'hypothèse de mélange qui est faite et aussi d'éviter la concordance exacte des variables entre les différentes populations considérées. De ce point de vue, il s'agit donc, à notre connaissance, du 1er modèle générique de classification dite simultanée.

Ce modèle possède *a priori* deux inconvénients. Il est d'une part assez peu parcimonieux puisqu'il impose peu de contraintes sur l'espace des paramètres par rapport au modèle général sans égalité entropique. D'autre part, il est difficile de maximiser la vraisemblance par un algorithme EM traditionnel, ce qui oblige à recourir à une version approchée qui est l'algorithme ẼM. Cependant, malgré son manque de parcimonie, les exemples numériques ont montré que la contrainte entropique pouvait être suffisamment pertinente pour améliorer les résultats, tant d'un point de vue du critère BIC que du taux d'erreur. En outre, l'algorithme ẼM a l'avantage d'être

réellement indépendant du modèle de mélange retenu (la contrainte est reportée de l'étape M à l'étape E) et son comportement pratique est plutôt encourageant même s'il faut être prudent dans le cas de composantes peu séparées et de faibles tailles d'échantillons comme le laissent penser les données simulées.

Les extensions possibles de ce travail sont nombreuses.

Tout d'abord, il serait instructif d'essayer ẼM sur des jeux de données où les variables ne concordent pas strictement d'un point de vue sémantique. L'hypothèse d'entropie homogène resterait probablement pertinente dans la partie 4.1 si, par exemple, la hauteur du bec était mesurée au niveau de la narine dans un échantillon d'oiseaux et à la base du bec dans l'autre échantillon, les variables étant de même nature.

Dans les applications numériques de la partie 4, les échantillons sont modélisés par des mélanges appartenant à la même famille de lois. Il est cependant possible d'envisager ẼM lorsque les populations appartiennent à des familles distinctes. Par exemple, on pourrait envisager de déterminer simultanément le sexe de deux échantillons d'oiseaux lorsque l'un d'eux est décrit par des variables continues et l'autre par des variables discrètes.

Par ailleurs, il est facile de combiner le modèle d'entropie homogène avec d'autres modèles de classification simultanée (par exemple ceux de [13] ou de [14]) grâce au caractère générique d'ẼM.

Les trois extensions précédentes sont directement réalisables avec l'algorithme ẼM décrit dans la partie 3.2. L'extension suivante en revanche nécessite de généraliser ẼM.

Dans ce travail, le nombre K de classes recherchées est le même dans chaque échantillon et l'on suppose que l'entropie de ces K classes est identique d'un échantillon à l'autre. On pourrait envisager que le nombre de classes recherchées varie selon l'échantillon (K^h pour l'échantillon h) et supposer qu'il existe dans chaque échantillon, K classes parmi K^h ($K \leq \min\{K^1, \ldots, K^H\}$) dont l'entropie ne dépend pas de h. Ce modèle revient à lier K des composantes de chaque mélange par la contrainte entropique, et à laisser libre les $K^h - K$ composantes restant. Cette extension élargit considérablement le contexte de la classification simultanée basée sur une hypothèse d'entropie homogène, mais elle implique une difficulté supplémentaire. Le nombre de combinaisons possibles dans le choix des K classes d'entropie homogène croît très vite avec H d'une part et avec le nombre K^h de groupes recherchés dans chaque échantillon d'autre part.

Références

[1] J. D. BANFIELD et A. E. RAFTERY : Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

[2] C. BIERNACKI et G. CASTELLAN : A Data-Driven Bound on Variances for Avoiding Degeneracy in Univariate Gaussian Mixtures. *Pub. IRMA Lille*, 71, 2011.

[3] C. BIERNACKI, G. CELEUX et G. GOVAERT : Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference*, pages 2991–3002, 2010.

[4] V. BRETAGNOLLE : Personal communication. source : Museum, 2007.

[5] G. CELEUX et G. GOVAERT : Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8(2):157—176, 1991.

[6] G. CELEUX et G. GOVAERT : Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

Soumis au Journal de la Société Française de Statistique
File: lourme_biernacki.tex, compiled with jsfds, version : 2009/12/09
date: 19 novembre 2011

[7] A. P. DEMPSTER, N. M. LAIRD et D. B. RUBIN : Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, B 39:1–38, 1977.

[8] P. DU JARDIN et E. SÉVERIN : Dynamic analysis of the business failure process : a study of bankruptcy trajectories. *In Portuguese Finance Network*, Ponte Delgada, Portugual, 2010.

[9] L. A. GOODMAN : Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215—231, 1974.

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

[11] D. KARLIS : An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, 30:63–77, 2002.

[12] E. LEBARBIER et T. MARY-HUARD : Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la SFdS*, 147(1):39–58, 2006.

[13] A. LOURME et C. BIERNACKI : Simultaneous Gaussian model-based clustering for samples of multiples origins. Pub. IRMA 70-VII, University Lille 1, Lille, 2010.

[14] A. LOURME et C. BIERNACKI : Simultaneous t-model-based clustering for data differing over time period : Application for understanding companies financial health. *Case Studies in Business, Industry and Government Statistics (CSBIGS)*, 2011. in press.

[15] K. S. MARDIA et P. E. JUPP : *Directional Statistics*. Wiley, New York, 2000.

[16] G. J. MCLACHLAN et D. PEEL : *Finite Mixture Models*. Wiley, New York, 2000.

[17] T.B. MURPHY et D. MARTIN : Mixtures of distance-based models for ranking data. *Comput. Statist. Data Anal.*, 41(3-4):645–655, 2003.

[18] N.J. SCHORK et B. THIEL : Mixture distributions in human genetics. *Statistical Methods in Medical Research*, 39:155–178, 1996.

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

[20] J.C. THIBAULT, V. BRETAGNOLLE et C. RABOUAM : Cory's shearwater calonectris diomedea. *Birds of Western Paleartic Update*, 1:75–98, 1997.

Soumis au Journal de la Société Française de Statistique
File: lourme_biernacki.tex, compiled with jsfds, version : 2009/12/09
date: 19 novembre 2011