

REVUE DE STATISTIQUE APPLIQUÉE

G. CELEUX

Classification et modèles

Revue de statistique appliquée, tome 36, n° 4 (1988), p. 43-57

http://www.numdam.org/item?id=RSA_1988__36_4_43_0

© Société française de statistique, 1988, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

CLASSIFICATION ET MODÈLES

G. CELEUX

*INRIA Domaine de Voluceau,
78153 Le Chesnay Cedex France*

RÉSUMÉ

On établit les liens entre des critères classiques de classification et des modèles de statistique inférentielle. Le critère d'inertie interclasse pour des données quantitatives est associé aux mélanges gaussiens. Le critère d'information mutuelle et le critère du χ^2 pour des données qualitatives sont associés au modèle des classes latentes. Cette présentation inférentielle de critères largement répandus de classification permet une analyse précise de leurs propriétés, de leurs limites et de leurs champs d'application.

Mots clés : Classification, Inertie interclasse, Mélange gaussien, Information mutuelle, χ^2 , Modèle des classes latentes.

ABSTRACT

We establish the relations between some well-known criteria for clustering and some statistical models. The criteria for continuous data are associated with Gaussian mixture. The Kullback information criterion and the χ^2 criterion for discrete data are associated with the latent classes model. This presentation of clustering criteria in an inferential framework allows us to study in a precise way the features and limits of many techniques of numerical classification.

Keywords: Clustering, Inertia, Gaussian mixture, Kullback information, χ^2 criterion, Latent classes model.

1. Introduction

Les méthodes d'analyse des données ont été conçues dans un cadre géométrique sans référence, en général, à des modèles probabilistes. Depuis quelque temps, un nombre croissant d'auteurs s'intéressent aux rapports de ces méthodes à des modèles probabilistes. Le numéro spécial "Lancaster-Toulouse" de la R.S.A. Vol.35 n° 3 (1987) témoigne bien de cette tendance. Cette orientation de la recherche permet bien souvent de préciser de manière rigoureuse les champs d'application de techniques largement répandues et constitue une voie privilégiée pour l'évaluation de la stabilité et de la validité d'analyses statistiques multidimensionnelles, une question elle aussi d'actualité.

Dans cet article, nous établissons les liens étroits existants entre les critères les plus classiquement utilisés en classification et des critères de vraisemblance utilisés

pour l'identification de mélanges de distributions de probabilité. Précisément, nous donnons la signification du critère d'inertie interclasse, utilisé pour la classification d'individus décrits par des variables quantitatives, pour le modèle de mélanges gaussiens, et la signification du critère du χ^2 , utilisé pour la classification d'individus décrits par des variables qualitatives, pour le modèle des classes latentes. Cette présentation inférentielle de critères traditionnels de classification permet une analyse précise des propriétés, des limites et des problèmes de protection des méthodes optimisant ces critères.

2. Identification d'un mélange fini par la classification

Soit un échantillon $E = (x_1, \dots, x_N)$ d'une variable aléatoire (v.a.) X à valeurs dans \mathbb{R}^d , dont la loi admet la densité :

$$f(x) = \sum_{k=1}^K p_k f(x, a_k)$$

avec

$$\forall k = 1, K \quad 0 < p_k < 1 \quad \text{et} \quad \sum_{k=1}^K p_k = 1$$

$f(\cdot, a)$ est une densité de probabilité appartenant à une famille paramétrée de densités dépendant du paramètre a de \mathbb{R}^s ($s \geq 1$).

p_k est la probabilité qu'un point de l'échantillon suive la loi de densité $f(\cdot, a_k)$. On appellera les $p_k, k = 1, K$ les proportions du mélange fini.

Le problème consiste à estimer le nombre K de composants et les paramètres inconnus ($q_k = (p_k, a_k), k = 1, K$) au vu de l'échantillon E . Il s'agit typiquement d'un problème d'estimation de paramètres. Les méthodes les plus efficaces pour le résoudre recherchent de manière itérative les estimateurs du maximum de vraisemblance des paramètres (cf. Everitt, Hand (1981), Celeux, Diebolt (1986), Celeux (1987)). Ici, nous ne traiterons pas de cette approche "estimation" du problème. Nous nous concentrerons sur l'approche "classification" de l'identification d'un mélange.

Cette approche (Scott, Symons(1971), Schroeder (1976)) consiste à redéfinir le problème de la manière suivante :

Rechercher une partition $P = (P_1, \dots, P_K), K$ étant supposé connu, telle que chaque classe P_k soit assimilable à un sous échantillon suivant la loi $f(\cdot, a_k)$.

Dans ce cadre, les algorithmes utilisés sont de type Nuées Dynamiques (Schroeder (1976)) et visent à maximiser le critère de vraisemblance classifiante

$$W(a, P) = \sum_{k=1}^K \ell_n L(P_k, a_k)$$

a désignant (a_1, \dots, a_K) et la fonction $L(P_k, a_k)$ étant la vraisemblance du sous échantillon P_k suivant la loi $f(\cdot, a_k)$.

L'algorithme se déroule ainsi. A partir d'une partition P° en K classes de l'échantillon, on applique successivement deux fonctions g et h jusqu'à obtention d'une partition stable.

g , appelée fonction de représentation, est définie par :

$$g(P) = g(P_1, \dots, P_K) = (a_1, \dots, a_K)$$

avec a_k : estimation du maximum de vraisemblance du paramètre de la densité associée au sous échantillon P_k .

h , appelée fonction d'affectation, est définie par :

$$h(a) = h(a_1, \dots, a_K) = (P_1, \dots, P_K)$$

avec

$$P_k = \{x \in E / f(x, a_k) \geq f(x, a_m) \text{ avec } k < m \text{ en cas d'égalité} \}.$$

On montre facilement que cet algorithme fait croître le critère à chaque itération. De plus, sous l'hypothèse que la famille de densités $f(x, a)$ est bornée supérieurement pour tout x et pour tout a , la suite de itérés (a^n, P^n) est stationnaire et atteint son état stationnaire en un nombre fini d'itérations. A la convergence, on obtient une partition P et une estimation des paramètres (a_1, \dots, a_K) . Il est à noter que cet algorithme ne s'attaque pas directement à l'estimation des proportions du mélange. Toutefois, à la convergence, il fournit comme sous-produit une estimation naturelle des $(p_k, k = 1, K)$ par les quantités $(\frac{\text{card } P_k}{N}, k = 1, K)$.

3. Mélanges gaussiens et classification de données quantitatives

Pour obtenir une partition en K classes d'une population de N individus (x_1, \dots, x_N) décrits par d variables quantitatives ($x_i \in \mathbb{R}^d, i = 1, N$), la pratique la plus courante consiste à minimiser l'inertie intraclasse de la partition, l'espace \mathbb{R}^d étant muni d'une métrique euclidienne M , par un algorithme itératif (Nuées Dynamiques, algorithme de transferts, ...). Suivant le choix de M , le critère à optimiser prend une forme particulière. Dans ce paragraphe, nous allons voir que les critères d'inertie les plus utilisés peuvent s'interpréter comme des vraisemblances classifiantes induites par des mélanges finis de lois gaussiennes.

Dans le cas d'un mélange gaussien, les paramètres $(a_k, k = 1, K)$ s'écrivent $a_k = (m_k, \Gamma_k)$ avec m_k : espérance du composant numéro k
 Γ_k : matrice variance du composant numéro k .

Le critère de vraisemblance classifiante s'écrit :

$$W(a, P) = \text{Cte} - \frac{1}{2} \sum_{k=1}^K \sum_{x_i \in P_k} \left[(x_i - m_k)' \Gamma_k^{-1} (x_i - m_k) + \ell n |\Gamma_k| \right]$$

$|\Gamma_k|$ désignant le déterminant de Γ_k .

Donc maximiser $W(a, P)$ revient à minimiser le critère :

$$C(a, P) = \sum_{k=1}^K \sum_{x_i \in P_k} \left[(x_i - m_k)' \Gamma_k^{-1} (x_i - m_k) + \ell n |\Gamma_k| \right]$$

que nous considérons dans la suite.

Nous allons voir ce que devient ce critère selon différentes hypothèses sur les matrices variance des composants du mélange.

Premier cas

$\Gamma_k = \Gamma, \forall k = 1, k; \Gamma$ étant supposé connu. Le critère à minimiser devient, à une constante près :

$$C(a, P) = \sum_{k=1}^K \sum_{x_i \in P_k} (x_i - m_k)' \Gamma^{-1} (x_i - m_k).$$

Γ étant symétrique, il existe une matrice T telle que $\Gamma = TT'$. Par le changement de variable $y = T^{-1}x$, on peut se ramener au cas où $\Gamma = I$. Donc le critère peut s'écrire :

$$C(a, P) = \sum_{k=1}^K \sum_{x_i \in P_k} (x_i - m_k)' (x_i - m_k).$$

La partition étant fixée, l'estimateur du maximum de vraisemblance de m_k est la moyenne empirique \hat{m}_k de la classe P_k (son centre de gravité) et le critère s'écrit alors :

$$C(a, P) = \sum_{k=1}^K \sum_{x_i \in P_k} (x_i - \hat{m}_k)' (x_i - \hat{m}_k).$$

Il s'agit du critère de la version la plus simple et la plus utilisée des Nuées Dynamiques (algorithmes des "centres mobiles" ou de "réallocation-recentrage").

Cette présentation explique clairement le fait, constaté expérimentalement, que cet algorithme ait tendance à donner des classes sphériques de même volume.

Deuxième cas

$\Gamma_k = \Gamma, \forall k = 1, K; \Gamma$ supposé inconnu. Dans ce cas, la partition étant fixée, l'estimateur du maximum de vraisemblance de Γ est $\frac{W}{N}$ où :

$$W = \sum_{k=1}^K \sum_{x_i \in P_k} (x_i - \hat{m}_k)(x_i - \hat{m}_k)'$$

Le critère à minimiser devient :

$$C(a, P) = \sum_{k=1}^K \sum_{x_i \in P_k} \left[N(x_i - \hat{m}_k)' W^{-1} (x_i - \hat{m}_k) + \ell n \left| \frac{W}{N} \right| \right]$$

$$C(a, P) = N \operatorname{tr} (W W^{-1}) + N \ell n \left(\frac{|W|}{N^d} \right).$$

Il se réduit donc à une constante près (et puisque $\ln(x)$ est une fonction croissante de x) à :

$$C(a, P) = |W|.$$

Ce critère de classification a été proposé par Friedman, Rubin (1967) et Govaert (1975) sans aucune référence au modèle gaussien. Le but de ces auteurs était précisément de proposer un algorithme de type Nuées Dynamiques capable de reconnaître des classes ayant le même type de dispersion mais possédant des directions d'allongement inconnues, ces classes n'étant pas nécessairement assimilables à des sphères. Pour ce faire, ils utilisent à chaque itération la métrique $M = |W|^{1/p} W^{-1}$, en justifiant ce choix par des arguments géométriques.

Troisième cas

Les Γ_k sont différents entre eux et inconnus. A chaque étape, la partition étant fixée, les estimateurs du maximum de vraisemblance des Γ_k sont les $V_k = \frac{W_k}{\text{Card } P_k}$ avec :

$$W_k = \sum_{x_i \in P_k} (x_i - \hat{m}_k)(x_i - \hat{m}_k)'$$

Le critère à minimiser devient :

$$C(a, P) = \sum_{k=1}^K \sum_{x_i \in P_k} \left[(x_i - m_k)' V_k^{-1} (x_i - m_k) + \ln |V_k| \right]$$

$$C(a, P) = \sum_{k=1}^K \text{tr} (\text{card } P_k V_k V_k^{-1}) + \sum_{k=1}^K \text{card } P_k \ln |V_k|$$

il se réduit donc à une constante près à :

$$C(a, p) = \sum_{k=1}^K \text{card } P_k \ln |V_k|.$$

Par ailleurs, le critère de l'algorithme des distances adaptatives (Govaert (1975)) élaboré dans un cadre géométrique pour permettre de reconnaître des classes de "formes" différentes est :

$$C'(a, P) = \sum_{k=1}^K \sum_{x_i \in P_k} (x_i - \hat{m}_k)' |V_k|^{1/d} V_k^{-1} (x_i - \hat{m}_k).$$

En effet, à chaque classe P_k cet algorithme associe la métrique $|W_k|^{1/d} W_k^{-1}$ pour laquelle on mesure la distance de chaque point à son centre de gravité. Le critère peut s'écrire :

$$C'(a, P) = \sum_{k=1}^K \text{card } P_k |V_k|^{1/d} \text{tr} (V_k V_k^{-1})$$

il devient à une constante multiplicative près

$$C'(a, P) = \sum_{k=1}^K \text{card } P_k |V_k|^{1/d} = \sum_{k=1}^K |W_k|^{1/d}.$$

Les deux critères ne sont pas identiques, mais sont très analogues. En pratique, les deux méthodes donnent des résultats quasi identiques (cf. Govaert (1975)).

Nous venons de voir que les algorithmes de partitionnement utilisant un critère d'inertie peuvent se présenter comme des méthodes pour identifier un mélange gaussien à l'aide d'une classification, ce qui a permis de mettre en évidence leurs principales caractéristiques. Un autre point mérite d'être noté. Le critère de vraisemblance classifiante ne tient pas compte des différentes proportions p_k . En conséquence, il amène à traiter les composantes à parts égales quelles que soient les valeurs des vrais proportions et a ainsi tendance à produire des classes d'effectifs égaux. Il s'agit, en pratique, d'une des caractéristiques la plus nette de ce type d'algorithme.

Cette caractéristique a incité Symons (1981) à proposer un critère différent de la vraisemblance classifiante pour l'identification d'un mélange gaussien par classification. Partant du fait que le problème de classification revient à estimer les N paramètres inconnus $(o(i), i = 1, N)$ avec $o(i) =$ numéro du composant dont x_i est issu, il écrit la vraisemblance de l'échantillon sachant $q = ((p_k, m_k, \Gamma_k), k = 1, K)$ et $o = (o(i), i = 1, N)$, (au facteur $(2\pi)^{-N/2}$ près) :

$$L(x_1, \dots, x_N \mid q, o) = \prod_{k=1}^K (p_k)^{n_k} \mid \Gamma_k \mid^{-\frac{1}{2}n_k} \exp \left[-\frac{1}{2} \sum_{k=1}^K \sum_{x_i \in P_k} (x_i - m_k)' \Gamma_k^{-1} (x_i - m_k) \right]$$

avec $n_k = \text{card } P_k$.

A partition fixée, il s'ensuit que les estimateurs du maximum de vraisemblance des paramètres q sont :

$$\begin{aligned} \hat{p}_k &= \frac{\text{card } P_k}{N} \\ \hat{m}_k &= \frac{1}{\text{card } P_k} \sum_{x_i \in P_k} x_i \\ \hat{\Gamma}_k &= \frac{W_k}{\text{card } P_k}. \end{aligned}$$

Suivant les trois cas distingués précédemment, il est facile de voir que maximiser la vraisemblance revient à minimiser les critères suivants :

Premier cas : $\Gamma_k = \Gamma, \forall k = 1, K; \Gamma$ connu; on supposera, sans restreindre la généralité, que Γ est égale à I .

$$S(q, P) = \sum_{k=1}^K \left[\sum_{x_i \in P_k} (x_i - \hat{m}_k)' (x_i - \hat{m}_k) - 2 \text{card } P_k \ln(\text{card } P_k) \right]$$

Deuxième cas : $\Gamma_k = \Gamma, \forall k = 1, K; \Gamma$ inconnu

$$S(q, P) = N \ln |W| - 2 \sum_{k=1}^K \text{card } P_k \ln(\text{card } P_k).$$

Troisième cas : les Γ_k sont différents entre eux et inconnus

$$S(q, P) = \sum_{k=1}^K \text{card } P_k \ell n |W_k| - 2 \sum_{k=1}^K \text{card } P_k \ell n (\text{card } P_k).$$

Fondamentalement, ces critères différents des précédents par l'introduction du terme $-2 \sum_{k=1}^K \text{card } P_k \ell n (\text{card } P_k)$.

En pratique, Symons note que ces critères ont moins tendance que les précédents à fournir des classes d'effectifs égaux mais ont tendance à surestimer l'effectif des grandes classes. En conclusion de son étude, il s'interdit de recommander ce type de critère plutôt que le critère de vraisemblance classifiante et invite les praticiens à faire leur choix sur la base de comparaisons de résultats.

4. Modèles de classes latentes et classification de données qualitatives

Pour obtenir une partition en K classes d'une population de N individus décrits par des variables qualitatives, une méthode classique consiste à raisonner à partir du tableau de données mis sous forme disjonctive complète et à chercher la partition des individus qui, croisée avec les variables, maximise le χ^2 de contingence. Cela revient à rechercher la partition d'inertie interclasse maximum pour la métrique du χ^2 . Dans cette section, nous allons voir que cette méthode peut s'interpréter comme une méthode d'identification du modèle des classes latentes. Pour mettre ce fait en évidence, nous étudierons au préalable les liens entre le modèle des classes latentes et la classification par maximisation d'un critère d'information.

En 4.1., nous fixons les notations. En 4.2., nous présentons le modèle des classes latentes en nous focalisant sur l'approche classification de ce modèle. En 4.3., nous établissons l'équivalence entre cette approche et la classification par maximisation de l'information mutuelle. En 4.4. nous en déduisons les liens avec la classification par maximisation du critère du χ^2 et commentons ces résultats.

4.1. Notations

Soit N individus décrits par d variables qualitatives. Soit m_j le nombre de modalités d'une variable j ; soit m le nombre total de modalités ($m = \sum_{j=1}^d m_j$).

Soit X le tableau des données mises sous forme disjonctive complète, à N lignes et m colonnes.

On note $x_i^{j\ell}$ le terme générique du tableau X . $x_i^{j\ell}$ vaut 1 si l'individu i présente la modalité ℓ de la variable j et 0 sinon. On pose :

$$x^{j\ell} = \sum_{i=1}^N x_i^{j\ell}; \quad x_i = \sum_{j=1}^d \sum_{\ell=1}^{m_j} x_i^{j\ell} = d \quad \text{et} \quad s = \sum_{i=1}^N \sum_{j=1}^d \sum_{\ell=1}^{m_j} x_i^{j\ell} = Nd.$$

De plus, si P_k est une classe d'une partition $P = (P_1, \dots, P_K)$ des individus, on note

$$x_k^{j\ell} = \sum_{i \in P_k} x_i^{j\ell} \quad \text{et} \quad x_k = \sum_{j=1}^d \sum_{\ell=1}^{m_j} x_k^{j\ell} = dn_k \quad \text{où} \quad n_k = \text{card } P_k.$$

4.2. Le modèle des classes latentes

L'hypothèse du modèle des classes latentes (cf. Goodman (1974), Everitt (1984)) est la suivante : il existe une variable qualitative ("cachée") à K modalités, telle que conditionnellement à la connaissance de l'une de ces modalités les d variables soient mutuellement indépendantes.

Les paramètres de ce modèle sont les fréquences relatives ($p_k, k = 1, K$) des K modalités de la variable cachée ou latente et les probabilités ($a_k^{j\ell}, \ell = 1, m_j; j = 1, d; k = 1, K$) que $x_i^{j\ell} = 1$ sachant que l'individu i présente la modalité k de la variable latente. On a bien sûr $\sum_{\ell=1}^{m_j} a_k^{j\ell} = 1$ pour tout $j = 1, d$ et pour tout $k = 1, K$.

Ce modèle revient à supposer (cf. Everitt (1984)) que les N vecteurs binaires à m coordonnées décrivant les individus sont un échantillon du mélange de densités :

$$f(x) = \sum_{k=1}^K p_k f(x, a_k)$$

avec

$$f(x, a_k) = \prod_{j=1}^d \prod_{\ell=1}^{m_j} (a_k^{j\ell})^{x_i^{j\ell}}$$

où

$$a_k = (a_k^{j\ell}, \ell = 1, m_j; j = 1, d)$$

avec

$$\sum_{\ell=1}^{m_j} a_k^{j\ell} = 1.$$

$f(x, a_k)$ est la densité d'une loi multinomiale multivariée de paramètre a_k .

Ainsi l'approche naturelle pour ajuster le modèle consiste à estimer les paramètres du mélange. A ce propos, notons que l'intitulé du modèle est trompeur : les paramètres du modèle des classes latentes ne sont pas des classes. Les classes que l'on peut déduire de l'estimation des paramètres du mélange en sont un sous-produit.

Par contre, pour l'approche classification à laquelle nous nous intéressons ici, ce sont les estimés de ces paramètres qui sont un sous-produit de la décomposition en classes de l'échantillon.

Le critère de vraisemblance classifiante s'écrit :

$$C(a, P) = \sum_{k=1}^K L_k \quad \text{avec} \quad L_k = \sum_{x_i \in P_k} \ln f(x_i, a_k),$$

$$\begin{aligned} C(a, P) &= \sum_{k=1}^K \sum_{x_i \in P_k} \sum_{j=1}^d \sum_{\ell=1}^{m_j} x_i^{j\ell} \ln(a_k^{j\ell}) \\ &= \sum_{k=1}^K \sum_{j=1}^d \sum_{\ell=1}^{m_j} x_k^{j\ell} \ln(a_k^{j\ell}). \end{aligned}$$

A partition P fixée, la recherche des a_k maximisant L_k se ramène, du fait que les d lois multinomiales sont mutuellement indépendantes, à la recherche des $a_k^j = (a_k^{j_1}, \dots, a_k^{j_{m_j}})$ qui maximise $L_{kj} = \sum_{\ell=1}^{m_j} x_k^{j\ell} \ln(a_k^{j\ell})$ sous la contrainte $\sum_{\ell=1}^{m_j} a_k^{j\ell} = 1$.

Le lagrangien de ce problème s'écrit :

$$\begin{aligned} \text{Lag} &= L_{kj} - \lambda \left(\sum_{\ell=1}^{m_j} a_k^{j\ell} - 1 \right) \\ \frac{\partial \text{Lag}}{\partial a_k^{j\ell}} &= \frac{x_k^{j\ell}}{a_k^{j\ell}} - \lambda \\ \frac{\partial \text{Lag}}{\partial a_k^{j\ell}} = 0 &\Leftrightarrow \lambda = \frac{x_k^{j\ell}}{a_k^{j\ell}} \quad \text{pour tout } \ell = 1, m_j. \end{aligned}$$

De $\sum_{\ell=1}^{m_j} a_k^{j\ell} = 1$ on tire $\lambda = n_k \Rightarrow a_k^{j\ell} = \frac{x_k^{j\ell}}{n_k}$.

Autrement dit a_k est le centre de gravité de P_k .

Il s'ensuit que le critère de vraisemblance classifiante s'écrit :

$$C(a, P) = \sum_{k=1}^K \sum_{j=1}^d \sum_{\ell=1}^{m_j} x_k^{j\ell} \ln(x_k^{j\ell}) - d \sum_{k=1}^K n_k \ln(n_k).$$

4.3. Classification par maximisation de l'information.

L'information mutuelle du tableau X s'écrit par définition (cf. Benzécri (1973)) :

$$H(X) = \sum_{i=1}^N \sum_{j=1}^d \sum_{\ell=1}^{m_j} \frac{x_i^{j\ell}}{s} \log_2 \left(\frac{x_i^{j\ell} s}{x_i x_i^{j\ell}} \right).$$

Elle se simplifie en (du fait que $x_i^{j\ell} = 0$ ou 1)

$$H(X) = \log_2 N - \frac{1}{s} \sum_{j=1}^d \sum_{\ell=1}^{m_j} x^{j\ell} \log_2 x^{j\ell}.$$

Soit $P = (P_1, \dots, P_K)$ une partition des individus. Considérer P comme un résumé de l'information apportée par les données revient à substituer au tableau

$$X = (x_i^{j\ell}, i = 1, N; j = 1, d; \ell = 1, m_j)$$

le tableau

$$(x_k^{j\ell}, k = 1, K; j = 1, d; \ell = 1, m_j).$$

L'information conservée par P s'écrit donc :

$$H(P) = \sum_{k=1}^K \sum_{j=1}^d \sum_{\ell=1}^{m_j} \frac{x_k^{j\ell}}{s} \log_2 \left(\frac{x_k^{j\ell} s}{x_k x_k^{j\ell}} \right).$$

Dans ce cadre, il est naturel (cf. Benzécri (1973)) de rechercher la partition qui maximise l'information $H(P)$.

Par des transformations géométriques simples, on obtient que $H(P)$ s'écrit :

$$H(P) = \log_2 N + \sum_{k=1}^K \sum_{j=1}^d \sum_{\ell=1}^{m_j} \frac{x_k^{j\ell}}{s} \log_2 x_k^{j\ell} - \sum_{k=1}^K \frac{n_k}{N} \log_2 n_k - \sum_{j=1}^d \sum_{\ell=1}^{m_j} \frac{x_k^{j\ell}}{s} \log_2 x_k^{j\ell}.$$

Il s'ensuit que maximiser $H(P)$ revient à maximiser

$$\sum_{k=1}^K \sum_{j=1}^d \sum_{\ell=1}^{m_j} x_k^{j\ell} \log_2 x_k^{j\ell} - d \sum_{k=1}^K n_k \log_2 n_k.$$

Or, à la constante multiplicative $\frac{1}{\ell n 2}$ près, cette quantité est la vraisemblance classifiante $C(a, P)$ du paragraphe 4.2.

On voit donc que rechercher la partition P en K classes maximisant le critère d'information mutuelle revient à identifier un modèle de K classes latentes sous l'approche "classification".

4.4. Classes latentes et critère du χ^2

Le critère d'information mutuelle est peu utilisé en classification malgré son caractère naturel, on lui préfère le critère du χ^2 qui s'écrit pour une partition $P = (P_1, \dots, P_K)$

$$\chi^2(P) = \sum_{k=1}^K \sum_{j=1}^d \sum_{\ell=1}^{m_j} \frac{(sx_k^{j\ell} - x_k x_k^{j\ell})^2}{x_k x_k^{j\ell} s}$$

Benzécri (1973) en exprime très bien la raison : l'expérimentation montre que l'emploi du critère d'information ou du critère du χ^2 conduit pratiquement aux mêmes classifications et le deuxième critère induit des calculs beaucoup plus simples. Ce comportement quasi identique des deux critères a également été souligné par Govaert (1983). Il a introduit dans son programme de classification simultanée des lignes et des colonnes d'un tableau de nombres positifs par maximisation du χ^2 , le calcul de l'information conservée à chaque itération. Sur les nombreux exemples qu'il a analysés, il a constaté que l'information croissait systématiquement.

D'après les résultats du paragraphe 4.3, on peut donc affirmer que, pratiquement, rechercher la partition maximisant la vraisemblance classifiante associée au modèle des classes latentes revient à rechercher la partition maximisant le critère du χ^2 .

Cette analogie appelle plusieurs commentaires :

1) Le modèle des classes latentes exige l'estimation d'un grand nombre de paramètres ($K \sum_{j=1}^d (m_j - 1) + K - 1$). Pour cette raison, il est surtout utilisé pour des tableaux comportant peu de variables ou sur des tableaux de données binaires. Dans ce dernier cas, le modèle sous-jacent des classes latentes pour la classification

revient à travailler sur le tableau binaire dédoublé, pratique courante qui a l'intérêt de rendre les résultats indépendants du codage initial en 0 et 1 (cet intérêt devient un défaut pour l'analyse des données de présence-absence, quand seule la présence apporte de l'information).

2) Chercher une partition de χ^2 maximum revient, implicitement, à rechercher à mettre à jour une structure en classes telle qu'à l'intérieur de chaque classe les variables observées soient indépendantes (hypothèse forte que Benzécri (1976) juge peu acceptable) et suivent une loi multinomiale.

3) Les auteurs qui se sont intéressés au modèle des classes latentes étaient, de fait, dans un cadre qui leur a permis d'analyser la qualité de leurs résultats. Leurs travaux peuvent être mis à profit pour proposer des outils d'évaluation d'une partition obtenue par un algorithme de classification maximisant le critère du χ^2 .

Notons tout d'abord qu'il est raisonnable de demander un nombre de classes $K \leq \left(\prod_{j=1}^d m_j + 1 \right) / (m - d + 1)$. Il s'agit en effet d'une condition nécessaire pour que le modèle des classes latentes soit identifiable : le nombre des valeurs possibles doit être supérieur au nombre de paramètres à estimer.

D'autre part, les outils d'évaluation du modèle proposés par Goodman (1974) s'adaptent sans modification. Ainsi, on peut étudier si les estimés des paramètres déduits de la partition constituent une solution localement optimale (c'est-à-dire une solution unique dans un certain voisinage). Pour cela, il suffit que le Jacobien du modèle soit de plein rang pour cette solution.

Pour choisir entre plusieurs partitions proposées par l'algorithme de classification (à nombres de classes différents par exemple), on peut utiliser la statistique du χ^2 suivante basée sur le rapport de vraisemblance :

$$\chi^2 = 2N \sum_{\ell_1, \dots, \ell_d} f_{\ell_1, \dots, \ell_d} \ln(f_{\ell_1, \dots, \ell_d} / F_{\ell_1, \dots, \ell_d})$$

formule où $f_{\ell_1, \dots, \ell_d}$ est la fréquence observée de l'évènement $x_i^{1\ell_1} = 1, \dots, x_i^{d\ell_d} = 1$ et $F_{\ell_1, \dots, \ell_d}$ représente la fréquence déduite de l'estimation du modèle pour le même évènement.

Sous l'hypothèse que la structure latente estimée est la bonne cette statistique suit une loi du χ^2 à $\prod_{j=1}^d m_j - K \sum_{j=1}^d (m_j - 1)$ degrés de liberté.

Il s'agit d'une mesure de la qualité d'ajustement du modèle induit par la partition à évaluer. Son bon usage est nous semble-t-il le suivant (cf. Goodman (1974)) : on retient la partition de plus petit nombre de classes possible qui fournit une bonne qualité d'ajustement. En effet plus on augmente le nombre de classes plus l'ajustement aura tendance à être bon, mais plus le modèle sera compliqué et difficile à exploiter.

4) Aitkin, Francis et Raynal (1987) avaient mis en évidence à partir de données réelles la similitude des résultats des Nuées Dynamiques sur variables qualitatives avec le critère du χ^2 et le modèle des classes latentes. Notons toutefois que, pour identifier le modèle des classes latentes, ils avaient directement estimé les paramètres

du modèle par l'algorithme EM (Dempster, Laird, Rubin (1977)) au lieu de se placer comme nous l'avons fait ici sous l'approche classification. Nous revenons sur cet aspect dans la section suivante.

5. Limites de l'approche classification pour les mélanges

Dans ce qui précède, nous nous sommes placés dans le cadre des mélanges pour évaluer les caractéristiques de différents critères de classification. Inversement, quelles performances peut-on attendre d'algorithmes de classification pour l'estimation des paramètres d'un mélange ?

Nous avons vu que l'approche classification du problème de mélanges ne s'attaque pas directement à l'estimation des paramètres du mélange.

D'un point de vue théorique, l'approche classification recherche des estimations des paramètres du mélange et des paramètres de classification ($o(i)$, $i = 1, N$). Ainsi le nombre de paramètres à estimer augmente indéfiniment avec la taille de l'échantillon. En conséquence, les estimateurs de ces paramètres ne sont pas convergents. De plus le biais, induit par cette approche, pour les paramètres (a_k , $k = 1, K$) peut être important. Cela est dû au fait que les algorithmes de partitionnement construisent des classes dont les enveloppes convexes sont disjointes et de ce fait estiment les paramètres (a_k , $k = 1, K$) sur la base de sous échantillons tronqués des composants du mélange. De plus, un théorème (Bryant, Williamson (1978)) montre que ce biais subsiste asymptotiquement.

Théorème :

Notons $a = (a_1, \dots, a_k)$ et $k(a, x) = \min\{k, 1 \leq k \leq K / f(x, a_k) \geq f(x, a_j)\}$ pour tout $j = 1, K$.

Sous des hypothèses de régularité et l'hypothèse sensible suivante :

Soit $M(a) = E(\ln f(X, a_{k(a,x)}))$, alors $M(a)$ a un unique maximum en $a = a_0$. Alors la suite des itérés de l'algorithme des Nuées Dynamiques a^n converge p.s. vers a_0 lorsque N tend vers l'infini.

Les auteurs indiquent que, en toute généralité, a_0 est différent des vrais paramètres a^* du mélange.

Il est clair que l'hypothèse d'un unique maximum a_0 n'est certainement pas vérifiée pour de nombreux mélanges. Cependant, dans le même article, des exemples où cette hypothèse s'applique sont données, ce qui permet de bien mettre en évidence le biais de l'approche classification.

Afin de réduire ce biais, des tentatives d'utilisation de l'approche classification avec le critère de vraisemblance classifiante ont été faites en utilisant les probabilités a posteriori ($t_k(x_i)$, $k = 1, K$; $i = 1, N$) d'appartenance des points de l'échantillon aux composants du mélange plutôt que d'affecter unilatéralement ces points à l'un des composants (cf. Cazes (1976)).

En fait l'étape de calcul des $t_k(x_i)$, les a_k étant fixés, conduit à prendre, pour tout $i = 1, N$, $t_k(x_i) = 1$ ou 0. C'est donc à tort que Aitkin, Francis et Raynal (1987) affirment qu'une version des Nuées Dynamiques prenant en compte les probabilités

d'affectation des points aux classes est similaire à l'algorithme EM pour l'estimation des paramètres d'un mélange.

En effet, le critère à maximiser s'écrit alors :

$$W(a, t) = \sum_{k=1}^K \sum_{i=1}^N t_k(x_i) \ln f(x_i, a_k)$$

où

$$t = (t_k(x_i), k = 1, K; i = 1, N).$$

L'étape de calcul des $t_k(x_i)$ consiste à maximiser, pour tout $i = 1, N$, l'expression

$$\sum_{k=1}^K t_k(x_i) \ln f(x_i, a_k)$$

or

$$\sum_{k=1}^K t_k(x_i) \ln f(x_i, a_k) \leq \ln f(x_i, a_m)$$

avec m vérifiant : $\forall k \neq m \quad f(x_i, a_k) \leq f(x_i, a_m)$.

La borne de $\sum_{k=1}^K t_k(x_i) \ln f(x_i, a_k)$ est donc atteinte pour $t_k(x_i) = 1$ si $k = m$, 0 sinon.

Maintenant quelle est la nature de ce biais? Il est très difficile de répondre à cette question en toute généralité. Dans le cas des mélanges gaussiens, qui nous intéressent particulièrement ici, on possède des éléments de réponse assez précis.

– Marriott (1975) a montré que les différences entre moyennes ont tendance à être surestimées, les variances sousestimées et les différences entre proportions sousestimées.

– Nous avons montré (Celeux (1987)) que asymptotiquement, pour un mélange gaussien réel à deux composants de mêmes variances, l'écart positif entre la probabilité de mauvaise classification par l'approche classification et la probabilité de mauvaise classification par l'approche estimation croît lorsque p_1 (ou p_2) s'éloigne de 1/2.

– Bryant et Williamson (1985) ont montré, à partir de simulations, que pour que le biais soit supportable il faut d'une part que les composants du mélange soient bien séparés (en particulier que les moyennes soient bien différentes) et d'autre part que les proportions du mélange soient du même ordre.

En ce qui concerne les mélanges associés au modèle des classes latentes, nous n'avons, à l'heure actuelle, guère d'élément particulier pour l'évaluation du biais. Il est vraisemblable que les conditions énoncées ci-dessus pour que le biais soit supportable restent valables. De plus, par analogie avec le cas gaussien, on pourrait penser que l'approche classification ait tendance à surestimer les différences entre les probabilités multinomiales voire à exagérer les probabilités extrêmes (proches de 0 ou de 1). L'exemple suivant ne confirme pas cette intuition.

Nous avons repris des données traitées par Goodman (1974) par l'algorithme EM. Il s'agit de 216 individus caractérisés par 4 variables binaires. Les fréquences observées sont :

1111 : 42	0111 : 1
1110 : 23	0110 : 4
1101 : 6	0101 : 1
1100 : 25	0100 : 6
1011 : 6	0011 : 2
1010 : 24	0010 : 9
1001 : 7	0001 : 2
1000 : 38	0000 : 20

Goodman propose une structure latente à 2 composants dont les paramètres sont :

p_1	a_1^{11}	a_1^{21}	a_1^{31}	a_1^{41}	p_2	a_2^{11}	a_2^{21}	a_2^{31}	a_2^{41}
0.279	0.993	0.940	0.927	0.769	0.721	0.714	0.330	0.354	0.132

Par les Nuées Dynamiques avec la distance du χ^2 , la partition en deux classes la meilleure induit la structure latente suivante :

p_1	a_1^{11}	a_1^{21}	a_1^{31}	a_1^{41}	p_2	a_2^{11}	a_2^{21}	a_2^{31}	a_2^{41}
0.298	0.938	0.781	0.797	1.000	0.702	0.728	0.381	0.397	0.130

Les résultats sont assez similaires. La différence la plus sensible concerne l'estimation de la probabilité a_1^{41} (0.769 contre 1.000). Effectivement, les Nuées Dynamiques ont exagéré la valeur de cette probabilité; mais ce n'est pas le cas, bien au contraire, pour les probabilités a_1^{21} et a_1^{31} .

Références

- M. AITKIN, B. FRANCIS, et N. RAYNAL (1987) . — Une étude comparative d'analyses des correspondances ou de classifications et des modèles de variables latentes ou de classes latentes, *R.S.A.* Vol. 35 n° 3.
- J.P. BENZÉCRI (1973) . — Théorie de l'information et classification d'après un tableau de contingence. *L'Analyse des données*, tome 1, Dunod.
- J.P. BENZÉCRI (1976) . — Histoire et préhistoire de l'analyse des données, *Cahiers de l'Analyse des données*, Vol. 1, n° 1 à 4.
- P. BRYANT et J.WILLIAMSON (1978) . — Asymptotic behaviour of classification ML estimates. *Biometrika*, 65.
- P. CAZES (1976) . — Décomposition d'un histogramme en composantes gaussiennes. *R.S.A.*, Vol. 24, n° 1.
- G. CELEUX et J. DIEBOLT (1986) . — L'algorithme SEM : un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. *R.S.A.*, Vol. 34, n° 2.

- G. CELEUX (1987) . — Reconnaissance de mélanges de densité de probabilités et applications à la validation des résultats en classification. Thèse d'état. Université Paris 9-Dauphine.
- A. DEMPSTER, N. LAIRD et D. RUBIN (1977) . — Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *JRSS*, B. 39.
- B. EVERITT (1984) . — An introduction to latent variable models. Chapman and Hall.
- B. EVERITT, D. HAND (1981) . — Finite mixture distributions . Chapman and Hall.
- H. FRIEDMAN et J. RUBIN (1967) . — On some invariant criterion for grouping data. *JASA*, 62.
- L. GOODMAN (1974) . — Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika*, 61.
- G. GOVAERT (1975) . — Classification avec distances adaptatives. Thèse 3^e cycle. Université Paris 6.
- G. GOVAERT (1983) . — Classification croisée. Thèse d'état. Université Paris 6.
- F. MARRIOTT (1975) . — Separating mixtures of normal distributions, *Biometrics*, 62.
- A. SCHROEDER (1976) . — Analyse d'un mélange de distribution de probabilité de même type. *R.S.A.*, Vol. 24 n° 1.
- A. SCOTT et M. SYMONS (1971) . — Clustering methods based on likelihood ratio criteria. *Biometrics* 27.