RAIRO. RECHERCHE OPÉRATIONNELLE

SAÏD LABRECHE
HERVÉ POULARD

Caractéristiques des partitions optimales

RAIRO. Recherche opérationnelle, tome 32, n° 4 (1998), p. 373-387

<http://www.numdam.org/item?id=RO_1998__32_4_373_0>

Recherche opérationnelle/Operations Research
(vol. 32, n° 4, 1998, p. 373-387)

CARACTÉRISTIQUES DES PARTITIONS OPTIMALES (*)

par Saïd LABRECHE et Hervé POULARD

Communiqué par Bernard VAN CUTSEM

Résumé. – *Dans la première partie de cet article, une présentation factorielle du problème de classification par le critère d'inertie inter-classes sera faite pour tout choix de métrique dans l'espace des individus. On en déduit que cette inertie possède une borne supérieure dépendant du nombre de classes et des résultats d'une Analyse en Composantes Principales. Dans la deuxième partie, on montre un ordre strict jusqu'à un certain rang sur tout ensemble de partitions optimales. Un coefficient contrôlant la qualité de toute approximation d'une partition optimale sera proposé ainsi qu'une procédure heuristique pour résoudre le problème, ouvert, du choix du nombre de classes.* © Elsevier, Paris

Mots clés : Analyse en composantes principales, classification, analyse factorielle.

Abstract. – *The first part of this work gives a factorial presentation of the clustering problem for any choice of the metric in the space of individuals. We deduce that the variance between classes has an upper bound, which depends on the number of classes and the outcome of a Principal Component Analysis. In the second part, it is shown that this variance induces an order (strict up to some rank, then equality holds) over every set of optimal partitions. A coefficient that controls the quality of approximations to exact solutions is presented, together with a heuristic procedure that solves the problem of choosing the number of classes.* © Elsevier, Paris

Keywords: Principal component analysis, clustering, factor analysis.

1. INTRODUCTION

Bien des méthodes de classification automatique non hiérarchique consistent en la recherche d'une partition, avec un nombre de classes fixé, qui optimise un critère mathématique bien défini [5, 6, 11, 12, 15]. Le critère le plus utilisé est le critère d'inertie inter-classes [3, 7] également appelé critère de variance inter-classes pour sa signification en statistique [9, 13]. Il sera question de ce critère dans cet article où deux objectifs sont poursuivis.

La présentation factorielle du problème de classification a tenté plusieurs auteurs [1, 13]. Une extension de la présentation due à Lerman, dans le cas

(*) Reçu en janvier 1996.

Laboratoire d'Analyse et d'Architecture des Systèmes, 7, avenue Colonel Roche, 31077 Toulouse Cedex, France.

Recherche opérationnelle/Operations Research, 0399-0559/98/04/

où la matrice associée à la métrique choisie dans l'espace des individus est diagonale, sera faite pour tout autre choix. De cette présentation, on déduit que la variance inter-classes d'une partition possède une borne supérieure. Cette borne dépend du nombre de classes et des résultats d'une Analyse en Composantes Principales (A.C.P.). Ce résultat nous permet de proposer un coefficient pour contrôler la qualité de l'approximation d'une partition optimale par toute partition de même nombre de classes.

Nous montrerons alors que le critère d'inertie inter-classes induit un ordre strict, jusqu'à un certain rang, sur tout ensemble de partitions optimales.

A partir de ces deux constats et des résultats significatifs obtenus sur plusieurs jeux de données (concrètes ou simulées), nous proposons une méthode permettant de résoudre partiellement le problème, qui reste ouvert, du choix du nombre de classes adapté aux données analysées.

2. PRÉSENTATION FACTORIELLE DE LA CLASSIFICATION

2.1. Notations et rappels

Les données sont sous la forme d'une matrice Individus×Variables X de dimensions (n, p) et de rang s. Les variables sont supposées centrées. On note $N_x = \{x_i\}_{i=1}^n$ le nuage des points, défini par les vecteurs lignes de X, dans l'espace E des individus. Soit M la matrice de la métrique choisie dans E et D la matrice diagonale des poids des individus définissant la métrique dans l'espace F des variables. Ces dernières sont supposées centrées.

Si on note μ_i ($\mu^i = D(i, i)$) le poids du i-ème individu, ${}^t x$ la transposée d'un vecteur ou d'une matrice x et V_x la matrice des variances covariances, l'inertie totale du nuage N_x est alors définie par

$$I[N_x] = \sum_{i=1}^{n} \mu_i \|x_i\|_M^2 = \sum_{i=1}^{n} \mu_i \, {}^t x_i \, M \, x_i$$
$$= \text{trace}\,[{}^t X D X M] = \text{trace}\,[V_x M] \qquad (1)$$

On rappelle que les vecteurs principaux de l'A.C.P. du triplet $(N_x.M, D)$ sont les vecteurs propres M-orthonormés de la matrice $V_x M$.

L'inertie inter-classes d'une partition P en q classes de l'ensemble I des individus est définie par

$$B(P) = \sum_{k=1}^{q} B[P(k)] = \sum_{k=1}^{q} \mu(k) \|g_k\|_M^2 = \sum_{k=1}^{q} \mu(k) \, {}^t g_k \, M \, g_k \qquad (2)$$

g_k [resp. $\mu(k)$] est le centre de gravité [resp. poids] de la classe $P(k)$ et $B[P(k)] = \mu(k)\,\|g_k\|^2_M$ est la contribution de la classe $P(k)$ à $B(P)$.

On note Y, la matrice de dimensions (n, q) et de rang q, des indicatrices $\{Y^k\}_{k=1}^q$ des classes, on a alors

$$\begin{aligned} B(P) &= \text{Trace}\,[({}^tYDY)^{-1}\,{}^tYDXM\,{}^tXDY] \\ &= \text{Trace}\,[({}^tYDY)^{-1}\,V_{yx}\,MV_{xy}]. \end{aligned} \tag{3}$$

La matrice diagonale ${}^tYDY$ est celle des poids des classes et la matrice $V_{yx}\,(= {}^tV_{xy})$ est celle des covariances des variables $\{X^j\}_{j=1}^p$ et des indicatrices $\{Y^k\}_{k=1}^q$. La k-ième colonne du produit $V_{xy}\,({}^tYDY)^{-1}$ est le centre de gravité de la classe $P(k)$.

On cherche, dans l'ensemble $\mathcal{P}_q$ des partitions de I en q classes, une partition P_q qui maximise $B(P)$.

$$B(P_q) = \max_{P\in\mathcal{P}_q} B(P) \tag{4}$$

La partition P_q sera dite **optimale**. On notera que ce problème peut avoir plusieurs solutions distinctes.

2.2. Définition d'un nuage de points dans F

La matrice M étant symétrique et définie positive, il existe une matrice carrée M_1 de rang p telle que $M = M_1\,{}^tM_1$. La matrice M_1 peut être obtenue soit par une décomposition spectrale de M soit par la décomposition de Doolitte-Cholesky [2].

Soient Z le produit XM_1 et $N_z = \{Z^j\}_{j=1}^p$ le nuage des points dans F défini par les vecteurs colonnes de Z. On pondère chaque point de N_z par $\frac{1}{p}$ et on note D_p la matrice de ces poids. La matrice M_1 étant de plein rang, on déduit que le nuage N_z est dans le sous-espace F_x de F engendré par les vecteurs colonnes de X.

De l'égalité (1), on déduit que les inerties des nuages N_z et N_x sont proportionnelles. En effet

$$\begin{aligned} I[N_z] &= \sum_{j=1}^p \frac{1}{p}\,\|Z^j\|^2_D = \text{trace}\,[{}^tZD_p\,ZD] = \text{trace}\,[D_p\,ZD\,{}^tZ] \\ &= \frac{1}{p}\,\text{trace}\,[{}^tM_1\,{}^tXDXM_1] = \frac{1}{p}\,\text{trace}\,[{}^tXDXM_1\,{}^tM_1] \\ &= \frac{1}{p}\,\text{trace}\,[V_x\,M] = \frac{1}{p}\,I[N_x]. \end{aligned}$$

On note R le produit de matrices $ZDp\,{}^tZD$. Cette matrice est de rang s ($s = \text{rang}\,(X)$), de plus sa j-ième valeur propre non nulle λ_z^j est égale, au coefficient $\frac{1}{p}$ près, au j-ième moment principal λ_x^j de l'A.C.P. du triplet (N_x, M, D). La variable principale associée à λ_x^j est un vecteur propre de R associé à λ_z^j.

2.3. Projection de N_z sur F_y

L'objectif ici est de montrer que $B\,(P)$ est proportionnelle à l'inertie du nuage N_z projeté sur le sous-espace F_{yc} de dimension $q-1$ engendré par les indicatrices $\{Y_c^k\}_{k=1}^q$ (centrées) des classes $\{P\,(k)\}_{k=1}^q$.

Le sous-espace F_y, engendré par les indicatrices $\{Y^k\}_{k=1}^q$, se décompose comme suit :

$$F_y = \Delta_{\mathbf{1}_n} \oplus F_{yc}$$

L'axe $\Delta_{\mathbf{1}_n}$ des constantes (engendré par le vecteur $\mathbf{1}_n$ de composantes toutes égales à 1) est D-orthogonal à la fois à F_{yc} et à F_x. Le projecteur OP_y sur F_y est la somme des projecteurs OP_{yc} et $OP_{\mathbf{1}_n}$ respectivement sur F_{yc} et sur $\Delta_{\mathbf{1}_n}$. Comme N_z est dans F_x, on déduit que l'image $\hat{N}_z$ de N_z par OP_y est dans F_{yc}.

En considérant les expressions de OP_y et de OP_{yc} (*voir* par exemple [2, p. 316]) et en tenant compte des propriétés de la trace des matrices, l'inertie du nuage $\hat{N}_z$ s'écrit :

$$\begin{aligned} I\,[\hat{N}_z] &= \frac{1}{p}\sum_{j=1}^{p} \|OP_y\,(Z^j)\|_D^2 = \frac{1}{p}\sum_{j=1}^{p} \|Y\,({}^tYDY)^{-1}\,{}^tYDZ^j\|_D^2 \\ &= \frac{1}{p}\,\text{Trace}\,[Y\,({}^tYDY)^{-1}\,{}^tYDXM\,{}^tXD] \\ &= \frac{1}{p}\,\text{Trace}\,[({}^tYDY)^{-1}\,{}^tY\,DXM\,{}^tXDY] \\ &= \frac{1}{p}\,B\,(P) \quad \big(\text{d'après}\,(2)\big). \end{aligned}$$

Comme $\hat{N}_z$ est dans F_{yc}, on a

$$I[\hat{N}_z] = \frac{1}{p}\sum_{j=1}^{p} \|OP_{yc}(Z^j)\|_D^2 = \frac{1}{p}\sum_{j=1}^{p} \|Y_c({}^tY_cDY_c)^+ {}^tY_cDZ^j\|_D^2$$
$$= \frac{1}{p}\,\mathrm{Trace}\,[Y_c({}^tY_cDY_c)^+ {}^tY_cDXM\,{}^tXD]$$
$$= \frac{1}{p}\,\mathrm{Trace}\,[V_y^+ V_{yx} M V_{xy}] = \frac{1}{p}B(P).$$

La matrice V_y^+ est une inverse généralisée [18] de la matrice V_y des variances-covariances des indicatrices $\{Y_c^k\}_{k=1}^q$.

2.4. Autre expression de $I[\hat{N}_z]$

Soit $\{U_k\}_{k=1}^{q-1}$ une base D-orthonormée de F_{yc}. On a une nouvelle expression de $I[\hat{N}_z]$.

$$I[\hat{N}_z] = \frac{1}{p}\sum_{j=1}^{p} \left\| \sum_{j=1}^{p} U_k\,{}^tU_k\,DZ^j \right\|_D^2 = \sum_{k=1}^{q-1} {}^tU_k\,DZD_p\,{}^tZDU_k$$
$$= \sum_{k=1}^{q-1} {}^tU_k\,DRU_k = \sum_{k=1}^{q-1} D(RU_k, U_k) = \frac{1}{p}B(P). \qquad (5)$$

Chercher une partition P_q qui maximise $B(P)$, revient donc à chercher $q-1$ vecteurs de F, D-orthonormés, tels que l'inertie du nuage N_z projeté sur le sous-espace qu'ils engendrent soit maximale. Ces vecteurs doivent engendrer l'un des sous-espaces de F associés à l'ensemble des indicatrices (centrées) des classes des éléments de $\mathcal{P}_q$.

La classification peut donc être présentée comme une A.C.P. du triplet (N_z, D, D_p) avec des contraintes sur les vecteurs axiaux principaux.

Remarques :

• Lorsque M est une matrice diagonale (resp. identité), la présentation précédente est équivalente à la présentation de Lerman [13] (resp. Mirkin [17])

• L'introduction de la métrique relationnelle (Schektman [20]) a permis de présenter la classification comme une A.C.P. sous contraintes dans l'espace des individus [12].

Pour plus de simplicité, si on note, A le produit DR, dans l'égalité (5), on montre la propriété suivante :

PROPRIÉTÉ 1: *L'inertie inter-classes de toute partition P en q classes est telle que :*

$$B(P) \leq \sum_{j=1}^{s_q} \lambda_x^j.$$

avec $s_q = \min(q-1, s)$ *et s étant le rang de X.*

Preuve : La matrice A est symétrique et positive. Les valeurs propres $\{\lambda_z^j\}_{j=1}^{s}$ sont des racines non nulles de l'équation

$$\det(A - \lambda D) = \det(D) \cdot \det(R - \lambda I_n)$$

où det, I_n et λ désignent respectivement le déterminant, la matrice identité de rang n et un réel.

Comme les vecteurs $\{U_k\}_{k=1}^{q-1}$ sont deux à deux D-orthonormés, d'après l'inégalité (2.10) donnée par Rao [19, p. 331], on déduit que :

$$B(P) = p\, I[\hat{N} z] = p \sum_{k=1}^{q-1} {}^t U_k A U_k \leq p \sum_{j=1}^{s_q} \lambda_z^j = \sum_{j=1}^{s_q} \lambda_x^j$$

La propriété est ainsi démontrée.

Une propriété semblable est démontrée par Marchotorchino dans le cadre de l'Analyse Factorielle Relationnelle [14, 15]. Sa démonstration est basée sur le théorème Hoffman-Wielandt [14, p. 87-88].

2.5. Proposition : Évaluer la qualité de l'approximation

Le problème de recherche des partitions optimales est un problème combinatoire car le cardinal de tout ensemble $\mathcal{P}_q$ est fini [4, p. 81]. Dans la pratique, ces partitions difficiles à obtenir, sont approchées par des algorithmes. Une littérature abondante est consacrée aux méthodes d'approximation [3, 4, 5, 6].

Dans les différents articles et ouvrages traitant de la classification que nous avons consultés, nous n'avons pas trouvé de critère pour contrôler la qualité de l'approximation d'une partition optimale par une partition ayant le même nombre de classes. Notre objectif est donc de proposer un coefficient à cet effet.

De la propriété 1, on déduit que pour toute partition P en q classes, le rapport $B(P)/B(P_q)$ est minoré par le rapport $B(P)/\sum_{j=1}^{s_q} \lambda_x^j$. Nous proposons donc d'évaluer la qualité de l'approximation de P_q par P à l'aide du coefficient EQA (Evaluation de la Qualité de l'Approximation) défini par :

$$EQA(P) = 1 - B(P)/\sum_{j=1}^{s_q} \lambda_x^j$$

On dira alors que la partition P est une **bonne approximation** d'une partition optimale P_q si $EQA(P)$ est proche de 0.

Remarques :

• Si le nombre de classes est supérieur ou égal à $s+1$, on a alors $\sum_{j=1}^{s_q} \lambda_x^j = I[N_x]$. Le coefficient EQA est équivalent au coefficient classique $Q(P) = B(P)/I[N_x]$ qui « mesure » la qualité de la partition P [3, p. 155]. Dans ce cas, on a $Q(P) = 1 - EQA(P)$.

• Lorsque le nombre de classes est inférieur à s, les deux coefficients sont différents. S'ils sont faibles pour une partition P, alors P est une bonne approximation de P_q mais P_q résume mal les données (grande perte d'inertie). Ce cas de figure n'est pas rare, pour l'exemple qui sera présenté plus loin, en choisissant la métrique de Mahalabobis ($M = V_x^{-1}$), la partition P en 2 classes obtenue par l'algorithme de réallocation est telle que $EQA(P) = .007$ et $Q(P) = 0.093$.

3. ORDRE SUR DES PARTITIONS OPTIMALES

La finalité de ce paragraphe est de montrer que l'inertie inter-classes induit un ordre, strict jusqu'à un certain rang, sur tout ensemble $\mathcal{PO} = \{P_q\}_{q=1}^{n}$ de partitions optimales.

Pour toute partition P en q classes, on note P^i une partition en $q+1$ classes obtenue en isolant l'élément i de sa classe $P(k)$ ($P(k) \neq \{i\}$). La j-ième classe de P^i est définie par :

$$P^i(j) = \begin{cases} P(j) & \text{si} \quad j \neq k,\ j \neq q+1 \\ P(j) - \{i\} & \text{si} \quad j = k \\ \{i\} & \text{si} \quad j = q+1 \end{cases}$$

D'après le théorème de Huyghens (*voir* par exemple [7, p. 59]), on a :

$$B(P^i) = B(P) + \mu_i \|X_i - g_k\|_M^2 + \mu^i(k) \| G_k^i - g_k \|_M^2, \qquad (6)$$

G_k^i [resp. $\mu^i(k)$] est le centre de gravité [resp. poids] de la classe $P^i(k)$. Les partitions P et P^i sont telles que $B(P^i) \geq B(P)$. Cette inégalité sera à la base des démonstrations des propriétés et du théorème de ce paragraphe.

PROPRIÉTÉ 2 : *Pour tout ensemble PO, la suite* $\{B(P_q)\}$ *est telle que* :

$$0 = B(P_1) \leq B(P_2) \leq \ldots \leq B(P_q) \leq \ldots \leq B(P_n) = I[N_x]$$

Preuve : Les inégalités $B(P_1) = 0$ et $B(P_n) = I[N_x]$ sont évidentes.

Pour toute partition P_q, $1 < q < n$, il existe au moins un élément $\{i\}$ dont la classe ne se réduit pas à cet élément. La partition P_q^i est donc telle que

$$B(P_q) \leq B(P_q^i) \leq B(P_{q+1})$$

La propriété est ainsi démontrée.

THÉORÈME 1 : *Une partition P de I en n_1 classes ($n_1 \leq n$) est telle que $B(P) = I[N_x]$ si et seulement si tous les points de chaque classe de P sont confondus avec le centre de gravité de leur classe. De plus $n_1 > s$.*

Démonstration :

• La condition nécessaire étant évidente, on montre par l'absurde la condition suffisante. L'inertie $B(P)$ étant égale à $I[N_x]$, s'il existe au moins un point X_i non confondu avec le centre de gravité de sa classe, alors P^i est telle que :

$$I[N_x] = B(P) < B(P^i) \leq I[N_x].$$

C'est-à-dire $I[N_x] < I[N_x]$, ce qui est impossible.

• L'entier n_1 est strictement supérieur à s. En effet, si tel n'est pas le cas ($n_1 \leq s$), alors $\dim F_x \leq s - 1$, car les variables sont supposées centrées.

Ce qui fait une contradiction avec $\dim F_x = \operatorname{rang} X = s$.

Du théorème précédent et de la propriété 2, on démontre la propriété suivante :

PROPRIÉTÉ 3 : *Pour tout couple de métriques* (M, D), *il existe un entier* n_1 *tel que* :

$$0 = B(P_1) < B(P_2) < \ldots < B(P_{s+1}) < \ldots < B(P_{n_1}) = \ldots = B(P_n) = I[N_x]$$

Preuve : Soit P une partition vérifiant le théorème précédent, et dont le nombre de classes n_1 est le plus petit possible (s'il n'existe pas de points confondus, alors $n_1 = n$). On déduit $P = P_{n_1}$ et les égalités suivantes :

$$B(P_{n_1}) = \ldots = B(P_n) = I[N_x]$$

Montrons par l'absurde que $\forall q$, si $q < n_1$, on a alors $B(P_q) < B(P_{q+1})$.

Pour tout q, on considère le sous-ensemble $\overline{\mathcal{P}}_{q+1}$ de $\mathcal{P}_{q+1}$, défini par

$$\overline{\mathcal{P}}_{q+1} = \{P_q^i \in \mathcal{P}_{q+1} / i \in P_q(k) \text{ et } P_q(k) \neq \{i\}\}$$

Il existe au moins une partition $P_q^{i_1} \in \overline{\mathcal{P}}_{q+1}$ telle que $B(P_q^{i_1}) > B(P_q)$, car sinon (égalité (6)) :

$$\forall i, \quad \|X_i - g_k\|_M^2 = 0.$$

Tout point X_i serait donc confondu avec g_k et d'après le théorème $B(P_q) = I[N_x]$. On aurait donc nécessairement $q \geq n_1$. Ce qui fait une contradiction avec l'hypothèse $(q < n_1)$.

Des propriétés 1 et 3, on déduit la conséquence suivante :

Conséquence : Les suites $\{h_1(q)\}$ et $\{h_2(q)\}$ définies par :

- $h_1(1) = 0$ et $h_1(q) = \sum_{j=1}^{s_q} \lambda_j^x$ pour $q > 1$.
- $h_2(q) = B(P_q)$

sont croissantes. Pour tout q, on a $h_1(q) \geq h_2(q)$.

Le terme général $h_2(q)$ est l'inertie expliquée par le (s_q)-ième sous-espace principal de l'A.C.P. du triplet (N_x, M, D).

Pour se faire une idée sur l'allure des croissances de ces suites ainsi que sur l'écart entre deux éléments $h_1(q)$ et $h_2(q)$, on les représente sur un même graphique.

Dans le cas où on désire comparer les inerties expliquées par une partition et par un sous-espace principal, les définitions des suites précédentes permettent de choisir la dimension appropriée de ce sous-espace.

EXEMPLE D'ILLUSTRATION

L'un des algorithmes utilisés pour rapprocher les partitions optimales est l'algorithme de réallocation par transfert. Pour se faire une idée sur les valeurs du coefficient EQA prises pour les approximations données par cet algorithme, nous avons considéré plusieurs jeux de données concrètes et simulées. Dans tous les exemples traités, ces valeurs sont en moyenne assez petites ($\leq .2$). Ces résultats prouvent l'efficacité de cet algorithme sur les données utilisées.

Nous présentons ici quelques résultats obtenus sur un exemple. La matrice X de dimensions (50, 10) et de rang 10 est extraite des mémoires de l'école Modulad (Strasbourg 1987). La matrice des variances-covariances V_x est choisie pour M et chaque individu est pondéré par $\frac{1}{50}$. Les approximations $\{\tilde{P}_q\}_{i=1}^{25}$ des 25 premières partitions optimales sont telles que

$$0 = B(\tilde{P}_1) < B(\tilde{P}_2) < \ldots < B(\tilde{P}_{24}) < B(\tilde{P}_{25})$$

A défaut de connaître les valeurs de la suite $\{h_2(q)\}$, on a approché une fonction g par une fonction croissante d'interpolation de l'ensemble des points $\{\{(q, B(\tilde{P}_q))\}_{q=1}^{25}, (50, I[N_x])\}$. Cette fonction d'approximation et une fonction f sont représentées sur la figure 1.

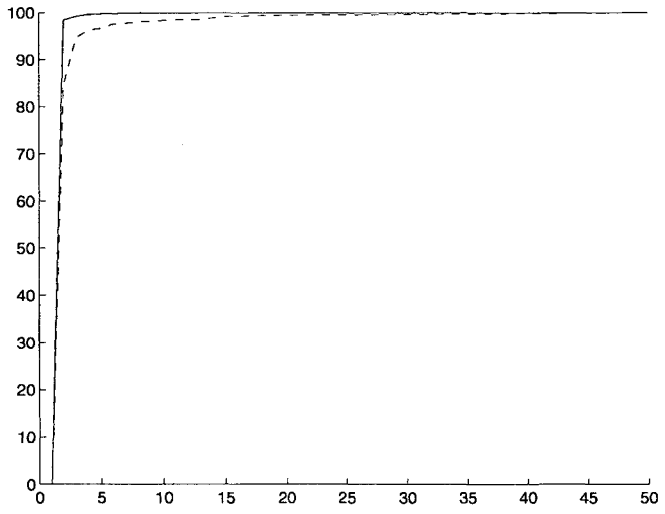

Figure 1. – Fonctions d'interpolation.

Ce graphique montre que les inerties inter-classes des partitions optimales en q classes sont très proches des inerties expliquées par les sous-espaces principaux de dimensions s_q.

Les valeurs du coefficient EQA prises pour les partitions $\{\tilde{P}_q\}_{q=2}^{25}$, sont toutes inférieures à .05 sauf pour la partition en 2 classes où on a $EQA(P_2) = .15$.

4. PROBLÈME DU CHOIX DU NOMBRE DE CLASSES

L'importance et la difficulté du problème du choix du nombre de classes sont notées par tous les auteurs [3, 5, 6, 8, 10, 16] qui se sont intéressés à ce problème. Everitt résume les différentes méthodes envisagées pour tenter de le résoudre.

La définition de la suite $\{h_2(q)\}$ montre que toute partition optimale est meilleure que ses précédentes au sens du critère d'inertie inter-classes. Ce problème n'est donc pas mathématique mais statistique. Sa difficulté est due essentiellement aux deux objectifs antagonistes poursuivis en classification :

La recherche d'une partition dont le nombre de classes est aussi petit *que possible et dont l'inertie inter-classes est aussi* grande *que possible.*

Des définitions des suites $\{h_1(q)\}$ et $\{h_2(q)\}$, on déduit qu'une partition optimale vérifiant le deuxième objectif est telle que :

$$\frac{B(P_q)}{B(P_{(q+1)})} \simeq \frac{h_1(q)}{h_1(q+1)} \simeq 1$$

Le nombre de classes q_1 le mieux adapté aux données traitées serait donc le plus petit possible tel que le rapport $h_1(q_1)/h_1(q_1+1)$ est proche de 1. Il correspond au point à partir duquel la fonction f reste presque constante. Il peut être déterminé en examinant soit le graphique de f soit l'ensemble des rapports $\{h_1(q)/h_1(q+1)\}_{i=1}^{s+1}$.

Le nombre q_1 correspond aussi à la plus petite dimension s_{q_1} des sous-espaces principaux les plus significatifs. Cette proposition heuristique rejoint donc le choix classique du nombre de classes que résume Lerman [13, p. 331], par la phrase suivante :

« *En général q est sensiblement plus grand que le nombre de facteurs interprétables* ».

Pour une étude plus détaillée des données, il serait utile d'examiner les approximations des partitions optimales dont le nombre de classes est inférieur à q_1 [resp. égal à $q_1 + 1$].

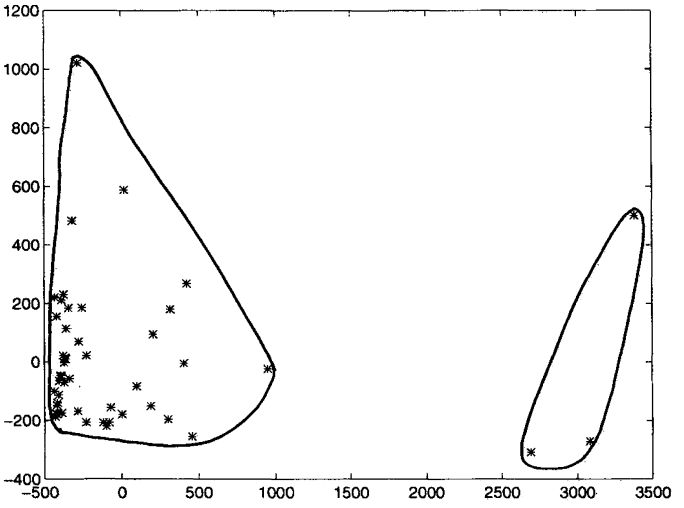

Figure 2. – Partition en deux classes.

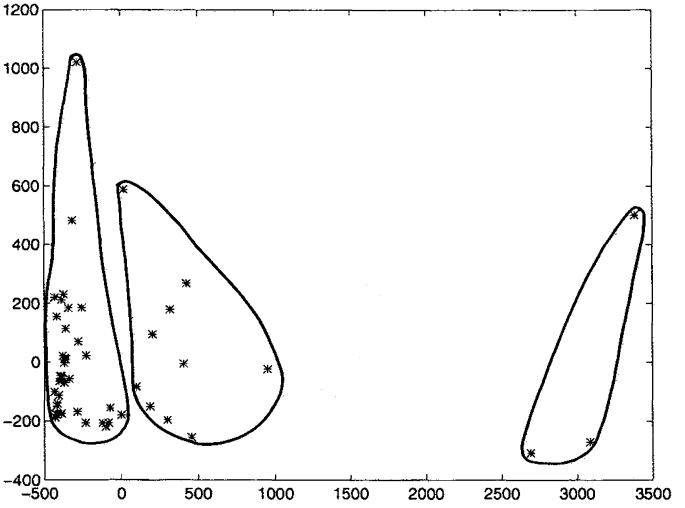

Figure 3. – Partition en trois classes.

Application : Pour l'exemple présenté plus haut, la suite des rapports $h_1(q)/h_1(q+1)$ est

$$\{.9922, .9960, .9987, .9989, .9990, .9995, .9998, .9999, \simeq 1, 1, ..., 1\}.$$

Bien que les éléments de cette suite sont tous proches de 1, on remarque qu'à partir de l'indice 3 les variations sont très faibles. Le nombre de

classes adapté aux données traitées et suggéré par la proposition heuristique pécédente est donc 4. La partition $\tilde{P}_4$ est représentée sur le plan principal (1, 2) (*fig.* 4) qui explique 99.22 % de l'inertie totale du nuage N_x. Les partitions $\tilde{P}_2$, $\tilde{P}_3$, et $\tilde{P}_5$ sont aussi représentées sur ce plan principal (1, 2).

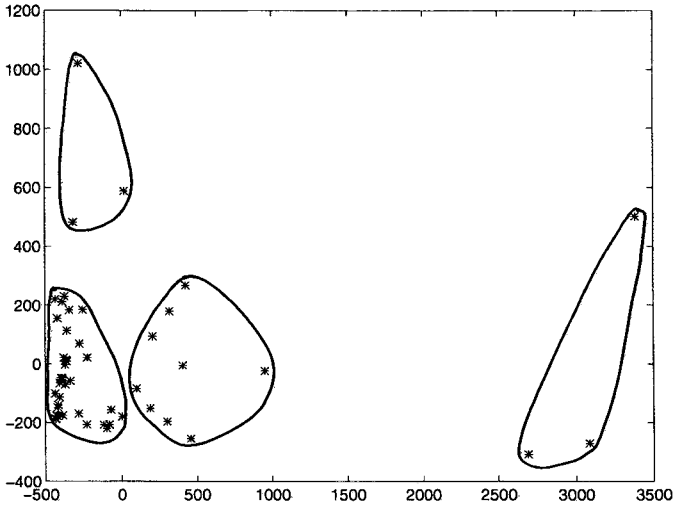

Figure 4. – Partition en quatre classes.

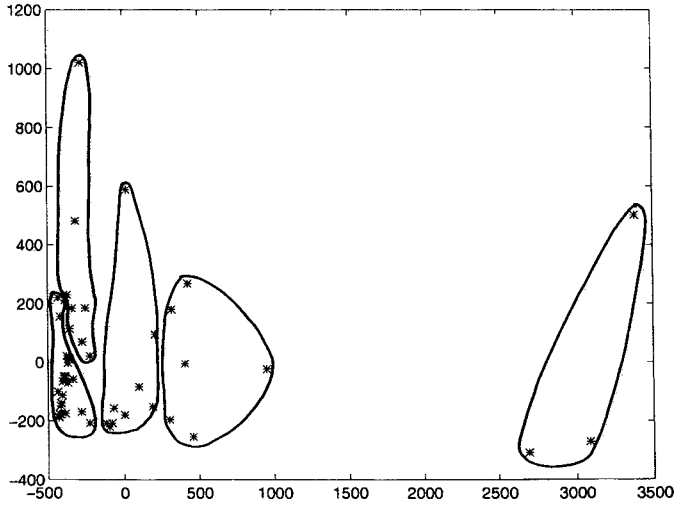

Figure 5. – Partition en cinq classes.

Les pourcentages d'inertie expliquée par ces quatre partitions sont très élevées, ils valent respectivement 84.71, 94.71, 96.33 et 96.66. On remarque que sur le plan principal (1, 2), leurs classes sont bien séparées. Elles permettraient donc une bonne interprétation des données.

Remarque : Le problème du nombre de classes ne se pose pas dans le cadre de l'Analyse Factorielle Relationnelle, car il est montré que la partition triviale, $q = n$, n'est pas nécessairement optimale [14]. L'un des avantages de cette méthode comme la méthode proposée par Mirkin [17] est que le nombre de classes de la partition recherchée est l'un des résultats de l'algorithme.

5. CONCLUSION

Les petites valeurs du coefficient EQA montrent que, sur plus de 300 exemples de données simulées ou concrètes traitées, l'inertie expliquée par un $s_q^{\text{ème}}$ sous-espace principal est proche de l'inertie inter-classes d'une partition optimale P_q. Cette obbservation et les résultats obtenus sont-ils plus généraux, c'est-à-dire indépendants des données et des métriques, ou bien spécifiques aux choix faits jusqu'à présent? La réponse à cette question, qui concerne le lien entre les résultats de l'A.C.P. et ceux de la classification, semble difficile. La difficulté tient au fait que le premier problème possède une solution formelle (analytique) et le deuxième est un problème combinatoire. Pour répondre partiellement à cette question, les recherches présentées dans cet article doivent être approfondies et complétées, en particulier en considérant d'autres jeux de données, d'autres métriques et en choisissant des algorithmes différents, plus ou moins efficaces pour approcher les partitions optimales.

REMERCIEMENTS

Les auteurs remercient vivement le Professeur I. C. Lerman pour ses conseils et ses encouragements, sans lesquels ce travail n'aurait pas été soumis à publication.

RÉFÉRENCES

1. J. P. Benzecri, *Analyse des données,* Dunod, Paris, 1973.
2. F. Cailliez et J. P. Pages, *Introduction à l'analyse des données,* SMASH, Paris, 1976.
3. G. Celeux et co-auteur, *Classification automatique des données,* Dunod-Informatique, Paris, 1989.

4. J. L. CHANDON et S. PINSON, *Analyses typologiques, théories et applications,* Masson, Paris, 1981.
5. M. CORMACK, *A review of classification,* Royal Journal Statistiqual Society, series A, 1971, *134*, n° 3, p. 321-367.
6. E. DIDAY et co-auteurs, *Optimisation en classification automatique,* INRIA, Rocquencourt, 1980.
7. E. DIDAY et co-auteurs, *Éléments d'analyses de données,* Dunod, Paris, 1985.
8. B. S. EVERITT, *Unresolved problems in cluster analysis,* Biometrics, 1979, *35*, p. 169-181.
9. H. P. FRIEDMAN et J. RUBIN, *One some invariant criteria for grouping data,* Journal of the American Statistique Association, 1967, *62*, p. 1159-1178.
10. J. C. GOWER, *Classification Problems,* Bulletin de l'Institut international de statistique, Actes de la 39^{e} session, 1973, Vienne.
11. J. C. GOWER, *Maximal predictive classification,* Biometrics, 1974, *30*, p. 643-654.
12. A. IBRAHIM et Y. SCHEKTMAN, *Principal cluster analysis, Classification as a tool for research,* W. GAUL and M. SCHADER, Elsevier Sc. Publ., North-Holland, p. 217-233, 1986.
13. I. C. LERMAN, *Les présentations factorielles de la classification,* RAIRO, 1979, *13*, n° 2, p. 107-128 et n° 3, p. 227-251.
14. F. MARCHOTORCHINO, *L'analyse factorielle-relationnelle Parties I et II,* Étude MAP-003, Centre Européen de Mathématiques Appliquées, IBM France, Décembre 1991.
15. F. MARCHOTORCHINO et C. B. EDECARRAX, *Le critère de différence de profils,* Congrès International sur l'analyse en Distance DISTANCIA'92, Rennes, France, 1992.
16. F. H. C. MARRIOTT, *Practical problems in a method of cluster analysis,* Biometrics, september 1971, n° 27, p. 501-514.
17. B. G. MIRKIN, *Additive clustering and qualitative factor analysis: Methods for similarity matrix,* Journal of classification, 1987, n° 1, Springer Verlag New-York, p. 3-27.
18. Z. NASHED, *Generalized inverses and applications,* Academic press, London, 1976.
19. C. R. RAO, *The use and interpretation of principal component analysis in applied research,* Sankhya, series A, 1965, *26*, p. 329-358.
20. Y. SCHEKTMAN, *Contribution à la mesure en facteurs dans les sciences expérimentales et à la mise en œuvre automatique dans les calculs statistiques,* Thèse d'État, Toulouse, 1978.