

REVUE DE STATISTIQUE APPLIQUÉE

DENIS RANDON

Classification non hiérarchique et optimisation. Une application : la décomposition d'échantillons non homogènes

Revue de statistique appliquée, tome 22, n° 3 (1974), p. 23-41

http://www.numdam.org/item?id=RSA_1974__22_3_23_0

© Société française de statistique, 1974, 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 NON HIÉRARCHIQUE ET OPTIMISATION. UNE APPLICATION : LA DÉCOMPOSITION D'ÉCHANTILLONS NON HOMOGENES ⁽¹⁾

Denis RANDON

Laboratoire de Statistique, Université Paul Sabatier, Toulouse

I – INTRODUCTION

La résolution d'un problème de classification considéré comme problème de regroupement optimal d'objets en classes peut être abordée de deux façons différentes : représentation hiérarchique de l'ensemble Z des objets à classer, représentation fournissant indirectement des classifications, soit plus directement par des méthodes itératives "améliorant" à chaque étape la classification obtenue à l'étape précédente. Nous considérons seulement ce dernier type de méthodes. Celles-ci supposent que l'on donne un sens précis au mot "améliorer", donc que l'on ait défini une (ou plusieurs) applications de l'ensemble des classifications dans R , critères permettant de juger de la valeur d'une classification. Nous verrons qu'il faut aussi introduire d'autres concepts qu'ignore la classification hiérarchique. Il nous paraît indispensable de faire maintenant quelques remarques d'ordre général.

Remarque 1

Considérons un échantillon qui comprendrait, pour simplifier, uniquement les observations des classes 1 et 2 (fig. 1). Si nous n'avons à notre disposition que les trois objets notés z_1, z_2, z_3 , nous proposerions, en admettant que le nombre de classes soit deux, la classification $C_1 = \{z_1, z_2\}, C_2 = \{z_3\}$. Mais si nous disposons de l'ensemble des observations des classes 1 et 2, nous préférons regrouper z_1 et z_3 dans une même classe, z_2 dans l'autre classe. Une méthode de classification doit donc éviter de considérer séparément chacune des dissimilarités $D(z_i, z_j)$. Elle doit les considérer globalement et utiliser toute l'information disponible.

Remarque 2

Il résulte de ce qui précède qu'il n'existe sans doute pas un algorithme meilleur que les autres dans tous les cas. On ne doit pas traiter de la même façon un échantillon de points de R^N et un tableau de données constitué de signes + où -. Il faut donc s'attacher à étudier des méthodes de construction d'algorithmes et fournir des théorèmes permettant d'établir aisément les conditions de convergence, plutôt que chercher à construire des algorithmes généraux venant s'ajouter à ceux qui existent déjà.

(1) Article remis le 18/6/73. Révisé le 28/1/74

Remarque 3

Toute méthode de classification devra s'efforcer de résoudre les difficultés posées par l'existence de corrélations entre les diverses composantes d'un objet z_n , corrélations pouvant varier beaucoup d'une classe à l'autre. Plus prosaïquement il faudra s'efforcer de déterminer la "forme" d'une classe. La figure 1 permet de se rendre compte qu'il s'agit d'un problème fondamental.

Nous avons proposé en [10] quelques éléments d'une théorie des méthodes itératives de classification, théorie qui tiendrait compte des remarques ci-dessus. Nous nous bornerons, dans les pages qui suivent, à présenter les principaux résultats obtenus en les appliquant à la résolution d'un problème très particulier : celui d'observations de R^P , observations considérées comme réalisations de variables aléatoires.

On trouvera un exposé plus précis, les démonstrations des théorèmes et quelques exemples supplémentaires dans le texte cité.

II CAS PARTICULIER

Décomposition d'un échantillon non homogène.

II.1 – Introduction ; généralités.

Ce problème déjà ancien, mais bénéficiant actuellement d'un regain d'intérêt, peut être considéré comme un problème de classification particulier. Divers heuristiques ont été proposés, par exemple [5], [7], [11]. Tous, à notre connaissance supposant l'égalité des matrices de covariance. C'est là une restriction importante.

Dans le modèle proposé ci-dessous, nous introduisons une matrice de covariance distincte pour chaque classe. Ces matrices servent à définir autant de distances de Mahalanobis et par suite de résoudre le problème des corrélations entre composantes des diverses observations évoqué dans la remarque 3.

II.2 – Notations –

Soient E un ensemble quelconque.

N, K, P des ensembles finis d'indices

$Z = \{z_n \mid z_n \in E, n \in N\}$ l'ensemble des observations
(échantillon) à classer.

$$S_K = \{x \mid x = (x^1, x^2, \dots) \in \{0,1\}^K \text{ vérifiant } \sum_{k \in K} x^k = 1\}$$

$$P_K = \{x \mid x = (x^1, x^2, \dots); x^k \geq 0, \forall k \in K, \text{ et } \sum_{k \in K} x^k = 1\}$$

Nous noterons :

$$S_K^N = (S_K)^N, P_K^N = (P_K)^N$$

$|A|$ désigne le cardinal de l'ensemble A .

${}^t M$ désigne la transposée de la matrice.

Les diverses classes seront notées $C_k, k \in K$.

II.3 – Modèle statistique –

II.3a – Généralités.

Nous faisons l'hypothèse que chaque observation $z_n \in Z$ est une réalisation d'une variable aléatoire parmi $|K|$ v.a. réelles distinctes. Nous supposons, de plus, que chacune des v.a. (notées X_k , $k \in K$) admet une densité p_{θ_k} , $\theta_k \in \Theta$ ensemble donné. θ_k est fixé mais en général inconnu. Notre propos est d'une part de déterminer ces paramètres θ_k et d'autre part de regrouper dans la classe C_k , celles des observations z_n qui sont des réalisations d'une même v.a. X_k .

Pour cela, nous introduisons des variables c_n^k valant 1 si z_n est une réalisation de X_k , valant 0 si tel n'est pas le cas. La densité de probabilité de z_n peut alors s'écrire :

$$p(z_n) = \sum_{k \in K} c_n^k p_{\theta_k}(z_n)$$

Le problème consiste alors à déterminer "au mieux" les paramètres c_n^k, θ_k , $k \in K, n \in N$.

Exprimons d'abord ce problème d'une manière plus formelle en introduisant un modèle statistique. Nous avons choisi un modèle du type maximum de vraisemblance, l'on peut construire, tout aussi bien, un modèle du type moindres carrés.

II.3.b – Modèle statistique et vraisemblance associée.

Soit $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ une famille de lois de probabilité sur un espace probabilisable (E, \mathcal{A}) . Soit $(E, \mathcal{A}, \mathcal{P})$ la structure statistique correspondante. Nous supposons cette structure dominée par une mesure μ , positive, σ – finie sur (E, \mathcal{A}) .

Soit p_θ une détermination de la densité $dP_\theta/d\mu$. Soit \mathcal{L} la fonction de vraisemblance, c'est-à-dire la fonction définie sur $E \times \Theta$ par

$$\mathcal{L}(z; \theta) = p_\theta(z)$$

A partir de la famille \mathcal{P} nous définissons la famille \mathcal{P}_K :

$$\mathcal{P}_K = \{P_{c_o, T} \mid P_{c_o, T} = \sum_{k \in K} c_o^k P_{\theta_k}; c_o \in S_K; T \in \Theta^K\}$$

qui est une famille de lois de probabilité sur (E, \mathcal{A}) , dominée par la mesure μ . La densité de $P_{c_o, T}$ s'écrit

$$p_{c_o, T} = \frac{dP_{c_o, T}}{d\mu} = \sum_{k \in K} c_o^k p_{\theta_k}$$

La fonction de vraisemblance devient, pour la famille \mathcal{P}_K

$$\mathcal{L}_K(z; c_o, T) = \sum_{k \in K} c_o^k p_{\theta_k}(z)$$

La structure statistique associée à la réalisation de l'échantillon Z est :

$$\{E^N, \otimes_{n \in N} \mathcal{A}_n, \{ \otimes_{n \in N} P_{c_n, T}; c_n \in S_k, \forall n \in N; T \in \Theta^K \} \}$$

La fonction de vraisemblance de l'échantillon s'écrit :

$$\begin{aligned} L_K(z_1, \dots, z_N; c_1, \dots, c_N, T) &= L_K(z; c, T) \\ &= L_K(z_1; c_1, T) L_K(z_2; c_2, T) \dots L_K(z_N; c_N, T) \\ &= \prod_{n \in N} \left(\sum_{k \in K} c_n^k p_{\theta_k}(z_n) \right) \end{aligned}$$

Le problème de la détermination des paramètres devient :

Trouver $\hat{c} \in S_K^N, \hat{T} \in \Theta^K$ tels que :

$$\begin{aligned} \prod_{n \in N} \sum_{k \in K} c_n^k p_{\theta_k}(z_n) &\geq \prod_{n \in N} \sum_{k \in K} c_n^k p_{\theta_k}(z_n); \\ \forall c \in S_K^N, \forall T \in \Theta^K. \end{aligned}$$

II. 4 – Le problème d'optimisation ; problème équivalent –

Comme la fonction $x \rightarrow -\text{Log}(x)$ est strictement décroissante pour $x > 0$, le problème précédent est équivalent au problème P1.

$$\text{Posons} \quad W(c, T) = \sum_{n \in N} -\text{Log} \left(\sum_{k \in K} c_n^k p_{\theta_k}(z_n) \right) \quad (1)$$

Problème P1 – (problème discret)

Trouver

$$\hat{c} \in S_K^N, \hat{T} \in \Theta^K \text{ tels que}$$

$$W(\hat{c}, \hat{T}) \leq W(c, T); \forall c \in S_K^N, \forall T \in \Theta^K$$

La condition $c \in S_K^N$ peut s'exprimer par le système de contraintes :

$$\begin{aligned} c_n^k &\in \{0, 1\} \\ &\forall n \in N, \forall k \in K \\ \sum_{k \in K} c_n^k &= 1 \end{aligned} \quad (2)$$

Problème P2 – (problème continu sous-jacent à P1)

Trouver

$$\hat{c} \in P_K^N, \hat{T} \in \Theta^K \text{ tels que}$$

$$W(\hat{c}, \hat{T}) \leq W(c, T); \forall c \in P_K^N, \forall T \in \Theta^K$$

Nous démontrons dans [10] que le problème P_2 admet les mêmes solutions que le problème P_1 et que, pour tout optimum local de P_2 , c_n^k vaut 0 ou 1. Il est donc possible de résoudre indifféremment le problème P_1 ou le problème P_2 et d'utiliser dans ce dernier cas les propriétés de continuité et de différentiabilité de W .

II.5 – Cas particulier :

Décomposition d'un échantillon en composantes gaussiennes –

II.5.a Généralités

Examinons maintenant le cas particulier où $E = \mathbf{R}^p$ muni de la tribu de ses boréliens $\mathcal{B}_{\mathbf{R}^p}$. Les v.a. X_k sont supposées gaussiennes non dégénérées dans \mathbf{R}^p

On pose

$$f_{nk} = f(z_n ; \mu_k, \Gamma_k) = \frac{\det(\Gamma_k^{-1})}{\sqrt{2\pi}^{|\mathbf{P}|}} e^{-\frac{1}{2} \Delta^2(z_n ; \mu_k, \Gamma_k)}$$

avec $\mu_k \in \mathbf{R}^p$

Γ_k matrice de covariance, appartient à l'ensemble \mathcal{G}_p des matrices carrées $|\mathbf{P}| \times |\mathbf{P}|$, symétriques, définies positives et bornées.

$$\theta_k = (\mu_k, \Gamma_k) \in \Theta = \mathbf{R}^p \times \mathcal{G}_p$$

$$T = (\theta_k, k \in K) \in \Theta^K$$

$$\Delta^2(z_n ; \mu_k, \Gamma_k) = {}^t(z_n - \mu_k) \Gamma_k^{-1} (z_n - \mu_k)$$

II.5.b. Algorithme de minimisation dans un espace produit (M.E.P.)

Il s'agit d'un algorithme de résolution du problème à variables discrètes. Ce type d'algorithme se rencontre très fréquemment en théorie de l'optimisation. L'idée de base est extrêmement simple.

Soit à trouver $\inf_{x \in D} f(x)$

$$x = (x_1, \dots, x_q, \dots, x_Q) \in D = D_1 \times D_2 \dots \times D_q \dots \times D_Q$$

L'algorithme procède par itérations successives. A chaque itération l'on construit, à partir d'une valeur x de D , une nouvelle valeur x' telle que $f(x') < f(x)$ en résolvant Q problèmes d'optimisation dans chacun des ensembles D_q .

On cherche d'abord x'_1 tel que

$$f(x'_1, x_2, \dots, x_Q) \leq f(y, x_2, x_3, \dots, x_Q), \forall y \in D_1$$

Puis x'_2 tel que

$$f(x'_1, x'_2, x_3, \dots, x_Q) \leq f(x'_1, y, x_3, \dots, x_Q), \forall y \in D_2$$

et ainsi de suite pour chacune des Q variables.

Cette méthode est utilisée en programmation mathématique (Méthodes de décomposition de programmes), en analyse numérique (Résolution des grands systèmes linéaires par la méthode de Gauss-Seidel), (Méthodes de pas fractionnaires).

Nous pouvons l'utiliser à la résolution du problème P_1 , puisque le domaine défini par les contraintes de P_1 est le produit de $|N| + |K|$ ensembles convexes.

Considérons une variable $c_{n_o} \in S_K$, le minimum c'_{n_o} sur S_K de l'application $c_{n_o} \mapsto W(c, T)$ est obtenu en posant $c'^{k_o}_{n_o} = 1$ pour l'indice k_o tel que $f_{n_o k_o}$ soit maximum (*)

Il faut également déterminer les valeurs $\mu'_k \in \mathbb{R}^p$, $\Gamma'_k \in \mathcal{G}_p$, qui rendent minimum la quantité.

$$W(c, T) = \sum_{n \in C'_k} - \text{Log } f(z_n; \mu_k, \Gamma_k)$$

avec

$$C'_k = \{n \mid n \in N \text{ et } c'^{k}_{n} > 0\}$$

Ces valeurs sont uniques. Ce sont les estimations habituelles(**)

$$\mu'_k = \frac{1}{|C'_k|} \sum_{n \in C'_k} z_n; \Gamma'_k = \frac{1}{|C'_k|} \sum_{n \in C'_k} (z_n - \mu_k)^t (z_n - \mu_k) \quad (3)$$

L'algorithme M.E.P.

– Initialement on se donne $c^{(o)}$ arbitraire dans S_K^N et l'on débute en b) ou bien l'on se donne $T^{(o)}$ arbitraire dans Θ^K et l'on débute en a).

a) Pour chaque valeur de n , calculer les valeurs $f_{n k}$, rechercher k_o tel que

$$f_{n k_o} \geq f_{n k}, \forall k \in K$$

et poser

$$c'^{k_o}_{n} = 1, c'^k_{n} = 0 \text{ pour } k \in K - \{k_o\}$$

b) Calculer les nouvelles valeurs μ'_k et Γ'_k , $k \in K$ à l'aide de (3)

– Répéter a) et b) jusqu'à ce que $c' = c$

Convergence de l'algorithme M.E.P.

L'algorithme converge en un nombre fini d'itérations, bien qu'il y ait un nombre infini de couples (c, T) possibles.

Proposition.

Soit $u^{(i)} = (c^{(i)}, T^{(i)})$ le résultat obtenu à la fin de la i -ème itération

$$\exists i_o \text{ tel que } u^{(i)} = u^{(i_o)}, \forall i \geq i_o \text{ et } W(u^{(i+1)}) < W(u^{(i)}), \forall i < i_o$$

(*) Nous supposons ici que l'indice k_o est unique(***)

(**) Nous admettons qu'aucune des classes C_k n'est vide (***)

(***) Ces hypothèses sont introduites dans le seul but de simplifier la présentation de l'algorithme ; [10] traite le cas général.

Nous dirons que $(c^{(i_0)}, T^{(i_0)})$ est un "point terminal" de l'algorithme ; $u^{(i_0)}$ n'est pas toujours l'optimum souhaité. Toutefois les classifications obtenues par M.E.P. sont généralement satisfaisantes. La solution optimale de P_1 est obtenue après quelques essais effectués avec des classifications initiales distinctes. Il est possible de modifier cet algorithme de nombreuses façons, en obtenant toujours un algorithme convergent.

II. 5.c – Algorithme de résolution des équations de vraisemblance – Conditions nécessaires vérifiées par tout optimum de P_2 .

Kuhn et Tucker ont établi [8] des conditions nécessaires (Elles seraient suffisantes si la fonction W était convexe) pour qu'un point soit un minimum sur un domaine D d'une fonction continument différentiable. Leurs résultats s'adaptent immédiatement au problème P_2 sous la forme :

Théorème – Pour que \hat{c}, \hat{T} soit solution du problème P_2 , il est nécessaire qu'il existe $\hat{v} = \{v_n | n \in N\} \in R^N$ tel que $(\hat{c}, \hat{T}, \hat{v})$ soit un point-selle du Lagrangien :

$$\Phi(c, T, v) = W(c, T) + \sum_{n \in N} v^n \left(\sum_{k \in K} c_n^k - 1 \right)$$

C'est à dire qu'il vérifie, pour tout $n \in N$, pour tout $k \in K$ et pour tout $j \in J$ ensemble d'indices, les relations suivantes :

$$\begin{aligned} \text{a) } \left(\frac{\partial \Phi}{\partial c_n^k} \right)_{(\hat{c}, \hat{T}, \hat{v})} &\geq 0 & \text{b) } \hat{c}_n^k \left(\frac{\partial \Phi}{\partial c_n^k} \right)_{(\hat{c}, \hat{T}, \hat{v})} &= 0 \\ \text{c) } c_n^k &\geq 0 & \text{d) } \left(\frac{\partial \Phi}{\partial T^j} \right)_{(\hat{c}, \hat{T}, \hat{v})} &= 0 & \text{e) } \left(\frac{\partial \Phi}{\partial v^n} \right)_{(\hat{c}, \hat{T}, \hat{v})} &= 0 \end{aligned} \quad (4)$$

Remarque : T^j désigne la j -ème composante de la variable $T \in \Theta^K \subset R^J$

Tous calculs faits, les relations (4) se réduisent dans le cas particulier du problème P_2 à :

$$\begin{aligned} \hat{c}_n^k &= \frac{\hat{c}_n^k \hat{f}_{nk}}{\sum_{k \in K} \hat{c}_n^k \hat{f}_{nk}} & \hat{c}_n^k &\geq 0 & \forall n \in N, \\ & & \sum_{k \in K} \hat{c}_n^k &= 1 & \forall k \in K. \end{aligned} \quad (5)$$

$$\hat{\mu}_k = \frac{1}{c_{\cdot k}} \sum_{n \in N} \hat{c}_n^k z_n$$

$$\hat{\Gamma}_k = \frac{1}{\hat{c}_{\cdot k}} \sum_{n \in N} \hat{c}_n^k (z_n - \hat{\mu}_k)^t (z_n - \hat{\mu}_k)$$

($\hat{c}_{\cdot k}$ désigne la somme $\sum_n \hat{c}_n^k$).

Ce système d'équations sera appelé système des équations de vraisemblance.

Soit le domaine $D = P_K^N \times (R^p \times \mathcal{G}_p)^K$. Soit \hat{D} l'intérieur topologique de D . Les équations (5) forment un système non linéaire écrit sous la forme

$$u = \mathcal{A}(u)$$

où u désigne le couple (c, T) et \mathcal{A} est une application de \hat{D} dans \hat{D} ; (Le cas où u appartient à la frontière de D est sans intérêt).

Nous disposons d'une méthode très simple pour trouver une solution de ce système : au départ, on se donne une valeur $u^{(0)}$; on calcule $u^{(i+1)} = \mathcal{A}(u^{(i)})$. Décrivons cet algorithme de manière plus détaillée.

Algorithme de résolution des équations de vraisemblance – R.E.V.

– Initialement on se donne des valeurs arbitraires $c_n^k \in]0, 1]$ telles que $\sum_{k \in K} c_n^k = 1$, des points μ_k de R^p , des matrices Γ_k symétriques définies positives.

a) on calcule $c_n^k = \frac{c_n^k f_{nk}}{\sum_{k \in K} c_n^{k'} f_{nk}}$

b) on calcule $\mu_k = \frac{1}{c_n^k} \sum_{n \in N} c_n^k z_n$

c) on calcule $\Gamma_k = \frac{1}{c_n^k} \sum_{n \in N} c_n^k (z_n - \mu_k)^t (z_n - \mu_k)$

on répète a), b), c) jusqu'à ce que

$$\|u' - u\| \leq \epsilon \text{ nombre positif fixé.}$$

Proposition – Si $c^{(0)} \in P_K^N$ et si Z engendre R^p , l'algorithme R.E.V. fournit une suite d'itérés $\{u^{(i)} \mid i \in N \text{ ensemble des entiers naturels}\}$ telle que :

a) $W(u^{(i+1)}) < W(u^{(i)})$

b) Tout point d'accumulation de la suite $u^{(i)}$ est solution des équations de vraisemblance –

Notons que la suite admet des points d'accumulation, car infinie dans P_K^N compact.

II. 5. d Méthode de directions admissibles –

Le problème P_2 est un problème de recherche d'optimum posé sous une forme familière, mais il présente des caractéristiques particulières.

- Très grand nombre de variables.
- Critère de forme complexe, impossibilité d'utiliser les dérivées secondes ; temps machine élevé pour chaque calcul de la valeur de W
- Les contraintes ont une forme extrêmement simple.

En utilisant cette dernière particularité, il est possible d'adapter des méthodes classiques de recherche d'optimum à la résolution de P_2 . Nous avons étudié dans un cadre beaucoup plus général une adaptation des méthodes de directions admissibles (Zoutendijk – 1960) et de gradient réduit (Wolfe).

Cette méthode possède de bonnes propriétés théoriques. Les calculs bien qu'assez lourds, s'effectuent sans difficulté majeure. Par contre, elle s'est révélée inférieure en efficacité aux méthodes précédentes.

Aussi nous bornerons nous à présenter ici, d'une manière tout à fait sommaire, l'idée directrice de cette méthode. Voir [3], [10].

La procédure est itérative.

Initialement, l'on se donne $u^{(0)} = (C^{(0)}, T^{(0)}) \in \hat{D}$; (\hat{D} désignant l'intérieur du domaine défini par les contraintes).

Chaque itération s'effectue en deux étapes bien distinctes.

a) Recherches d'une direction h et d'un nombre $\lambda_M > 0$ tels que :

$$\begin{aligned} u - \lambda h \in D \\ W(u - \lambda h) < W(u) \end{aligned} \quad ; \quad \lambda \in [0, \lambda_M]$$

b) Recherche du point "optimal" $\hat{\lambda}$ sur $[0, \lambda_M]$

Le même calcul se répète à partir de la nouvelle valeur

$$u' = u - \hat{\lambda} h$$

La direction h est généralement construite à partir du gradient de W .

On peut prendre $\hat{\lambda}$ tel que $W(u - \hat{\lambda} h) \leq W(u - \lambda h)$; $\forall \lambda \in [0, \lambda_M]$

Nous disposons du résultat suivant :

Proposition

L'algorithme fournit une suite d'itérés $\{u^{(i)} \mid i \in \mathbf{N}\}$ telle que :

a) ou bien cette suite est stationnaire à partir d'un certain rang et sa limite est solution des équations de vraisemblance.

b) ou bien $W(u^{(i+1)}) < W(u^{(i)})$, $\forall i \in \mathbf{N}$ et tout point d'accumulation de la suite $\{u^{(i)} \mid i \in \mathbf{N}\}$ est solution des équations de vraisemblance.

II.6 – Conclusion –

Les algorithmes précédents restent des heuristiques. La classification obtenue n'est pas en général l'optimum cherché. Cette classification finale dépend de la partition choisie initialement. En appliquant l'algorithme à des valeurs de départ distinctes, il est possible de découvrir l'optimum cherché, le plus souvent après un petit nombre d'essais.

L'algorithme M.E.P. est le plus simple et celui qui demande le moins de place mémoire et de temps de calcul.

L'algorithme R.E.V., très simple lui aussi, exige toutefois que l'on conserve en mémoire toutes les quantités c_n^k . Il est bon d'appliquer à la classification obtenue par R.E.V., l'algorithme M.E.P.

Quant au dernier algorithme, bien que l'exposé complet de la méthode soit assez lourd, la logique de l'algorithme obtenue est simple. Cet algorithme nous paraissait intéressant, parce que, contrairement aux précédents, il modifie simultanément toutes les variables et non une variable après l'autre. Mais il demande de conserver beaucoup trop de résultats intermédiaires. Pratiquement il converge lentement et pose des problèmes de précision numérique.

Nous avons appliqué ces algorithmes à des échantillons fictifs engendrés à l'aide de nombres pseudo-aléatoires. La classification "exacte" était donc connue à l'avance. Nous pouvions donc calculer les moyennes et matrices de covariance exactes et empiriques. Dans tous ces essais la classification débutait avec des matrices Γ_k diagonales ayant de très grandes valeurs positives sur la diagonale. Les valeurs initiales de μ_k étaient les seules à varier d'un essai à l'autre. Elles étaient tirées au hasard par le programme. Dans les algorithmes de résolution de P_2 nous avons pris systématiquement $c_n^k = \frac{1}{|K|}$, $\forall n, \forall k$.

Nous pensons que l'on obtient de meilleurs résultats en effectuant, à partir de conditions initiales du type de celles que nous venons de décrire, quelques itérations (5 par exemple) par l'algorithme R.E.V., puis en appliquant l'algorithme M.E.P. à la classification ainsi obtenue. En procédant de cette façon, la plupart des classifications obtenues présentaient une valeur du critère W inférieure à la valeur calculée à partir de la classification exacte, donc étaient totalement satisfaisantes bien que non optimales.

Ces méthodes permettent de traiter sans difficulté les grands tableaux de données.

Par contre, il existe une limitation liée à la nature même du modèle. En effet, le nombre de paramètres à estimer devient vite très grand. D'où la nécessité de définir les objets par un petit nombre de composantes. On peut penser utiliser, préalablement à toute classification, une analyse en composantes principales. Mais ceci peut conduire à des résultats erronés. En effet nous ne savons pas si les axes principaux sont les mêmes dans toutes les classes. Si le nombre de paramètres est trop grand, il vaut mieux utiliser un modèle du type moindres carrés.

Nous n'avons pas abordé l'étude des propriétés statistiques des estimations obtenues. S'il est exclu que les valeurs estimées pour c_n^k convergent, l'on peut par contre supposer que les estimations de μ_k et Γ_k sont asymptotiquement sans biais et convergent presque sûrement.

Le problème du meilleur choix du nombre de classes reste entier. Ce choix ne peut résulter d'une propriété d'optimalité. (il faudrait prendre $|K| = |N|$). Dans certains cas particuliers, on pourra faire un test de l'hypothèse de normalité de chacune des $|K|$ classes. En cas de rejet, il convient d'augmenter le nombre de classes et de recommencer la classification.

III – GENERALISATION A D'AUTRES PROBLEMES DE CLASSIFICATION

Nous avons établi les résultats qui précèdent dans un cadre beaucoup plus général. Nous allons présenter rapidement le problème d'optimisation associé à un problème de classification. Nous indiquerons ensuite quelques unes de ses propriétés.

III.1 Les hypothèses H – Problème.

. Soit Θ un ensemble non vide (espace des paramètres ou de toutes les "formes" possibles).

. Soient des applications D_k données pour tout $k \in K$,

$$D_k : (z, \theta) \in E \times \Theta \mapsto D_k(z, \theta) \in G \subset \mathbb{R}$$

Pratiquement D_k est une mesure de la proximité de l'objet z à la forme θ). G est un intervalle.

. Pour tout n , Ψ_n désigne une application de G dans \mathbb{R} . Nous supposons les fonctions Ψ_n toutes strictement monotones de même sens.

. Nous supposons que le critère mesurant la "valeur" d'une classification peut se mettre sous la forme :

$$W(c, T) = \sum_{n \in N} \Psi_n \left(\sum_{k \in K} c_n^k D_k(z_n, \theta_k) \right)$$

et que W est défini pour $c \in \mathbb{P}_K^N$, $\theta_k \in \Theta$

Nous supposons de plus que l'ensemble

$$W_f = \{ W(c, T) \mid c \in \mathbb{P}_K^N, T \in \Theta^K \}$$

est borné inférieurement et que $\inf W_f \in W_f$

Le problème d'optimisation –

Le problème de la recherche simultanée des $|K|$ "formes" de Θ représentant le mieux l'échantillon Z et du meilleur regroupement d'objets en classes peut s'exprimer :

Problème 1 :

Trouver, sous les hypothèses H, $\hat{c} \in \mathbb{S}_K^N$, $\hat{T} \in \Theta^K$
tels que $W(\hat{c}, \hat{T}) \leq W(c, T)$, $\forall c \in \mathbb{S}_K^N$, $\forall T \in \Theta^K$

III. 2 – Exemples –

. Exemple 1. Recherche de la partition à variance intra-classe minimale. Si $E = \mathbb{R}^p$, si Θ est l'ensemble des moyennes de points de l'échantillon Z ou ce qui revient au même, si $\Theta = \mathbb{R}^p$, trouver la partition rendant minimum la variance intra-classe, revient à résoudre le problème P_1 en prenant pour critère

$$W_1(c, T) = \sum_{k \in K} \sum_{n \in C_k} \|z_n - \bar{z}_k\|^2$$

où

$$C_k = \{ n \mid n \in N \text{ et } c_n^k > 0 \} \quad (6)$$

\bar{z}_k désigne la moyenne empirique de C_k ,
 $\|x\|$ la norme euclidienne usuelle.

. Exemple 2 –

C'est l'exemple étudié ci-dessus

$$W_2(c, T) = \sum_{n \in N} - \text{Log} \left(\sum_{k \in K} c_n^k p_{\theta_k}(z_n) \right)$$

. Exemple 3 –

E ensemble quelconque, $\theta_k = C_k$
d est un indice de dissimilarité sur $E \times E$

$$\text{On obtient } W_3(c, T) = \sum_{\substack{k \in K(n, n') \in C_k \times C_{k'} \\ n < n'}} d(z_n, z_{n'})$$

$$\text{en prenant } D_k(z_n, \theta_k) = \frac{1}{2} \sum_{n' \in C_k} d(z_n, z_{n'})$$

Le critère W_3 peut aussi s'écrire

$$W_3(c, T) = \frac{1}{2} \sum_{k \in K} \sum_{(n, n') \in N^2} c_n^k c_{n'}^k d(z_n, z_{n'})$$

Il existe des algorithmes permettant théoriquement de trouver le maximum absolu dans ce cas particulier important.

. Exemple 4 –

E ensemble quelconque, θ_k partie de Z ayant n_k éléments, n_k entier positif fixé. Soit D une application de $E \times \Theta$ dans R .

On prend le critère

$$W_4(c, T) = \sum_{k \in K} \sum_{n \in C_k} D(z_n, \theta_k)$$

III. 3 – Le problème à variables continues – Recouvrements

Dans le cas où G est un intervalle, le critère W est défini pour $c \in P_K^N$. (6) définit alors un recouvrement de Z ou plus exactement une famille de "sous-ensembles flous" recouvrant Z. Le problème du recouvrement optimal peut s'écrire :

Problème 2.

Sous les hypothèses H,

trouver $\hat{c} \in P_K^N$, $\hat{T} \in \Theta^K$ tels que

$$W(\hat{c}, \hat{T}) \leq W(c, T), \forall c \in P_K^N, \forall T \in \Theta^K$$

En fait l'optimum ne sera pas un recouvrement à proprement parler, mais une partition :

Proposition

Sous les hypothèses H, si D_2 désigne $P_K^N \times \Theta^K$, Θ étant un espace topologique, pour tout minimum local \hat{c}, \hat{T} de W sur D_2 , il existe $\hat{c}^* \in S_K^N$ tel que $W(\hat{c}^*, \hat{T}) = W(\hat{c}, \hat{T})$.

Ce résultat subsiste pour tout point $(\hat{c}, \hat{T}) \in D_2$ vérifiant les conditions de Kuhn-Tucker (4), même si ce point n'est pas un minimum local.

Dans le cas général, il est donc possible de plonger le problème 1 dans un problème plus vaste ayant mêmes solutions et pouvant être plus facile à résoudre.

III.4 Algorithmes généraux

III. 4 a) L'algorithme M.E.P. s'écrit :

– étape 1 – Recherche pour chaque $n \in N$ de l'indice k_o tel que :

$$\Psi_n(D_{k_o}(z_n, \theta_{k_o})) \text{ soit minimum}$$

On pose $c_n^{k_o} = 1$.

– étape 2 – on détermine les nouvelles valeurs θ'_k en résolvant les problèmes

$$\inf_{\theta \in \Theta} W_k(c', \theta)$$

avec

$$W_k(c', \theta) = \sum_{n \in C'_k} \Psi_n \circ D_k(z_n, \theta)$$

La démonstration de la convergence de M.E.P. est faite dans le cas général.

III.4.b.)

L'algorithme de directions admissibles peut se généraliser au cas où W est continument différentiable lorsque Θ est un espace vectoriel de dimension finie.

III. 5 Remarque

La formulation mathématique donnée à notre problème permet de le rapprocher du problème de localisation-transport. [4].

Soit $D(z_n, \theta_k)$ le coût du transport d'une unité d'un certain produit de la "source" θ_k vers la "destination" z_n . Soit c_n^k la quantité de produit envoyé de θ_k à z_n .

On se propose de déterminer l'emplacement optimal des $|K|$ "sources" et les quantités c_n^k de telle sorte que le coût global du transport d'un certain produit soit minimum. La formulation mathématique de ce problème se ramène, dans certains cas particuliers à notre problème 1. Les algorithmes proposés en [4] sont des variantes de M.E.P.

IV – EXEMPLES –

IV. 1 – Exemple 1 –

Il s'agit d'un échantillon fictif de 200 observations, généré et classé comme il a été indiqué en II.6.

Caractéristiques de chaque distribution –

Classe n°	Nombre d'observations	Moyenne Exacte	Matrice de covariance exacte
1	80	(0. , 0.)	$\begin{pmatrix} 50.5 & 49.5 \\ 49.5 & 50.5 \end{pmatrix}$
2	60	(0. , 20.)	$\begin{pmatrix} 200 & 100 \\ 100 & 100 \end{pmatrix}$
3	60	(10. , 10.)	$\begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}$

Cet échantillon est représenté sur la figure 1. Les observations sont notées en utilisant un symbole différent pour chaque classe

Pour la classification exacte la valeur du critère W est :

$W = 934$. si l'on utilise les moyennes et matrices de covariance exactes.

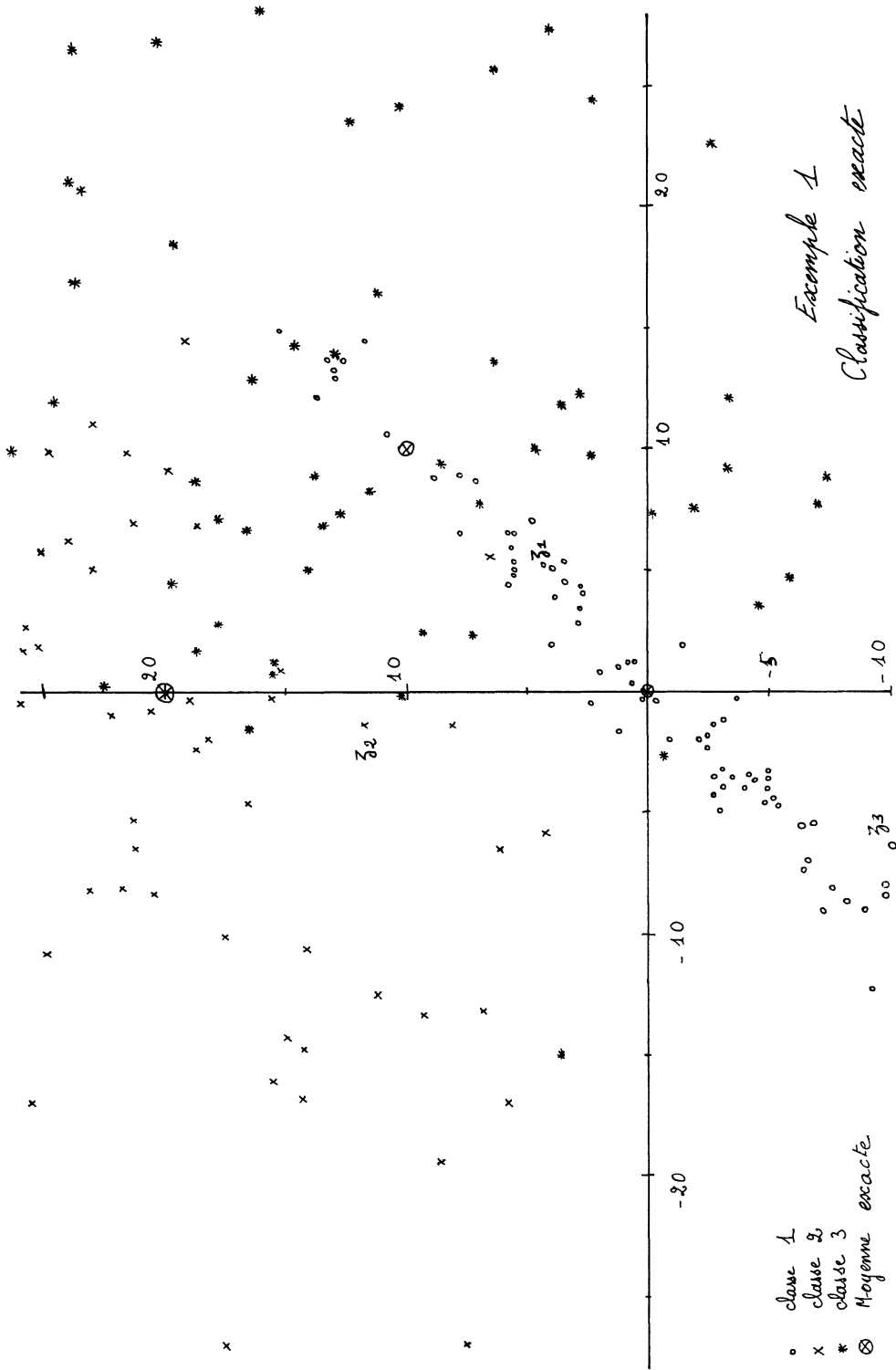
$W = 926.5$ si l'on utilise les moyennes et matrices de covariance estimées à partir des valeurs de l'échantillon..

Nous avons effectué trois classifications de cet échantillon en procédant comme il a été indiqué en II.6. Le nombre de classes restait fixé à trois. Les trois essais ont été effectués avec des moyennes initiales différentes à chaque essai. Ces essais ont donné la même classification finale. Celle ci est représentée sur la figure 2. Remarquons que la classe 1, nuage allongé selon la première bisectrice est très bien retrouvé.

Caractéristiques de la classification obtenue –

Classe n°	Nombre d'observations	Moyenne empirique	Matrice de covariance empirique
1	76	(-0.8, - 1.0)	$\begin{pmatrix} 34.8 & 33.8 \\ 33.8 & 34.8 \end{pmatrix}$
2	67	(-1.5, 21.5)	$\begin{pmatrix} 151 & 84.8 \\ 84.8 & 90.6 \end{pmatrix}$
3	57	(13.4 , 10.8)	$\begin{pmatrix} 70.7 & 28.9 \\ 28.9 & 76.4 \end{pmatrix}$

La valeur du critère W est 891.2



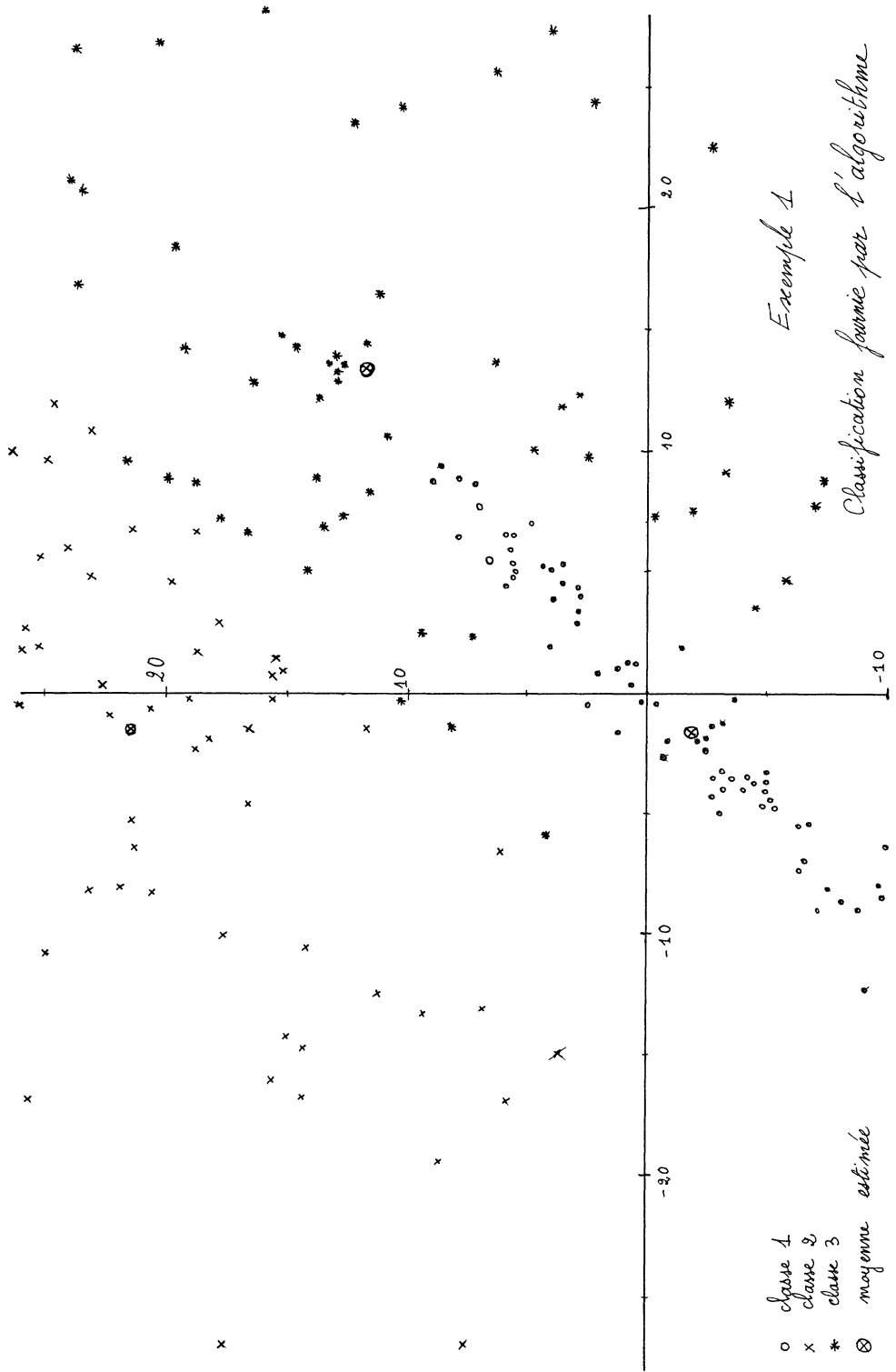


figure 2

Les résultats apparaissent nettement sur la figure 2. Notons seulement que pour des raisons évidentes les moyennes présentent un biais relativement important et que la comparaison des matrices de covariance est assez difficile à réaliser. Il faudrait comparer d'une part les valeurs propres des matrices respectives, d'autre part les vecteurs propres. Notons enfin que, même si nous ne sommes pas certains que la classification obtenue soit optimale, il n'est pas très utile d'en rechercher une meilleure.

Il est commode, pour comparer globalement deux classifications, de représenter les divers échanges sous la forme d'un tableau à 2 entrées dans le terme général n_{ij} désigne le nombre d'observations appartenant à la classe i dans la première classification et placées dans la classe j pour l'autre classification. Dans ce qui suit les lignes correspondent à la classification exacte.

Tableau des échanges de l'exemple 1 –

72	0	8	80
1	54	5	60
3	13	44	60
76	67	57	200

IV.2 – Exemple 2 –

- . Echantillon de 100 observations de R^5
- . 3 variables aléatoires

Les trois nuages d'observations ne sont pas nettement dissociés ; Les matrices de covariance sont très différentes. La difficulté, dans cet exemple, provient du très grand nombre de paramètres à estimer simultanément.

Caractéristiques du nuage d'observations:

Numéro de la variable	Nombre d'observations	Moyenne Exacte	Moyenne empirique	Matrice de covariance Exacte				
1	30	0.	0.	1.0	0.	0.	0.	0.
		0.	-0.1	0.	100.	0.	0.	0.
		0.	-1.6	0.	0.	25.	0.	0.
		0.	-2.4	0.	0.	0.	100.	0.
		0.	3.4	0.	0.	0.	0.	100.
2	30	0.	2.0	100.	0.	0.	0.	0.
		20.	20.1	0.	100.	0.	0.	0.
		0.	0.	0.	0.	100.	0.	0.
		0.	0.	0.	0.	0.	100.	0.
		0.	0.	0.	0.	0.	0.	100.
3	40	10.	10.1	1.	0.	0.	0.	0.
		10.	10.3	0.	9.	0.	0.	0.
		0.	0.8	0.	0.	25.	0.	0.
		0.	0.1	0.	0.	0.	100.	0.
		0.	0.4	0.	0.	0.	0.	100.

Valeur du critère W : 939.6 (avec paramètres exacts)

: 911.5 (avec paramètres estimés)

Nous avons réalisé cinq classifications en utilisant à chaque fois des valeurs de μ_k différentes.

Meilleure classification obtenue :

$$W = 904.5$$

Moyennes

classe 1 : (0., -0.3, -1.5, -2.4, 3.3)

classe 2 : (2.2, 20., 0., 0., 0.1)

classe 3 : (10.1, 10.3, 0.8, 0.1, 0.3)

Tableau des échanges

29	1	0	30
1	29	0	30
0	1	39	40
30	31	39	100

V CONCLUSION

Ramener une classification à un problème d'optimisation fournit donc un moyen commode pour construire des algorithmes utilisant au mieux les particularités du problème et étudier la convergence des algorithmes ainsi construits, ceci en utilisant tout l'acquis de l'analyse numérique.

Les recherches effectuées sur la minimisation de critères à variables binaires nous fournissent la possibilité de résoudre complètement certains problèmes particuliers (par exemple III.2 – Exemple 3). Mais étant donné le nombre élevé de variables c_n^k , nous ne pensons pas que ces méthodes soient directement applicables à de gros problèmes.

Outre l'introduction d'un critère, toute méthode itérative nécessite l'utilisation d'un espace Θ espace de "paramètres" dont le rôle est de "résumer au mieux" l'information obtenue sur les classes C_k au cours des itérations précédentes. L'introduction de ces paramètres augmente aussi les performances de l'algorithme ; il n'est plus nécessaire de stocker les quantités $d(z_i, z_j)$. Mais elle a une signification plus profonde :

. Il est interdit dans le cas général – et ceci pour des raisons fondamentales – d'espérer retrouver la classification exacte puisque les variables c_n^k ne convergent pas nécessairement.

. Par contre il peut exister un choix de Θ tel que θ_k converge vers sa "vraie valeur" quand le nombre d'observations croît indéfiniment ; Le sens de l'expression ci dessus est à préciser en fonction des particularités du problème. Une fois les paramètres θ_k déterminés, l'utilisateur a la possibilité de choisir une classification, partition ou recouvrement de Z .

L'étude de la convergence, lorsque $|N|$ tend vers l'infini, des paramètres θ reste à faire. Les résultats classiques relatifs au maximum de vraisemblance sont inapplicables ici. Le seul résultat que nous connaissions est le théorème établi par Mac Queen [9] dans le cas d'observations de \mathbf{R}^p .

Des recherches en ce sens devraient permettre de rattacher plus étroitement la théorie de la classification à la statistique mathématique.

Note :

Cette étude a été réalisée au Laboratoire de Statistique de l'université Paul Sabatier à Toulouse.

Nous tenons à exprimer nos remerciements à Monsieur le professeur Caussinus pour toutes les remarques et suggestions qu'il a bien voulu faire, remarques qui nous ont beaucoup aidé à préciser nos idées premières.

REFERENCES

- [1] BARRA – “Notions fondamentales de statistique mathématique” Dunod – Paris (1971)
- [2] BEALE “Euclidean cluster analysis” Bulletin ISI 43,2 pages 92–94 (1969)
- [3] J. CEA “Optimisation – Théorie et algorithmes” Dunod – Paris (1971)
- [4] L. COOPER – “The Transportation-location problem” *Operations Research* 20–1 p. 94 (1972)
- [5] DAY – “Estimating the components of a mixture of normal distributions” *Biometrika* 56 p. 463–474 (1969)
- [6] DIDAY – “Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques” *Rev. Stat. Appli.* XIX–2 p.19–23 (1971)
- [7] GREGOR – “An algorithm for the decomposition of a distribution into gaussian components” *Biometrics* 25–1 (1969)
- [8] KUHN–TUCKER – “Non linear programming” 2d Berkely Symposium (1951)
- [9] MAC QUEEN – “Some methods for classification and analysis of multivariate observations” 5 th Berkeley Symposium (1965)
- [10] RANDON – “Problèmes de classification et théorie de l'optimisation” Thèse de 3eme cycle – Toulouse – (1972)
- [11] SCOTT–SYMONS – “Clustering methods based on likelihood – ratio criteria” *Biometrics* 27–2. (1971)