Méthodes de classification non supervisée

Jean-Marie Monnez

COURS D'ANALYSE DES DONNEES ET APPRENTISSAGE
Méthodes de classification non supervisée

Jean-Marie MONNEZ
Université de Lorraine, CNRS, Inria, Institut Elie Cartan de Lorraine, F-54000 Nancy, France

April 29, 2021

Soit un ensemble d'individus $I = \{1, \cdots, n\}$ munis de poids $\{p_1, \cdots, p_n\}$ sur lesquels on a observé des caractères quantitatifs ou qualitatifs. Le but de la *classification non supervisée* (ou simplement *classification*) est de constituer des classes d'éléments de I telles que :

1) chaque classe soit composée d'individus semblables vis-à-vis des caractères (homogénéité intra-classes) ;

2) les classes soient hétérogènes entre elles vis-à-vis des caractères (hétérogénéité inter-classes).

De façon générale :

Soit un ensemble d'objets $I = \{1, \cdots, n\}$ (individus, classes d'individus, modalités de caractères qualitatifs, caractères quantitatifs, points d'un espace euclidien,....).
Soit $\{p_1, \cdots, p_n\} = P$ un ensemble de poids attribués aux objets tels que :

$$0 \leq p_i \leq 1 \ ; \ \sum_{i=1}^{n} p_i = 1.$$

Soit d une dissimilarité, application de $I \times I$ dans $\mathbb{R}^+$ telle que :

$$\cdot \ d(i_1, i_2) = d(i_2, i_1) \quad \text{(axiome de symétrie)}$$
$$\cdot \ d(i_1, i_1) = 0 \quad \text{(axiome de séparation, partie directe)}$$

On a la structure (I, P, d).
Le but de la classification est de construire des classes d'éléments de I homogènes à l'intérieur et hétérogènes entre elles au sens de la dissimilarité d.

On va étudier trois familles de méthodes de classification :

1) des méthodes de construction d'une *hiérarchie* de parties de I en effectuant des regroupements successifs de parties par l'algorithme de *classification ascendante hiérarchique* (CAH) ; on obtient à chaque pas de l'algorithme une classification de I ;
2) des méthodes de partitionnement en un nombre fixé a priori de classes : *méthodes des nuées dynamiques* (MND) ;
3) des méthodes basées sur l'utilisation de *modèles probabilistes de mélange.*

Première partie :
La classification ascendante hiérarchique (CAH)

.

1 Algorithme de CAH

1.1 Hiérarchie indicée et dendrogramme (arbre de classification)

Soit un ensemble I. Soit $\pi(I)$ l'ensemble des parties de I.

Définition *Un sous-ensemble $\mathcal{H}$ de $\pi(I)$ est appelé une hiérarchie si :*
1) $I \in \mathcal{H}$
2) $\forall i \in I, \{i\} \in \mathcal{H}$
3) $\forall\, H, H' \in \mathcal{H} \quad (H \cap H' \neq \emptyset) \Rightarrow (H \subset H' \text{ ou } H' \subset H)$.

Exemple
$I = \{1, 2, 3, 4, 5, 6, 7\}$
$\mathcal{H} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{1, 2, 4\}, \{3, 5\}, \{3, 5, 6\}, \{3, 5, 6, 7\}, \{1, 2, 3, 4, 5, 6, 7\}\}$.

Définition *Une hiérarchie $\mathcal{H}$ de parties de I est dite indicée (respectivement indicée au sens large) s'il existe une application $f : \mathcal{H} \to \mathbb{R}^+$ appelée indice telle que :*
1) $f(H) = 0 \Leftrightarrow \exists\, i \in I : H = \{i\}$
2) $\forall\, H, H' \in \mathcal{H} \quad (H \subset H', H \neq H') \Rightarrow (f(H) < f(H'))$ $\big(resp.\ f(H) \leq f(H')\big)$.

Exemple
$f\big(\{1\}\big) = \cdots = f\big(\{7\}\big) = 0$
$f\big(\{1, 2, 4\}\big) = 2$
$f\big(\{3, 5\}\big) = 1$
$f\big(\{3, 5, 6\}\big) = 1, 5$
$f\big(\{3, 5, 6, 7\}\big) = 3$
$f\big(\{1, 2, 3, 4, 5, 6, 7\}\big) = 4$

On peut représenter cette hiérarchie indicée par un dendrogramme ou arbre de classification.

1.2 Algorithme

Soit la structure (I, P, d). A partir de d, on définit une distance entre classes Δ ; on étudiera dans la suite la définition de plusieurs distances entre classes.

On considère alors le triplet (I, P, Δ).

$\underline{1^{er} \text{ pas}}$: On considère la partition $\mathcal{P}_0 = \big\{\{1\}, \{2\}, ..., \{n\}\big\}$. On calcule, pour tous les couples (i_1, i_2) d'éléments de I, $\Delta\big(\{i_1\}, \{i_2\}\big)$. On regroupe en une seule classe les deux classes $\{l\}$ et $\{m\}$ pour lesquelles cette distance Δ est minimale.

On considère la partition $\underbrace{\mathcal{P}_1 = \big\{\{l, m\}, \{1\}, ..., \{n\}\big\}}_{\text{sauf } \{l\} \text{ et } \{m\}}$ à $(n-1)$ éléments.

$\underline{\text{r}^{\text{ième}} \text{ pas}}$: On part de la partition $\mathcal{P}_{r-1}$ à $n-(r-1)$ éléments. On calcule les distances Δ entre les couples d'éléments de cette partition ; en fait, il ne reste à calculer que les distances entre la classe contituée au pas $(r-1)$ par réunion de deux classes et les autres classes.

On regroupe les deux classes à distance minimale et on constitue ainsi la partition $\mathcal{P}_r$ à $(n-r+1)-1 = n-r$ éléments.

$\underline{(n-1)^{\text{ème}} \text{ pas}}$: On constitue la partition $\mathcal{P}_{n-1} = \{I\}$.

Notons I_r la classe constituée au pas r de la CAH. L'ensemble des classes $\big\{\{1\}, \cdots, \{n\}, I_1, I_2, \cdots, I_{n-1} = I\big\}$ est par construction une hiérarchie de parties de I.

1.3 Indice de la hiérarchie dans le cas de la réductibilité

Définition *On dit que la distance entre classes Δ est réductible si, quelles que soient trois parties I_1, I_2, I_3 de I deux à deux disjointes, quel que soit $\rho > 0$,*

$$\big(\Delta(I_3, I_1) \geq \rho,\ \Delta(I_3, I_2) \geq \rho,\ \Delta(I_1, I_2) \leq \rho\big) \Rightarrow \big(\Delta(I_3, I_1 \cup I_2) \geq \rho\big).$$

Proposition *Si la distance entre classes Δ est réductible, les distances entre les classes de la partition P_r constituée au pas r de la CAH sont supérieures ou égales à la distance entre les classes réunies à ce pas.*

Démonstration

Au pas r, on part de la partition $\mathcal{P}_{r-1}$ dont les éléments sont notés dans cette démonstration $I_1, I_2, I_3, \cdots, I_{n-r+1}$. Supposons que l'on réunisse I_1 et I_2 en une classe pour constituer

$$\mathcal{P}_r = \{I_1 \cup I_2, I_3, \cdots, I_{n-r+1}\}.$$

Soit l et $m > 2$. On a $\Delta(I_l, I_m) \geq \Delta(I_1, I_2)$, car I_1 et I_2 sont à distance minimale.

Considérons $\Delta(I_l, I_1 \cup I_2)$.

Prenons $\rho = \Delta(I_1, I_2)$. On a $\Delta(I_l, I_1) \geq \rho,\ \Delta(I_l, I_2) \geq \rho,\ \Delta(I_1, I_2) = \rho$, donc comme Δ est réductible, $\Delta(I_l, I_1 \cup I_2) \geq \rho = \Delta(I_1, I_2)$.■

Conséquence

Dans ce cas, on peut prendre pour indice de la classe $I_1 \cup I_2$ constituée au pas r de la CAH le nombre $\Delta(I_1, I_2)$, car la distance entre les classes réunies au pas $r+1$ sera supérieure ou égale à $\Delta(I_1, I_2)$.

2 Cas où les objets sont représentés par des points d'un espace euclidien

Cadre d'étude :

$$\begin{cases} I = \{1, 2, \cdots, n\} \\ i \mapsto A_i \in (\mathbb{R}^p, M) \\ P = \{p_1, p_2, \cdots, p_n\} \\ d = d_M \end{cases}$$

I est un ensemble d'objets, A_i le point représentatif de l'objet i dans $\mathbb{R}^p$, P un ensemble de poids attribués aux objets, d la distance euclidienne associée à la métrique M.

2.1 Exemples d'application

2.1.1 Tableaux (individus × caractères quantitatifs)

On a observé sur n individus numérotés de 1 à n des caractères quantitatifs $x^1, ..., x^p$ que l'on centre. On a le tableau des mesures centrées des caractères :

	x^1	.	.	x^j	.	.	x^p
1	.	.	.	.	.	.	.
.	.			.			.
.	.			.			.
.	.			.			.
i	.	.	.	x_i^j	.	.	.
.	.			.			.
.	.			.			.
.	.			.			.
n	.	.	.	.	.	.	.

a) Classification des individus.

$$I = \{1, \cdots, n\}$$
$$i \longmapsto A_i(x_i^1, \cdots, x_i^P) \in (\mathbb{R}^P, M)$$

(M peut être par exemple la métrique de l'ACP normée ou de l'ACG ou de l'AFM).

b) Classification des caractères.

$$I = \{x^1, \cdots, x^p\}, \text{ ensemble des caractères centrés}$$
$$Q = \{q^1, \cdots, q^p\}, \text{ ensemble de poids attribués aux caractères.}$$

Le caractère x^j est représenté par le point $B^j(x_1^j, \cdots, x_n^j\} \in \mathbb{R}^n$. Si on munit $\mathbb{R}^n$ de la métrique D des poids des individus, on a :

$$d(B^j, B^{j'}) = \sqrt{\sum_{i=1}^{n} p_i(x_i^j - x_i^{j'})^2} = s_{x^j - x^{j'}} = \sqrt{s_{x^j}^2 - 2s_{x^j x^{j'}} + s_{x^{j'}}^2}.$$

Si les caractères sont réduits, on a :

$$d(B^j, B^{j'}) = \sqrt{2 - 2\, r_{x^j x^{j'}}}.$$

Dans ce cas :

$$\text{si} \quad r_{x^j x^{j'}} = +1,\ d(B^j, B^{j'}) = 0\ ;$$
$$\text{si} \quad r_{x^j x^{j'}} = -1,\ d(B^j, B^{j'}) = 2.$$

2.1.2 Tableaux de contingence

On a observé sur n individus deux caractères qualitatifs x, de modalités $a_1, ..., a_r$, et y, de modalités $b_1, ..., b_s$. On a le tableau de contingence :

$x \backslash y$	b_1	.	.	b_j	.	.	b_s	
a_1	.	.	.	.	.	.	.	
.	.			.			.	
.	.			.			.	
.	.			.			.	
a_i	.	.	.	n_{ij}	.	.	.	$n_{i.}$
.	.			.			.	
.	.			.			.	
.	.			.			.	
a_r	.	.	.	.	.	.	.	
				$n_{.j}$				n

On peut classifier les modalités de x ou de y :

$$\begin{cases} I = \{a_1, \cdots, a_r\} \\ a_i \mapsto A_i\left(\frac{f_{ij}}{f_{i\cdot}}, j = 1, \cdots, s\right) \in \mathbb{R}^s \\ P = \{f_{1\cdot}, \cdots, f_{r\cdot}\} \\ \quad d \text{ est la distance du khi-deux} \end{cases}$$

2.1.3 Tableaux (individus × indicatrices de modalités de caractères qualitatifs)

On a observé sur n individus numérotés de 1 à n des caractères qualitatifs $x^1, ..., x^r$. On a le tableau des valeurs des indicatrices des modalités des caractères :

	x^{11}	.	.	.	x^{1m_1}	.	x^{jk}	.	x^{r1}	.	.	.	x^{rm_r}
1	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.						.						.
.	.						.						.
.	.						.						.
i	.	.	.	.	.	.	x_i^{jk}	.	.	.	.	.	.
.	.						.						.
.	.						.						.
.	.						.						.
n	.	.	.	.	.	.	.	.	.	.	.	.	.

On peut classifier les individus ou les modalités de caractères qualitatifs en traitant formellement ce tableau comme un tableau de contingence.

2.2 Critère de partitionnement basé sur l'inertie

Soit le cadre d'étude :

$$\begin{cases} I = \{1, \cdots, n\}, \text{ ensemble d'objets partitionné en classes } I_1, \cdots, I_r \\ i \mapsto A_i \in (\mathbb{R}^p, M) \\ P = \{p_1, \cdots, p_n\} \\ d = d_M \end{cases}$$

Notons G_k le barycentre des points (A_i, p_i), $i \in I_k$. Notons $P_k = \sum_{i \in I_k} p_i$.

Notons G le barycentre des points (A_i, p_i), $i \in I$.

On a la formule de HUYGENS :

$$\underbrace{\sum_{i \in I} p_i \, d^2(A_i, G)}_{\text{inertie totale}} = \underbrace{\sum_{k=1}^{r} \sum_{i \in I_k} p_i \, d^2(A_i, G_k)}_{= \text{inertie intra}} + \underbrace{\sum_{k=1}^{r} P_k \, d^2(G_k, G)}_{+ \text{inertie inter}}$$

Notons $\mathcal{G}$ l'ensemble des barycentres. La formule s'écrit :

$$\mathcal{J}_I(G) = \sum_{k=1}^{r} \mathcal{J}_{I_k}(G_k) + \mathcal{J}_{\mathcal{G}}(G).$$

On peut définir le critère de partitionnement suivant : *déterminer la partition en r classes d'inertie intra minimale ou, de façon équivalente, d'inertie inter maximale.*

Ceci est en pratique irréalisable, car le nombre de partitions est en général trop élevé.

Notons P_{nk} le nombre de partitions en k classes d'un ensemble de n éléments. On a :

$$P_{nk} = P_{n-1,k-1} + kP_{n-1,k}.$$

En effet : soit i un élément fixé de I ; deux cas sont possibles :

a) i constitue à lui seul une classe ; il y a $P_{n-1,k-1}$ partitions de ce type ;

b) i n'est pas seul dans une classe ; $I - \{i\}$ est alors divisé en k classes : il y a $P_{n-1,k}$ partitions de $I - \{i\}$; or i peut faire partie de n'importe laquelle de ces k classes : il y a donc $kP_{n-1,k}$ partitions de ce type.

On démontre alors par récurrence que :

$$P_{nk} = \sum_{i=0}^{k} (-1)^{k-i} \frac{i^n}{i!(k-i)!}$$

Pour $n = 50$ et $k = 10$, on a $\quad P_{nk} = 2,61.10^{43}$.

2.3 La distance de WARD

Définition *Soit I_1 et I_2 deux classes d'éléments de I, de poids respectifs P_1 et P_2. Soit G_1 et G_2 les barycentres respectifs des points (A_i, p_i) pour $i \in I_1$ et $i \in I_2$. Soit $G_{1,2}$ le barycentre des points (A_i, p_i) pour $i \in I_1 \cup I_2$. La distance entre classes de WARD est :*

$$\begin{aligned} \Delta(I_1, I_2) &= \frac{P_1 P_2}{P_1 + P_2}\, d^2(G_1, G_2) && (1)\\ &= P_1 d^2(G_1, G_{1,2}) + P_2 d^2(G_2, G_{1,2}) && (2)\\ &= \mathcal{J}_{I_1 \cup I_2}(G_{1,2}) - \mathcal{J}_{I_1}(G_1) - \mathcal{J}_{I_2}(G_2) && (3) \end{aligned}$$

Démonstration
On démontre facilement que (1) = (2).
Montrons que (2) = (3).
Ecrivons la formule de Huygens pour l'ensemble $I_1 \cup I_2$:

$$\mathcal{J}_{I_1 \cup I_2}(G_{1,2}) = \underbrace{\mathcal{J}_{I_1}(G_1) + \mathcal{J}_{I_2}(G_2)} + \underbrace{P_1 d^2(G_1, G_{1,2}) + P_2 d^2(G_2, G_{1,2})}.$$

On en déduit que (2) = (3).■

Propriété *Soit I_1, I_2, I_3 trois classes d'éléments de I deux à deux disjointes. On a :*

$$\Delta(I_3, I_1 \cup I_2) = \frac{1}{P_1 + P_2 + P_3}\big((P_1 + P_3)\Delta(I_1, I_3) + (P_2 + P_3)\Delta(I_2, I_3) - P_3 \Delta(I_1, I_2)\big).$$

On peut utiliser cette formule à chaque pas de la CAH pour calculer la distance entre une classe et la nouvelle classe constituée par réunion de 2 classes.

Propriété *La distance entre classes Δ est réductible.*

Démonstration
Soit $\rho > 0$.

$$\big(\Delta(I_1, I_3) \geq \rho, \quad \Delta(I_2, I_3) \geq \rho, \quad \Delta(I_1, I_2 \leq \rho)\big) \Rightarrow$$
$$\Rightarrow \Delta(I_3, I_1 \cup I_2) \geq \frac{1}{P_1 + P_2 + P_3}\big((P_1 + P_3)\rho + (P_2 + P_3)\rho - P_3 \rho\big) = \rho.\blacksquare$$

On prend alors pour indice de la classe $I_1 \cup I_2$ obtenue par réunion de deux classes à un pas de l'algorithme de CAH la distance $\Delta(I_1, I_2)$ entre ces deux classes.

Propriété d'optimalité de la distance de Ward

Soit la partition $\mathcal{P}_r$ obtenue au pas r de la CAH ; on note :

$$\mathcal{P}_r = \{I_1, I_2, \cdots, I_{n-r}\}.$$

On note $\mathcal{J}_{\mathcal{G}_r}(G)$ son inertie inter-classes.
On suppose qu'au pas $(r+1)$, ce sont les classes I_1 et I_2 qui sont réunies :

$$\mathcal{P}_{r+1} = \{I_1 \cup I_2, I_3, \cdots, I_{n-r}\}.$$

On note $\mathcal{J}_{\mathcal{G}_{r+1}}(G)$ l'inertie inter-classes de la partition $\mathcal{P}_{r+1}$.

Proposition *Soit Δ la distance de Ward. On a : $\mathcal{J}_{\mathcal{G}_r}(G) - \mathcal{J}_{\mathcal{G}_{r+1}}(G) = \Delta(I_1, I_2)$.*

Démonstration
On écrit la formule de Huygens :

$$\begin{aligned} \mathcal{J}_I(G) &= \mathcal{J}_{I_1}(G_1) + \mathcal{J}_{I_2}(G_2) + \cdots + \mathcal{J}_{I_{n-r}}(G_{n-r}) + \mathcal{J}_{\mathcal{G}_r}(G) && (\text{pas } r)\\ \mathcal{J}_I(G) &= \mathcal{J}_{I_1 \cup I_2}(G_{1,2}) + \mathcal{J}_{I_3}(G_3) + \cdots + \mathcal{J}_{I_{n-r}}(G_{n-r}) + \mathcal{J}_{\mathcal{G}_{r+1}}(G) && (\text{pas } r+1)\\ &\Longrightarrow \mathcal{J}_{\mathcal{G}_r}(G) - \mathcal{J}_{\mathcal{G}_{r+1}}(G) = \mathcal{J}_{I_1 \cup I_2}(G_{1,2}) - \mathcal{J}_{I_1}(G_1) - \mathcal{J}_{I_2}(G_2) = \Delta(I_1, I_2).\blacksquare \end{aligned}$$

Conséquence

Lorsqu'on utilise la distance de WARD comme distance entre classes, l'inertie inter-classes diminue de façon minimale du pas r au pas $(r+1)$ de la CAH.

La partition P_r étant fixée, la partition P_{r+1} obtenue par réunion de deux classes est la meilleure au sens du critère d'inertie inter classes maximale lorsqu'on utilise la distance de Ward.

2.4 Autres distances entre classes

$$\Delta_1(I_1, I_2) = \frac{P_1 P_2}{P_1 + P_2}\, d^2(G_1, G_2) = \mathcal{J}_{I_1 \cup I_2}(G_{1,2}) - \mathcal{J}_{I_1}(G_1) - \mathcal{J}_{I_2}(G_2)$$

$$\Delta_2(I_1, I_2) = \frac{\Delta_1(I_1, I_2)}{P_1 + P_2}$$

$$\Delta_3(I_1, I_2) = \mathcal{J}_{I_1 \cup I_2}(G_{1,2})$$

$$\Delta_4(I_1, I_2) = \frac{\Delta_3(I_1, I_2)}{P_1 + P2}$$

$$\Delta_5(I_1, I_2) = d(G_1, G_2)$$

3 Cas général où l'on a une dissimilarité entre objets

Soit la structure (I, P, d). Soit le tableau de dissimilarités :

	1	.	.	.	j	.	.	.	n
1	.	.	.	.	.	.	.	.	.
.	.				.				.
.	.				.				.
i	.	.	.	.	$d(i,j)$	.	.	.	.
.	.				.				.
.	.				.				.
.	.				.				.
.	.				.				.
n	.	.	.	.	.	.	.	.	.

3.1 Définition de distances entre classes

1) *Distance du lien minimal (single linkage)*

$$\Delta_1(I_1, I_2) = \min\big(d(i_1, i_2),\ i_1 \in I_1,\ i_2 \in I_2\big)$$

2) *Distance du lien maximal (complete linkage)*

$$\Delta_2(I_1, I_2) = \max\big(d(i_1, i_2),\ i_1 \in I_1,\ i_2 \in I_2\big)$$

3) *Distance du lien moyen (average linkage)*

$$\Delta_3(I_1, I_2) = \frac{1}{P_1 P_2} \sum_{i_1 \in I_1} \sum_{i_2 \in I_2} p_{i_1} p_{i_2} d(i_1, i_2)$$

3.2 Calcul de $\Delta(I_3, I_1 \cup I_2)$

On démontre facilement que :

$$\begin{aligned}
\Delta_1(I_3, I_1 \cup I_2) &= \min\big(\Delta_1(I_3, I_1), \Delta_1(I_3, I_2)\big)\ ; \\
\Delta_2(I_3, I_1 \cup I_2) &= \max\big(\Delta_2(I_3, I_1), \Delta_2(I_3, I_2)\big)\ ; \\
\Delta_3(I_3, I_1 \cup I_2) &= \frac{1}{P_1 + P_2}\big(P_1 \Delta_3(I_3, I_1) + P_2 \Delta_3(I_3, I_2)\big).
\end{aligned}$$

Propriété *Les trois distances entre classes définies sont réductibles.*

4 Ultramétrique associée à une hiérarchie indicée.

4.1 Définition d'une distance ultramétrique

Définition *On appelle distance ultramétrique toute application* $u : I \times I \to \mathbb{R}^+$ *telle que :*
1) $u(i_1, i_2) = u(i_2, i_1)$ *(axiome de symétrie)*
2) $u(i_1, i_2) = 0 \Leftrightarrow i_1 = i_2$ *(axiome de séparation)*
3) $u(i_1, i_2) \leq \max\big(u(i_1, i_3), u(i_3, i_2)\big)$ *(inégalité ultramétrique).*

Remarque Une ultramétrique est une distance :

$$u(i_1, i_2) \leq \max\big(u(i_1, i_3), i(i_3, i_2)\big) \leq u(i_1, i_3) + u(i_3, i_2).$$

Propriété *Tout triangle* (i_1, i_2, i_3) *est isocèle et le troisième côté est au plus égal aux deux côtés égaux.*

Démonstration
Soit $u(i_1, i_2)$ le plus petit des trois côtés.

$$u(i_1, i_3) \leq \max\big(u(i_1, i_2), u(i_2, i_3)\big) = u(i_2, i_3)$$
$$u(i_2, i_3) \leq \max\big(u(i_2, i_1), u(i_1, i_3)\big) = u(i_1, i_3)$$

Donc : $u(i_1, i_3) = u(i_2, i_3) \geq u(i_1, i_2)$.■

4.2 Bijection entre ultramétrique et hiérarchie indicée

4.2.1 Construction d'une hiérarchie indicée à partir d'une ultramétrique

On utilise l'algorithme de CAH pour construire une hiérarchie indicée.

Au premier pas de la CAH, on a $\mathcal{P}_0 = \{\{1\}, ..., \{n\}\}$. On définit la distance entre classes $\Delta\big(\{i_1\}, \{i_2\}\big) = u(i_1, i_2)$.

Montrons par récurrence sur r que, étant donné deux classes I_1 et I_2 obtenues par l'algorithme de CAH, la distance entre un élément de I_1 et un élément de I_2 est toujours la même : on prend alors cette valeur commune pour définir $\Delta(I_1, I_2)$.

Soit la partition $\mathcal{P}_r = \{I_1, I_2, \cdots, I_{n-r}\}$; on suppose l'hypothèse de récurrence vérifiée pour $\mathcal{P}_r$.

Soit la partition $\mathcal{P}_{r+1} = \{I_1 \cup I_2, I_3, \cdots, I_{n-r}\}$. Montrons que la distance entre un point de $I_1 \cup I_2$ et un point de $I_l,\ l > 2$, est toujours la même.

Soit $i_1 \in I_1, i_2 \in I_2, i_l \in I_l$. Considérons le triangle (i_1, i_2, i_l).

Par hypothèse de récurrence :

$$u(i_1, i_2) = \Delta(I_1, I_2)\ ;\ u(i_1, i_l) = \Delta(I_1, I_l)\ ;\ u(i_2, i_l) = \Delta(I_2, I_l).$$

Or :

$$\Delta(I_1, I_2) \leq \Delta(I_1, I_l)\ ;\ \Delta(I_1, I_2) \leq \Delta(I_2, I_l).$$

Donc : $u(i_1, i_2) \leq u(i_1, i_l) = u(i_2, i_l)$.

Par conséquent :

$$\Delta(I_1, I_l) = \Delta(I_2, I_l).$$

On définit alors $\Delta(I_1 \cup I_2, I_l) = \Delta(I_1, I_l) = \Delta(I_2, I_l) \geq \Delta(I_1, I_2)$.

On peut prendre pour indice de la classe $I_1 \cup I_2$ la distance $\Delta(I_1, I_2)$.

On note g l'application qui associe à l'ultramétrique u la hiérarchie indicée $(\mathcal{H}, f)$.

4.2.2 Construction d'une ultramétrique à partir d'une hiérarchie indicée

Soit une hiérarchie indicée $(\mathcal{H}, f)$ de parties de I. Soit $H(i_1, i_2)$ la plus petite partie de la hiérarchie contenant i_1 et i_2.

On définit $u(i_1, i_2) = f\big(H(i_1, i_2)\big)$.

Montrons que u est une ultramétrique.

1) On a : $H(i_1, i_2) = H(i_2, i_1)$; donc $u(i_1, i_2) = u(i_2, i_1)$.

2) $u(i_1, i_2) = 0 = f\big(H(i_1, i_2)\big) \Leftrightarrow H(i_1, i_2)$ a un seul élément $\Leftrightarrow i_1 = i_2$.

3) Soit i_1, i_2, i_3 trois éléments de I. On considère les classes $H(i_1, i_2), H(i_1, i_3)$ et $H(i_2, i_3)$.

· Supposons que : $H(i_1, i_2) \subset H(i_1, i_3)$
On a alors : $H(i_2, i_3) = H(i_1, i_3)$
Donc : $H(i_1, i_2) \subset H(i_1, i_3) = H(i_2, i_3)$
$f\big(H(i_1, i_2)\big) \leq f\big(H(i_1, i_3)\big) = f\big(H(i_2, i_3)\big)$
$u(i_1, i_2) \leq u(i_1, i_3) = u(i_2, i_3)$
$u(i_1, i_2) \leq \max\big(u(i_1, i_3), u(i_3, i_2)\big)$

· Supposons que : $H(i_1, i_3) \subset H(i_1, i_2)$
Alors : $H(i_2, i_3) = H(i_2, i_1)$
$H(i_1, i_3) \subset H(i_1, i_2) = H(i_2, i_3)$
$f\big(H(i_1, i_3)\big) \leq f\big(H(i_1, i_2)\big) = f\big(H(i_2, i_3)\big)$
$u(i_1, i_3) \leq u(i_1, i_2) = u(i_2, i_3)$
$u(i_1, i_2) = \max\big(u(i_1, i_3), u(i_3, i_2)\big)$.■

On établit le théorème suivant :

Théorème *L'application g est une bijection de l'ensemble des ultramétriques sur I dans l'ensemble des hiérarchies indicées de parties de I.*

Conclusion

Lorsqu'on dispose d'une distance ultramétrique entre éléments de I et que l'on utilise l'algorithme de CAH, le choix de la distance Δ entre classes est naturel : c'est la distance commune entre un élément d'une classe et un élément d'une autre classe. Lorsque l'on dispose d'une dissimilarité, on peut définir différentes distances entre classes ; on peut alors poser le problème suivant : déterminer l'ultramétrique "la plus proche" en un certain sens de la dissimilarité donnée.

4.3 Propriété d'optimalité des distances du lien minimal et du lien maximal

4.3.1 Cas de la distance du lien minimal

$$\Delta_1(I_1, I_2) = \min\big(d(i_1, i_2), i_1 \in I_1, i_2 \in I_2\big)$$

En utilisant l'algorithme de CAH, on construit une hiérarchie indicée $(\mathcal{H}_1, f_1)$. A cette hiérarchie indicée est associée par la bijection g^{-1} définie dans le paragraphe précédent une ultramétrique u_1.

$H(i_1, i_2)$, plus petite partie contenant i_1 et i_2, est la réunion d'une classe I_1 contenant i_1 et d'une classe I_2 contenant i_2.

$$\begin{array}{rcl} \Delta_1 & \longmapsto & (\mathcal{H}_1, f_1) \longmapsto u_1 \\ u_1(i_1, i_2) & = & f_1\big(H(i_1, i_2)\big) = f_1(I_1 \cup I_2) \quad (i_1 \in I_1, i_2 \in I_2) \\ & = & \Delta_1(I_1, I_2) \leq d(i_1, i_2) \end{array}$$

Donc $u_1 \leq d$. On établit la proposition suivante :

Proposition *L'ultramétrique u_1 est l'enveloppe supérieure des ultramétriques inférieures ou égales à d :*

$$u_1(i_1, i_2) = \sup\big(u(i_1, i_2),\ u \leq d\big).$$

C'est donc l'ultramétrique la plus proche inférieurement de la dissimilarité d.

4.3.2 Cas de la distance du lien maximal

$$\begin{aligned} \Delta_2(I_1, I_2) &= \max\big(d(i_1, i_2),\ i_1 \in I_1, i_2 \in I_2\big) \\ \Delta_2 &\longmapsto (\mathcal{H}_2, f_2) \longmapsto u_2 \\ u_2(i_1, i_2) &= f_2\big(H(i_1, i_2)\big) = f_2(I_1 \cup I_2) = \Delta_2(I_1, I_2) \geq d(i_1, i_2). \end{aligned}$$

Donc : $u_2 \geq d$. On établit la proposition suivante :

Proposition *L'ultramétrique u_2 est un élément minimal dans l'ensemble des ultramétriques supérieures ou égales à d.*

Remarque Il peut exister plusieurs éléments minimaux dans l'ensemble des ultramétriques supérieures ou égales à une dissimilarité d. Par exemple :

$I = \{a, b, c\}$; $d(a, b) = 3$; $d(a, c) = 2$; $d(b, c) = 1$.
Il y a deux éléments minimaux u_1 et u_2 dans l'ensemble des ultramétriques supérieures ou égales à d :
$u_1(a, b) = 3$; $u_1(a, c) = 2$; $u_1(b, c) = 3$;
$u_2(a, b) = 3$; $u_2(a, c) = 3$; $u_2(b, c) = 1$.■

5 Algorithme accéléré pour grands tableaux : la méthode des voisinages réductibles

On suppose la distance entre classes Δ réductible.

Lorsqu'on effectue une CAH, le nombre de calculs à effectuer peut être très important. On cherche par un algorithme accéléré à diminuer ce nombre.

Algorithme
• $1^{ère}$ étape :
On fixe un nombre ρ_1.
1^{er} pas : On considère l'ensemble $\mathcal{P}_0(\rho_1)$ des classes à un élément telles qu'il existe au moins une autre classe à une distance inférieure à ρ_1 (propriété (a)). On regroupe les deux classes les plus proches et on constitue l'ensemble $\mathcal{P}_1(\rho_1)$.

$2^{ème}$ pas : On élimine de $\mathcal{P}_1(\rho_1)$ les classes qui ne vérifient plus (a). On regroupe les deux classes les plus proches et on constitue l'ensemble $\mathcal{P}_2(\rho_1)$.

etc ..., jusqu'à ce que l'ensemble des classes soit vide.

• $2^{ème}$ étape :
On fixe un nombre $\rho_2 > \rho_1$.
On utilise le même algorithme que dans la première étape en constituant $\mathcal{P}_0(\rho_2)$ à partir de l'ensemble de toutes les classes éliminées à la première étape.

• etc ..., jusqu'à la construction de la classe I.■

On établit la proposition suivante :

Proposition *Lorsque la distance entre classes Δ est réductible, la hiérarchie obtenue par cet algorithme est la même que celle obtenue par l'algorithme habituel.*

Le nombre de calculs à effectuer est moins important, car on élimine à chaque étape des classes pour lesquelles on ne fait aucun calcul ultérieur dans cette étape.

6 Partition en classes et stabilité

A partir d'une hiérarchie indicée, on peut obtenir une partition de I en coupant à un certain niveau le dendrogramme (ce qui revient à retenir une partition obtenue à un pas de l'algorithme). Une règle empirique est de retenir une partition obtenue à un pas précédant une croissance forte de l'indice.

Dans le logiciel SPAD, l'aide à l'interprétation des classes permet d'en définir les caractéristiques (procédure DECLA). Par exemple, soit des classes d'individus sur lesquels on a mesuré des caractères quantitatifs ; on dispose alors pour chaque classe d'une liste des caractères dont la moyenne dans la classe s'écarte beaucoup au sens d'un certain critère de la moyenne générale.

Pour vérifier si l'ensemble de données dont on dispose se prête effectivement à une classification, on peut procéder à plusieurs classifications en utilisant différentes distances entre classes. Si les données sont effectivement classifiables, il ne doit pas y avoir trop de différences entre les partitions à petit nombre de classes. Les objets toujours classés ensemble constituent les parties vraiment homogènes de I.

Deuxième partie :
La méthode des nuées dynamiques (MND)

7 Méthode d'agrégation autour de centres mobiles

7.1 Cadre d'étude

$$\begin{cases} I = \{1, \cdots, n\} \\ i \mapsto A_i(\underline{x}_i) \quad \in (\mathbb{R}^p, M) \\ P = \{p_1, \cdots, p_n\} \\ d = d_M \end{cases}$$

I est un ensemble d'objets, A_i le point représentatif de l'objet i dans $\mathbb{R}^p$, P un ensemble de poids attibués aux objets, d la distance euclidienne associée à la métrique M.

L'objectif de la méthode est de faire une partition de I en r classes (r est fixé a priori).

7.2 Algorithme

Initialisation :

On choisit r points de $\mathbb{R}^p$ (par exemple, au hasard parmi les points A_i) : $\underline{g}_1^0, \cdots, \underline{g}_r^0$.

Autour de ces points, on constitue r classes en affectant chaque objet à la classe correspondant au point dont il est le plus proche :

$$pour\ k = 1, \cdots, r, \quad I_k^0 = \left\{ i \in I \ : \ \forall j = 1, \cdots, r, \quad d(\underline{x}_i, \underline{g}_k^0) \leq d(\underline{x}_i, \underline{g}_j^0) \right\}.$$

(en cas d'égalité, on prend une convention d'affectation).

Remarque L'initialisation peut être une partition de I.

$l^{ième}$ pas :

On détermine les barycentres des classes I_k^{l-1} constituées au pas précédent : $\underline{g}_1^l, \cdots, \underline{g}_r^l$.

Autour de ces points, on constitue r classes en utilisant la règle d'affectation précédente : $I_1^l, \cdots, I_r^l$.

Convergence :

On montre qu'à chaque pas, l'inertie intra-classes $\displaystyle\sum_{k=1}^{r} \sum_{i \in I_k^{l-1}} p_i \, d^2(\underline{x}_i, \underline{g}_k^l)$ décroît ; donc, elle converge vers un minimum local qui dépend de $\underline{g}_1^0, \cdots, \underline{g}_r^0$. En outre, la suite des partitions de I, $\left(\{I_1^l, \cdots, I_r^l\}\right)$, converge vers une forme stable qui dépend du choix de $\underline{g}_1^0, \cdots, \underline{g}_r^0$.

Remarque Minimiser l'inertie intra équivaut à maximiser l'inertie inter. Cet algorithme donne donc une solution locale au problème de la recherche d'une partition d'inertie inter maximale.

Remarque On peut utiliser cet algorithme pour améliorer (on dit également "consolider") une partition obtenue par coupure d'un dendrogramme obtenu par une CAH : à partir de cette partition, on construit par l'algorithme une partition d'inertie intra plus faible. C'est ce que fait le logiciel SPAD.

7.3 Les formes fortes

On peut appliquer plusieurs fois cet algorithme en faisant différentes initialisations.

Les objets toujours classés ensemble constituent les parties vraiment homogènes de I : on les appelle les *formes fortes*.

8 Méthode des nuées dynamiques.

8.1 Cadre d'étude

Dans l'algorithme des centres mobiles, on constate que :

1) les éléments de I ont une représentation euclidienne ;

2) on représente une classe d'objets par un point (le barycentre des points représentatifs des objets de la classe).

Dans l'algorithme des nuées dynamiques :

1) On considère la structure générale (I, P, d) (d : dissimilarité) ;

2) une classe d'éléments de I est représentée par un *noyau*, comme par exemple :

· pour des données euclidiennes : un point, un sous-espace affine ;

· pour des données quelconques : un sous-ensemble de q éléments de la classe.

8.2 Critère à optimiser

8.2.1 Structure de représentation de classe

Définition : *C'est la donnée d'un ensemble* $\mathbb{L}$ *appelé espace de représentation, dont les éléments sont appelés noyaux, et d'une application* D' *de* $I \times \mathbb{L}$ *dans* $\mathbb{R}^+$ *appelée mesure de dissemblance d'un objet à une représentation.*

Exemples

1) Agrégation autour de centres mobiles

L'objet i est représenté par le point $A_i(\underline{x}_i) \in (\mathbb{R}^p, M)$; $\mathbb{L} = \mathbb{R}^p$.
Un noyau est un point $L(\underline{l})$.
La mesure de dissemblance D' est définie par : $D'(i, L) = p_i\, d^2(\underline{x}_i, \underline{l})$

2) Analyse factorielle typologique

L'objet i est représenté par le point $A_i(\underline{x}_i) \in (\mathbb{R}^p, M)$
Un noyau L représentant une classe est un sous-espace affine de dimension q fixée de $\mathbb{R}^p$.
Soit Π_L l'opérateur de projection orthogonale sur L. La mesure de dissemblance D' est définie par : $D'(i, L) = p_i\, d^2(\underline{x}_i, \Pi_L \underline{x}_i)$.

3) Soit la structure (I, P, d). On suppose que l'on dispose d'un tableau de dissimilarités entre les objets.

On peut définir comme noyau représentant une classe un sous-ensemble $L = (i_1, \cdots, i_q)$ de q éléments de cette classe.
On peut définir la mesure de dissemblance D' par : $D'(i, L) = p_i \sum_{j=1}^{q} p_{i_j} d(i_j, i)$.■

8.2.2 Mesure de dissemblance d'une partie de I à une représentation

Définition *C'est une application* $D : \mathcal{P}(I) \times \mathbb{L}$ dans $\mathbb{R}^+$ *telle que pour toute partie* P *de* I *et toute représentation* L, $D(P, L) = \sum_{i \in P} D'(i, L)$.

Exemple
Dans la méthode des centres mobiles :

$$D(P, L) = \sum_{i \in P} p_i\, d^2(\underline{x}_i, \underline{l}) \quad \text{(inertie des points de } P \text{ par rapport au point } L).\blacksquare$$

8.2.3 Définition du critère

Considérons un r-uplet de parties de I constituant une partition de I :

$$\mathcal{P} = (P_1, P_2, \cdots, P_r)$$

Considérons un r-uplet de représentations :

$$\mathcal{L} = (L_1, L_2, \cdots, L_r)$$

On définit :

$$W(\mathcal{P}, \mathcal{L}) = \sum_{k=1}^{r} D(P_k, L_k)$$

Remarque Soit k_i le numéro de la classe à laquelle appartient l'objet i. On a :

$$W(\mathcal{P}, \mathcal{L}) = \sum_{k=1}^{r} \sum_{i \in P_k} D'(i, L_k) = \sum_{i=1}^{n} D'(i, L_{k_i}).$$

Exemple
Dans la méthode des centres mobiles :

$$W(\mathcal{P}, \mathcal{L}) = \sum_{k=1}^{r} \sum_{i \in P_k} p_i \, d^2(\underline{x}_i, \underline{l}_k). \blacksquare$$

On pose le problème d'optimisation suivant :

Déterminer $\mathcal{P}^*$ et $\mathcal{L}^*$ tels que $W(\mathcal{P}, \mathcal{L})$ soit minimal pour $\mathcal{P} = \mathcal{P}^*$ et $\mathcal{L} = \mathcal{L}^*$.

8.3 Algorithme des nuées dynamiques

8.3.1 Fonction d'affectation

Définition *C'est une application $f : \mathcal{L} = (L_1, L_2, \cdots, L_r) \mapsto \mathcal{P} = (P_1, P_2, \cdots, P_r)$ (partition de I) telle que :*

$$P_k = \big\{ i \in I \ : \ \text{pour } j = 1, 2, \cdots, r, \quad D'(i, L_k) \leq D'(i, L_j) \big\}.$$

En cas d'égalité, on prend une convention d'affectation.

Proposition *Comme $W(\mathcal{P}, \mathcal{L}) = \sum_{i=1}^{n} D'(i, L_{k_i})$, $\mathcal{P} = f(\mathcal{L})$ rend $W(\mathcal{P}, \mathcal{L})$ minimal à $\mathcal{L}$ fixé.*

Exemple
Dans la méthode des centres mobiles :

$$D'(i, L) = p_i \, d^2(\underline{x}_i, \underline{l})$$
$$P_k = \big\{ i \in I \ : \ \text{pour } j = 1, 2, \cdots, r, \quad d(\underline{x}_i, \underline{l}_k) \leq d(\underline{x}_i, \underline{l}_j) \big\}. \blacksquare$$

8.3.2 Fonction de représentation

Définition *C'est une application $g : \mathcal{P} = (P_1, P_2, \cdots, P_r) \mapsto \mathcal{L} = (L_1, L_2, \cdots, L_r)$ telle que, pour $k = 1, 2, \cdots, r$, L_k rende minimale $D(P_k, L)$.*

Hypothèse *On suppose l'existence et l'unicité de L_k.*

Proposition Comme $W(\mathcal{P}, \mathcal{L}) = \sum_{k=1}^{r} D(P_k, L_k)$, *$\mathcal{L} = g(\mathcal{P})$ rend $W(\mathcal{P}, \mathcal{L})$ minimal à $\mathcal{P}$ fixé.*

Exemple

Dans la méthode des centres mobiles :

$D(P_k, L) = \sum_{i \in P_k} p_i \, d^2(\underline{x}_i, \underline{l})$ est minimale pour $\underline{l} = \underline{g}_k$ barycentre de la classe P_k. En effet, $D(P_k, L)$ est l'inertie $\mathcal{J}_{P_k}(\underline{l})$ par rapport à L de l'ensemble des points A_i pour $i \in P_k$ et :

$$\mathcal{J}_{P_k}(\underline{l}) = \mathcal{J}_{P_k}(\underline{g}_k) + q_k \, d^2(\underline{l}, \underline{g}_k) \text{ } (q_k \text{ poids de } P_k).\blacksquare$$

8.3.3 Algorithme

C'est un algorithme d'optimisation alternée.

Initialisation : $\mathcal{L}^0$

$\mathcal{P}^0 = f(\mathcal{L}^0)$

1er pas : $\mathcal{L}^1 = g(\mathcal{P}^0)$

$\mathcal{P}^1 = f(\mathcal{L}^1)$

.

.

$l^{ème}$ pas : $\mathcal{L}^l = g(\mathcal{P}^{l-1})$

$\mathcal{P}^l = f(\mathcal{L}^l)$

.

.

Théorème

1) On a $W(\mathcal{P}^{l-1}, \mathcal{L}^{l-1} \geq W(\mathcal{P}^{l-1}, \mathcal{L}^l) \geq W(\mathcal{P}^l, \mathcal{L}^l) \geq W(\mathcal{P}^l, \mathcal{L}^{l+1})$. *Les suites* $\big(W(\mathcal{P}^{l-1}, \mathcal{L}^l)\big)$ *et* $\big(W(\mathcal{P}^l, \mathcal{L}^l)\big)$ *convergent en décroissant vers une limite qui dépend de* $\mathcal{L}^0$.

2) La suite $\big((\mathcal{P}^l, \mathcal{L}^l)\big)$ *est stationnaire à partir d'un certain rang, la limite dépendant de* $\mathcal{L}^0$.

Démonstration

1) On a : $\begin{cases} \mathcal{L}^l = g(\mathcal{P}^{l-1}) \\ \mathcal{P}^l = f(\mathcal{L}^l) \end{cases}$

$\mathcal{L}^l$ rend $W(\mathcal{P}^{l-1}, \mathcal{L})$ minimal ; donc $W(\mathcal{P}^{l-1}, \mathcal{L}^{l-1}) \geq W(\mathcal{P}^{l-1}, \mathcal{L}^l)$.

$\mathcal{P}^l$ rend $W(\mathcal{P}, \mathcal{L}^l)$ minimal ; donc $W(\mathcal{P}^{l-1}, \mathcal{L}^l) \geq W(\mathcal{P}^l, \mathcal{L}^l)$.

On a $W(\mathcal{P}^{l-1}, \mathcal{L}^l) \geq W(\mathcal{P}^l, \mathcal{L}^l) \geq W(\mathcal{P}^l, \mathcal{L}^{l+1})$; la suite $(u_l) = \big(W(\mathcal{P}^{l-1}, \mathcal{L}^l)\big)$ est décroissante minorée. Donc elle converge.

$W(\mathcal{P}^{l-1}, \mathcal{L}^{l-1}) \geq W(\mathcal{P}^l, \mathcal{L}^l)$; la suite $v_l = \big(W(\mathcal{P}^l, \mathcal{L}^l)\big)$ est décroissante minorée. Donc elle converge.

2) La suite $\big(W(\mathcal{P}^{l-1}, \mathcal{L}^l)\big)$ converge.

$W(\mathcal{P}^{l-1}, \mathcal{L}^l) = W\big(\mathcal{P}^{l-1}, g(\mathcal{P}^{l-1})\big)$.

Or l'ensemble des partitions de I en r classes est fini.

$W\big(\mathcal{P}^{l-1}, g(\mathcal{P}^{l-1})\big)$ ne peut avoir qu'un ensemble fini de valeurs. Donc, la limite de la suite $\big(W(\mathcal{P}^{l-1}, \mathcal{L}^l)\big)$ est atteinte. Soit N le rang à partir duquel elle est atteinte :

$$W(\mathcal{P}^{N-1}, \mathcal{L}^N) = W(\mathcal{P}^N, \mathcal{L}^{N+1}) = \cdots .$$

Donc d'après 1), $W(\mathcal{P}^N, \mathcal{L}^N) = W(\mathcal{P}^N, \mathcal{L}^{N+1})$.

Or, $\mathcal{L}^{N+1}$ qui rend minimal $W(\mathcal{P}^N, \mathcal{L})$ est unique.

Par conséquent, on a : $\begin{cases} \mathcal{L}^N = \mathcal{L}^{N+1} \\ \mathcal{P}^N = f(\mathcal{L}^N) = \mathcal{P}^{N+1} \end{cases}$.$\blacksquare$

Exemple

Dans la méthode des centres mobiles :

$W(\mathcal{P}^{l-1}, \mathcal{L}^l) = \sum_{k=1}^{r} \sum_{i \in I_k^{l-1}} p_i \, d^2(\underline{x}_i, \underline{g}_k^l)$ est l'inertie intra-classes de la partition $\mathcal{P}^{l-1}$; donc, la suite des inerties intra-classes converge en décroissant vers un minimum local et la suite des partitions de I converge vers une forme stable qui dépend de $(\underline{g}_1^0, \cdots, \underline{g}_r^0)$.$\blacksquare$

Troisième partie : Modèles probabilistes en classification

9 Modèle de mélange

9.1 Cadre d'étude

$$\begin{cases} I = \{1, 2, \cdots, n\} \\ i \mapsto A_i(\underline{x}_i) \quad \in (\mathbb{R}^p, M) \\ P = \{\frac{1}{n}, \cdots, \frac{1}{n}\} \quad \text{poids uniformes} \\ d = d_M \end{cases}$$

I est un ensemble d'individus, A_i le point représentatif de l'individu i dans $\mathbb{R}^p$, P l'ensemble des poids uniformes attribués aux individus, d la distance euclidienne associée à la métrique M.

9.2 Modélisation probabiliste

On fait l'hypothése que l'ensemble des vecteurs d'observations $(\underline{x}_1, \cdots, \underline{x}_n)$ est un échantillon i.i.d. d'un vecteur aléatoire $\underline{X}$ dans $\mathbb{R}^p$ défini sur un espace probabilisé $(\Omega, \mathcal{A}, P)$ partitionné en classes $\Omega_1, \Omega_2, \cdots, \Omega_r$.

On suppose que la densité de probabilité conditionnelle de $\underline{X}$ dans la classe Ω_k appartient à une famille paramétrée de densités $\mathcal{F} = \left\{f(\cdot\ ;\ \underline{a}),\ \underline{a} \in \mathbb{R}^s\right\}$; dans le cas où $\underline{X}$ est discret, $f(\underline{x}\ ;\ \underline{a}_k) = P(\underline{X} = \underline{x}|\Omega_k)$.

Notons $P(\Omega_k) = p_k$.

La densité de $\underline{X}$ est $g(\underline{x}\ ;\ \underline{a}_1, \cdots, \underline{a}_r) = \sum_{k=1}^{r} p_k f(\underline{x}\ ;\ \underline{a}_k)$: c'est une *densité de mélange.*

10 Estimation des paramètres du modèle et classification par la méthode du maximum de vraisemblance classifiante

10.1 Critère du maximum de vraisemblance classifiante

Considérons un échantillon P_k issu de la classe Ω_k. On définit la vraisemblance de cet échantillon :

$$L(P_k\ ;\ \underline{a}_k) = \prod_{i \in P_k} f(\underline{x}_i\ ;\ \underline{a}_k)$$

On définit la *vraisemblance classifiante* :

$$L(P_1, \cdots, P_r\ ;\ \underline{a}_1, \cdots, \underline{a}_r) = \prod_{k=1}^{r} L(P_k\ ;\ \underline{a}_k) = \prod_{k=1}^{r} \prod_{i \in P_k} f(\underline{x}_i\ ;\ \underline{a}_k)$$

On pose le problème suivant :

déterminer une partition $\mathcal{P}^* = \{P_1^*, \cdots, P_r^*\}$ et un r-uplet de paramètres $\mathcal{L}^* = \{\underline{a}_1^*, \cdots, \underline{a}_r^*\}$ qui rendent la vraisemblance $L(P_1, \cdots, P_r\ ;\ \underline{a}_1, \cdots, \underline{a}_r)$ maximale.

10.2 Algorithme des nuées dynamiques

On a à maximiser

$$lnL = \sum_{k=1}^{r} \sum_{i \in P_k} ln\ f(\underline{x}_i\ ;\ \underline{a}_k)$$

On définit pour mettre en oeuvre l'algorithme les fonctions d'affectation et de représentation.

1) *Fonction d'affectation f*

Elle est définie par :

$$f : \mathcal{L} = (\underline{a}_1, \cdots, \underline{a}_r) \mapsto \mathcal{P} = (P_1, \cdots, P_r)$$
$$\text{avec } P_k = \left\{ i \in I \ : \ \text{pour } j = 1, \cdots, r, \ f(\underline{x}_i, \underline{a}_k) \geq f(\underline{x}_i \ ; \ \underline{a}_j) \right\}$$

2) *Fonction de représentation g*
Elle est définie par :

$$g : \mathcal{P} = (P_1, \cdots, P_r) \mapsto \mathcal{L} = (\underline{a}_1, \cdots, \underline{a}_r)$$
$$\text{telle que pour } k = 1, \cdots, r, \ \underline{a}_k \text{ rende maximale } \sum_{i \in P_k} ln \ f(\underline{x}_i \ ; \ \underline{a}).$$

$\underline{a}_k$ est donc l'estimation du paramètre $\underline{a}$ au sens du maximum de vraisemblance à partir de l'échantillon P_k.

3) *Algorithme*
Au pas l de l'algorithme, on détermine $\left\{ \begin{array}{l} \mathcal{L}^l = g(\mathcal{P}^{l-1}) \\ \mathcal{P}^l = f(\mathcal{L}^l) \end{array} \right.$.

10.3 Cas particulier d'un mélange de lois multinormales

Dans ce cas, la densité conditionnelle de $\underline{X}$ dans Ω_k est

$$f(\underline{x} \ ; \ \underline{a}_k) = \frac{1}{\sqrt{\det \Sigma_k}(2\pi)^{p/2}} \ e^{-\frac{1}{2} \ (\underline{x} - \underline{m}_k)' \Sigma_k^{-1} (\underline{x} - \underline{m}_k)}, \ \underline{a}_k = (\underline{m}_k, \Sigma_k)$$

On a à maximiser $\sum_{k=1}^{r} \sum_{i \in P_k} ln \ f(\underline{x}_i \ ; \ \underline{a}_k)$ (1).

$$ln \ f(\underline{x}_i \ ; \ \underline{a}_k) = -\frac{1}{2} \ ln \ \det \ \Sigma_k - \frac{p}{2} \ ln \, 2\pi - \frac{1}{2} (\underline{x}_i - \underline{m}_k)' \Sigma_k^{-1} (\underline{x}_i - \underline{m}_k)$$
$$(1) \iff \sum_{k=1}^{r} \sum_{i \in P_k} \left(ln \ \det \ \Sigma_k + (\underline{x}_i - \underline{m}_k)' \Sigma_k^{-1} (\underline{x}_i - \underline{m}_k) \right) \quad \min$$

On obtient une règle quadratique.

Plaçons-nous dans le cas particulier : $\Sigma_1 = \cdots = \Sigma_r = \Sigma$.

Le critère s'écrit alors : $\sum_{k=1}^{r} \sum_{i \in P_k} (\underline{x}_i - \underline{m}_k)' \Sigma^{-1} (\underline{x}_i - \underline{m}_k) = \sum_{k=1}^{r} \sum_{i \in P_k} ||\underline{x}_i - \underline{m}_k||^2_{\Sigma^{-1}}$ min.

On suppose la matrice de covariance Σ connue. Pour déterminer un minimum local, on utilise l'algorithme des nuées dynamiques avec les fonctions d'affectation et de représentation suivantes.

Fonction d'affectation

$$\begin{array}{rcl} & & f : \mathcal{L} = (\underline{m}_1, \cdots, \underline{m}_r) \mapsto \mathcal{P} = (P_1, \cdots, P_r) \\ P_k & = & \left\{ i \in I \ : \ \text{pour } j = 1, \cdots, r \ , \ f(x_i, \underline{m}_k) \geq f(\underline{x}_i \ ; \ \underline{m}_j) \right\} \\ & = & \{ i \in I \ : \ \text{pour } j = 1, \cdots, r, ||\underline{x}_i - \underline{m}_k||_{\Sigma^{-1}} \leq ||\underline{x}_i - \underline{m}_j||_{\Sigma^{-1}} \}. \end{array}$$

Fonction de représentation

$$g : \mathcal{P} = (P_1, \cdots, P_r) \mapsto \mathcal{L} = (\underline{m}_1, \cdots, \underline{m}_r)$$

Pour $k = 1, \cdots, r$, $\underline{m}_k$ rend minimale $\sum_{i \in P_k} ||\underline{x}_i - \underline{m}||^2_{\Sigma^{-1}}$; $\underline{m}_k$ est le centre de gravité des points $\left(\underline{x}_i, \frac{1}{n} \right)$ pour $i \in P_k$.

Algorithme
Au pas l de l'algorithme, on détermine : $\left\{ \begin{array}{l} \mathcal{L}^l = g(\mathcal{P}^{l-1}) \\ \mathcal{P}^l = f(\mathcal{L}^l) \end{array} \right.$.

On retrouve l'algorithme des centres mobiles.

11 Estimation des paramètres du modèle par la méthode du maximum de vraisemblance et classification

Ω est partitionné en classes $\Omega_1, \cdots, \Omega_r$. D'après le théorème de Bayes, on a :

$$\mathbb{P}(\Omega_k | \underline{X} = \underline{x}_i) = \frac{P(\Omega_k) f(\underline{x}_i \ ; \ \underline{a}_k)}{\sum_{j=1}^{r} P(\Omega_j) f(\underline{x}_i \ ; \ \underline{a}_j)}.$$

11.1 Equations de la vraisemblance

La densité de $\underline{X}$ est $g(\underline{x} \ ; \ \underline{a}_1, \cdots, \underline{a}_r) = \sum_{k=1}^{r} p_k f(\underline{x} \ ; \ \underline{a}_k)$.

Soit $(\underline{x}_1, \cdots, \underline{x}_n)$ un échantillon i.i.d. de $\underline{X}$. Sa vraisemblance est :

$$L(\underline{x}_1, \cdots, \underline{x}_n \ ; \ p_1, \cdots, p_r, \ \underline{a}_1, \cdots, \underline{a}_r) = \prod_{i=1}^{n} \left(\sum_{k=1}^{r} p_k \ f(\underline{x}_i \ ; \ \underline{a}_k) \right)$$

$$lnL = \sum_{i=1}^{n} ln \left(\sum_{k=1}^{r} p_k \, f(\underline{x}_i ; \ \underline{a}_k) \right)$$

On recherche p_k^0 et $\underline{a}_k^0$, $k = 1, \cdots, r$ qui rendent maximale lnL. On note a_{kj} la $j^{ième}$ composante de $\underline{a}_k$, $j = 1, ..., s$.

Proposition *Le système des équations de la vraisemblance s'écrit :*

$$\begin{cases} p_k = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} P(\Omega_k | \underline{X} = \underline{x}_i), \quad k = 1, \cdots, r \\ \displaystyle\sum_{i=1}^{n} P(\Omega_k | \underline{X} = \underline{x}_i) \ \frac{\partial ln \, f(\underline{x}_i ; \underline{a}_k)}{\partial a_{kj}} = 0, \quad k = 1, ..., r, \quad j = 1, ..., s. \end{cases}$$

Démonstration
On a à maximiser

$$\sum_{i=1}^{n} ln \left(\sum_{k=1}^{r} p_k \ f(\underline{x}_i \ ; \ \underline{a}_k) \right) \quad \text{sous la contrainte} \ \sum_{k=1}^{r} p_k = 1.$$

On utilise pour cela la méthode des multiplicateurs de Lagrange. Soit la fonction

$$\Phi(p_1, \cdots, p_r \ , \ \underline{a}_1, \cdots, \underline{a}_r \ ; \ \lambda) = \sum_{i=1}^{n} ln \left(\sum_{k=1}^{r} p_k \ f(\underline{x}_i \ ; \ \underline{a}_k) \right) - \lambda \left(\sum_{k=1}^{r} p_k - 1 \right).$$

$$\frac{\partial \Phi}{\partial p_k} = \sum_{i=1}^{n} \frac{f(\underline{x}_i \ ; \ \underline{a}_k)}{\sum_{l=1}^{r} p_l \ f(\underline{x}_i \ ; \ \underline{a}_l)} - \lambda = 0, \quad k = 1, \cdots, r.$$

$$\sum_{k=1}^{r} p_k \ \frac{\partial \Phi}{\partial p_k} = \sum_{i=1}^{n} \frac{\sum_{k=1}^{r} p_k \ f(\underline{x}_i \ ; \ \underline{a}_k)}{\sum_{l=1}^{r} p_l \ f(\underline{x}_i \ ; \ \underline{a}_l)} - \lambda \sum_{k=1}^{r} p_k = n - \lambda = 0 \ ; \ \text{donc}, \ \lambda = n.$$

$$\frac{\partial \Phi}{\partial p_k} = 0 \iff \frac{1}{n}\sum_{i=1}^{n} \frac{f(\underline{x}_i;\underline{a}_k)}{\sum_{l=1}^{r} p_l\, f(\underline{x}_i;\underline{a}_l)} = 1$$

$$\iff \frac{1}{n}\sum_{i=1}^{n} \frac{p_k\ f(\underline{x}_i;\underline{a}_k)}{\sum_{l=1}^{r} p_l f(\underline{x}_i;\underline{a}_l)} = p_k$$

$$\iff \frac{1}{n}\sum_{i=1}^{n} P(\Omega_k|X=\underline{x}_i) = p_k.$$

$$\frac{\partial \Phi}{\partial a_{kj}} = \sum_{i=1}^{n} \frac{p_k \frac{\partial}{\partial a_{kj}} f(\underline{x}_i\ ;\ \underline{a}_k)}{\sum_{l=1}^{r} p_l\, f(\underline{x}_i\ ;\ \underline{a}_l)} = \sum_{i=1}^{n} \frac{p_k\, f(\underline{x}_i\ ;\ \underline{a}_k)\ \frac{\frac{\partial}{\partial a_{kj}} f(\underline{x}_i\ ;\ \underline{a}_k)}{f(\underline{x}_i\ ;\ \underline{a}_k)}}{\sum_{l=1}^{r} p_l\, f(\underline{x}_i\ ;\ \underline{a}_l)} =$$

$$= \sum_{i=1}^{n} P(\Omega_k|X=x_i)\, \frac{\partial}{\partial a_{kj}} ln\, f(\underline{x}_i\ ;\ \underline{a}_k) = 0.\blacksquare$$

11.2 Algorithme EM

On utilise une méthode itérative de résolution du système des équations de la vraisemblance.

Au pas m, on obtient une estimation p_k^m de p_k et $\underline{a}_k^m$ de $\underline{a}_k,\ k = 1, \cdots, r$. Au pas $m+1$, on effectue deux étapes :

1) *Etape E (estimation)*

On estime $P(\Omega_k|\underline{X} = \underline{x}_i) = \dfrac{P(\Omega_k) f(\underline{x}_i\ ;\ \underline{a}_k)}{\sum_{l=1}^{r} P(\Omega_l) f(\underline{x}_i\ ;\ \underline{a}_l)}$ par

$$t_k^m(\underline{x}_i) = \frac{p_k^m\, f(\underline{x}_i\ ;\ \underline{a}_k^m)}{\sum_{l=1}^{r} p_l^m\, f(\underline{x}_i\ ;\ \underline{a}_l^m)}.$$

On estime p_k par $p_k^{m+1} = \dfrac{1}{n}\sum_{i=1}^{n} t_k^m(\underline{x}_i), \quad k = 1, \cdots, r.$

2) *Etape M (maximisation)*

On résout, pour $k = 1, \cdots, r$, le système d'équations :

$$\sum_{i=1}^{n} t_k^m(\underline{x}_i)\, \frac{\partial ln\, f(\underline{x}_i\ ;\ \underline{a}_k)}{\partial a_{kj}} = 0, \quad j = 1, \cdots, s.$$

On obtient alors $\underline{a}_k^{m+1}$.

Cas particulier : mélange de lois multinormales

On a dans ce cas $\underline{a}_k = (\underline{\mu}_k, \Sigma_k)$.

$$\underline{\mu}_k^{m+1} = \frac{\sum_{i=1}^{n} t_k^m(\underline{x}_i)\underline{x}_i}{\sum_{i=1}^{n} t_k^m(\underline{x}_i)}, \Sigma_k^{m+1} = \frac{\sum_{i=1}^{n} t_k^m(\underline{x}_i)(\underline{x}_i - \underline{\mu}_k^{m+1})(\underline{x}_i - \underline{\mu}_k^{m+1})'}{\sum_{i=1}^{n} t_k^m(\underline{x}_i)}.$$

11.3 Convergence

Cet algorithme converge vers un maximum local de la vraisemblance sous certaines conditions. En pratique, on le lance plusieurs fois à partir d'initialisations différentes et on retient la solution qui donne la vraisemblance la plus grande.

11.4 Classification

On affecte l'individu i à la classe Ω_k telle que l'estimation de la probabilité conditionnelle $P(\Omega_k|\underline{X} = \underline{x}_i)$ soit maximale.

12 Algorithme des k-means séquentiel de MacQueen

12.1 Critère de classification

On note I_k la variable aléatoire indicatrice de la classe Ω_k : $I_k(\omega) = \begin{cases} 1 \text{ si } \omega \in \Omega_k \\ 0 \text{ sinon} \end{cases}$.

On définit le critère de partitionnement de Ω suivant :

déterminer une partition $(\Omega_1, \cdots, \Omega_r)$ *de* Ω *et un* r*-uplet de points* $\left(\underline{g}_1, \cdots, \underline{g}_r\right)$ *tels que* $E\left[\sum_{k=1}^{r} I_k ||\underline{X} - \underline{g}_k||^2\right]$ *soit minimale.*

Si $\underline{g}_1, \cdots, \underline{g}_r$ sont déterminés, alors, pour $k = 1, ..., r$:

$$\Omega_k = \left\{\omega \in \Omega : \left\|\underline{X}(\omega) - \underline{g}_k\right\| \leq \left\|\underline{X}(\omega) - \underline{g}_j\right\|,\ j = 1, ..., r\right\} ;$$

donc : $E\left[\sum_{k=1}^{r} I_k ||\underline{X} - \underline{g}_k||^2\right] = E\left[\min_k ||\underline{X} - \underline{g}_k||^2\right]$.

On cherche alors $\underline{g}_1, \underline{g}_2, \cdots, \underline{g}_r$ qui rendent $E\left[\min_k ||\underline{X} - \underline{g}_k||^2\right]$ minimale.

Pour estimer une solution locale de ce problème, on peut utiliser un algorithme séquentiel d'approximation stochastique : au pas l, on introduit une observation $\underline{x}_l$ de $\underline{X}$ et on actualise l'estimation $\underline{g}_k^l$ de $\underline{g}_k$, $k = 1, \cdots, r$.

Remarque Un algorithme d'approximation stochastique est récursif. Un exemple élémentaire est le suivant. Soit un échantillon i.i.d. $(x_1, \cdots, x_n, \cdots)$ d'une variable aléatoire réelle X. On estime l'espérance $E[X]$ à partir des n premières observations par

$$\overline{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

et à partir des $n + 1$ premières observations par

$$\begin{aligned} \overline{x}_{n+1} &= \frac{x_1 + \cdots + x_n + x_{n+1}}{n+1} = \frac{n}{n+1}\,\overline{x}_n + \frac{1}{n+1}\,x_{n+1} \\ &= \overline{x}_n - \frac{1}{n+1}(\overline{x}_n - x_{n+1}). \end{aligned}$$

On a ainsi défini un algorithme récursif d'approximation stochastique de l'espérance de X.■

12.2 Algorithme

Initialisation :

On choisit r points $\underline{g}_1^0, \cdots, \underline{g}_r^0$ de $\mathbb{R}^p$.

1^{er} pas :

On introduit $\underline{x}_1$.

On calcule $d(\underline{x}_1, \underline{g}_k^0),\ k = 1, \cdots, r$.

Soit k_0 la valeur de k pour laquelle cette distance est minimale. On définit :

$$\begin{aligned} w_{k_{\grave{a}}}^l &= 2, \underline{g}_{k_0}^1 = \underline{g}_{k_0}^0 - \frac{1}{w_{k_0}^1}(\underline{g}_{k_0}^0 - \underline{x}_1), \\ &\left(\text{on a : } \underline{g}_{k_0}^1 = \frac{\underline{g}_{k_0}^0 + \underline{x}_1}{2}\right) \\ w_k^1 &= 1, \underline{g}_k^1 = \underline{g}_k^0, \text{ pour } k \neq k_0. \end{aligned}$$

$l^{ième}$ pas :

On introduit $\underline{x}_l$.

On calcule $d(\underline{x}_l, \underline{g}_k^{l-1}), \quad k = 1, \cdots, r$.

Soit k_0 la valeur de k pour laquelle cette distance est minimale. On définit :

$$\begin{aligned} w_{k_0}^l &= w_{k_0}^{l-1} + 1, \underline{g}_{k_0}^l = \underline{g}_{k_0}^{l-1} - \frac{1}{w_{k_0}^l}(\underline{g}_{k_0}^{l-1} - \underline{x}_l), \\ w_k^l &= w_k^{l-1}, \underline{g}_k^l = \underline{g}_k^{l-1}, \quad \text{pour } k \neq k_0. \end{aligned}$$

Remarques

1) $\underline{g}_k^l$ n'est autre que le barycentre des points affectés à la $k^{ième}$ classe jusqu'au $l^{ième}$ pas.

2) A chaque pas, on affecte une seule observation à une classe. Dans la méthode des centres mobiles (ou algorithme des k-means non séquentiel), on affecte à chaque pas chacune des observations à une classe.

12.3 Convergence

On démontre que, sous certaines hypothèses, $E\big[\min_k ||\underline{X} - \underline{g}_k^l||^2\big]$ converge vers un minimum local de $E\big[\min_k ||\underline{X} - \underline{g}_k||^2\big]$.

Une fois la convergence obtenue, on dispose d'estimations des points $\underline{g}_1, \underline{g}_2, \cdots, \underline{g}_r$. On constitue r classes en réaffectant chaque point $\underline{x}_n$ à la classe correspondant au point $\underline{g}_k$ de l'estimation duquel il est le plus proche.