Cours 4 : Classification non-supervisée

Essentiel du cours 3

Définitions

- Covariance :

$$\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

- Coefficient de corrélation :

$$\mathrm{cor}(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\mathrm{var}(\mathbf{x})\,\mathrm{var}(\mathbf{y})}}.$$

- Critère des moindres carrés :

$$C(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- Critère du R^2 : $\widetilde{y}_i = a + bx_i$

$$R^2(a, b) = \left(\mathrm{cor}(\widetilde{\mathbf{y}}, \mathbf{y})\right)^2.$$

Propositions

- Bilinéarité : $u_i = a + bx_i$ et $v_i = c + dy_i$ pour $1 \leqslant i \leqslant n$

$$\mathrm{cov}(\mathbf{u}, \mathbf{v}) = bd\,\mathrm{cov}(\mathbf{x}, \mathbf{y}).$$

- Inégalité de Cauchy-Schwartz :

$$\langle \mathbf{u}, \mathbf{v} \rangle^2 \leqslant \|\mathbf{u}\|^2 \|\mathbf{v}\|^2.$$

- Encadrement :

$$\left(\mathrm{cov}(\mathbf{x}, \mathbf{y})\right)^2 \leqslant \mathrm{var}(\mathbf{x})\,\mathrm{var}(\mathbf{y})$$

- Minimisation de $C(a, b)$:

$$b^* = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{\mathrm{var}(\mathbf{x})}, \qquad a^* = \overline{y} - b^* \overline{x}.$$

- R^2 des moindres carrés :

$$R^2(a^*, b^*) = \left(\mathrm{cor}(\mathbf{x}, \mathbf{y})\right)^2.$$

Exemple

Budget de l'état français

- $n = 24$ années, de 1872 à 1971
- x_i = budget répartition (en %) du budget de l'état lors de la i-ème année, réparti en $p = 11$ postes :

PVP =	pouvoirs publics	AGR =	agriculture
CMI =	commerce et industrie	TRA =	travail
LOG =	logement	EDU =	éducation
ACS =	action sociale	ANC =	anciens combattants
DEF =	défense	DET =	remboursement de la dette
DIV =	divers		

Année	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
1872	18.0	0.5	0.1	6.7	0.5	2.1	2.0	0.0	26.4	41.5	2.1
1880	14.1	0.8	0.1	15.3	1.9	3.7	0.5	0.0	29.8	31.3	2.5
1890	13.6	0.7	0.7	6.8	0.6	7.1	0.7	0.0	33.8	34.4	1.7
1900	14.3	1.7	1.7	6.9	1.2	7.4	0.8	0.0	37.7	26.2	2.2
1903	10.3	1.5	0.4	9.3	0.6	8.5	0.9	0.0	38.4	27.2	3.0
1906	13.4	1.4	0.5	8.1	0.7	8.6	1.8	0.0	38.5	25.3	1.9

Peut-on définir une typologie des années (*périodes historiques*) en fonction de la politique budgétaire ?

On mesure p variables ($1 \leqslant j \leqslant p$) continues sur n *individus* ($1 \leqslant i \leqslant n$).
On note :

- x_{ij} la valeur de la j-ème variable pour le i-ème individu ;
- $\mathbf{x}_i$ le vecteur des mesures faites sur le i-ème individu :

$$\mathbf{x}_i^T = [x_{i1} \ \dots \ x_{ij} \ \cdots \ x_{ip}] ;$$

- $\mathbf{X}$ la matrice à n lignes et p colonnes contenant l'ensemble des mesures :

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_n \end{bmatrix} .$$

Représentation géométrique. Observation i = point de $\mathbb{R}^p$ de coordonnées $\mathbf{x}_i$

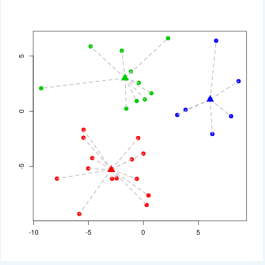
Définition d'une typologie

Objectif : Partant des observations faites sur les n individus contenues dans $\mathbf{X}$, on veut définir k groupes ($1 \leqslant g \leqslant k$) d'individus qui soient à la fois :

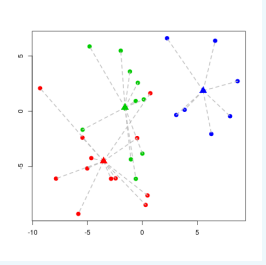
- bien homogènes et
- bien distincts les uns des autres.

Données fictives

Groupes bien homogènes



Groupes peu homogènes



Il s'agit donc de définir une *partition* de l'ensemble $\{1, 2, \ldots n\}$ des individus en k parties.

Définition 1 (Parties d'un ensemble)

On note $\mathcal{P}(\mathcal{E})$ l'ensemble des parties de l'ensemble $\mathcal{E}$:

$$\mathcal{P}(\mathcal{E}) = \{\mathcal{A} : \mathcal{A} \subset \mathcal{E}\}.$$

Exemple

$$\mathcal{E} = \{a, b, c\},$$

$$\mathcal{P}(\mathcal{E}) = \{\varnothing, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.$$

Proposition 1 (Nombre de parties)

Le nombre de parties d'un ensemble $\mathcal{E}$ comprenant n éléments est $|\mathcal{P}(\mathcal{E})| = 2^n$.

Nombre de parties

Partition (1/2)

Rappel. L'objectif est de répartir les n individus en k groupes (ou parties de $\{1, 2, \dots n\}$) de telle façon que chaque individu appartiennent à un groupe et un seul.

Définition 2 (Partition)

Soit un ensemble $\mathcal{C}$ de k éléments de $\mathcal{P}(\mathcal{E})$: $\mathcal{C} = \{C_1, C_2, \dots C_k\}$. $\mathcal{C}$ est une partition de l'ensemble $\mathcal{E}$ (supposé non vide) en k parties si et seulement si

1. toutes les parties qu'elle contient sont non vides :

$$\forall 1 \leqslant g \leqslant k : \quad C_g \neq \varnothing,$$

2. les parties qu'elle contient sont toutes disjointes deux à deux :

$$\forall 1 \leqslant g \neq \ell \leqslant k : \quad C_g \cap C_\ell = \varnothing,$$

3. la réunion des parties qui la composent donne $\mathcal{E}$ tout entier :

$$\mathcal{E} = \bigcup_{g=1}^{k} C_g \qquad = (C_1 \cup C_2 \cup \dots \cup C_k)\,.$$

Exemple

En prenant $C_1 = \{b\}$ et $C_2 = \{a, c\}$,

$$\mathcal{C} = \{C_1, C_2\} \qquad \text{est une partition de} \qquad \mathcal{E} = \{a, b, c\}.$$

Remarques.

1. La définition implique que chaque élément de $\mathcal{E}$ appartient à une partie et une seule. *(La réunion des parties contient $\mathcal{E}$ et leurs intersections deux à deux sont vides)*
2. Une partition est un ensemble de parties, au sein duquel celles-ci ne sont pas ordonnées :

$$\mathcal{C} = \{C_1, C_2, C_3\} = \{C_3, C_1, C_2\} = \{C_2, C_3, C_2\} = \dots$$

On peut permuter les indices g des parties sans changer la partition : ces indices n'ont pas de sens en eux-mêmes.

Proposition 2 (Nombre de partitions)

Soit $S(n,k)$, le nombre partitions d'un ensemble à n éléments en k parties. $S(n,k)$ vérifie :

1. $n \geqslant 1$: $S(n,0) = 0$;
2. $n \geqslant 1$: $S(n,1) = 1$;
3. $n \geqslant 1$ *et* $k > n$: $S(n,k) = 0$;
4. $n \geqslant 2$ *et* $1 \leqslant k \leqslant n$, $S(n,k) = S(n-1,k-1) + kS(n-1,k)$.

Remarques.

- $S(n,k)$ est le *nombre de Stirling (de deuxième espèce)*.
- Une forme *explicite* de $S(n,k)$ sera démontrée en TD.

Quelques valeurs du nombre de Stirling

n	k	$S(n,k)$
10	3	9330
20	4	$4.52\ 10^{10}$
50	5	$7.40\ 10^{32}$
100	10	$2.76\ 10^{93}$

Remarques.

- Pour mémoire : nombre d'atomes dans l'univers $\simeq 10^{80}$.
- Exploration systématique des toutes les partitions impossible dès que $n >$ quelques dizaines.

Nombre de partitions (2/2)

Définition 3 (Point central de l'ensemble des données)

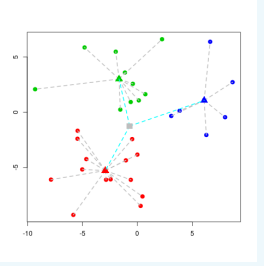
$$\bar{\bar{\mathbf{x}}} = \left[\bar{\bar{x}}_1 \ \ldots \bar{\bar{x}}_j \ \ldots \ \bar{\bar{x}}_p\right]^T, \qquad \bar{\bar{x}}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}.$$

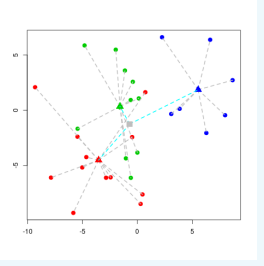
Définition 4 (Point central d'un groupe)

Soit $C_g \subset \{1, \ldots, n\}$ un groupe d'observations de cardinal $|C_g| = n_g$, on note $\bar{\mathbf{x}}_g$ son point central :

$$\bar{\mathbf{x}}_g = \left[\bar{x}_{g1} \ \ldots \bar{x}_j \ \ldots \ \bar{x}_{gp}\right]^T, \qquad \bar{x}_{gj} = \frac{1}{n_g}\sum_{i \in C_g} x_{ij}.$$

Données fictives





Définition 5 (Norme d'un vecteur)

La norme d'un vecteur $\mathbf{x}$, notée $\|\mathbf{x}\|$ est la racine carrée de son produit scalaire avec lui-même :

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{\sum_{j=1}^{p} x_i^2}.$$

Définition 6 (Distance)

La distance entre les points de coordonnées $\mathbf{x}$ et $\mathbf{y}$ est la norme du vecteur $\|\mathbf{x} - \mathbf{y}\|$

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}.$$

Notamment : $d(\mathbf{x}, \mathbf{y})^2 = \|\mathbf{x} - \mathbf{y}\|^2 = \sum_{j=1}^{p} (x_j - y_j)^2$.

$\rightarrow$ chaque norme au carré $\|\cdot\|^2$ est une somme sur les coordonnées $j = 1, \ldots, p$

Dispersion inter-groupes et intra-groupes (1/2)

Définition 7 (Dispersion inter-groupes)

Dispersion des points centraux autour du point central de l'ensemble des données :

$$\sum_{g=1}^{k} n_g \|\overline{\mathbf{x}}_g - \overline{\overline{\mathbf{x}}}\|^2.$$

Définition 8 (Dispersion intra-groupes)

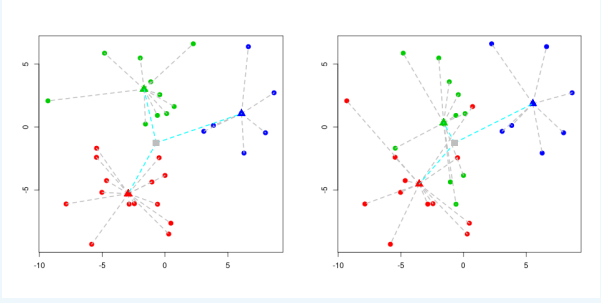
Dispersion des éléments qui composent chaque groupe autour de son point central :

$$\sum_{g=1}^{k} \sum_{i \in C_g} \|\mathbf{x}_i - \overline{\mathbf{x}}_g\|^2$$

Dispersion inter-groupes et intra-groupes (2/2)

Données fictives

$$\text{Dispersion inter} = \sum_{g=1}^{k} n_g \|\bar{\mathbf{x}}_g - \bar{\bar{\mathbf{x}}}\|^2, \qquad \text{dispersion intra} = \sum_{g=1}^{k} \sum_{i \in C_g} \|\mathbf{x}_i - \bar{\mathbf{x}}_g\|^2,$$



	Dispersion inter	Dispersion intra
Bon cas	795	379
Mauvais cas	601	573

Critères d'optimalité. On cherche une partition $\{C_1, \ldots, C_k\}$ présentant

- une dispersion *inter-groupes forte* et
- une dispersion *intra-groupes faible*.

Proposition 3 (Décomposition de la dispersion)

La dispersion totale du jeu de données autour de son point central se décompose en

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \overline{\overline{\mathbf{x}}}\|^2 = \sum_{g=1}^{k} n_g \|\overline{\mathbf{x}}_g - \overline{\overline{\mathbf{x}}}\|^2 + \sum_{g=1}^{k} \sum_{i \in C_g} \|\mathbf{x}_i - \overline{\mathbf{x}}_g\|^2$$

Conséquence. Puisque la dispersion totale $\sum_{i=1}^{n} \|\mathbf{x}_i - \overline{\overline{\mathbf{x}}}\|^2$ ne dépend pas de la partition $\{C_1, \ldots, C_k\}$,

$$\text{maximiser} \sum_{g=1}^{k} n_g \|\overline{\mathbf{x}}_g - \overline{\overline{\mathbf{x}}}\|^2 \quad \Leftrightarrow \quad \text{minimiser} \sum_{g=1}^{k} \sum_{i \in C_g} \|\mathbf{x}_i - \overline{\mathbf{x}}_g\|^2$$

Critère(s) d'optimalité

Le critère d'optimalité est fondé sur des distances $\|\mathbf{x} - \mathbf{y}\|$, $\|\overline{\mathbf{x}}_g - \overline{\overline{\mathbf{x}}}\|$, $\|\mathbf{x}_i - \overline{\mathbf{x}}_g\|$.
Hypothèse implicite : Les variables associées à chacune des coordonnées $x_1, \ldots, x_j, \ldots, x_p$ sont comparables, c'est-à-dire :

- elles sont toutes exprimées dans la même unités ou
- elles varient dans les mêmes ordres de grandeurs.

Définition 9 (Données centrées et réduites)

Pour chaque observation $1 \leqslant i \leqslant n$ et chaque variable $1 \leqslant j \leq n$, on définit la valeur *centrée-réduite*

$$\widetilde{x}_{ij} = \frac{x_{ij} - \overline{\overline{x}}_j}{\sqrt{\mathsf{var}(x_j)}} \qquad \Leftrightarrow \qquad x_{ij} = \overline{\overline{x}}_j + \widetilde{x}_{ij} \sqrt{\mathsf{var}(x_j)}.$$

Par construction, les variables $\widetilde{x}_1, \ldots \widetilde{x}_j, \ldots \widetilde{x}_p$:

- sont toutes sans dimension et
- varient dans les mêmes ordres de grandeurs.

Données centrées et réduites : exemple

Budget de l'état

Données originales (%) :

Année	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
1872	18.0	0.5	0.1	6.7	0.5	2.1	2.0	0.0	26.4	41.5	2.1
1880	14.1	0.8	0.1	15.3	1.9	3.7	0.5	0.0	29.8	31.3	2.5
1890	13.6	0.7	0.7	6.8	0.6	7.1	0.7	0.0	33.8	34.4	1.7
1900	14.3	1.7	1.7	6.9	1.2	7.4	0.8	0.0	37.7	26.2	2.2
1903	10.3	1.5	0.4	9.3	0.6	8.5	0.9	0.0	38.4	27.2	3.0
1906	13.4	1.4	0.5	8.1	0.7	8.6	1.8	0.0	38.5	25.3	1.9

Moyennes, écarts-types (%) :

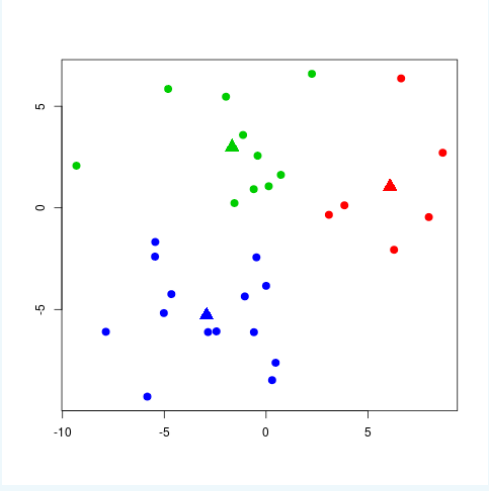
	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
moyenne	12.2	2.0	3.9	8.3	4.0	9.9	4.8	4.3	30.3	19.1	1.2
écart type	2.2	1.6	4.5	2.5	4.2	5.2	3.4	4.2	7.3	12.2	1.0

Données centrées réduites :

	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
1872	2.64	-0.91	-0.86	-0.66	-0.83	-1.50	-0.83	-1.03	-0.53	1.83	0.89
1880	0.86	-0.73	-0.86	2.83	-0.49	-1.19	-1.27	-1.03	-0.06	1.00	1.28
1890	0.63	-0.79	-0.72	-0.62	-0.80	-0.54	-1.21	-1.03	0.48	1.25	0.50
1900	0.95	-0.18	-0.50	-0.58	-0.66	-0.49	-1.18	-1.03	1.02	0.58	0.99
1903	-0.87	-0.30	-0.79	0.40	-0.80	-0.28	-1.15	-1.03	1.11	0.66	1.77
1906	0.54	-0.36	-0.77	-0.09	-0.78	-0.26	-0.88	-1.03	1.13	0.51	0.70

Données fictives

3e classification des observations autour des points centraux $\mathcal{C}^{(2)} \rightarrow \mathcal{C}^{(3)}$



aucune observation ne change de groupe : l'algorithme s'arrête

L'algorithme des *k-means* détermine alternativement

- une partition $\mathcal{C} = \{C_1, \ldots, C_k\}$ et
- k points centraux $\mathbf{m}_1, \ldots, \mathbf{m}_k$ qu'on réunira dans la matrice $(k \times p)$ $\mathbf{M}$

$$\mathbf{M} = \begin{bmatrix} \mathbf{m}_1^T \\ \vdots \\ \mathbf{m}_k^T \end{bmatrix} = \begin{bmatrix} m_{11} & \cdots & m_{1j} & \cdots & m_{1p} \\ \vdots & & \vdots & & \vdots \\ m_{k1} & \cdots & m_{kj} & \cdots & m_{kp} \end{bmatrix}$$

Notation pour les itérations. $\mathcal{C}$ et $\mathbf{M}$ sont modifiées itérativement : on note $\mathcal{C}^{(h)}$ et $\mathbf{M}^{(h)}$ leur valeur à la h-ème itération.
De même, $c_i^{(h)}$ = numéro du groupe auquel l'observation i est affectée à la h-ème itération :

$$C_g^{(h)} = \{i : c_i^{(h)} = g\}.$$

Définition 10 (Algorithme des *k-means*)

Initialisation : Choisir k points centraux $\mathbf{m}_g^{(0)}$ $(1 \leqslant g \leqslant k)$ (réunis dans $\mathbf{M}^{(0)}$).

Itération $h \geqslant 1$:

Classification : chaque observation i est affectée au groupe g dont le centre $\mathbf{m}_g^{(h)}$ lui est le plus proche :

$$c_i^{h+1} = g \qquad \Leftrightarrow \qquad \|\mathbf{x}_i - \overline{\mathbf{m}}_g^{(h)}\|^2 = \min_{1 \leqslant \ell \leqslant k} \|\mathbf{x}_i - \overline{\mathbf{m}}_\ell^{(h)}\|.$$

Mise à jour : le point central $\overline{\mathbf{m}}_g^{(h)}$ de chaque groupe $C_g^{(h+1)}$ nouvellement formé est remplacé par le point moyen du groupe

$$m_{gj}^{(h+1)} = \frac{1}{n_k^{(h)}} \sum_{i \in C_g^{(h+1)}} x_{ij}.$$

Arrêt : Si aucun point ne change de groupe lors de l'étape de classification.

Définition 11 (Dispersion intra-groupes)

On note $D(\mathcal{C}, \mathbf{M})$ la dispersion de la partition $\mathcal{C} = \{C_1, \dots C_k\}$ autour des points $\mathbf{M} = [\mathbf{m}_1^T \dots \mathbf{m}_k^T]^T$:

$$D(\mathcal{C}, \mathbf{M}) = \sum_{g=1}^{k} \sum_{i \in C_g} \|\mathbf{x}_i - \mathbf{m}_g\|^2.$$

Proposition 4 (Convergence de l'algorithme des *k-means*)

L'algorithme des k-means converge vers un minimum (local) du critère de dispersion $D(\mathcal{C}, \mathbf{M})$ en un nombre fini d'étapes.

k-means = *heuristique* **:** convergence garantie seulement vers un minimum local.

Convergence vers un minimum local : démonstration

Nous devons montrer que :

1. à chaque itération h, $D(\mathcal{C}^{(h+1)}, \mathbf{M}^{(h+1)}) \leqslant D(\mathcal{C}^{(h)}, \mathbf{M}^{(h)})$;
2. l'algorithme s'interrompt au bout d'un nombre fini d'étapes.

$$D(\mathcal{C}, \mathbf{M}) = \sum_{g=1}^{k} \sum_{i \in C_g} \|\mathbf{x}_i - \mathbf{m}_g\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}_{c_i}\|^2.$$

1. Décroissance :
   - classification
   $$\|\mathbf{x}_i - \mathbf{m}^{(h)}_{c_i^{(h+1)}}\|^2 \leqslant \|\mathbf{x}_i - \mathbf{m}^{(h)}_{c_i^{(h)}}\|$$
   donc sauf arrêt :
   $$D(\mathcal{C}^{(h+1)}, \mathbf{M}^{(h)}) < D(\mathcal{C}^{(h)}, \mathbf{M}^{(h)})$$
   - mise à jour
   $$\sum_{i \in C_g^{(h+1)}} \|\mathbf{x}_i - \mathbf{m}_g^{(h+1)}\|^2 \leqslant \sum_{i \in C_g^{(h+1)}} \|\mathbf{x}_i - \mathbf{m}_g^{(h)}\|^2 \text{ (Huygens)}$$
   donc :
   $$D(\mathcal{C}^{(h+1)}, \mathbf{M}^{(h+1)}) \leqslant D(\mathcal{C}^{(h+1)}, \mathbf{M}^{(h)}).$$
2. Finitude : Au plus $S(n, k)$ valeurs possibles pour $D(\mathcal{C}^{(h)}, \mathbf{M}^{(h)})$

Exemple (1/2)

Budget de l'état : $k = 4$ groupes

Convergence en 3 itérations :

Disp. totale	Disp. inter	Disp. intra
264	176.4	87.6

Points centraux (données centrées-réduites $\widetilde{m}_{gj}$) :

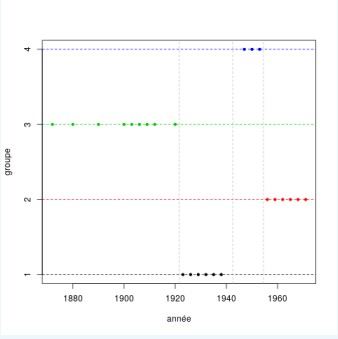
	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
C_1	-1.19	-0.54	-0.61	-0.59	-0.58	-0.24	-0.16	1.48	-0.17	0.59	-0.24
C_2	0.01	1.57	0.96	-0.44	0.70	1.35	1.46	0.07	-0.99	-1.09	-0.54
C_3	0.63	-0.58	-0.77	0.38	-0.70	-0.64	-0.89	-0.98	0.80	0.73	0.74
C_4	0.47	-0.32	1.62	0.91	1.88	-0.31	0.05	-0.17	-0.08	-1.19	-0.67

Points centraux (données originales $m_{gj} = \overline{\overline{x}}_j + \sqrt{\mathsf{var}(x_j)}\ \widetilde{m}_{gj}$) :

	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
C_1	9.6	1.1	1.2	6.9	1.5	8.7	4.3	10.4	29.1	26.4	0.9
C_2	12.2	4.6	8.2	7.2	6.9	17.0	9.8	4.6	23.0	5.9	0.6
C_3	13.6	1.0	0.5	9.3	1.0	6.6	1.8	0.2	36.1	28.0	1.9
C_4	13.2	1.5	11.2	10.6	11.8	8.3	5.0	3.6	29.7	4.7	0.5

Exemple (2/2)

Budget de l'état : $k = 4$ groupes



$C_1 = \{1923, 1926, 1929, 1932, 1935, 1938\}$

$C_2 = \{1956, 1959, 1962, 1965, 1968, 1971\}$

$C_3 = \{1872, 1880, 1890, 1900, 1903, 1906, 1909, 1912, 1920\}$

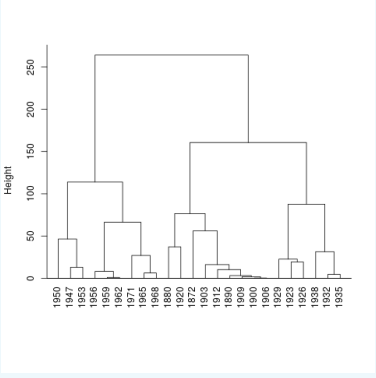
$C_4 = \{1947, 1950, 1953\}$

Points centraux (données originales) :

	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
C_1	9.6	1.1	1.2	6.9	1.5	8.7	4.3	10.4	29.1	26.4	0.9
C_2	12.2	4.6	8.2	7.2	6.9	17.0	9.8	4.6	23.0	5.9	0.6
C_3	13.6	1.0	0.5	9.3	1.0	6.6	1.8	0.2	36.1	28.0	1.9
C_4	13.2	1.5	11.2	10.6	11.8	8.3	5.0	3.6	29.7	4.7	0.5

Autre heuristique : agréger itérativement les observations les plus proches en constituant des groupes faits de paires, triplets, quadruplets, etc d'observations.

Budget de l'état



Principe général de la CAH

Principe général. La CAH démarre avec une partition en n groupes (1 groupe = 1 individu) et réunit deux groupes à chaque étape. On a donc

- $k = n$ groupes à l'étape initiale $(h = 0)$,
- $k = n - 1$ groupes à la première étape $(h = 1)$,
- $k = n - h$ groupes à l'étape h.

Notations. Comme pour l'algorithme des *k-means*, on repère toutes les quantités avec l'indice (h) de l'itération :

- $\mathcal{C}^{(h)} = \{C_1^{(h)}, \ldots, C_k^{(h)}, \ldots, C_{n-h}^{(h)}\}$
- $n_g^{(h)}$ = nombre d'observations dans le groupe $C_g^{(h)}$ à l'itération h ;
- $\bar{\mathbf{x}}_g^{(h)}$ = point central du groupe $C_g^{(h)}$.

Partitions successives. On construit la partition $\mathcal{C}^{(h+1)}$ en réunissant deux éléments (deux groupes, disons $C_a^{(h)}$ et $C_b^{(h)}$) de la partition $\mathcal{C}^{(h)}$.

$$\mathcal{C}^{(h+1)} = \{C_1^{(h)}, \ldots, C_{a-1}^{(h)}, C_{a+1}^{(h)}, \ldots, C_{b-1}^{(h)}, C_{b+1}^{(h)}, \ldots, C_{n-h}^{(h)}, C_a^{(h)} \cup C_b^{(h)}\}$$

Décomposition de la dispersion.

$$\underbrace{\sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\bar{\mathbf{x}}}\|^2}_{\text{dispersion totale}} = \underbrace{\sum_{g=1}^{k} n_g \|\bar{\mathbf{x}}_g - \bar{\bar{\mathbf{x}}}\|^2}_{\text{dispersion inter-groupes}} + \underbrace{\sum_{g=1}^{k} \sum_{i \in C_g} \|\mathbf{x}_i - \bar{\mathbf{x}}_g\|^2}_{\text{dispersion intra-groupes}}.$$

Critère d'optimalité de Ward. Pour une partition $\mathcal{C} = \{C_1, \ldots, C_k, \ldots, C_{n-h}\}$, on définit donc le critère :

$$D(\mathcal{C}) = \sum_{g=1}^{n-h} \sum_{i \in C_g} \|\mathbf{x}_i - \bar{\mathbf{x}}_g\|^2.$$

Proposition 5 (Réunion de deux groupes)

Si, partant de la partition $\mathcal{C}^{(h)}$ on réunit les groupes $C_a^{(h)}$ et $C_b^{(h)}$ pour obtenir la partition $\mathcal{C}^{(h+1)}$, en notant $\bar{\mathbf{x}}_{g\ell}$ le point central de groupe $C_g^{(h)} \cup C_\ell^{(h)}$ on a

$$D(\mathcal{C}^{(h+1)}) - D(\mathcal{C}^{(h)}) = n_g^{(h)} \|\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_{g\ell}\|^2 + n_\ell^{(h)} \|\bar{\mathbf{x}}_\ell - \bar{\mathbf{x}}_{g\ell}\|^2$$

(Cette propriété sera démontrée en TD)

Définition 12 (Algorithme de la CAH (méthode de Ward))

Initialisation : Définir la partition initiale en $k = n$ groupes (1 groupe = 1 observation) :

$$\mathcal{C}^{(0)} = \{C_1^{(0)} = \{1\}, C_2^{(0)} = \{2\}, \ldots, C_i^{(0)} = \{i\}, \ldots C_n^{(0)} = \{n\}\}.$$

Étape $h+1$ Partant de partition $\mathcal{C}^{(h)}$, choisir les groupes $C_g^{(h)} \cup C_\ell^{(h)}$ qui minimisent

$$n_g^{(h)} \|\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_{g\ell}\|^2 + n_\ell^{(h)} \|\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}_{g\ell}\|^2$$

(dit *critère de Ward*).

Arrêt : à la $n-1$ itération ($k = 1$ groupe)

La réunion de deux groupes entraîne toujours une augmentation de la dispersion intra-groupes :

$$D(\mathcal{C}^{(h+1)}) \geqslant D(\mathcal{C}^{(h)}).$$

Les partitions $(\mathcal{C}^{(k)})_{n \geqslant k \geqslant 1}$ vérifient donc

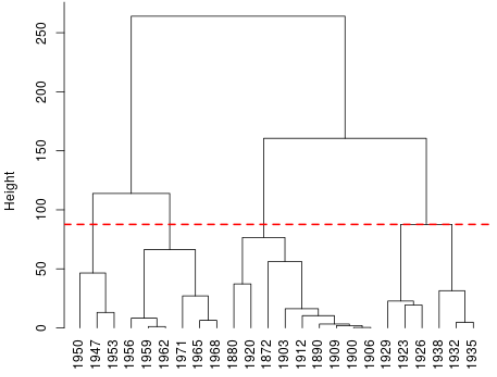
$$(0 =) \qquad D(\mathcal{C}^{(0)}) \leqslant D(\mathcal{C}^{(1)}) \leqslant \cdots \leqslant D(\mathcal{C}^{(h)}) \leqslant \cdots \leqslant D(\mathcal{C}^{(n-1)}).$$

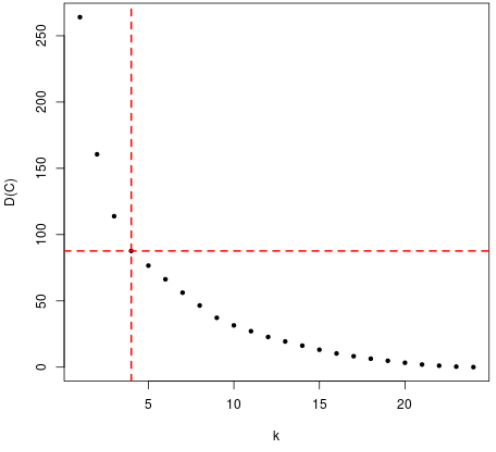
Définition 13 (Dendrogramme)

On peut retracer l'historique de l'algorithme dans un arbre enraciné dont

- les feuilles sont les n observations initiales ($\mathcal{C}^{(0)}$),
- la racine est la classification finale de toutes les observations en un seul groupe ($\mathcal{C}^{(n-1)}$),
- chaque noeud interne correspond à une fusion de deux groupes,
- la hauteur de chaque noeud interne vaut $D(\mathcal{C}^{(h)})$.

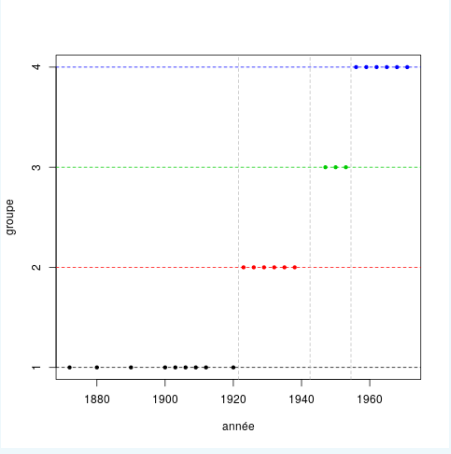
Budget de l'état





Budget de l'état

Pour $k = 4$ groupes :



	CAH1	CAH2	CAH3	CAH4
Kmeans1	0	0	0	6
Kmeans2	0	6	0	0
Kmeans3	9	0	0	0
Kmeans4	0	0	3	0

$\rightarrow$ Mêmes groupes avec les deux méthodes

Comparaison avec les *k-means* (2/2)

Budget de l'état

Pour $k = 6$ groupes :

	CAH1	CAH2	CAH3	CAH4	CAH5	CAH6
Kmeans1	0	0	0	0	3	0
Kmeans2	0	0	0	3	0	0
Kmeans3	6	1	0	0	0	0
Kmeans4	1	1	0	0	0	0
Kmeans5	0	0	0	0	0	6
Kmeans6	0	0	3	0	0	0

Les années non concordantes sont 1872, 1880, et 1920.

Essentiel du cours 4

Définitions

- partition :
  parties non-vides, 2 à 2 disjointes et dont la réunion forme l'espace entier
- *k-means* :
  1. tirage au sort de k points centraux
  2. affectation de chaque point au point central le plus proche $\rightarrow$ groupes
  3. moyenne des points de chaque groupe $\rightarrow$ nouveaux points centraux
  4. arrêt si les groupes n'ont pas changé sinon retour en 2.
- CAH par la méthode de Ward :
  1. chaque point constitue un groupe
  2. on fusionne les groupes g et ℓ si $n_g \|\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_{g\ell}\|^2 + n_\ell \|\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}_{g\ell}\|^2$ est maximal
  3. arrêt lorsqu'on a k groupes sinon retour en 2.

Propositions

- Soit $\mathcal{E}$ un ensemble fini. Si $|\mathcal{E}| = n$ alors $|\mathcal{P}(\mathcal{E})| = 2^n$.
- Décomposition de la dispersion :

$$\underbrace{\sum_{i=1}^{n} \|\mathbf{x}_i - \overline{\overline{\mathbf{x}}}\|^2}_{\text{dispersion totale}} = \underbrace{\sum_{g=1}^{k} n_g \|\overline{\mathbf{x}}_g - \overline{\overline{\mathbf{x}}}\|^2}_{\text{dispersion inter-groupes}} + \underbrace{\sum_{g=1}^{k} \sum_{i \in Cg} \|\mathbf{x}_i - \overline{\mathbf{x}}_g\|^2}_{\text{dispersion intra-groupes}}.$$

  avec
  - $\mathbf{x}_i$: i-ème point
  - $\overline{\overline{\mathbf{x}}}$: point central de tous les points
  - $\overline{\mathbf{x}}_g$: point central du groupe g.
- *k-means* converge en un nombre fini d'étapes vers un minimum local de la dispersion intra-groupes.
- À chaque étape, la méthode de Ward augmente la dispersion intra-groupes de façon minimale.