École Polytechnique de l'Université de Tours
64, Avenue Jean Portalis
37200 TOURS, FRANCE
Tél. +33 (0)2 47 36 14 14
www.polytech.univ-tours.fr

Departement Informatique
4emeannee
2016-2017

Cours - Partie 2

Analyse de données

Encadrants

Jean-Yves RAMEL
jean-yves.ramel@univ-tours.fr

Université Francois-Rabelais, Tours

Auteurs

Jean-Yves RAMEL
jean-yves.ramel@univ-tours.fr
+ Eleves de la promo 2008-2009

Version du 15 mars 2017

Table des matières

Table des figures

Plan

PARTIE I

— **Introduction**
Qu'est-ce que l'analyse de données ?
Rappels sur les outils.

— **ACP : Analyse en composantes principales**.
→ Méthodes adaptées aux problèmes descriptifs :

PARTIE II

— **AFC : Analyse factorielle des correspondances** : lien entre ligne et colonnes
→ Méthodes adaptées aux problèmes descriptifs :

— **Analyse discriminante** : reconnaissance des formes / aide au diagnostic.

— **Méthode de classification** : faire des groupes (classes).

CHAPITRE 1

Analyse factorielle des correspondances binaires (AFCB)

1.1 Généralités et notations

1.1.1 Présentation des objectifs de la méthode

Binaires : deux variables qualitatives représentées dans un tableau.

Correspondances : liens, liaisons, dépendances entre ces deux variables.

Exemple 1.1

Existe-t'il un lien entre la qualité d'un produit et la machine utilisée pour obtenir le produit ?

Les variables doivent être qualitatives $\Rightarrow$ on a des *modalités* (rouge, vert, ...).
Sinon, si on a une variable quantitative, on la transforme en créant des classes.

Pour répondre à la question posée, on a observé un certain nombre d'individus (produits) et construit l'histogramme conjoint des deux variables. C'est un tableau :

Machine \ Qualité	A	B	C	Total
M_1	191	22	13	226
M_2	154	26	15	195
M_3	157	32	11	200
M_4	170	14	9	193
Total	672	94	48	814

$n \downarrow$ (sur la colonne Total) ; $\overrightarrow{p}$ (sur la ligne Total)

Nombre d'individus : 814 produits
Nombre de variables : 2 ($\underbrace{\text{n° de machine}}_{\text{Quatre modalités } M_1,\ M_2,\ M_3,\ M_4}$ + $\underbrace{\text{qualité}}_{\text{Trois modalités } A,\ B,\ C}$).
Un individu appartient à une classe.

On appelera :

— I l'ensemble des modalités de la variable correspondant aux lignes i de 1 à n (nombre de lignes, 4 dans l'exemple).

— J est l'ensemble des modalités de la variable correspondant aux colonnes j de 1 à p (nombre de colonnes, 3 dans l'exemple).

I \ J	1	$\dots$	j	$\dots$	p	Total
1			$\vdots$			
$\vdots$						
i		$\dots$	k_{ij}	$\dots$		$k_{i.}$
$\vdots$						
n			$\vdots$			
Total			$k_{.j}$			k

effectifs marginaux

k_{ij} = effectif = nombre d'individus qui ont la modalité i pour I et j pour J

— $k_{i.}$: somme des individus ayant i pour I

$$k_{i.} = \sum_{j=1}^{p} k_{ij}$$

— $k_{.j}$

$$k_{.j} = \sum_{i=1}^{n} k_{ij}$$

— k : effectif total

$$k = \sum_{j=1}^{p} k_{.j} = \sum_{i=1}^{n} k_{i.} = \sum_{i=1}^{n} \sum_{j=1}^{p} k_{ij}$$

On utilise rarement le tableau de contingence brut car les valeurs fluctuent trop selon la population étudiée, on utilise plutÃ‡t le tableau des fréquences relatives :

$$f_{ij} = \frac{k_{ij}}{k} \quad \in [0;1]$$

fréquences marginales $\begin{cases} f_{i.} = \sum_{j=1}^{p} f_{ij} = \frac{k_{i.}}{k} \\ f_{.j} = \sum_{i=1}^{n} f_{ij} = \frac{f_{.j}}{k} \end{cases}$

En fait, il est encore plus intéressant d'étudier la distribution de la population modalité par modalité (variable par variable). On est alors amené à étudier deux tableaux différents : le tableau des *profils lignes* puis le tableau des *profils colonnes*.

Tableau des profils lignes

	1	j	p	
1		$\vdots$		1
i	$\dots$	$\frac{f_{ij}}{f_{i.}}$	$\dots$	1
n		$\vdots$		1
		$f_{.j}$		

$\frac{f_{ij}}{f_{i.}} = \frac{k_{ij}}{k_{i.}}$ = Probabilité (%) d'avoir la modalité j pour J sachant que j'ai i pour I

Tableau des profils colonnes

		j		p	
		$\vdots$			
i	$\dots$	$\frac{f_{ij}}{f_{.j}}$	$\dots$		
n		$\vdots$			
	1 $\dots$	1	$\dots$	1	

On a des tableaux qui peuvent être vus comme des ensembles individus / variables.

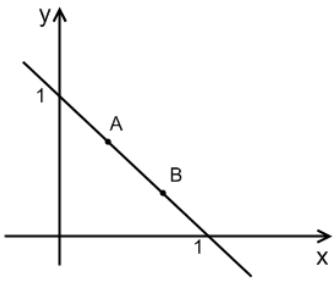
— Un individu (profil ligne) $= \begin{cases} \text{une ligne} \\ \text{un vecteur } \underline{x_i} \in \mathbb{R}^p \end{cases}$

— Un individu (profil colonne) $= \begin{cases} \text{une colonne} \\ \text{un vecteur } \underline{x_j} \in \mathbb{R}^n \end{cases}$

Ces individus sont tous situés sur un hyperplan d'équation $\sum x_j = 1$

Exemple 1.2

	x	y	
A	$1/3$	$2/3$	1
B	$2/3$	$1/3$	1



Pour étudier les tableaux profils lignes et colonnes, il est donc possible d'utiliser une ACP. Les deux études ne sont pas similaires car :

Profils lignes : on travaille dans $\mathbb{R}^p$ avec n individus. Les composantes de chaque individu sont $\left(\frac{f_{ij}}{f_{i.}}\right)$.
Chaque ligne ne pèse pas le même poids dans la population générale.
Chaque ligne pèse $f_{i.}$: on donne un poids aux individus en fonction des effectifs utilisés pour calculer les variables.
Différent de l'ACP classique.

Profils colonnes : on travaille dans $\mathbb{R}^n$ avec p individus. Les composantes de chaque individu sont $\left(\frac{f_{ij}}{f_{.j}}\right)$.
Chaque colonne pèse $f_{.j}$

1.1.2 Remarque : indice de liaison

Pour évaluer le lien existant entre deux variables, il existe une autre méthode bien connue. Indice de liaison $=$ indice du χ^2.

$$\text{Indice du } \chi^2 = n\left[\sum_{i=1 \text{ à } n,\, j=1 \text{ à } p}\left(\frac{(k_{ij})^2}{k_{i.} \times k_{.j}}\right) - 1\right]$$

$\chi^2 > 0$: plus la valeur est grande, plus la liaison est forte.
Problème du χ^2 : sa valeur dépend de n.

1.2 Distance entre individus

On veut étudier le lien existant entre deux variables indépendamment des effectifs considérés. Pour cela, il est possible de calculer des distances entre individus.

Soit deux individus

$$V_i \begin{pmatrix} \frac{f_{i1}}{f_{i.}} \\ \frac{f_{i2}}{f_{i.}} \\ \vdots \\ \frac{f_{ip}}{f_{i.}} \end{pmatrix} \qquad \text{et} \qquad V_i' \begin{pmatrix} \frac{f_{i'1}}{f_{i'.}} \\ \vdots \\ \frac{f_{i'p}}{f_{i'.}} \end{pmatrix}$$

$$d^2\left(V_i, V_i'\right) = \sum_{j=1}^{p}\left[\frac{1}{f_{.j}}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2\right]$$

On associe également un poids aux variables, le poids associé correspond à l'utilisation de la métrique du χ^2. Pour les profils lignes, le poids associé aux variables est $\frac{1}{f_{.j}}$.

$$\begin{aligned} d^2_{\chi^2}\left(V_i, V_i'\right) &= \sum_{j=1}^{p}\left[\frac{1}{f_{.j}}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2\right] \\ &= \sum_{j=1}^{p}\left[\underbrace{\frac{f_{ij}}{f_i \times \sqrt{f_{.j}}}}_{y_{ij}} - \underbrace{\frac{f_{i'j}}{f_{i'.}\sqrt{f_{.j}}}}_{y_{i'j}}\right]^2 \end{aligned}$$

On intègre la pondération des variables dans les coordonnées elles même pour revenir à une distance euclidienne classique (lors du calcul de la distance).
ACP classique possible sur les profils lignes modifiés.

1.3 ACP sur les tableaux des profils (lignes / colonnes)

1.3.1 Analyse des profils lignes

On travaille sur n points dans $\mathbb{R}^p$ et on veut faire une ACP qui utilise une distance euclidienne classique. On a vu qu'il fallait utiliser les coordonnées des points modifiés $\left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}\right)_{j=1 \text{ à } p}$ pour introduire la métrique du χ^2.
Les coordonnées étant déjà bornées entre $[0;1]$ (normalisées), une ACP *centrée* est suffisante.
Pour faire une ACP centrée, il faut les moyennes des variables $\Rightarrow$ calcul du centre de gravité.

Soit g_j la coordonnée j du centre de gravité G. Chaque individu étant pondéré par la fréquence marginale associée, on obtient

$$g_j = \frac{1}{\sum_{i=1}^{n} f_{i.}} \left(\sum_{i=1}^{n} \left(f_{i.} \times \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} \right) \right) = \sqrt{f_{.j}} \quad [1]$$

G : centre de gravité des profils lignes

$$G\left(\sqrt{f_{.1}}, \sqrt{f_{.2}}, \ldots, \sqrt{f_{.j}}, \ldots, \sqrt{f_{.p}}\right)$$

ACP centrée signifie que l'on va travailler sur la matrice des variances – covariances.

$$\begin{aligned} \text{Cov}(j,j') &= \sum_{i=1}^{n} \left[f_{i.} \left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \sqrt{f_{.j}} \right) \left(\frac{f_{ij'}}{f_{i.}\sqrt{f_{.j'}}} - \sqrt{f_{.j'}} \right) \right] \qquad = E\left((X - E(X))\,(Y - E(Y))\right) \\ &= \sum_{i=1}^{n} \left[\left(\frac{f_{ij}}{\sqrt{f_{i.}}\sqrt{f_{.j}}} - \sqrt{f_{i.}}\sqrt{f_{.j}} \right) \left(\frac{f_{ij'}}{\sqrt{f_{i.}}\sqrt{f_{.j}}} - \sqrt{f_{i.} f_{.j'}} \right) \right] \end{aligned}$$

Sous forme matricielle, on retrouve $V = {}^tXX$. Dans X, on a

$$\left(\frac{f_{ij} - f_{i.} f_{.j}}{\sqrt{f_{i.} \times f_{.j}}} \right) \quad \text{avec} \quad \left| \begin{array}{l} i = 1 \text{ à } n \\ j = 1 \text{ à } p \end{array} \right.$$

- On a montré que faire l'ACP revient à résoudre $V \cdot u_\alpha = \lambda_\alpha u_\alpha$ pour trouver les axes de projections $\vec{u_\alpha}$ optimaux.
- Cela revient à chercher les axes minimisant le déplacement des points lors de la projection $\Rightarrow$ maximiser les abscisses des projections sur les axes.

Quelques propriétés

1. Le vecteur $\vec{OG}$ est l'un des vecteurs propres de V et $\|\vec{OG}\| = 1 = \sum_{j=1}^{p} \left(\sqrt{f_{.j}}\right)^2$
   On peut montrer que la valeur propre associée à $\vec{OG}$ est $\boxed{\lambda_\alpha = 0}$ $(V \cdot u_{OG} = \lambda \cdot u_\alpha = 0)$
   L'axe factoriel associé à $\vec{OG}$ est inutile car lors d'une ACP, on s'intéresse aux valeurs propres élevées.

1. cf. TD

2. $\vec{OG}$ est orthogonal au sous-espace de dimension $p-1$ qui contient les observations centrées $\vec{OG} \cdot \vec{GI} = 0$
   I est le point représentant l'individu i.
3. Tous les vecteurs propres de V sont orthogonaux à $\vec{OG}$: $\vec{u_\alpha} \cdot \vec{OG} = 0$

$$\sum_{j=1}^{p} \left(\sqrt{f_{.j}} \cdot u_{\alpha j} \right) = 0$$

   On en déduit qu'il est inutile de centrer les données. On peut alors utiliser la matrice X^* au lieu de X avec

$$X* = \left(\frac{f_{ij}}{\sqrt{f_{i.} \times f_{.j}}} \right)_{j=1 \text{ à } p, i=1 \text{ à } n}$$

   On cherche ensuite les valeurs propres et vecteurs propres de $V^* = {}^tX^* \cdot X^*$. V et V^* ont les mêmes valeurs propres et vecteurs propres.
   Par contre, lorsque l'on utilise V^* au lieu de V, $\vec{OG}$ est associé à la valeur propre $\lambda = 1$.
4. Les valeurs propres λ_α sont dans $[0;1]$ $\forall \alpha$.

1.3.2 Analyse des profils colonnes

Démarches et calculs identiques en remplaÃğant les lignes par les colonnes.

1.3.3 Résumé

	lignes	colonnes
espace	$\mathbb{R}^p$	$\mathbb{R}^n$
individus	$\left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} \right)_{j=1 \text{ à } p}$	$\left(\frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} \right)_{i=1 \text{ à } n}$
poids	$f_{i.}$	$f_{.j}$
centre de gravité	$(\sqrt{f_{.j}})_{j=1 \text{ à } p}$	$(\sqrt{f_{i.}})_{i=1 \text{ à } n}$
matrice à étudier	matrice des variances – covariances $V_{(p\times p)} = \text{Cov}(j,j')$ ou V^* (on ne centre pas)	$W_{(n\times n)} = \text{Cov}(i,i')$ ou W^* (on ne centre pas)

cf. feuille distribuée.

1.3.4 Relation entre les deux études (profils lignes – profils colonnes)

— Comme en ACP, V^* et W^* ont les mêmes valeurs propres.
— Comme en ACP, on obtient les formules de transition suivantes :

$$V_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X^* \cdot u_\alpha$$

V_α : *vecteur propre de* W^*
u_α : *vecteur propre de* V^*

$$u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} {}^tX^* \cdot V_\alpha$$

— Comme en ACP, il est possible de passer des abscisses des projections des profils lignes (sur u_α) aux abscisses des projections des profils colonnes (sur V_α).

$$\Psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i.}} \times \Phi_{\alpha j} \right)$$

$\Psi_{\alpha i}$: *abscisses des points lignes sur* u_α

$$\Phi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{.j}} \times \Psi_{\alpha i} \right)$$

Φ_α : *abscisses des points colonnes sur* $\vec{v_\alpha}$

Remarque : Les abscisses $(\Psi_{\alpha i})$ sont équivalentes aux coordonnées du pseudo barycentre des projections des points colonnes $\Phi_{\alpha j}$ affectées du poids f_{ij} et au coefficient $\left(\frac{1}{\sqrt{\lambda_\alpha} f_{i.}}\right)$ près.

1.4 Aide à l'interprétation des résultats obtenus

1.4.1 Contributions

On distingue deux aspects :
— La *contribution absolue* d'un individu. C'est la part prise par un élément dans l'inertie (ou la variance) expliquée par un facteur. Elle s'exprime pas

$$\mathrm{Ca}^\alpha(i) = \frac{f_{i.} \times (\Psi_{\alpha i})^2}{\lambda_\alpha} \quad \text{(Profils lignes)}$$

$$\mathrm{Ca}^\alpha(j) = \frac{f_{.j} \times (\Phi_{\alpha j})^2}{\lambda_\alpha} \quad \text{(Profils colonnes. On se focalise sur un axe } \alpha\text{)}$$

$$\sum_{i=1}^{n} \mathrm{Ca}^\alpha(i) = 1$$

— La *contribution relative* à un facteur (axe). C'est la part prise par un facteur dans l'expression (l'explication) de la disposition d'un élément (individu) dans le nouvel espace de représentation fourni par l'ACP.

$$\mathrm{Cr}^\alpha(X_i) = \frac{(\Psi_i^\alpha)^2}{\underbrace{\sum_{k=1}^{q} \left(\Psi_i^k\right)^2}_{\text{nouveau repère d'origine } G}} = \underbrace{\frac{(\Psi_i^\alpha)^2}{d^2(G, X_i)}}_{\text{repère initial}}$$

q est le nombre de valeurs propres non nulles obtenues (souvent, $q = p - 1$).

$$\sum_{\alpha=1}^{q} \mathrm{Cr}^{\alpha}(X_i) = 1$$

La contribution relative est comparable à ce que l'on appelait *indice ponctuel de qualité* en ACP.

1.4.2 Représentations graphiques

— L'AFCB permet comme l'ACP de représenter les deux nuages profils lignes et profils colonnes sur le même graphique.
— Si des lignes ont des projections regroupées et proches de certains colonnes, c'est que ces lignes sont semblables et qu'elles se ressemblent surtout pour ce qui concerne les colonnes correspondantes.
— Un point loin de l'origine se distingue du profil moyen.
— L'interprétation de la proximité entre deux points X_i (ligne) et Y_j (colonne) est beaucoup plus dangereuse et donc à réaliser avec précaution (surtout pour les points *proches de l'origine*).

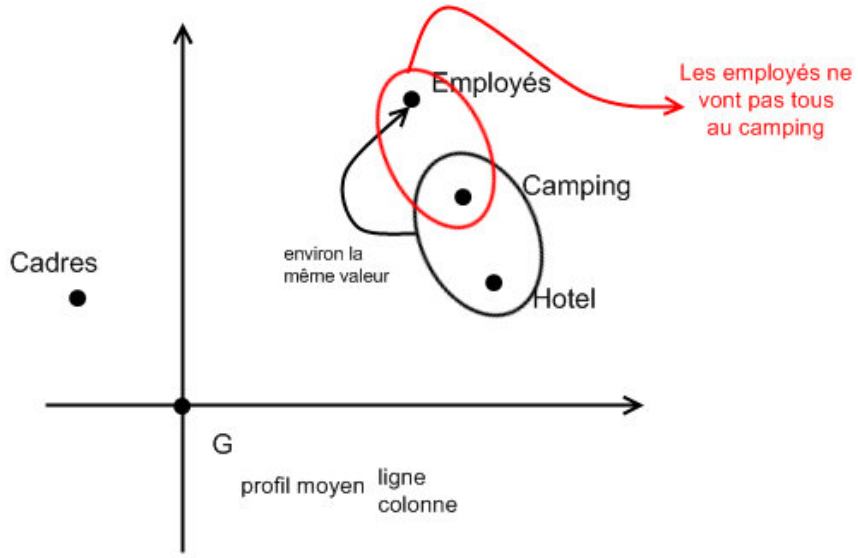


FIGURE 1.1 – Représentation graphique de l'ACFB

1.4.3 Formes classiques des nuages

1.4.3.1 Points abérrants

Les points abérrants sont des points isolés du nuage principal. Ils correspondent à des individus dont la structure est marginale. Il est préférable de refaire l'étude sans ces points.

1.4.3.2 Deux paquets de points

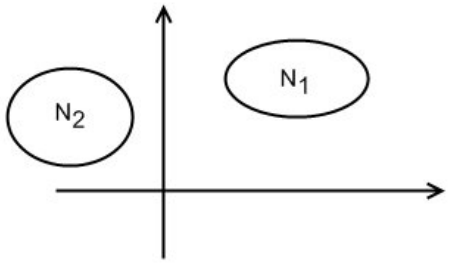


FIGURE 1.2 – Deux paquets de points

Deux groupes d'individus distincts sont mis en valeur. Il y a de grandes chances que l'on puisse réordonner les lignes ou les colonnes pour obtenir un tel tableau :

	J_1	J_2
I_1	effectifs importants	effectifs faibles
I_2	effectifs faibles	effectifs importants

1.4.3.3 Effet Guttman

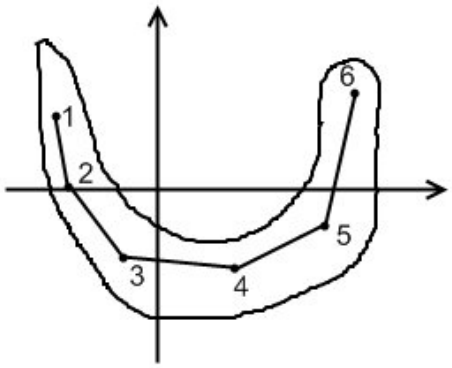


FIGURE 1.3 – Effet Guttman. Le nuage a une forme de parabole

L'axe 1 oppose les modalités à effectifs faibles aux modalités à effectifs forts.

Là aussi, il est souvent possible de réordonner les lignes et/ou colonnes du tableau pour un obtenir un tel tableau :

effectifs importants			faible
	effectifs importants		
faible			effectifs importants

Effectifs importants sur la diagonale.

Cela se produit fréquemment lorsque l'une des deux variables est ordinales (classe d'Ãćges, niveau, . . .). Il peut alors être intéressant de relier les points de cette variable pour faire apparaÃőtre la relation d'ordre.

CHAPITRE 2

Analyse factorielle discriminante (AFD)

2.1 Introduction

Les données que l'on va traiter ici ne sont plus homogènes mais au contraire on étudie une population d'individus *partitionnée* en classes, c'est-à-dire en k sous population. Pour chaque individu, on connait un ensemble de variables quantitatives. Dans la population totale, on a toujours n individus.

	p variables quantitatives
n individus	classe 1
	.....
	classe 2
	.....
	⋮
	.....

On peut avoir deux objectifs pour l'AFD :

- **But descriptif :**
  - On cherche parmi la liste des variables celles qui caractérisent le mieux la répartition des individus dans les différentes classes.
  - On cherche également à construire une combinaison linéaire des variables initiales qui permette de séparer au mieux les différentes classes.

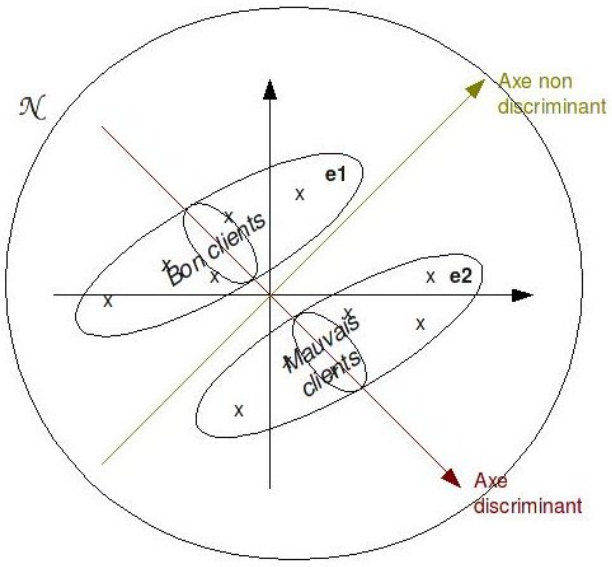


FIGURE 2.1 – Illustration de l'analyse factorielle discriminante

- **But décisionnel :**
  On cherche à affecter une classe à un individu, qui n'est pas dans le tableau de départ, pour lequel on connait les valeurs de ses p variables (mais pas sa classe).
  $\rightarrow$ Prédire la valeur d'une variable qualitative (la classe) à partir de p variables quantitatives (prédicteurs ou descripteurs). *Reconnaissance des formes.*

2.2 Principes de l'AFD

On veut projeter un nuage de points qui sont dans un espace $\mathbb{R}^p$ sur un sous espace de dimensions inférieures et dans lequel les classes sont clairement séparées.
$\rightarrow$ Il va falloir définir un critère de séparation des classes faisant intervenir la dispersion des points dans la population totale et dans les sous populations (les classes).

2.2.1 Etude des variances

On dispose d'une population globale dont on connait la répartition.

$$\mathcal{N} = \bigcup_{k=1}^{K} \mathcal{N}_k \quad \text{(K classes)}$$

$$n = \sum_{k=1}^{K} \mathcal{N}_k$$

Si on s'intéresse à une sous population :

$$\text{Centre de gravité} = \bar{x}^k = \frac{1}{N_k} \sum_{i=1}^{N_k} \left(x_i^k\right)$$

$$\Rightarrow \quad \bar{x}_j^k = \sum_{i=1}^{N_k} \left(x_{ij}^k\right) \times \frac{1}{N_k} \quad (j \text{ varie de 1 à } p)$$

Si on s'intéresse à tous le nuage $\mathcal{N}$:

$$\text{Centre de gravité} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{N_k} x_{ij}^k$$

$$\text{Centre de gravité de } \mathcal{N} = \bar{x}_j = \frac{1}{n} \sum_{i=1}^{K} \left(N_k \bar{x}_j^k\right)$$

$\bar{x}_j^k$ *: centre de gravité de* $\mathcal{C}_k$

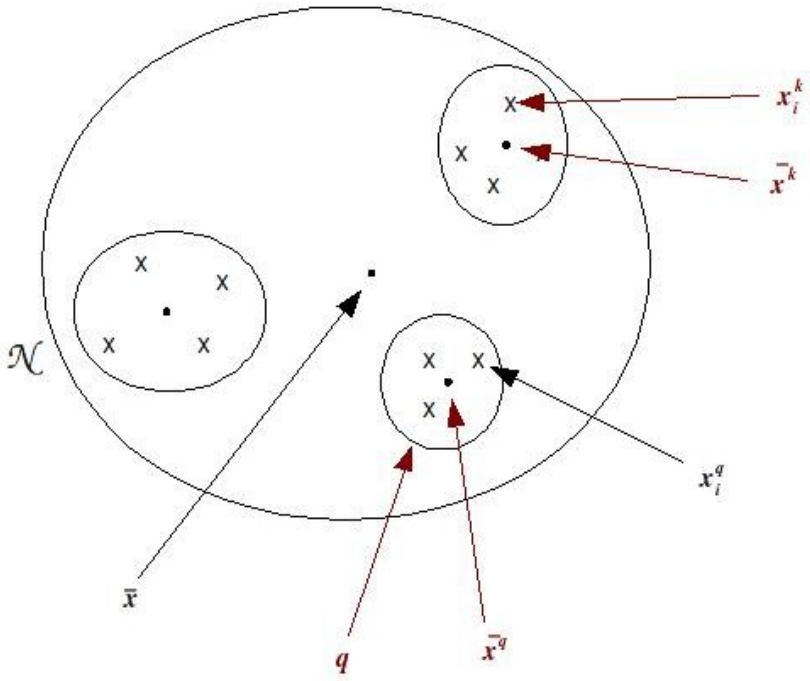


FIGURE 2.2 – Nuage $\mathcal{N}$

$$\begin{array}{r} \mathcal{N}\left(\bar{x};T\right) \\ \mathcal{N}_1\left(\bar{x}^1;T_1\right) \\ \mathcal{N}_2\left(\bar{x}^2;T_2\right) \end{array}$$

$$\begin{array}{cc} & \begin{array}{c} p \\ \leftrightarrow \end{array} \\ i=1 \text{ à } n \updownarrow & \begin{pmatrix} \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} \begin{array}{l} \}T_1 \\ \}T_2 \\ \}T_3 \end{array} \\ \bar{x} & \begin{array}{ccc} 0 & & \end{array} \end{array}$$

$(n \times p)$

Le centre de gravité de $\mathcal{N}$ est le barycentre des centres de gravité de chacune des classes affecté d'un poids égale à la fréquence d'apparition de la classe $= \frac{N_k}{N}$.
On peut chercher à mettre en place une relation similaire entre la matrice des variances – covariances globale et pour chaque classe.

$$t_{ij} = \underbrace{\frac{1}{n}\sum_{k=1}^{n}\left[\left(x_{ki}-\bar{x}_i\right)\left(x_{kj}-\bar{x}_j\right)\right]}_{\text{covariance totale (sur } \mathcal{N})} = \text{Covariance entre les variables } x_i \text{ et } x_j$$

$$t_{ij} = \underbrace{\frac{1}{n}\sum_{q=1}^{K}\left(N_q t_{ij}^q\right)}_{\text{covariance intraclasse}} + \underbrace{\frac{1}{n}\sum_{q=1}^{K}\left[N_q\left(\bar{x}_i^q-\bar{x}_i\right)\left(\bar{x}_j^q-\bar{x}_j\right)\right]}_{\text{covariance interclasse}}$$

$\bar{x}_i$: *moyenne des variables* x_i
$\bar{x}_j$: *moyenne des variables* x_j
N_q : *nombre d'individus dans* $\mathcal{N}_q$
$\bar{x}_i^q$: *coordonnée du centre de gravité de* $\mathcal{N}_q$
$\bar{x}_i$: *coordonnée du centre de gravité de* $\mathcal{N}$
t_{ij}^q : *covariance entre* i *et* j *calculée sur* $\mathcal{N}_q$

cf. feuille jointe pour la démonstration.

$$T = \begin{array}{c} \\ i \end{array}\begin{array}{c} \begin{array}{ccc} & j & \end{array} \\ \begin{pmatrix} \cdots & \vdots & \cdots \\ & t_{ij}^k & \\ & \vdots & \end{pmatrix} \end{array} = \text{Matrice de variances – covariances}$$

$(p \times p)$

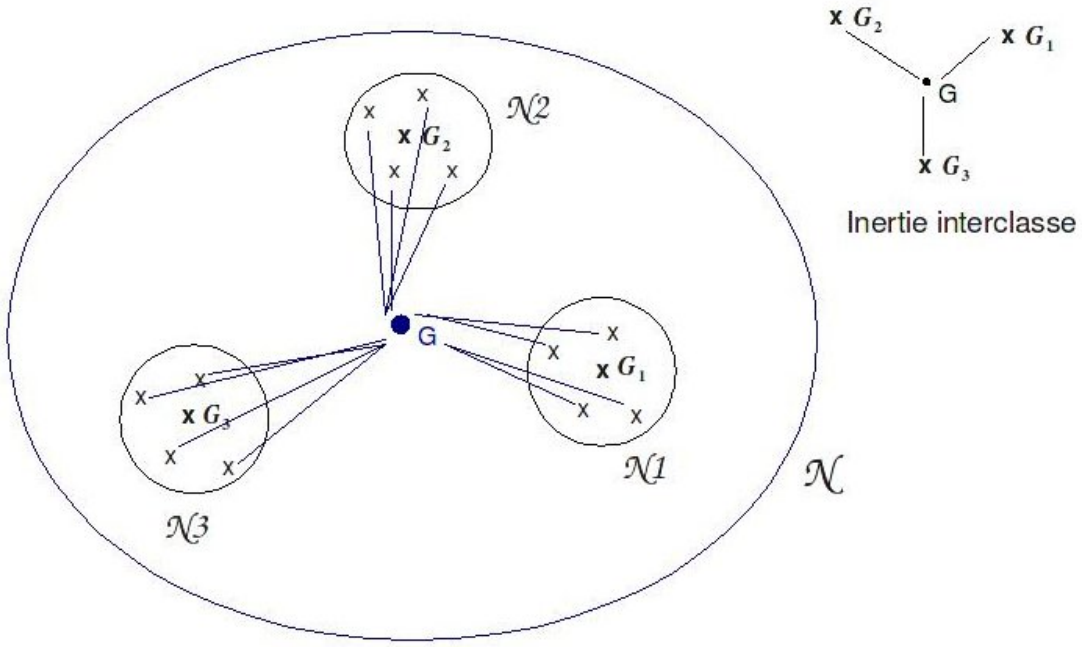


FIGURE 2.3 – Illustration de l'inertie interclasse

Dans la formule t_{ij} ci-dessus,

— le premier terme est représentatif des variances – covariances intraclasses ;

— le second terme est représentatif des variances – covariances interclasses (entre centres de gravité).

Si on note

— T la matrice des variances – covariances totales (calculées sur $\mathcal{N}$),

$$T = (t_{ij})_{i=1 \text{ à } p, j=1 \text{ à } p}$$

— T_k la matrice des variances – covariances intraclasses (calculées sur $\mathcal{N}_k$),

$$V = \sum_{k=1}^{K} \left(\frac{T_k \times N_k}{n} \right)$$

— W la matrice des variances – covariances interclasses.

On obtient la relation de Huygens :

$$T = V + W$$

La matrice des covariances d'un nuage de points réparties en K classes et égale à la somme des matrices de covariances intraclasses et interclasses.

Au niveau des variances, on peut noter

— Variance du nuage total :

$$t = \text{trace}(T)$$

— Variance intraclasse :

$$v = \text{trace}(V)$$

— Variance interclasse :

$$w = \text{trace}(W)$$

On montre également que l'on obtient

$$t = v + w$$

La variance intraclasse permet de mesurer la compacité à l'intérieur de chaque classe. Plus elle est faible, plus les points sont proches des centres de gravités.

La variance interclasse mesure la dispersion entre les classes. Plus elle est faible, plus les classes sont proches les une des autres.

Dans un problème de discrimination (AFD), on cherche

— une bonne compacité des classes $\Rightarrow$ v petit ;

— une grande dispersion entre les classes $\Rightarrow$ w grand.

$\rightarrow$ Le critère utilisé sera $\frac{w}{v}$ que l'on cherche à maximiser.

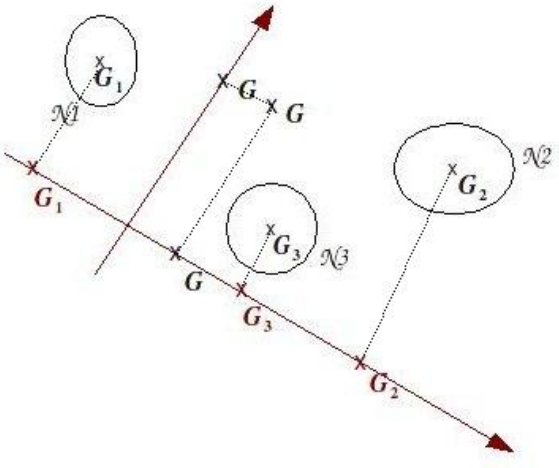


FIGURE 2.4 – Projection des centres de gravité

On a besoin de T, V et W pour faire une AFD.

- Matrice des données centrées :

$$T = {}^tRR \text{ avec } R_{(n\times p)} = \frac{1}{\sqrt{n}} \begin{pmatrix} (x_{11} - \bar{x}_1) & \cdots & (x_{1p} - \bar{x}_p) \\ \vdots & \ddots & \vdots \\ (x_{n1} - \bar{x}_1) & \cdots & (x_p - \bar{x}_p) \end{pmatrix} \updownarrow i = 1 \text{ à } n \\ \longleftrightarrow \\ j = 1 \text{ à } p$$

-

$$W = {}^tSS \text{ avec } S = \frac{1}{\sqrt{n}} \begin{pmatrix} \sqrt{N_1}(\bar{x}_1^1 - \bar{x}_1) & \cdots & \sqrt{N_1}(\bar{x}_p^1 - \bar{x}_p) \\ \vdots & \ddots & \vdots \\ \sqrt{N_K}(\bar{x}_1^K - \bar{x}_1) & \cdots & \sqrt{N_K}(\bar{x}_p^K - \bar{x}_p) \end{pmatrix} \updownarrow k = 1 \text{ à } K \\ \longleftrightarrow \\ j = 1 \text{ à } p$$

$$V = T - W$$

Le critère utilisé en AFD est

$$\underset{\vec{u}}{\text{Max}} \left(\frac{w_u}{v_u} \right)$$

w_u : *variance interclasse après projection*
v_u : *variance intraclasse après projection*

$$r_u = \frac{w_u}{v_u}$$

— Nuage des points projetés sur $\vec{u}$:

$$^tu \cdot x_i = \text{abscisse de } x_i \text{ sur } \vec{u}$$

x_i : *vecteur de l'individu* i

— Abscisse du centre de gravité de $\mathcal{N}$ sur $\vec{u}$:

$$\bar{x_u} = \frac{1}{n}\sum_{i=1}^{n}\left(^tu \cdot x_i\right) = \text{moyenne des abscisses des points après projection}$$
$$\bar{x_u} = {}^tu\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = {}^tu \cdot \bar{x} = \text{projection de } G \text{ (centre de gravité de } \mathcal{N}) \text{ sur } \vec{u}$$

Variance après projection (sur $\vec{u}$)

$$t_u = \frac{1}{n}\sum_{i=1}^{n}\left({}^tux_i - {}^tu\bar{x}\right)$$

En développant, on montre que

$$t_u = {}^tuTu \quad \text{sous forme matricielle}$$

De mÃłme,

$$v_u = {}^tuVu$$
$$\text{et} \quad w_u = {}^tuWu$$

On sait que $T = W + V$ d'oÃź $t_u = w_u + v_u$.

Propriété : Après projection sur un axe $\vec{u}$, on a toujours la relation (de Huygens) :

$$\text{variance totale} = \text{variance intraclasse} + \text{variance interclasse}$$

On peut réécrire

$$r_u = \frac{t_uWu}{t_uVu}$$

On cherche à maximiser r_u, la variable est u.
En cherchant à annuler la dérivée du rapport, on montre que r_u est maximum quand

$$Wu = \lambda Vu$$

Si V est inversible, on peut écrire

$$\underbrace{V^{-1}W}u = \lambda u$$

Le meilleur axe de projection est un des vecteurs propres de $V^{-1}W$.
De plus, on cherche

$$\begin{aligned}\max(r_u) &= \frac{\max\left({}^tuWu\right)}{\left({}^tuVu\right)} \qquad (Wu = \lambda Wu)\\ &= \frac{\max\left({}^tu\left(\lambda Vu\right)\right)}{{}^tuVu}\\ &= \lambda\end{aligned}$$

Il faut prendre le vecteur propre $\vec{u}$ qui correspond à la plus grande valeur propre $= \lambda$:

— on écarte les centres de gravité des sous nuages $\mathcal{N}_1$, $\mathcal{N}_2$, ... ;
— on compacte les projections.

Propriétés et remarques :

— V^{-1} n'est pas forcément symétrique : les vecteurs propres ne sont pas forcément orthogonaux les uns par rapport aux autres. On les représentera cependant comme orthogonaux dans les projections sur les plans factoriels.
— λ est une valeur propre de $V^{-1}W$ postitive car $\lambda = \max(r_u)$. r_u est un rapport de variance ≥ 0.
— On a maximisé

$$\left(\frac{{}^tuWu}{{}^tuVu}\right) = \frac{{}^tuWu}{{}^tuTu - {}^tuWu}$$

Cela revient à minimiser

$$\frac{{}^tuTu - {}^tuWu}{{}^tuWu} = \frac{{}^tuTu}{{}^tuWu} - 1$$

Cela revient à maximiser

$$\frac{({}^tuWu)}{({}^tuVu)} = \frac{\text{variance interclasse}}{\text{variance totale}}$$

On recherche cette fois les valeur propres et vecteurs propres de $(T^{-1}W)$.
— $(T^{-1}W)$ et $(V^{-1}W)$ ont les mÃłmes vecteurs propres et si on appelle μ les valeurs propres de $T^{-1}W$, alors on a la relation

$$\lambda = \frac{\mu}{1-\mu}$$

λ : *valeur propre de* $V^{-1}W$
μ : *valeur propre de* $T^{-1}W$

— On a toujours $0 \leq \mu \leq 1$

DÃłfinition : **Pouvoir discriminant**

Le *pouvoir discriminant* d'un axe est

$$\mu = \frac{(\text{variance interclasse})_u}{(\text{variance totale})_u}$$

Généralisation : Ce que l'on vient de dire pour un axe de projection se généralise pour q axes de projections.

Lors de la recherche de sous espaces discriminants, il est rarement utile de considérer un espace de dimension $q > K - 1$.

2.2.2 Résumé

AFD (tableau $n \times p$, individu $\times$ variable) :

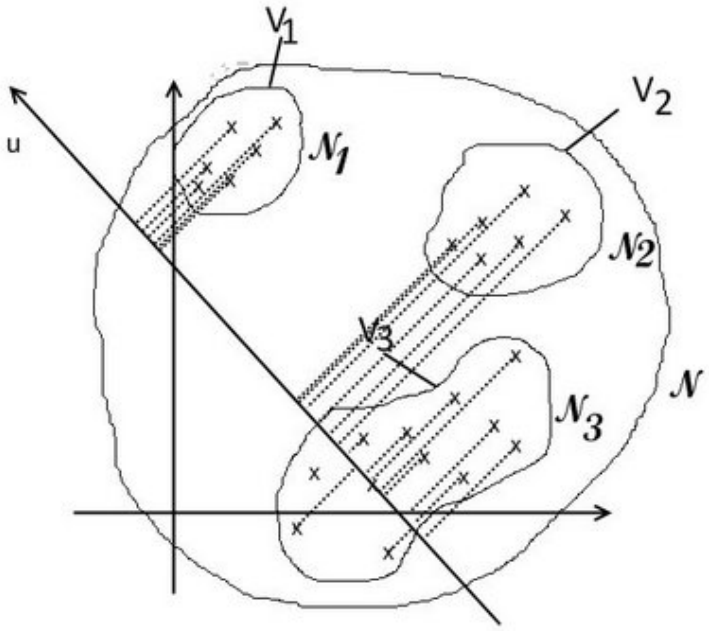


FIGURE 2.5 – Résumé du fonctionnement de l'AFD

1. On calcul de $T_{(p\times p)}$ (matrice des variances – covariances calculées sur $\mathcal{N}$) et $W_{(p\times p)}$ (calculées sur les G_k (il y en a K)).
2. On en déduit $V_{(p\times p)}$ (calculées sur chacun des sous nuages $\mathcal{N}_1$, puis $\mathcal{N}_2$, puis $\mathcal{N}_K$) $= T_{(p\times p)} - W_{(p\times p)}$
3. On calcul $T^{-1}W$
4. On cherche les valeurs propres et vecteurs propres de $T^{-1}W$
5. On classe les valeurs propres et vecteurs propres selon leur pouvoir discriminant μ.
6. On effectue les projections ${}^t u \cdot x_i$
7. On analyse, interprète, réfléchit.

2.3 Analyse discriminante à but décisionnel

2.3.1 Introduction

Problème : Soit $x \in \mathbb{R}^p$ un individu connu, on veut l'affecter à une classe.

Solution : Avant ou après projection, le critère pour classer x sera $d^2(x, G_k)$

$$x \in w_k \text{ si } \forall\, l \in [1;K], \quad d^2(x,G_k) \leq d^2(x,G_l)$$

Question : Quelles possibilités pour calculer $d^2(x, G_k)$?
On va poser (avec $\bar{x}_k = G_k$) :

$$d^2(x,\bar{x}_k) = {}^t(x-\bar{x}_k)\, Q\, (x-\bar{x}_k)$$

2.3.2 Choix de Q

- **Possibilité 1** : distance euclidienne classique :

$$Q = \text{matrice identité}$$

On parle alors d'inertie sphérique.

- **Possibilité 2** : distance de Mahalanebis :

$$Q = T^{-1}$$

Cela revient à effectuer une «normalisation» des variables. On tient compte de la forme du nuage initial pour faire l'affectation.
On modélise alors toutes les classes non plus par une sphère mais par une ellipsoÃŕde dont les caractéristiques dépendent de la forme de $\mathcal{N}$.

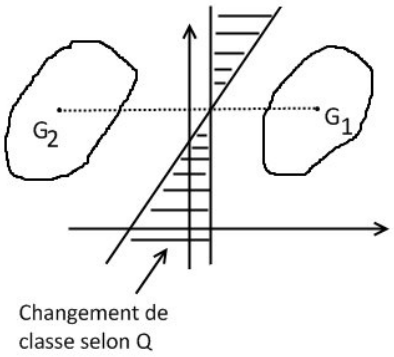


FIGURE 2.6 – Illustration de la distance de Mahalanebis

2.4 AFD décisionnelle quadratique

2.4.1 Introduction

On recherche plus des surfaces de séparation linéaire mais qui soient du second ordre pour adapter la distance à la forme de chacune des classes.

On adapte la formule de distance à chaque classe :

$$d^2(x, \bar{x}_k) = {}^t(x - \bar{x}_k)\, Q_k\, (x - \bar{x}_k)$$

2.4.2 Choix des Q_k

Q_k est une matrice carrée définie positive, et on va la choisir de manière à minimiser la moyenne des distances au carré à l'intérieur des classes. Pour cela, on va définir deux notions :

— *la similitude d'un individu* x avec un nuage $\mathcal{N}_k$:

$$s\,(x, \mathcal{N}_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} d^2(x, \bar{x}_k)$$

On pourra alors choisir d'affecter x à la classe qui minimise $s(x, \mathcal{N}_k)$

— *le coefficient d'agrégation d'une classe*, c'est-à-dire la moyenne des distances au carré entre tous les points d'une classe :

$$D_k^2 = \frac{1}{n_k(n_k-1)} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} d_k^2(x_i, x_j)$$

On obtient différentes approches, la plus utilisée étant l'approche de Sebesteyn [1].

2.5 Sélection de variables discriminantes

AFD : création de nouvelles *variables synthétiques* parfois difficiles à interpréter ($\rightarrow$ projection sur les axes factoriels).

Autre possibilité : sélection de caractéristiques = détecter les variables discriminantes dans la liste initiale de variables.

Question : Combien de variables choisir ? Lesquelles ?

Deux méthodes :

Méthode pas à pas : On cherche la meilleure variable (p choix) puis, à chaque étape, on ajoute une nouvelle variable ($p-1$).
Nombre total de choix :

$$\frac{p(p+1)}{2}$$

Méthode exacte : On teste toutes les combinaisons possibles de variables, on conserve la meilleure combinaison.

$$C_p^1 + C_p^2 + \ldots + C_p^p = 2^p \text{ choix possibles (coÃżt exponentiel)}$$

Critère de choix de la meilleure combinaison de variables

— Ensemble d'apprentissage $\Rightarrow$ un tableau Individus/Variables/Classes
— Critère d'affectation des individus aux différentes classes.
Souvent

$$d_Q^2(G_k, X)$$

G_k : *centre de gravité de la classe* k
X : *l'individu à classer*

On prend d_Q^2 minimum.
— Matrice de confusion. Cette matrice permet de comparer les résultats obtenus par différentes configurations de variables.

Classe / Classement	w_1	w_2	$\ldots$	w_K
Ω_1	**8**	0		
Ω_2	2	**9**		
$\vdots$	0			
Ω_K	0			
Total	10	10		

Somme des éléments sur la diagonale = nombre de bien classés.

1. cf. exercice II, TD AFD.

CHAPITRE 3

Classification automatique

3.1 Présentation

Le but de la classification automatique est de réduire le nombre d'individus dans un nuage en les regroupant en classes homogènes. Les individus sont représentés par p variables.

Deux types de méthodes :

Méthodes hiérarchiques : elles produisent un ensembles de partitions emboitées représentables sous forme d'arbre.

Méthodes non hiérarchiques : elles produisent directement une seuile partition des individus en un nombre fixé de classes.

Deux problèmes :

— Comment choisir le nombre de classes optimal ?
→ Un expert choisi (méthode supervisée), on trouve automatiquement (non supervisée).

— Comment faire les regroupements ?
→ Utilisation de mesures de similarité entre individus.

3.2 Classification ascendante hiérarchique (CAH)

3.2.1 Hiérarchie des partitions

L'ensemble des objets est représenté sous forme d'un arbre (appelé dendogramme) dont chaque section correspond à une partition de l'ensemble initial.

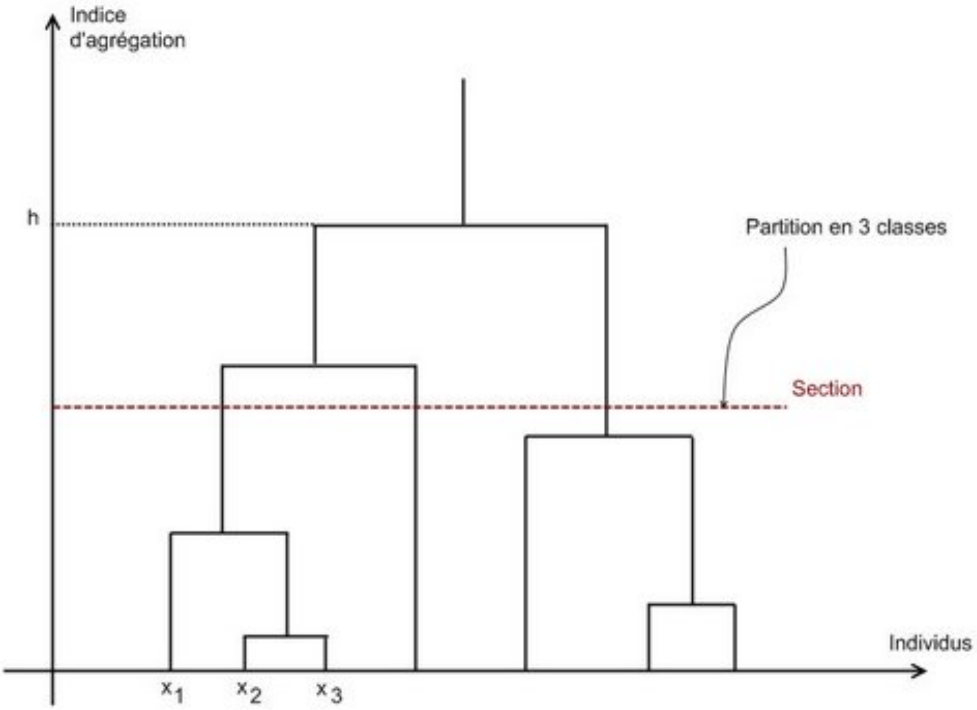


FIGURE 3.1 – Exemple de dendogramme

La mesure h d'un palier est l'indice d'agrégation qui donne une information sur la similarité qui existe entre les éléments regroupés.

DÃ©finition : Hiérarchie H

Etant donné un ensemble fini d'objet Ω, une hiérarchie H est un ensemble de partie Ω qui vérifie :

1. $\Omega \in H$
   Au sommet de la hiérarchie, tous les individus sont regroupés.
2. $\forall \omega \in \Omega, \{\omega\} \in H$
   En bas, on a tous les singletons d'individus.
3. $\forall h \in H$ et $h' \in H$, on a 3 possibilités : $\begin{cases} h \cap h' = \varnothing & \text{ou} \\ h \subset h' & \text{ou} \\ h' \subset h \end{cases}$
   Un individu ne peut pas appartenir simultanément à deux classes.

Hiérarchie binaire : chaque palier correspond à un regroupement de deux individus.

Hiérarchie indicée : notée (H, f)

— H est une hiérarchie.

— f associe un indice d'agrégation à chaque palier tel que f est définie dans $\mathbb{R}^*$ et $\left\| \begin{array}{l} f(h) = 0 \Rightarrow \text{card}(h) = 1 \\ \text{Si } h \subset h' \Rightarrow f'(h) \leq f(h') \end{array} \right.$

3.2.2 Algorithme

Données : un ensemble d'objets (individus) Ω à classer.

— On définit une mesure de similarité (ou dissimilarité) entre ces objets $\rightarrow$ Matrice de dissimilarité entre objets.

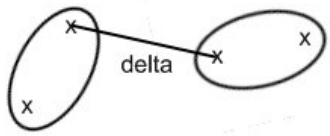
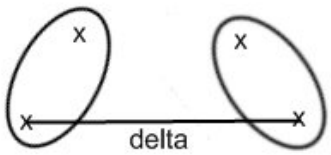
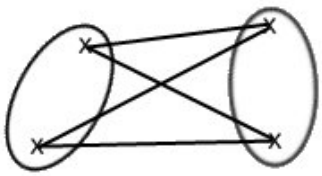
— On définit une mesure de similarité (ou dissimilarité) entre ces groupes (classes) d'objets $\rightarrow$ il existe différentes stratégies.

On boucle. *Critères d'arrÃłt :*
— $\underline{X}$ de taille $(1, 1)$, ou
— $\delta \geq$ seuil, ou
— nombre de classe $\geq$ seuil

1. On cherche le *minimum* dans $\underline{X}$ (attention : valable pour toute stratégie).
2. On agrège les deux objets (ou groupes) similaires en une nouvelle classe (on mémorise l'indice d'agrégation $= \delta$).
3. On recalcule $\underline{X}$ suivant *la stratégie* choisie $\rightarrow$ la taille de $\underline{X}$ diminue.

3.2.3 Stratégie d'agrégation

cf. TD.

Saut minimum	Saut maximum
	
Lien moyen	**Ward**
 Moyenne de toutes les distances	Liée à l'augmentation de l'inertie intra-classe de la classe créée.

Il y en a beaucoup d'autres.

3.3 Classification par partitionnement

Méthode non hiérarchique, on va étudier un exemple : méthode des centres mobiles ($\simeq K$means).

Données initiales :

— un ensemble d'objets (individus) $= \Omega, \text{card}(\Omega) = n$, on est dans $\mathbb{R}^*$.
— $K =$ nombre de classes désirées.

3.3.1 Algorithme [Forgy, 1965]

0. On choisit aléatoirement dans Ω K centres provisoires $C_1^0, C_2^0, \ldots, C_K^0$
1. On partitionne Ω tel que $\left[\begin{array}{l} \Omega = \Omega_1^\varnothing \cup \Omega_2^\varnothing \cup \ldots \cup \Omega_K^\varnothing \\ \Omega_i^\varnothing = \text{ ensemble des objets plus proches de } C_i^0 \text{ que de tous les autres.} \end{array}\right.$

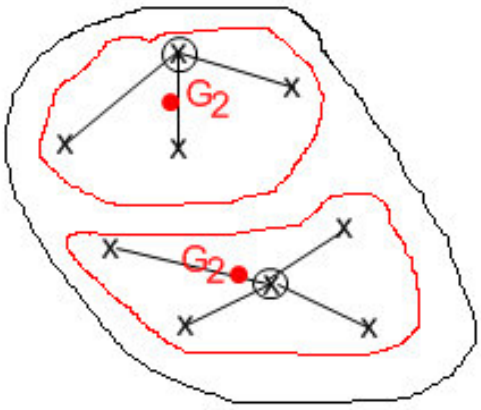


FIGURE 3.2 – Partitionnement de Ω

2. On boucle. *Critères d'arrÃłt :*
   — les centres deviennent immobiles, ou
   — nombre d'itérations, ou
   — variation de la variance intra-classe $<$ seuil

   ↗↖
   (a) Calcul des nouveaux centres mobiles (de gravité) $C_1^1 C_2^1 \ldots C_K^1$.
   (b) Regénération d'un nouveau partitionnement de $\Omega = \Omega_1^1 \cup \ldots \Omega_K^1$

Cet algorithme est *convergeant* car la variance intra-classe ne peut que décroitre à chaque étape ou *stagner* (d'où critère = nombre d'itération).

Le résultat dépent du tirage initial des centres.

Solutions :

— Faire plusieurs tirages aléatoires → analyse des résultats
   ↗ formes fortes (toujours la mÃłme classe)
   ↘ formes faibles (points non stables)
— Tirages non aléatoires.

— Au lieu de choisir un seul centre mobile par classe, on choisit plusieurs points représentants chaque classe ⇒ Noyau [1].

1. Méthodes des nuées dynamiques (Didey, 1972).

Index

Analyse de données

Departement Informatique
4emeannee
2016-2017

Cours - Partie 2

Résumé : Cours d'analyse de données

Mots clefs : analyse de données

Abstract:

Keywords:

Encadrants

Jean-Yves RAMEL
jean-yves.ramel@univ-tours.fr

Université Francois-Rabelais, Tours

Auteurs

Jean-Yves RAMEL
jean-yves.ramel@univ-tours.fr
+ Eleves de la promo 2008-2009