Fouille de données

Cours 3 - Exploration des données multidimensionnelles - Apprentissage non supervisé

NGUYỄN Thị Minh Huyền
Email: huyenntm@vnu.edu.vn
VNU University of Science, Hanoi

Plan

Analyse exploratoire

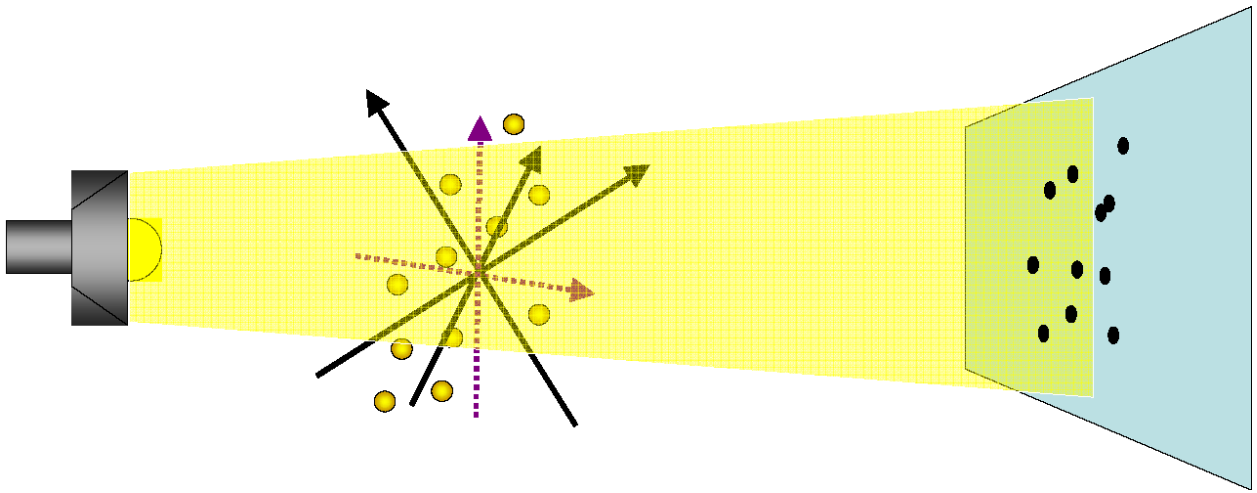
Méthodes de réduction

- Réduction du nombre de variables (dimension) : méthodes factorielles
- Réduction du nombre d'individus en regroupant par classes : méthodes de classification

Plan

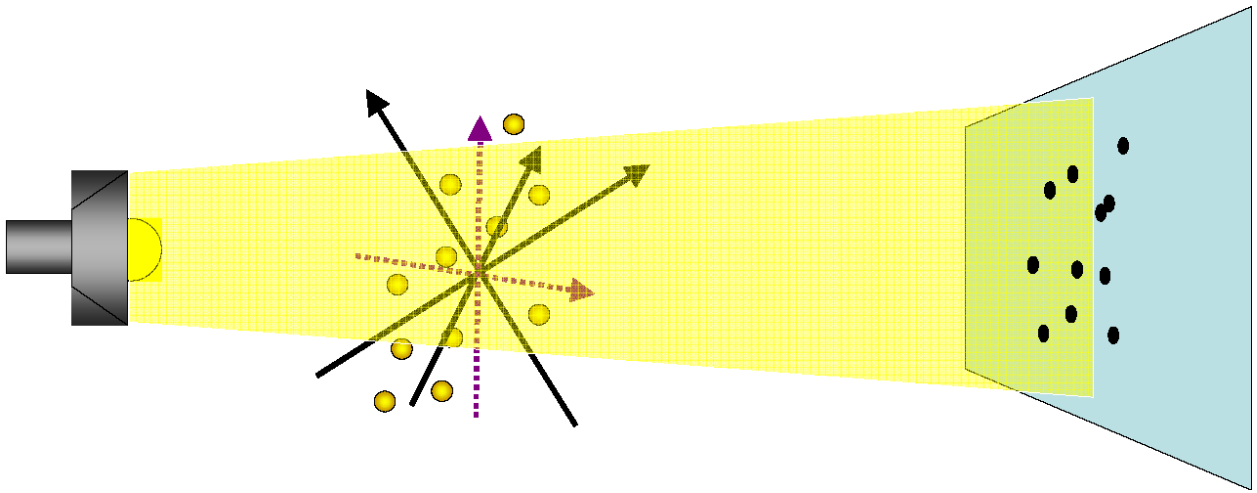
Méthodes factorielles

- Basées sur une représentation des données (individus) dans un espace de dimension inférieure
- Utilisent essentiellement des résultats d'algèbre linéaire : changement de base, diagonalisation de matrices, etc.
- Analyse en composantes principales - ACP / (*PCA - Principal Component Analysis*)
- Analyse factorielle des correspondances - AFC / (*CA - Correspondence Analysis*)



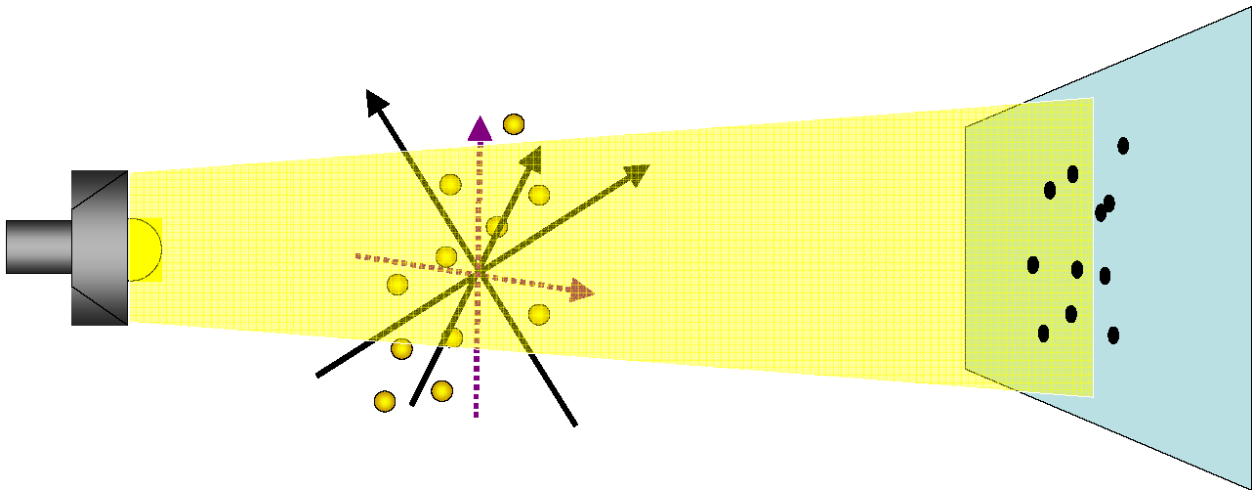
Méthodes factorielles

- Basées sur une représentation des données (individus) dans un espace de dimension inférieure
- Utilisent essentiellement des résultats d'algèbre linéaire : changement de base, diagonalisation de matrices, etc.
- Analyse en composantes principales - ACP / (*PCA - Principal Component Analysis*)
- Analyse factorielle des correspondances - AFC / (*CA - Correspondence Analysis*)



Méthodes factorielles

- Basées sur une représentation des données (individus) dans un espace de dimension inférieure
- Utilisent essentiellement des résultats d'algèbre linéaire : changement de base, diagonalisation de matrices, etc.
- Analyse en composantes principales - ACP / (*PCA - Principal Component Analysis*)
- Analyse factorielle des correspondances - AFC / (*CA - Correspondence Analysis*)



Objectifs de l'ACP

- Résumer un tableau individus $\times$ variables à l'aide d'un petit nombre de facteurs
- Visualiser le positionnement des individus les uns par rapport aux autres
- Visualiser les corrélations entre les variables
- Interpréter les facteurs

Visualisation des données

	X_1	$\cdots$	X_p
1	x_{11}	$\cdots$	x_{p1}
$\vdots$			
i	x_{1i}	$\cdots$	x_{pi}
$\vdots$			
n	x_{1n}	$\cdots$	x_{pn}

$\Rightarrow$

Y_1	Y_2	
y_{11}	y_{21}	
$\vdots$		
y_{1i}	y_{2i}	$\cdots$
$\vdots$		
y_{1n}	y_{2n}	

- X : les données
- Y : les composantes principales (non corrélées entre elles) :

$$Y_h = \sum_{j=1}^{p} u_{hj} X_j$$

Un exemple de positionnement de produits

- Fichier MS Excel/Open Office Calc auto_2004
- Caractéristiques de 24 modèles de voiture
- Résumé des données
- Tableau des corrélations

Nuage de points associé aux données

- Nuage de points

$$N = \{x_1, \cdots, x_i, \cdots, x_n\}$$

$x_i = (x_{1i}, \cdots, x_{pi})$, $i = \overline{1, n}$.

- Centre de gravité du nuage N :

$$g = \frac{1}{n}\sum_{i=1}^{n} x_i = (\bar{X}_1, \cdots, \bar{X}_p)$$

- Inertie totale du nuage N :

$$I(N, g) = \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, g) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ji} - \bar{X}_j)^2$$

$$= \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}(x_{ji} - \bar{X}_j)^2 = \sum_{j=1}^{p}\sigma_j^2$$

Réduction des données

- Pour faciliter le calcul on remplace les données d'origine par les données centrées-réduites de moyenne 0 et d'écart-type 1 :

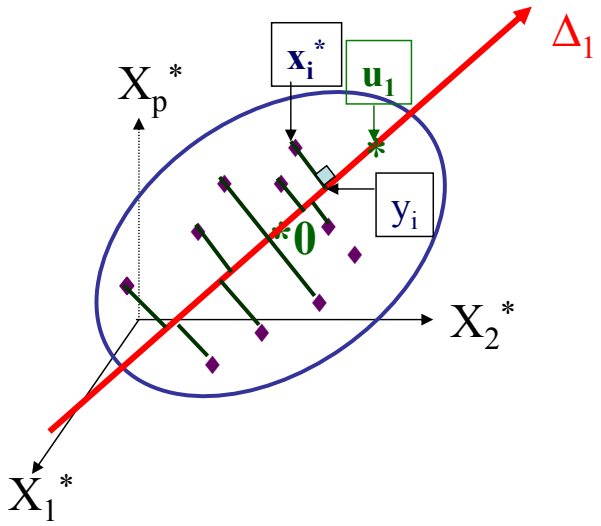
$$X_1^* = \frac{X_1 - \bar{X}_1}{\sigma_1}$$

$$\cdots$$

$$X_p^* = \frac{X_p - \bar{X}_p}{\sigma_p}$$

- Nuage de points associé aux données réduites $N^* = \{x_1^*, \cdots, x_i^*, \cdots, x_n^*\}$
- Centre de gravité $g^* = \vec{0}$
- Inertie totale $I(N^*, \vec{0}) = p$

Premier axe principal Δ_1

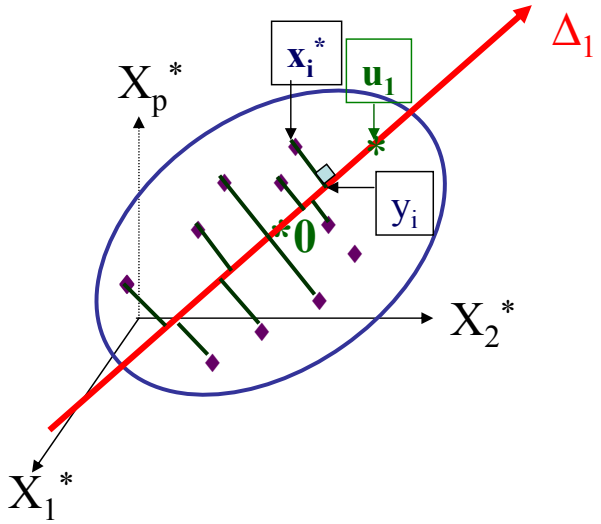


- Objectif 1: chercher l'axe Δ_1 passant le mieux possible au milieu du nuage N^*
  $\Rightarrow$ *minimiser* l'inertie du nuage N^* par rapport à l'axe Δ_1

$$I(N^*, \Delta_1) = \frac{1}{n} \sum_{i=1}^{n} d^2(x_i^*, y_i)$$

y_i - projection de x_i^* sur l'axe Δ_1

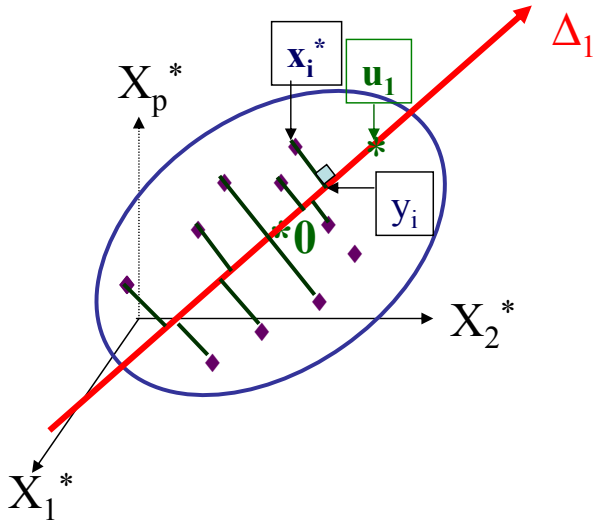
Premier axe principal Δ_1



- Objectif 2 : chercher l'axe d'allongement Δ_1 du nuage N^*
  $\Rightarrow$ *maximiser* l'inertie du nuage N^* projeté sur l'axe Δ_1

$$I(\{y_1, \cdots, y_n\}, \vec{0}) = \frac{1}{n} \sum_{i=1}^{n} d^2(y_i, \vec{0})$$

Les deux objectifs sont atteints simultanément



- De $d^2(x_i^*, \vec{0}) = d^2(y_i, \vec{0}) + d^2(x_i^*, y_i)$ on déduit

$$\frac{1}{n}\sum_{i=1}^{n} d^2(x_i^*, \vec{0}) = \frac{1}{n}\sum_{i=1}^{n} d^2(y_i, \vec{0}) + \frac{1}{n}\sum_{i=1}^{n} d^2(x_i^*, y_i)$$

$\Rightarrow$ Inertie totale p = Inertie expliquée par Δ_1 (à maximiser)
+ Inertie résiduelle (à minimiser)

Résultats

- L'axe Δ_1 passe par le centre de gravité 0 du nuage de points N^*.
- L'axe Δ_1 est engendré par le vecteur normé u_1, vecteur propre de la matrice des corrélations R associé à la plus grande valeur propre λ_1.
- L'inertie expliquée par l'axe Δ_1 est égal à λ_1.
- La part d'inertie expliquée par le premier axe principal Δ_1 est égal à λ_1/p.

Première composante principale Y_1

- Y_1 est une nouvelle variable définie pour chaque individu i par :

  $Y_1(i)$ = longueur algébrique du segment $0y_i$
  = coordonnées de y_i sur l'axe Δ_1
  = produit scalaire entre les vecteurs x_i^* et u_1
  = $\sum_{j=1}^{p} u_{1j} x_{ji}^*$

$$\Rightarrow \qquad Y_1 = \sum_{j=1}^{p} u_{1j} X_j^*$$

Propriétés de la première composante principale Y_1

- $Y_1 = u_{11}X_1^* + u_{12}X_2^* + \cdots + u_{1p}X_p^*$
- Moyenne de Y_1 = 0
- Variance de Y_1 = Inertie expliquée par $\Delta_1 = \lambda_1$
- $Cor(X_j, Y_1) = \sqrt{\lambda_1}u_{1j}$

$$\frac{1}{p}\sum_{j=1}^{p} Cor^2(X_j, Y_1) = \frac{\lambda_1}{p} \quad \text{est maximum}$$

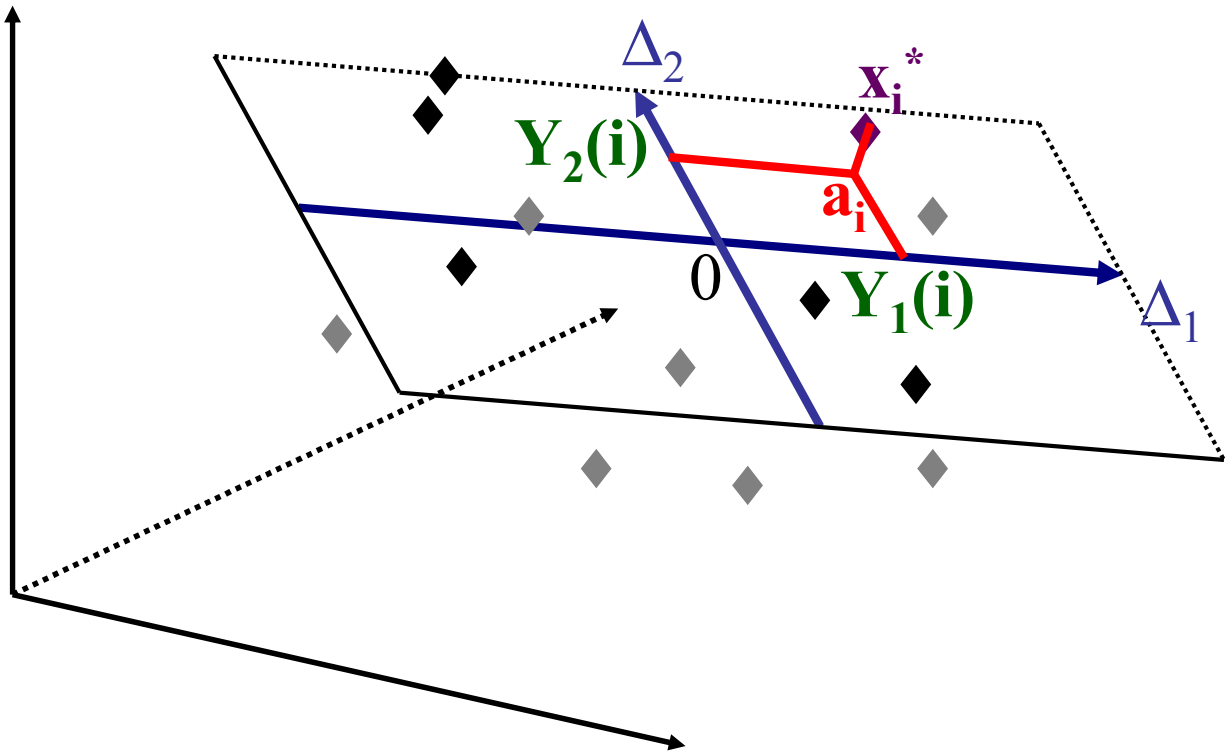
Qualité de la première composante principale

- Exemple auto_2004
- Inertie totale = 6
- Inertie expliquée par le premier axe principal = $\lambda_1 = 4.4113$
- Part d'inertie expliquée par le premier axe principal

$$\frac{\lambda_1}{p} = \frac{4.4113}{6} = 0.7352$$

- La première composante principale explique 73,5% de la variance totale

Deuxième axe principale Δ_2



Résultats

- On recherche le deuxième axe principal Δ_2 orthogonal à Δ_1 et passant le mieux possible au milieu du nuage.
- Il passe par le centre de gravité 0 du nuage de points et est engendré par le vecteur normé u_2, vecteur propre de la matrice des corrélations R associé à la deuxième plus grande valeur propre λ_2.
- La deuxième composante principale Y_2 est définie par projection des points sur le deuxième axe principal.
- La deuxième composante principale Y_2 est centrée, de variance λ_2, et non corrélée à la première composante principale Y_1

Qualité globale de l'analyse

- Inertie totale = variance totale = p
- Part de variance expliquée par la première composante principale

$$\frac{\lambda_1}{p}$$

- Part de variance expliquée par la deuxième composante principale

$$\frac{\lambda_2}{p}$$

- Part de variance expliquée par les deux premières composantes principales

$$\frac{\lambda_1 + \lambda_2}{p}$$

- Et ainsi de suite pour les autres dimensions...

Revenant à l'exemple - Tanagra

- auto_2004
- Tableau des valeurs propres
- Corrélations des variables actives avec les facteurs
- Tableau de vecteurs propres
- Données projetées sur les axes principaux
- Le premier plan principal (deux premiers axes)
- Le cercle des corrélations

Interprétation

- Cercle des corrélations et les données projetées sur le premier plan principal

Analyse factorielle des correspondances

- AFC simple : deux variables qualitatives
- AFC multiple : plus de deux variables qualitatives
- Objectifs : Représentation des variables qualitatives sur un espace multidimensionnel qui montre les corrélations entre ces variables.

AFC simple

- Données : Tableau de contingence
- Test du khi-deux d'indépendance
- ACP du tableau des profils-lignes
- ACP du tableau des profils-colonnes
- Représentation pseudo-barycentrique

Données

	1		j		p	
1	k_{11}		k_{1j}		k_{1p}	$k_{1.}$
i	k_{i1}		k_{ij}		k_{ip}	$k_{i.}$
n	k_{n1}		k_{nj}		k_{np}	$k_{n.}$
	$k_{.1}$		$k_{.j}$		$k_{.p}$	$k = \sum k_{ij}$

- Deux variables qualitatives X et Y
- Tableau de contingence : croisement de ces deux variables
- $k_{ij}/k_{i.}$: profils-lignes ; $k_{ij}/k_{.j}$: profils-colonnes.

AFC simple

- Données : Tableau de contingence
- Test du khi-deux d'indépendance
- ACP du tableau des profils-lignes
- ACP du tableau des profils-colonnes
- Représentation pseudo-barycentrique

Test du khi-deux d'indépendance (rappel)

- H_0 : Les variables X et Y sont indépendantes
- H_1 : Les variables X et Y sont liées
- Effectif attendu sous l'hypothèse d'indépendance $k_{i.}k_{.j}/k$

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(k_{ij} - k_{i.}k_{.j}/k)^2}{k_{i.}k_{.j}/k}$$

- On rejette H_0 au risque α de se tromper si $\chi^2 \geq \chi^2_{1-\alpha}[(n-1)(p-1)]$

Ce test ne donne pas d'information sur les associations par case, par ligne ou par colonne
$\Rightarrow$ Analyse factorielle des correspondances

Test du khi-deux d'indépendance (rappel)

- H_0 : Les variables X et Y sont indépendantes
- H_1 : Les variables X et Y sont liées
- Effectif attendu sous l'hypothèse d'indépendance $k_{i.}k_{.j}/k$

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(k_{ij} - k_{i.}k_{.j}/k)^2}{k_{i.}k_{.j}/k}$$

- On rejette H_0 au risque α de se tromper si $\chi^2 \geq \chi^2_{1-\alpha}[(n-1)(p-1)]$

Ce test ne donne pas d'information sur les associations par case, par ligne ou par colonne

$\Rightarrow$ Analyse factorielle des correspondances

AFC simple

- Données : Tableau de contingence
- Test du khi-deux d'indépendance
- ACP du tableau des profils-lignes
- ACP du tableau des profils-colonnes
- Représentation pseudo-barycentrique

ACP du tableau des profils-lignes

- Profil-ligne i : $f_J^i = \{k_{ij}/k_{i.}\}$
- Profil-ligne global : $f_J = \{f_{.j} = k_{.j}/k\}$
- Chaque ligne i a un poids $f_{i.}$
- f_J est le centre de gravité du nuage de points pondérés $\{f_J^i, f_{i.}\}$
- Distance du χ^2 entre les lignes i :

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

- Inertie totale (mesure de la dispersion des profils-lignes par rapport au centre de gravité)

$$\sum_{i=1}^{n} f_{i.} d^2(f_J^i, f_J) = \chi^2/k$$

- Les composantes principales F_1, F_2, ...

AFC simple

- Données : Tableau de contingence
- Test du khi-deux d'indépendance
- ACP du tableau des profils-lignes
- ACP du tableau des profils-colonnes
- Représentation pseudo-barycentrique

ACP du tableau des profils-colonnes

- Profil-colonne j : $f_I^j = \{k_{ij}/k_{.j}\}$
- Profil-colonne global : $f_I = \{f_{i.} = k_{i.}/k\}$
- Chaque colonne j a un poids $f_{.j}$
- f_I est le centre de gravité du nuage de points pondérés $\{f_I^j, f_{.j}\}$
- Distance du χ^2 entre les colonnes j :

$$d^2(f_I^j, f_I^{j'}) = \sum_{j=1}^{n} \frac{1}{f_{i.}} \left(\frac{k_{ij}}{k_{.j}} - \frac{k_{ij'}}{k_{.j'}}\right)^2$$

- Inertie totale (mesure de la dispersion des profils-colonnes par rapport au centre de gravité)

$$\sum_{j=1}^{p} f_{.j} d^2(f_I^j, f_I) = \chi^2/k$$

- Les composantes principales G_1, G_2, ...

AFC simple

- Données : Tableau de contingence
- Test du khi-deux d'indépendance
- ACP du tableau des profils-lignes
- ACP du tableau des profils-colonnes
- Représentation pseudo-barycentrique

Lien entre les deux analyses : les relations de transitions

- Les modalités en colonne sont au barycentre des modalités en ligne à $1/\sqrt{\lambda_h}$ près, où $\lambda_h = \text{Var}(F_h) = \text{Var}(G_h)$.

$$F_h(i) = \frac{1}{\sqrt{\lambda_h}} \sum_{j=1}^{p} \frac{k_{ij}}{k_{i.}} G_h(j)$$

- Les modalités en ligne sont au barycentre des modalités en colonne à $1/\sqrt{\lambda_h}$ près

$$G_h(j) = \frac{1}{\sqrt{\lambda_h}} \sum_{i=1}^{n} \frac{k_{ij}}{k_{.j}} F_h(i)$$

- Deux modalités qui vont plus souvent ensemble sont plus près l'une de l'autre.

Contributions des modalités i à la construction de F_1

- De

$$\text{Var}(F_1) = \sum_{i=1}^{n} f_{i.} F_1(i)^2 = \lambda_1$$

on déduit :

$$CTR_1(i) = \frac{f_{i.} F_1(i)^2}{\lambda_1}$$

$CTR_1(i)$ fort $\Leftrightarrow$

- Point fortement explicatif de F_1
- Point contribuant fortement à la construction de l'axe.

Contributions des modalités i à la construction de F_1

- De

$$\text{Var}(F_1) = \sum_{i=1}^{n} f_{i.} F_1(i)^2 = \lambda_1$$

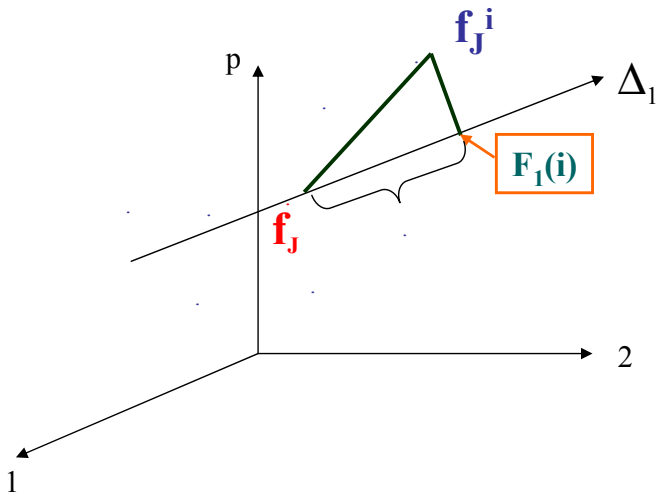
on déduit :

$$CTR_1(i) = \frac{f_{i.} F_1(i)^2}{\lambda_1}$$

$CTR_1(i)$ fort $\Leftrightarrow$

- Point fortement explicatif de F_1
- Point contribuant fortement à la construction de l'axe.

Qualité de représentation du point *i* sur le premier axe principal : Cosinus carré

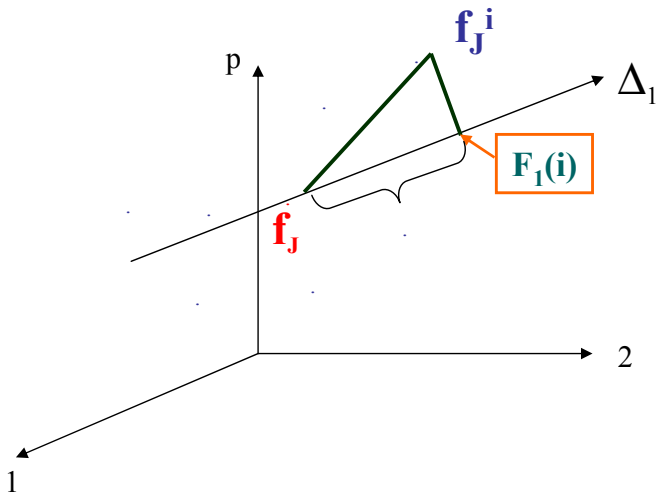


$$\cos_1^2(i) = \cos^2(f_J\vec{f}_J^i, \Delta_1) = \frac{F_1(i)^2}{d^2(f_J^i, f_J)}$$

$\cos_1^2(i)$ fort $\Leftrightarrow$

- Point fortement expliqué par l'axe Δ_1
- Point bien représenté sur l'axe Δ_1.

Qualité de représentation du point *i* sur le premier axe principal : Cosinus carré



$$\cos_1^2(i) = \cos^2(f_J\vec{f}_J^i, \Delta_1) = \frac{F_1(i)^2}{d^2(f_J^i, f_J)}$$

$\cos_1^2(i)$ fort $\Leftrightarrow$

- Point fortement expliqué par l'axe Δ_1
- Point bien représenté sur l'axe Δ_1.

AFCM

- Données
- Tableau de Burt
- Tableau des khi-deux
- Tableau disjonctif complet

$x_{ij_l} = 1$ si l'individu i possède la modalité l de la variable j
$= 0$ sinon

ANALYSE FACTORIELLE DES CORRESPONDANCES MULTIPLES DES VARIABLES x_1, x_2, ..., x_m
=
ANALYSE FACTORIELLE DES CORRESPONDANCES DU TABLEAU DISJONCTIF COMPLET

Exemple - Data

	Race	Taille	Poids	Vitesse	Intell.	Affect.	Agress.	Fonction
1	Beauceron	TA++	PO+	V++	INT+	AF+	AG+	Utilité
2	Basset	TA-	PO-	V-	INT-	AF-	AG+	Chasse
3	Berger-Allemand	TA++	PO+	V++	INT++	AF+	AG+	Utilité
4	Boxer	TA+	PO+	V+	INT+	AF+	AG+	Compagnie
5	Bull-Dog	TA-	PO-	V-	INT+	AF+	AG-	Compagnie
6	Bull-Mastiff	TA++	PO++	V-	INT++	AF-	AG+	Utilité
7	Caniche	TA-	PO-	V+	INT++	AF+	AG-	Compagnie
8	Chihuahua	TA-	PO-	V-	INT-	AF+	AG-	Compagnie
9	Cocker	TA+	PO-	V-	INT+	AF+	AG+	Compagnie
10	Colley	TA++	PO+	V++	INT+	AF+	AG-	Compagnie
11	Dalmatien	TA+	PO+	V+	INT+	AF+	AG-	Compagnie
12	Doberman	TA++	PO+	V++	INT++	AF-	AG+	Utilité
13	Dogue Allemand	TA++	PO++	V++	INT-	AF-	AG+	Utilité
14	Epagneul Breton	TA+	PO+	V+	INT++	AF+	AG-	Chasse
15	Epagneul Français	TA++	PO+	V+	INT+	AF-	AG-	Chasse
16	Fox-Hound	TA++	PO+	V++	INT-	AF-	AG+	Chasse
17	Fox-Terrier	TA-	PO-	V+	INT+	AF+	AG+	Compagnie
18	Grd Bleu de Gascogne	TA++	PO+	V+	INT-	AF-	AG+	Chasse
19	Labrador	TA+	PO+	V+	INT+	AF+	AG-	Chasse
20	Lévrier	TA++	PO+	V++	INT-	AF-	AG-	Chasse
21	Mastiff	TA++	PO++	V-	INT-	AF-	AG+	Utilité
22	Pékinois	TA-	PO-	V-	INT-	AF+	AG-	Compagnie
23	Pointer	TA++	PO+	V++	INT++	AF-	AG-	Chasse
24	Saint-Bernard	TA++	PO++	V-	INT+	AF-	AG+	Utilité
25	Setter	TA++	PO+	V++	INT+	AF-	AG-	Chasse
26	Teckel	TA-	PO-	V-	INT+	AF+	AG-	Compagnie

Exemple - Tableau de Burt

©TABLEAU DE BURT

	TA-	TA+	TA++	PO-	PO+	PO++	VE-	VE+	VE++	INT-	INT+	INT++	AF-	AF+	AG-	AG+	Comp	Chas	Util
TA-	7	0	0																
TA+	0	5	0																
TA0++	0	0	15																
PO-	7	1	0	8	0	0													
PO+	0	4	10	0	14	0													
PO++	0	0	5	0	0	5													
VE-	5	1	4	6	0	4	10	0	0										
VE+	2	4	2	2	6	0	0	8	0										
VE++	0	0	9	0	8	1	0	0	9										
IN-	3	0	5	3	3	2	4	1	3	8	0	0							
IN+	3	4	6	4	7	2	5	5	3	0	13	0							
IN++	1	1	4	1	4	1	1	2	3	0	0	6							
AF-	1	0	12	1	7	5	5	2	6	6	4	3	13	0					
AF+	6	5	3	7	7	0	5	6	3	2	9	3	0	14					
AG-	5	3	6	5	8	1	5	5	4	3	8	3	5	9	14	0			
AG+	2	2	9	3	6	4	5	3	5	5	5	3	8	5	0	13			
Comp	6	3	1	7	3	0	5	4	1	2	7	1	0	10	7	3	10	0	0
Chas	1	2	6	1	8	0	1	4	4	4	3	2	7	2	6	3	0	9	0
Util	0	0	8	0	3	5	4	0	4	2	3	3	6	2	1	7	0	0	8
	TA-	TA+	TA++	PO-	PO+	PO++	VE-	VE+	VE++	INT-	INT+	INT++	AF-	AF+	AG-	AG+	Comp	Chas	Util

Exemple - Tableau des khi-deux

	Poids	Vélocité	Intelligence	Affection	Agressivité	Fonction
Taille	25.3 (.000)	15.9 (.000)	3.6 (.46)	14.0 (.001)	2.1 (.36)	16.35 (.003)
Poids		18.4 (.001)	1.35 (.85)	9.5 (.008)	2.6 (.28)	24.41 (.000)
Vélocité			3.16 (.53)	3.0 (.23)	.57 (.75)	8.49 (.08)
Intelligence				3.9 (.14)	1.15 (.56)	4.14 (.39)
Affection					1.8 (.18)	14.76 (.000)
Agressivité						7.07 (.03)

AFCM

- Données
- Tableau de Burt
- Tableau des khi-deux
- Tableau disjonctif complet

$x_{ij_l} = 1$ si l'individu i possède la modalité l de la variable j
$= 0$ sinon

ANALYSE FACTORIELLE DES CORRESPONDANCES MULTIPLES DES VARIABLES x_1, x_2, ..., x_m
=
ANALYSE FACTORIELLE DES CORRESPONDANCES DU TABLEAU DISJONCTIF COMPLET

Notations

Les données

- n individus
- m variables qualitatives $x_1, \cdots, x_m$
- la variable x_j a p_j modalités $\{j_1, j_2, \cdots, j_{p_j}\}$
- $p = p_1 + p_2 + \cdots + p_m$
- la modalité j_l a une fréquence absolue n_{j_l} et une fréquence relative $n_{j_l}/(n \times m)$ dans le tableau disjonctif complet.

Résutats

- Inertie totale :
$$\frac{p}{m} - 1$$
- Dimension du nuage de points : $p - m$
- Moyenne des valeurs propres :
$$\frac{\frac{p}{m} - 1}{p - m} = \frac{1}{m}$$

$\Rightarrow$ On retient les axes associés à des valeurs propres supérieures à $1/m$.

Contribution des individus à la construction des axes

- Variance de F_h :

$$\text{Var}(F_h) = \frac{1}{n}\sum_{i=1}^{n} F_h(i)^2 = \lambda_h$$

- Contribution de l'individu i à la construction de l'axe h :

$$\text{Contribution}_h(i) = \frac{\frac{1}{n}F_h(i)^2}{\lambda_h}$$

Contribution des modalités à la construction des axes

- Variance de G_h :

$$\text{Var}(G_h) = \sum_{j=1}^{m} \sum_{l=1}^{p_j} \frac{n_{j_l}}{n \times m} G_h(j_l)^2 = \lambda_h$$

- Contribution de la modalité j_l à la construction de l'axe h :

$$\text{Contribution}_h(j_l) = \frac{\frac{n_{j_l}}{n \times m} G_h(j_l)^2}{\lambda_h}$$

Lien entre les deux analyses : les relations de transitions

- A $1/\sqrt{\lambda_h}$ près, chaque individu est représenté au barycentre de ses caractéristiques :

$$F_h(i) = \frac{1}{\sqrt{\lambda_h}} \frac{1}{m} \sum_{j=1}^{m} \sum_{l=1}^{p_j} x_{ij_l} G_h(j_l)$$

- A $1/\sqrt{\lambda_h}$ près, chaque modalité est représentée au barycentre de ses caractéristiques :

$$G_h(j_l) = \frac{1}{\sqrt{\lambda_h}} \frac{1}{n_{j_l}} \sum_{i=1}^{n} x_{ij_l} F_h(i)$$

- Conséquence : La représentation graphique des modalités d'une variable est centrée ($\sum_{l=1}^{p_j} n_{j_l} G_h(j_l) = 0$)

Lien entre les deux analyses : les relations de transitions

- A $1/\sqrt{\lambda_h}$ près, chaque individu est représenté au barycentre de ses caractéristiques :

$$F_h(i) = \frac{1}{\sqrt{\lambda_h}} \frac{1}{m} \sum_{j=1}^{m} \sum_{l=1}^{p_j} x_{ij_l} G_h(j_l)$$

- A $1/\sqrt{\lambda_h}$ près, chaque modalité est représentée au barycentre de ses caractéristiques :

$$G_h(j_l) = \frac{1}{\sqrt{\lambda_h}} \frac{1}{n_{j_l}} \sum_{i=1}^{n} x_{ij_l} F_h(i)$$

- Conséquence : La représentation graphique des modalités d'une variable est centrée ($\sum_{l=1}^{p_j} n_{j_l} G_h(j_l) = 0$)

Lien entre les deux analyses : les relations de transitions

- A $1/\sqrt{\lambda_h}$ près, chaque individu est représenté au barycentre de ses caractéristiques :

$$F_h(i) = \frac{1}{\sqrt{\lambda_h}} \frac{1}{m} \sum_{j=1}^{m} \sum_{l=1}^{p_j} x_{ij_l} G_h(j_l)$$

- A $1/\sqrt{\lambda_h}$ près, chaque modalité est représentée au barycentre de ses caractéristiques :

$$G_h(j_l) = \frac{1}{\sqrt{\lambda_h}} \frac{1}{n_{j_l}} \sum_{i=1}^{n} x_{ij_l} F_h(i)$$

- Conséquence : La représentation graphique des modalités d'une variable est centrée ($\sum_{l=1}^{p_j} n_{j_l} G_h(j_l) = 0$)

Excercice

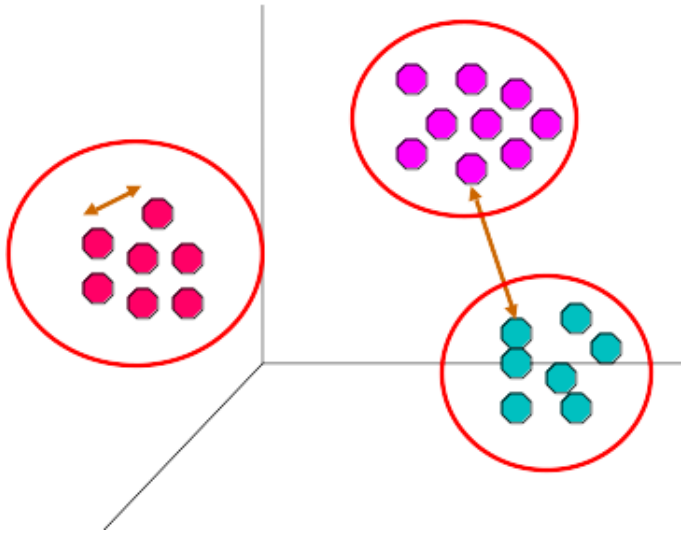
- Données
- ACP/AFCM

Plan

Objectif

Constituer des groupes d'objets *homogènes et différenciés*, *i.e.* des groupes d'objets tels que :

- les objets soient les plus similaires possibles au sein d'un groupe (critère de **compacité**),
- les groupes soient aussi dissemblables que possible (critère de **séparabilité**),
  Besoin d'une mesure pour la ressemblance ou la dissemblance sur l'ensemble des variables descriptives



Application de la classification

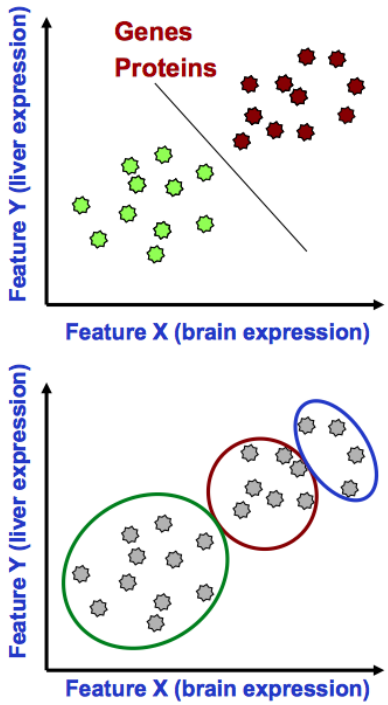
- Compréhension des données
  - Regrouper les documents liés
  - Regrouper des gènes et protéines ayant des fonctions similaires
  - Regrouper des stocks avec des fluctuations de prix similaires
  - ...
- Résumé des données
  - Réduction de la taille des grands jeux de données

Classification (*Clustering*) et classement (*Classification*)

- Classification : regrouper les objets similaires, **détecter la structure cachée** des données ;
- Classement : extraire des caractéristiques de données permettant **le classement de nouveaux individus** dans les classes prédéfinies.

Classification et classement

- Les objets sont décrits par un ou plusieurs caractéristiques
- **Classement (apprentissage supervisé)**
  - Certains points de données étiquetés
  - Besoin de règles pour étiqueter précisément des nouveaux points
  - Sous-problème : sélectionner des caractéristiques
  - Mesure : précision du classement
- **Classification (apprentissage non supervisé)**
  - Pas d'étiquettes préalables
  - Regrouper les points en basant sur leur "proximité"
  - Déterminer la structure des données
  - Mesure : des caractéristiques de validation indépendante



Types de classification

- Représentation des groupes
  - Partition
  - Arbre hiérachique
- Distinction de groupes :
  - Exclusive vs. non-exclusive : Chaque point appartient à un seul ou plusieurs groupes
  - Floue vs. non floue : Chaque point appartient à un groupe avec ou sans poids probabiliste
  - Hétérogène vs. homogène au niveau de taille, forme ou densité

Les étapes d'une classification automatique

1. Collecte des données
2. Calcul des ressemblances entre les n individus à partir du tableau initial
3. Choix d'un algorithme de classification et exécution
4. Interprétation des résultats :
   - évaluation de la qualité de la classification,
   - description des classes obtenues.

Chaque classification est établie à partir de variables et de méthodes choisies intentionnellement

Les étapes d'une classification automatique

1. Collecte des données
2. Calcul des ressemblances entre les n individus à partir du tableau initial
3. Choix d'un algorithme de classification et exécution
4. Interprétation des résultats :
   - évaluation de la qualité de la classification,
   - description des classes obtenues.

Chaque classification est établie à partir de variables et de méthodes choisies intentionnellement

Intérêts de la classification

- Les classes obtenues assurent une vue concise et structurée des données.
- Des regroupements inattendus apparaissent.
- Des regroupements attendus n'existent pas.
- Les classes significatives entraînent la définition de fonctions de décision permettant d'attribuer un nouvel individu à la classe dont il est le plus proche.

Dissimilarité - Agrégation

Toutes les méthodes nécessitent

- un critère de dissimilarité (distance) entre éléments (ou entre groupes - *clusters*)
- un critère d'agrégation des éléments (ou entre clusters)

Indice de dissimilarité

Mesurer la différence entre éléments

- Soit E l'ensemble des n objets à classer
- Une dissimilarité $d : E \times E \to R^+$

$$\begin{array}{lll} 1. & d(i,i) = 0 & \forall i \in E \\ 2. & d(i,i') = d(i',i) & \forall i, i' \in E \times E \end{array}$$

Une distance satisfait les propriétés d'un indice de dissimilarité

Indice de dissimilarité

Mesurer la différence entre éléments

- Soit E l'ensemble des n objets à classer
- Une dissimilarité $d : E \times E \to R^+$

$$\begin{array}{lll} 1. & d(i,i) = 0 & \forall i \in E \\ 2. & d(i,i') = d(i',i) & \forall i, i' \in E \times E \end{array}$$

Une distance satisfait les propriétés d'un indice de dissimilarité

Exemples pour des variables binaires (1/3)

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
P_1	1	0	0	1	1	0	1	1
P_2	1	1	0	1	0	1	0	1
P_3	1	1	1	1	0	0	1	1
P_4	0	1	0	0	0	1	0	1
P_5	0	0	1	1	0	1	0	1

n_{ij} = nombre d'accords positifs (11)

$n_{\overline{ij}}$ = nombre d'accords négatifs (00)

q_{ij} = nombre de désaccords (01) ou (10)

- Fonction de similarité : mesurer la ressemblance entre éléments
  - croissante avec les accords
  - décroissante avec les désaccords

$$\forall e_i, e_j \in E \times E \ : \qquad S(e_i, e_j) = f(n_{ij}, n_{\overline{ij}}, q_{ij})$$

Exemples pour des variables binaires (2/3)

$$S_\theta(e_i, e_j) = \frac{n_{ij}}{\theta n_{ij} + q_{ij}}$$

$\theta = 0$ – indice de Kulczinsky
$\theta = 1$ – indice de Jaccard
$\theta = 1/2$ – indice d'Anderberg
$\theta = 2$ – indice de Dice

$$S_{\alpha,\beta}(e_i, e_j) = \frac{n_{ij} - \alpha q_{ij} + n_{\overline{ij}}}{n_{ij} + \beta q_{ij} + n_{\overline{ij}}}$$

$\alpha = 0, \quad \beta = 2$ – indice de Rogers & Tanimoto
$\alpha = 0, \quad \beta = 1/2$ – indice de Sokal & Sneath
$\alpha = 1, \quad \beta = 1$ – indice de Hamman
$\alpha = 0, \quad \beta = 1$ – indice du simple matching

Exemples pour des variables binaires (2/3)

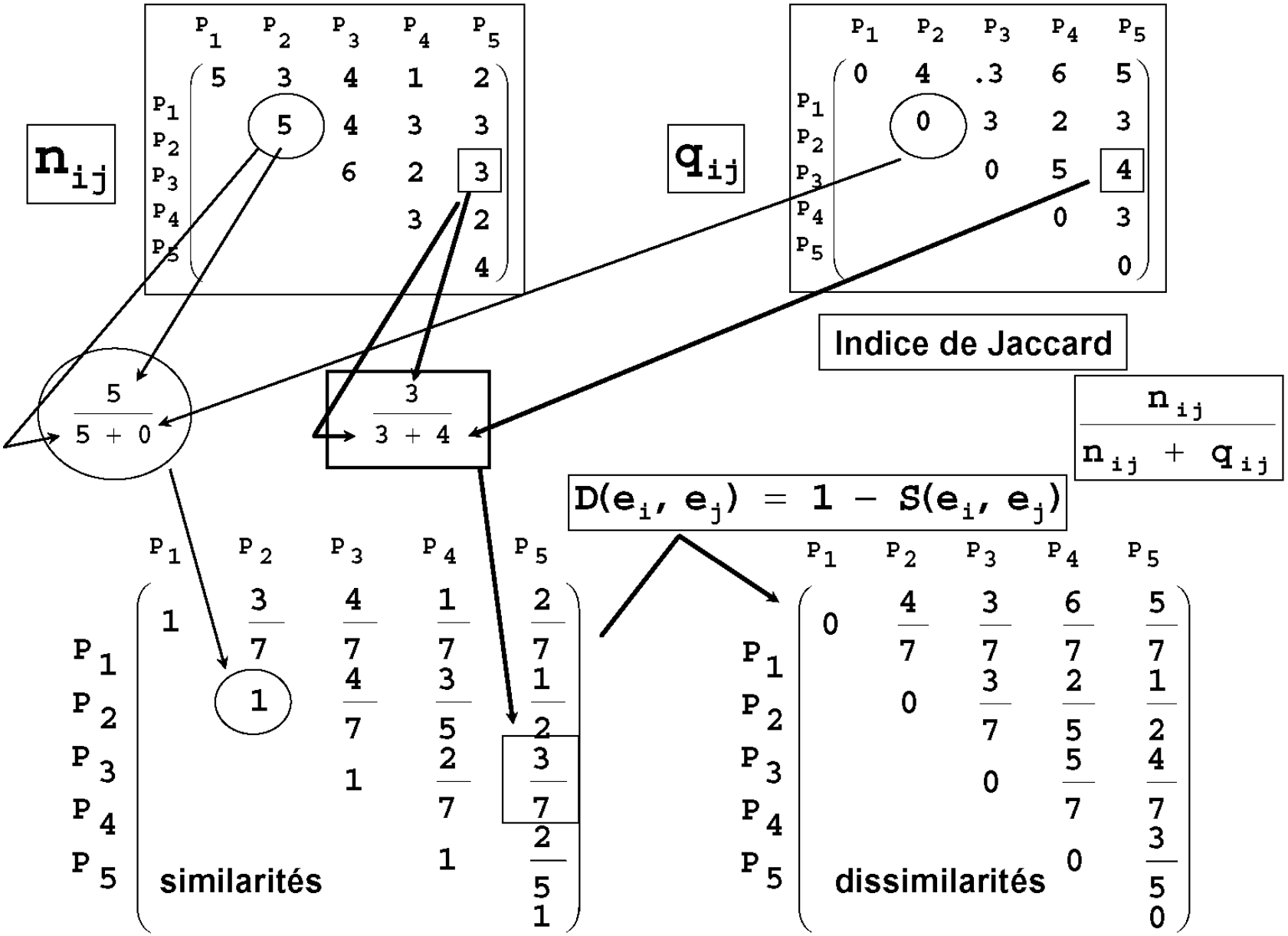
$$S_\theta(e_i, e_j) = \frac{n_{ij}}{\theta n_{ij} + q_{ij}}$$

$\theta = 0$ – indice de Kulczinsky
$\theta = 1$ – indice de Jaccard
$\theta = 1/2$ – indice d'Anderberg
$\theta = 2$ – indice de Dice

$$S_{\alpha,\beta}(e_i, e_j) = \frac{n_{ij} - \alpha q_{ij} + n_{\overline{ij}}}{n_{ij} + \beta q_{ij} + n_{\overline{ij}}}$$

$\alpha = 0, \quad \beta = 2$ – indice de Rogers & Tanimoto
$\alpha = 0, \quad \beta = 1/2$ – indice de Sokal & Sneath
$\alpha = 1, \quad \beta = 1$ – indice de Hamman
$\alpha = 0, \quad \beta = 1$ – indice du simple matching

Exemples pour des variables binaires (3/3)



Exemple pour des variables qualitatives

Distance du χ^2

$$D^2(e_i, e_j) = \sum_{k=1}^{p} \frac{1}{X_{.k}} \left(\frac{X_{ik}}{X_{i.}} - \frac{X_{jk}}{X_{j.}} \right)^2$$

- adaptée aux tableaux de contingence
- supprime l'effet de taille des lignes et colonnes

(C'est la distance retenue dans l'AFC)

Exemple pour des variables qualitatives

Distance du χ^2

$$D^2(e_i, e_j) = \sum_{k=1}^{p} \frac{1}{X_{.k}} \left(\frac{X_{ik}}{X_{i.}} - \frac{X_{jk}}{X_{j.}} \right)^2$$

- adaptée aux tableaux de contingence
- supprime l'effet de taille des lignes et colonnes

(C'est la distance retenue dans l'AFC)

Exemple pour des variables quantitatives

Distance de Minkowski

$$D(e_i, e_j) = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q \right]^{1/q}$$

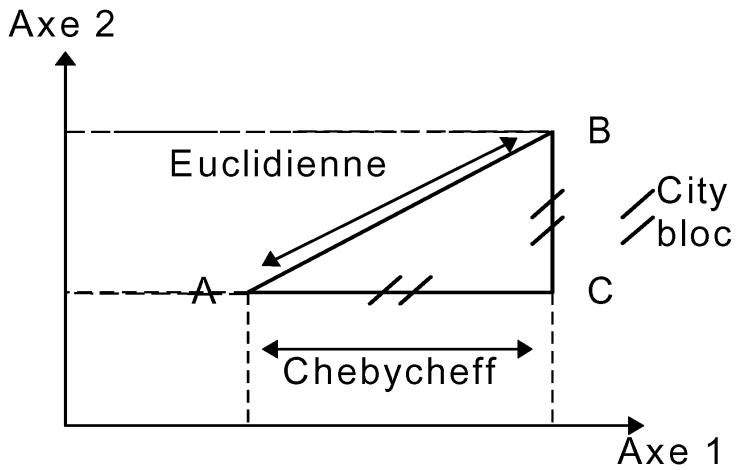
- $q = 2$: distance euclidienne, $q = 1$: distance du city-block
- $q \to +\infty$: distance de Chebycheff $= max_k(|x_{ik} - x_{jk}|)$

Exemple pour des variables quantitatives

Distance de Minkowski

$$D(e_i, e_j) = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q \right]^{1/q}$$

- $q = 2$: distance euclidienne, $q = 1$: distance du city-block
- $q \to +\infty$: distance de Chebycheff $= max_k(|x_{ik} - x_{jk}|)$



(changement avec rotation des axes pour city-block ou Chebycheff)

Critères d'agrégation (1/2)

- Lorsqu'on a regroupé 2 ou plusieurs individus, il faut recalculer la distance de ce groupe aux autres éléments.
- Pour cela, on a besoin d'un critère d'agrégation.

Critères d'agrégation (2/2)

- Saut minimum (simple lien)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \min\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Diamètre de la réunion (lien complet)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \max\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Critère de la moyenne (lien moyen)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\sum_{i \in a, i' \in a'} d(\{i\}, \{i'\})}{\text{card}(a) \times \text{card}(a')}$$

- Critère de Ward (inertie expliquée)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\text{card}(a) \times \text{card}(a')}{\text{card}(a) + \text{card}(a')} d^2(g_a, g_{a'})$$

Critères d'agrégation (2/2)

- Saut minimum (simple lien)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \min\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Diamètre de la réunion (lien complet)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \max\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Critère de la moyenne (lien moyen)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\sum_{i \in a, i' \in a'} d(\{i\}, \{i'\})}{\text{card}(a) \times \text{card}(a')}$$

- Critère de Ward (inertie expliquée)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\text{card}(a) \times \text{card}(a')}{\text{card}(a) + \text{card}(a')} d^2(g_a, g_{a'})$$

Critères d'agrégation (2/2)

- Saut minimum (simple lien)

$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \min\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$

- Diamètre de la réunion (lien complet)

$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \max\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$

- Critère de la moyenne (lien moyen)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\sum_{i \in a, i' \in a'} d(\{i\}, \{i'\})}{\text{card}(a) \times \text{card}(a')}$$

- Critère de Ward (inertie expliquée)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\text{card}(a) \times \text{card}(a')}{\text{card}(a) + \text{card}(a')} d^2(g_a, g_{a'})$$

Critères d'agrégation (2/2)

- Saut minimum (simple lien)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \min\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Diamètre de la réunion (lien complet)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \max\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Critère de la moyenne (lien moyen)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\sum_{i \in a, i' \in a'} d(\{i\}, \{i'\})}{\text{card}(a) \times \text{card}(a')}$$

- Critère de Ward (inertie expliquée)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\text{card}(a) \times \text{card}(a')}{\text{card}(a) + \text{card}(a')} d^2(g_a, g_{a'})$$

Critères d'agrégation (2/2)

- Saut minimum (simple lien)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \min\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Diamètre de la réunion (lien complet)

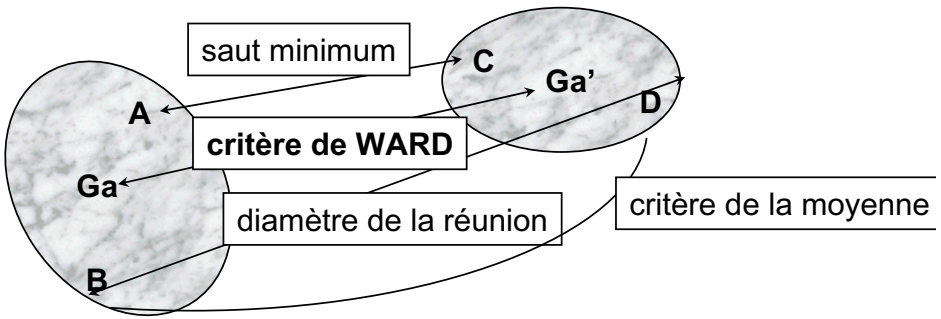
$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \max\{d(\{i\}, \{i'\}) | i \in a, i' \in a'\}$$

- Critère de la moyenne (lien moyen)

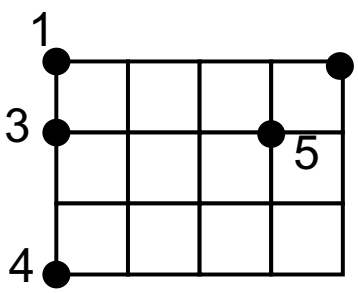
$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\sum_{i \in a, i' \in a'} d(\{i\}, \{i'\})}{\text{card}(a) \times \text{card}(a')}$$

- Critère de Ward (inertie expliquée)

$$\forall a \subset I, \forall a' \subset I \quad nv(a, a') = \frac{\text{card}(a) \times \text{card}(a')}{\text{card}(a) + \text{card}(a')} d^2(g_a, g_{a'})$$



Exemple d'application de la classification (1/2)



distances euclidiennes au carré

	1	2	3	4	5
1	0	16	**1**	9	10
2		0	17	25	2
3			0	4	9
4				0	13
5					0

Agrégation de 1 et 3 noeud 6

Critère d'agrégation : **saut minimum**

$$nv(6,\{2\}) = \inf\{d(\{1\},\{2\}), d(\{3\},\{2\})\} = \inf\{16,17\} = 16$$
$$nv(6,\{4\}) = \inf\{d(\{1\},\{4\}), d(\{3\},\{4\})\} = \inf\{9,4\} = 4$$
$$nv(6,\{5\}) = \inf\{d(\{1\},\{5\}), d(\{3\},\{5\})\} = \inf\{10,9\} = 9$$

	2	4	5	6
2	0	25	**2**	16
4		0	13	4
5			0	9
6				0

étape suivante agrégation de 2 et 5 noeud 7 $d^2=2$

$$nv(7,\{4\}) = \inf\{d(\{2\},\{4\}), d(\{5\},\{4\})\} = \inf\{25,13\} = 13$$
$$nv(7,6) = \inf\{d(\{2\},\{1\}), d(\{5\},\{1\}), d(\{2\},\{3\}), d(\{5\},\{3\})\} = \inf\{16,10,17,9\} = 9$$

	4	6	7
4	0	**4**	13
6		0	9
7			0

étape suivante agrégation de 4 et 6 noeud 8 $d^2=4$

$$nv(7,8) = \inf\begin{Bmatrix} d(\{2\},\{1\}), d(\{5\},\{1\}), \\ d(\{2\},\{3\}), d(\{5\},\{3\}), \\ d(\{2\},\{4\}), d(\{5\},\{4\}) \end{Bmatrix} = \inf\begin{Bmatrix} 16,10, \\ 17,9, \\ 25,13 \end{Bmatrix} = 9$$

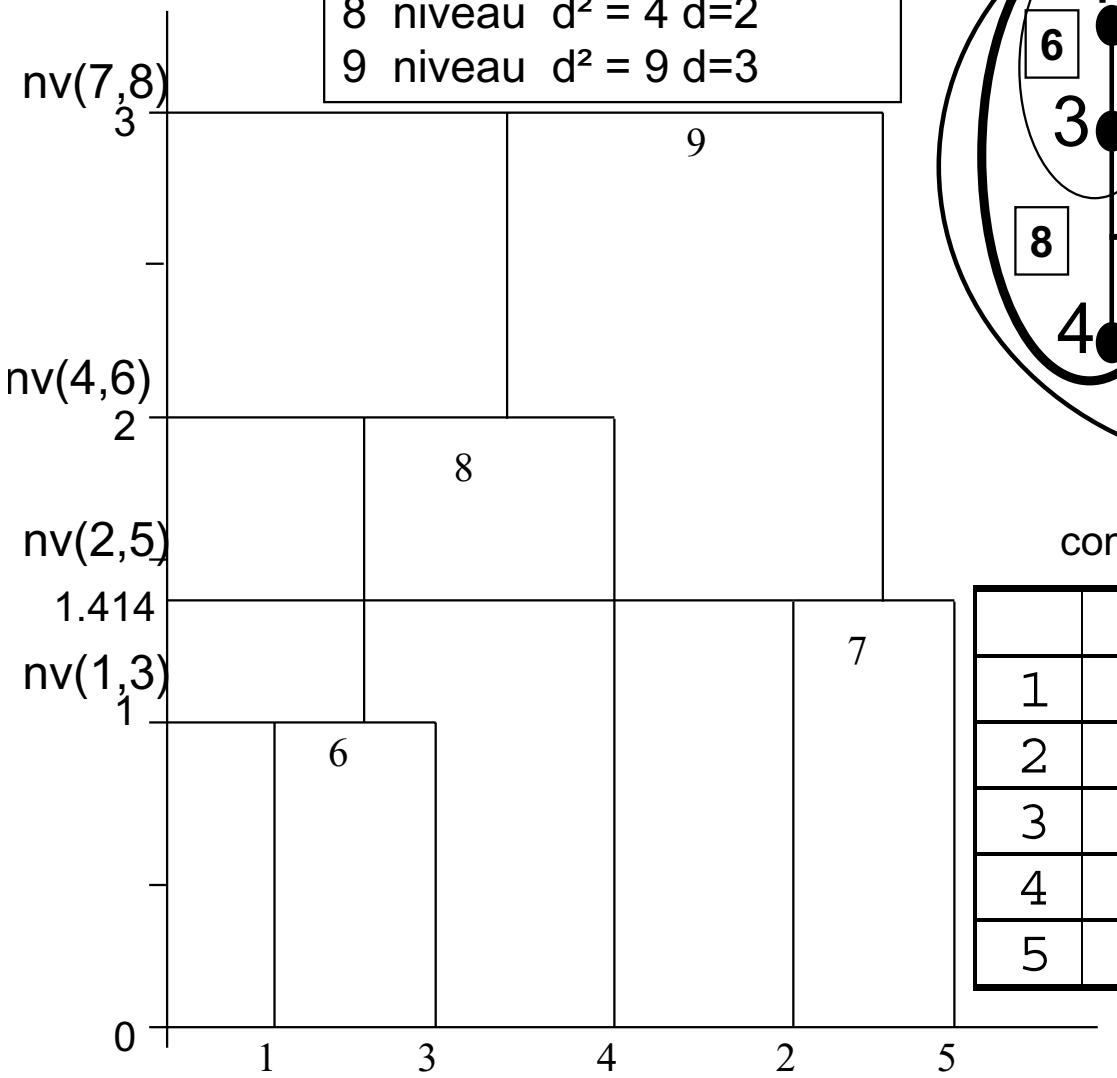
	7	8
7	0	9
8		0

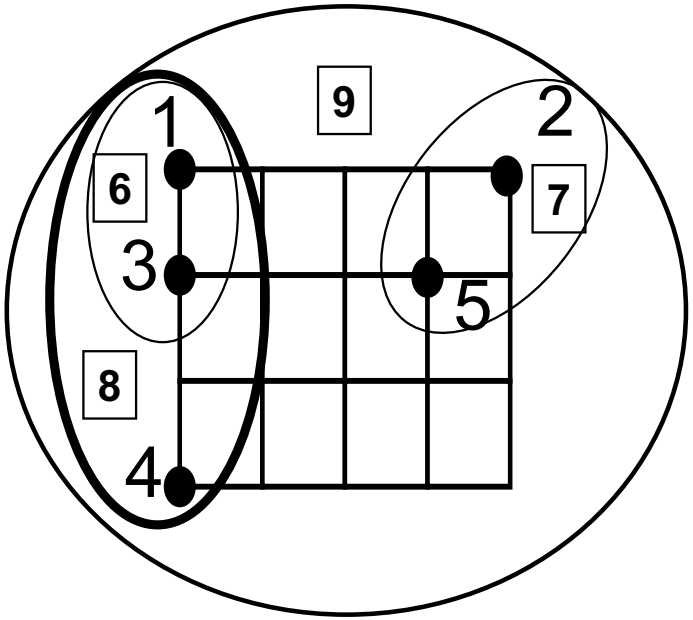
étape suivante agrégation de 7 et 8 noeud 9 $d^2=9$

Fin de la classification

Exemple d'application de la classification (2/2)

6 niveau $d^2 = 1$ d=1
7 niveau $d^2 = 2$ d=1.414
8 niveau $d^2 = 4$ d=2
9 niveau $d^2 = 9$ d=3





comparaison des distances

	1	2	3	4	5
1	0	16/9	1/1	9/4	10/9
2		0	17/9	25/9	2/2
3			0	4/4	9/9
4				0	13/9
5					0

Méthodes de classification

Différents algorithmes

- Classification hiérarchique
- Partitionnement autour de centres
- Analyse floue

Les deux derniers donnent directement des partitions

Mise en oeuvre en général

Principes

- Réduction de dimension (analyses factorielles)
- Puis clustering

Complexité de la classification

- Nombre de partitions de n objets est le nombre de Bell :

$$B_{n+1} = \sum_{k=0}^{n} C_n^k B_k = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k}$$

- Par exemple :
  pour $n = 4$, $B_n = 15$
  pour $n = 40$, $B_n = 8.47 * 10^{23}$.

On ne peut pas tout essayer !

Choix des paramètres et évaluation

- Nombre de classes ?
- Qualité de la classification ?

Critères

Une bonne classification doit :

- Détecter les structures présentes dans les données
- Fournir des classes bien différenciées
- Permettre de déterminer le nombre optimal de classes
- Fournir des classes stables
- Traiter tous les types de données
- Traiter efficacement les gros volumes de données

Inertie et qualité

- Inerties intra-classe (I_A) et inter-classe (I_R)

$$I_A = \sum_j \left(\sum_{i \in c_j} p_i (x_i - \bar{x}_j)^2 \right) \qquad I_R = \sum_j \left(\sum_{i \in c_j} p_i \right) (\bar{x}_j - \bar{x})^2$$

- Qualité :
  - Proportion de variance expliquée par les classes (inertie inter-classe) maximale : R^2
  - Autres critères : quantité d'information, entropie, vraisemblance, etc.

Partionnement direct (classes disjointes)

- Centres mobiles, k-means, MND (méthode des nuées dynamiques)
- K-modes, k medoids (PAM, CLARA, CLARANS, ...)
- Méthodes basées sur la densité
- Méthode neuronale (cartes de Kohonen)
- Mélange de distributions...

Les méthodes de partitionnement direct

- Opérer au sens d'un critère donné, le « meilleur » regroupement possible des individus en **un nombre choisi a priori de classes**.
- Principe de ces méthodes : constitution de k groupes (k étant un nombre choisi par l'analyste) à partir des n individus sur la base d'un algorithme itératif « Recentrage/Réallocation » en essayant d'optimiser un indice global mesurant la qualité de la classification.

Remarque : $(2^{n-1} - 1)$ partitions de n individus en 2 classes, $\Rightarrow$ exploration intelligente, appelee encore *heuristique*.

Méthode des centres mobiles

1. On choisit k individus comme centres initiaux des classes
2. On calcule les distances entre chaque individu et chaque centre c_i de l'étape précédente, et on affecte chaque individu au centre le plus proche, ce qui définit k classes
3. On remplace les k centres c_i par les barycentres des k classes définies à l'étape 2
4. On regarde si les centres sont restés suffisamment stables ou si un nombre fixé d'itérations a été atteint :
   - si oui, on arrête (en général, après au moins une dizaine d'itérations)
   - si non, on revient à l'étape 2

Variante de la méthode des centres mobiles

- k-means
  - le barycentre de chaque groupe est recalculé à chaque nouvel individu introduit dans le groupe, au lieu d'attendre l'affectation de tous les individus
  - la convergence est parfois possible en une seule itération $\Rightarrow$ plus grande rapidité
  - les résultats risquent de dépendre de l'ordre du fichier !
- Nuées dynamiques
  - chaque classe n'est plus représentée par son barycentre (éventuellement extérieur à la population), mais par un sous-ensemble de la classe, appelé noyau, qui, s'il est bien composé (des individus les plus centraux, par exemple), sera plus représentatif de la classe que son barycentre

Méthodes des *k*-medoids

PAM (*Partitioning Around Medoids*)

1. On choisit aléatoirement k individus représentatifs des classes, puis associe chaque individu à la classe ayant l'individu choisi le plus similaire.
2. Pour chaque paire de l'individu non-choisi h et l'individu choisi i, on calcule le coût total d'échange h et i: $TC_{ih} = \sum_{j \texttt{non choisi}} (d_{jh} - d_{ji})$
3. On choisit la paire i et h ayant le minimum TC
   - Si $TC_{ih} < 0$, i est remplacé par h;
   - Puis on attribue chaque individu non-choisi à l'individu représentatif le plus similaire
4. On répète les étapes 2 - 3 jusqu'à l'obtention des medoids stables.

Méthodes hiérarchiques

Deux classes disjointes ou bien l'une contient l'autre

- Agglomératives (CAH)
- Divisives (classification descendante)
- Méthodes mixtes : par exemple *fuzzy clustering*

Algorithme de la classification ascendante hiérarchique

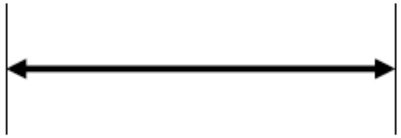
- Phase préalable : Calcul des ressemblances des objets 2 à 2
- Entrées : $n(n-1)/2$ ressemblances
- Jusqu'au regroupement de tous les objets en un seul groupe : $n-1$ étapes
  1. Regroupement des 2 éléments les plus proches
  2. Calcul d'une nouvelle matrice de ressemblances entre les éléments (objets isolés ou groupes) restants
- On obtient un arbre hiérachique indicé
- Dans cette catégorie : Cure, Rock, Birch, Cameleon

Données dans les espaces de grandes dimensions

- Adaptation des k-means
- Birch - Balanced Iterative Reducing and Clustering using Hierachies http://www.cs.sfu.ca/CourseCentral/459/han/papers/zhang96.pdf
- PDDP - Principal Direction Divisive Partitioning
- DBSCAN - Density-Based Spatial Clustering of Applications with Noise
- Piège : fléau de la dimension

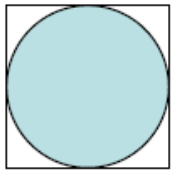
Piège

- une dimension



$1 / 1 = 1$

- deux dimensions



$3,14 \ / \ 2^2 = 0.798$

- Trois dimensions

$(4/3 \times 3,14) \ / \ 2^3 = 0.52$

- « p » dimensions

Volume de l'hyper-sphère / volume de l'hyper cube $\rightarrow 0$

… Les points deviennent équidistants …

Nombre de classes

- Dans un arbre hiérarchique, utilisation de l'indice de similarité
- Sinon, utilisation de critères liés à l'inertie
- Graphique 'silhouette'

Caractérisation des classes

- **Aides à l'interprétation d'une partition :**
  Une partition est considérablement enrichie par une description des classes à l'aide des individus et des variables

Interprétation par les individus

Pour chaque classe, on examine :

- son effectif,
- son diamètre (distance entre les 2 points les plus éloignés),
- la séparation (distance minimum entre la classe considérée et la classe la plus proche) et le numéro de la classe la plus proche,
- les identités des individus les plus proches du centre de gravité de la classe ou « parangons »,
- les identités des individus les plus éloignés du centre de gravité de la classe ou « extrêmes ».

Interprétation par les variables continues

Comparaison de la moyenne $\bar{x}_k$ et de l'écart-type s_k d'une variable X dans la classe k à la moyenne générale et à l'écart-type général.

Interprétation par les variables qualitatives

	Classe k	Autres classes	Population
Modalité j	n_{kj}	*	n_j
Autres modalités	*	*	*
Population	n_k	*	n

Pourcentage global $\Rightarrow n_j/n$

Pourcentage " mod / clas " $\Rightarrow n_{kj}/n_k$

Pourcentage " clas / mod " $\Rightarrow n_{kj}/n_j$

Valeur-test

- **Ces statistiques sur les variables peuvent être converties en un critère appelé “ valeur-test “.**
  La valeur-test permet de sélectionner les variables continues ou les modalités des variables nominales les plus caractéristiques de chaque classe.

Valeur-test pour les variables continues

La valeur-test est égale à l'écart entre la moyenne dans la classe et la moyenne générale exprimée en nombre d'écarts-types :

$$\text{v-test} = \frac{\bar{x}_k - \bar{x}}{s_k(X)}$$

avec

$$s_k^2(X) = \frac{n - n_k}{n - 1} \cdot \frac{s^2(X)}{n_k}$$

Valeur-test pour les variables nominales

Valeur-test de la modalité k de la variable j :

$$\text{v-test} = \frac{n_{jk} - n_k \cdot \frac{n_j}{n}}{\sqrt{n_k \cdot \frac{n - n_k}{n-1} \cdot \frac{n_j}{n} \cdot \left(1 - \frac{n_j}{n}\right)}}$$

Interprétation de la valeur-test

Si |v-test| > 2, la moyenne ou la proportion dans la population globale diffèrent significativement de celle dans la classe.

- Cette interprétation n'a de sens que pour les variables supplémentaires n'ayant pas participé à la construction des classes : il n'y a pas indépendance entre les classes d'une partition et une des variables ayant servi à définir la partition.
- Pour les variables actives, les valeurs-test constituent de simples mesures de similarité entre variables et classes.

Pratique de la classification (1/3)

- Pour une classification ascendante hiérarchique, on coupe l'arbre hiérarchique de façon à avoir des classes les plus homogènes possibles tout en étant bien séparées entre elles en se référant à l'histogramme des indices de niveau.
- La stratégie « Analyse factorielle + Classification » permet d'éliminer les fluctuations aléatoires et d'obtenir des classes plus stables, les axes factoriels étant très stables relativement à l'échantillonnage.

Pratique de la classification (2/3)

- La stratégie « Classification mixte », consistant à pratiquer une classification ascendante hiérarchique sur quelques dizaines de groupes homogènes obtenus par un algorithme d'agrégation autour de centres mobiles, est bien adaptée au partitionnement d'un ensemble comprenant un grand nombre d'individus (des milliers, voire des dizaines de milliers).
- L'homogénéité des classes obtenues peut être optimisée par une procédure de consolidation qui consiste à utiliser de nouveau une procédure d'agrégation autour des centres mobiles.

Pratique de la classification (3/3)

- La méthode de Ward s'allie efficacement avec les constructions de partition du type « Réaffectation / Recentrage » en fournissant une partition initiale de bonne qualité. L'exigence de variables quantitatives pour cette méthode peut être satisfaite grâce à un traitement préalable par analyse factorielle.

Complémentarité entre analyse factorielle et classification

La classification (dans l'espace entier) permet de " voir " au-delà du plan factoriel.

Classification sur variables

- La classification sur individus, afin de les regrouper en un nombre restreint de classes représentatives, est la plus utilisée, mais on peut aussi faire, après avoir transposé le fichier de données, une classification sur variables afin de réduire leur nombre et éventuellement, étudier leurs redondances.
- On peut utiliser le cosinus (corrélation), l'AVL (algorithme de vraisemblance du lien) de Lerman, la méthode de Bertin (graphique AMADO)

Quelques exemples

- Classification ascendante hiérarchique des variables (auto_2004.xls)
- Classification des voitures en 3 classes
- Caractérisation des classes par les variables continues

Plan

Modèle supermarché - panier de la ménagère

- Grands ensembles de données
  - Un grand ensemble d'items $A = \{A_1, A_2, \cdots, A_m\}$
    Ex : produits en vente dans un supermarché
  - Un grand ensemble de paniers $B = \{B_1, B_2, \cdots, B_n\}$
    - Chaque panier : petit sous-ensemble de l'ensemble d'items A
    - Ex : B_i - ensemble de produits achetés par un client dans une transaction
- Le compteur de support: sup(X) = nombre de paniers contenant le sous-ensemble d'items X
- Question : trouver des ensembles d'items qui apparaissent fréquemment dans les paniers
  - Donné un seuil de support s
  - Les ensembles d'items fréquents ayant sup(X) $\geq s$
  - A trouver tous les ensembles d'items fréquents

Exemple

- Ensemble d'items $A = \{\text{lait, coke, pepsi, bière, jus}\}$.
- 8 paniers:

  $B_1 = \{l, c, b\}$ $B_2 = \{l, p, j\}$
  $B_3 = \{l, b\}$ $B_4 = \{c, j\}$
  $B_5 = \{l, p, b\}$ $B_6 = \{l, c, b, j\}$
  $B_7 = \{c, b, j\}$ $B_8 = \{b, c\}$
- Seuil de support $s = 3$
- Ensembles d'items fréquents:
  $\{l\}, \{c\}, \{b\}, \{j\}, \{l, b\}, \{c, b\}, \{j, c\}$

Application 1 (Supermarchés)

- Vraix paniers de la ménagère
  - Les chaînes de magasins sauvegardent des teraoctets d'information permettant de découvrir des similitudes entre des produits achetés par des clients
    - Faire connaître les comportements d'achats des clients typiques
    - Améliorer les relations clients
    - Campagnes marketing. Ex : promotion sur hamburger en augmentant le prix de ketchup
    - . . .
- Il faut un grand support

Application 2 (Recherche d'information)

- Scénario 1
  - paniers = documents
  - items = mots dans ces documents
  - ensemble de mots fréquents = concepts reliés
- Scénario 2
  - paniers = phrases
  - items = documents contenant ces phrases
  - ensemble de documents fréquents = possibilité de plagiat

Application 3 (Fouille de web)

- Scénario 1
  - paniers = pages de web
  - items = pages reliées
  - pages avec beaucoup de références communes peuvent avoir le même thème
- Scénario 2
  - paniers = pages de web p
  - items = pages reliées à p
  - pages ayant beaucoup de liens communs peuvent être miroir ou avoir le même thème

Règles d'association

- Les règles "si ... alors" sur le contenu des paniers.
  - $\{A_1, A_2, \cdots, A_k\} \rightarrow A_j$
  - si le panier contient $X = \{A_1, \cdots, A_k\}$, alors il contient vraisemblablement A_j
  - cela signifie une co-occurrence, non une causalité
- Confiance - probabilité conditionnelle de A_j sachant $A_1, \cdots, A_k$

$$\text{conf}(X \rightarrow A_j) = \frac{\sup(X \cup \{A_j\})}{\sup(X)}$$

- Support (de la règle)

$$\sup(X \rightarrow A_j) = \sup(X \cup \{A_j\})$$

Exemple

- $B_1 = \{l, c, b\}$ $B_2 = \{l, p, j\}$
  $B_3 = \{l, b\}$ $B_4 = \{c, j\}$
  $B_5 = \{l, p, b\}$ $B_6 = \{l, c, b, j\}$
  $B_7 = \{c, b, j\}$ $B_8 = \{b, c\}$
- Règles d'association
  - $\{l, b\} \rightarrow c$
  - Support = 2
  - Confiance = 2/4 = 50%

Détection de règles d'association

- Objectif - trouver toutes les règles d'association satisfaisant
  - support $\geq s$
  - confiance $\geq c$
- $\Rightarrow$ trouver les ensembles d'items fréquents
  - Trouver tous les ensembles d'items fréquents X
  - Etant donné $X = \{A_1, \cdots, A_k\}$, génère tous les règles $X \backslash \{A_j\} \rightarrow A_j$
  - Confiance = $\sup(X)/\sup(X \backslash \{A_j\})$
  - Support = $\sup(X)$
- Remarque : $X \backslash \{A_j\}$ est également fréquent

Trouver des paires fréquentes

- Ensembles de 2 items fréquents
  - le problème le plus difficile est souvent de trouver les paires fréquentes
  - On concentre sur ce problème, puis discuter l'extension pour trouver les ensembles de k items
- Algorithme Naïve
  - Compteurs - tous les $m(m-1)/2$ paires d'items
  - Lecture de données une fois : lire tous les paniers
  - Taille de panier b - augmenter $b(b-1)/2$ compteurs
- Echec ?
  - taille de mémoire principale $< m(m-1)/2$
  - même avec $m = 100,000$ (milions en réalité)

Monotonicité

- Monotonicité
- Etant donné un ensemble d'items. Pour 2 ensembles X et $Y \subset X$

$$\sup(X) \geq s \Rightarrow \sup(Y) \geq s$$

- Pour les paires d'items :

$$\sup(A_i) < s \Rightarrow \sup(\{A_i, A_j\}) < s$$

Algorithme A-Priori

- A-Priori - une approche de deux phases limite le besoin de la mémoire principale
- Phase 1
  - m compteurs (les items candidats dans A)
  - Lecture des paniers b
  - Augmenter les compteurs pour chaque item dans b
- Marque fréquent pour f items qui apparaissent plus que s fois
- Phase 2
  - $f(f-1)/2$ compteurs (les paires candidats d'items fréquents)
  - Lecture des paniers b
  - Augmenter les compteurs pour chaque paires qui apparaissent dans b
- Echec si la mémoire $< m + f(f-1)/2$

Détection des ensembles d'items plus grands

- Objectif - extension aux ensembles de k items fréquents, $k > 2$
- Monotonicité
  Ensemble d'items X est fréquent seulement si $X - \{X_j\}$ fréquent pour tout X_j
- Idée
  - Trouver tous les ensembles de k items fréquents
  - Pas 1 - trouver tous les items fréquents
  - Pas k - compteurs pour tous les ensembles candidats de k items
  - Candidats - ensembles k items dont les sous ensembles (k-1) items sont fréquents
  - Coût total: nombre de pas = la taille la plus grande des ensembles d'items fréquents

D'autres algorithmes bien connus

- FP-growth: Utilisation de FP-tree (Frequent-pattern tree) - pour représenter des données

Résumé - Objectifs

Analyse factorielle

- Analyse en composantes principales
- Analyse factorielle des correspondances

Classification (clustering)

- Réduire en regroupant les individus :
  - méthodes non hiérarchiques,
  - méthodes hiérarchiques en arbre.

Règles d'association

- Algorithme "A priori "