Cours:
Analyse des données

Niveau : S6

Option : Economie

Pr: Zineb SAYL

Année universitaire : 2019/2020

Plan

Introduction
Chapitre I : Analyse en Composantes Principales
Chapitre II : Analyse Factorielle des Correspondances
Chapitre III : Classification Ascendante Hiérarchique

Introduction

1-Définition et objectifs

L' Analyse des données (ADD) : l'ensemble de méthodes **descriptives** ayant pour objectif de **résumer et visualiser l'information** contenue dans un grand tableau de données

➔ « l'analyse des données est un ensemble de techniques pour **découvrir** la structure, éventuellement compliquée, d'un tableau de nombres à plusieurs dimensions et de traduire par une structure plus simple et qui la **résume** au mieux. Cette structure peut le plus souvent, être représentée graphiquement'» (J-P. Fénelon)

Introduction

1-Définition et objectifs

Les principaux objectifs de l'Analyse des données :

- Répondre aux problèmes posés par des tableaux de grandes dimensions
- Résumer les informations contenues dans un grand tableau sous forme d'une matrice
- Organiser et visualiser les informations

Introduction

1-Définition et objectifs

Le développement des outils informatiques a fortement contribué au développement de nombreuses méthodes statistiques

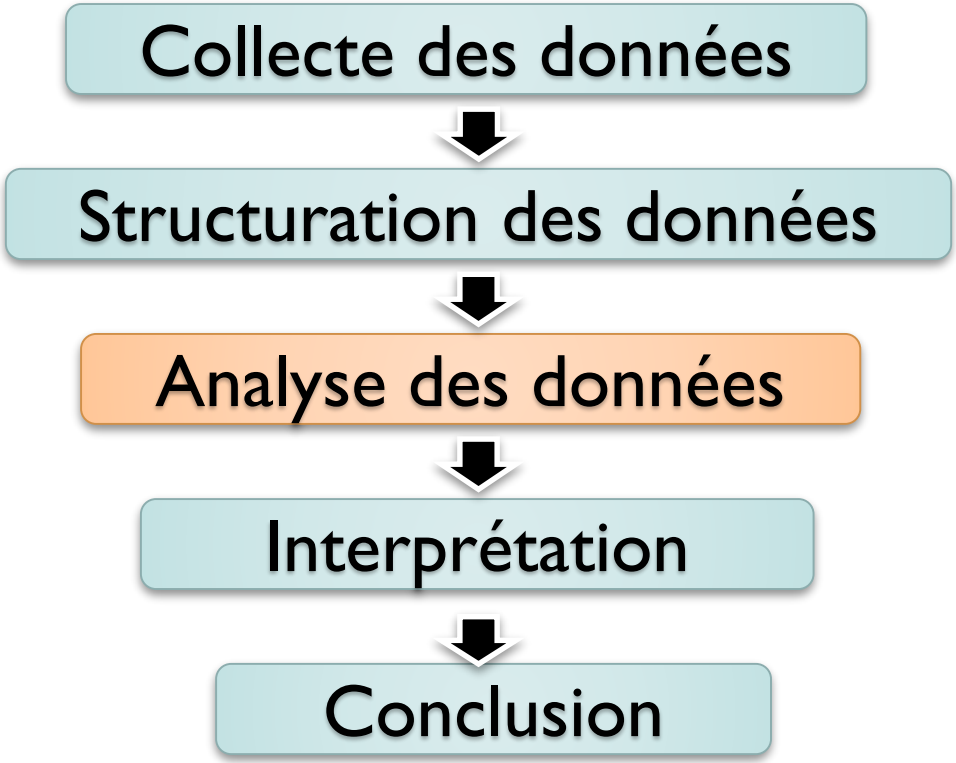
➔ Ce qui permet de traiter sans difficultés de vastes données des enquêtes et des investigations (grands tableaux de milliers de lignes et milliers de colonnes)

➔ Ex: SPSS, EVIEWS…

Introduction

2- Le processus de ADD

- Les principales étapes du processus d'analyse :



Introduction

2- Le processus de ADD

Collecte des données

On distingue deux types des données :

- **Les données primaires** sont spécialement collectées pour répondre à une étude statistique précise.

- **Les données secondaires:** sont des données qui existent déjà (Ex: statistiques officielles…)

Introduction

3- Les méthodes

On distingue plusieurs méthodes selon deux critères :

1- Selon l'objectif de la méthode

- Méthodes explicatives
- Méthodes descriptives

2- Selon type de mesure

- Mesure nominale
- Mesure ordinale
- Mesure métrique

Introduction

3- Les méthodes

Méthodes explicatives

Consiste à expliquer une variable au moyen d'une ou plusieurs variable:

- Variable à expliquer
- Variable explicative

Ex : $INVES = f(CROI, INF, INDH)$

$INVES = \alpha\, CROI + \beta\, INF + \gamma\, INDH$

Exemple :

- Problèmes de régression et de corrélation
- Analyse de la variance
- Analyse discriminatoire
- Régression logistique
- Corrélation canonique

Introduction

3- Les méthodes

Méthodes descriptives

Consiste à résumer, visualiser et synthétiser les informations.

Exemple :

- Analyse Factorielle des Correspondances
- Analyse en Composantes Principales
- Classification Ascendante Hiérarchique

Introduction

3- Les méthodes

2- Selon type de mesure

- Mesure nominale
- Mesure ordinale
- Mesure métrique

Introduction

3- Les méthodes

Mesure nominale

On utilise des chiffres sans aucune relation d'ordre, ni de distance, ni d'origine:

Femme (1)

Homme (2)

Exemple :

Sexe: Femme , Homme

Situation matrimoniale: marié, célibataire...

Méthode :

- Analyse Factorielle des Correspondances

Introduction

3- Les méthodes

Mesure ordinale

Les chiffre qui identifient la relation d'ordre entre les propriétés d'objet sans aucune relation de distance

Exemple :
Classement des goûts des clients selon un critère classe d'âge (15- 25), (26- 35)….
le rang

Méthode :
- Analyse Factorielle des Correspondances

Introduction

3- Les méthodes

Mesure métrique

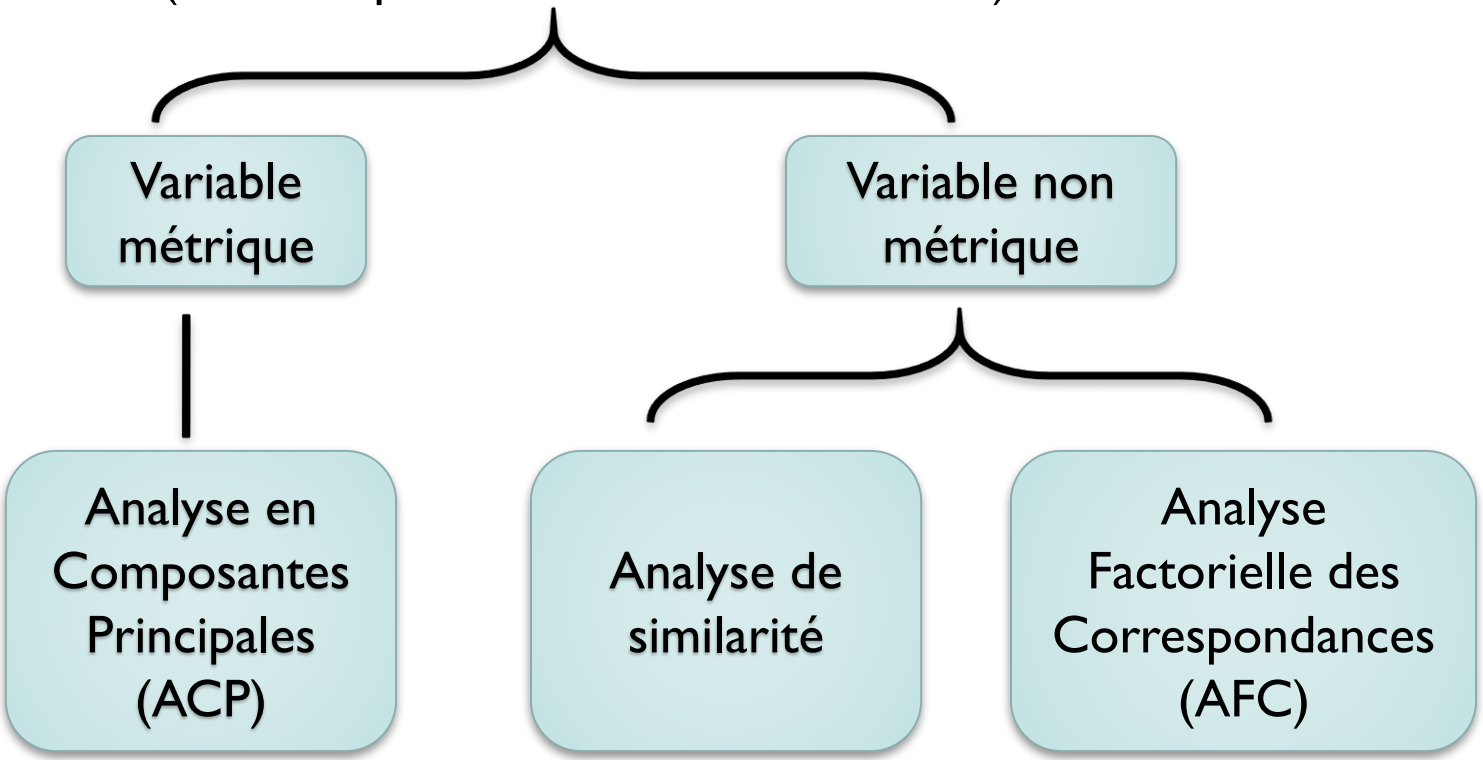
Variable quantitative dont les valeurs ont une relation d'ordre et de distance

Exemple :
Le nombre de points de vente d'une marque
La valeur d'investissement dans une ville

Méthode :
- Analyse en Composantes Principales

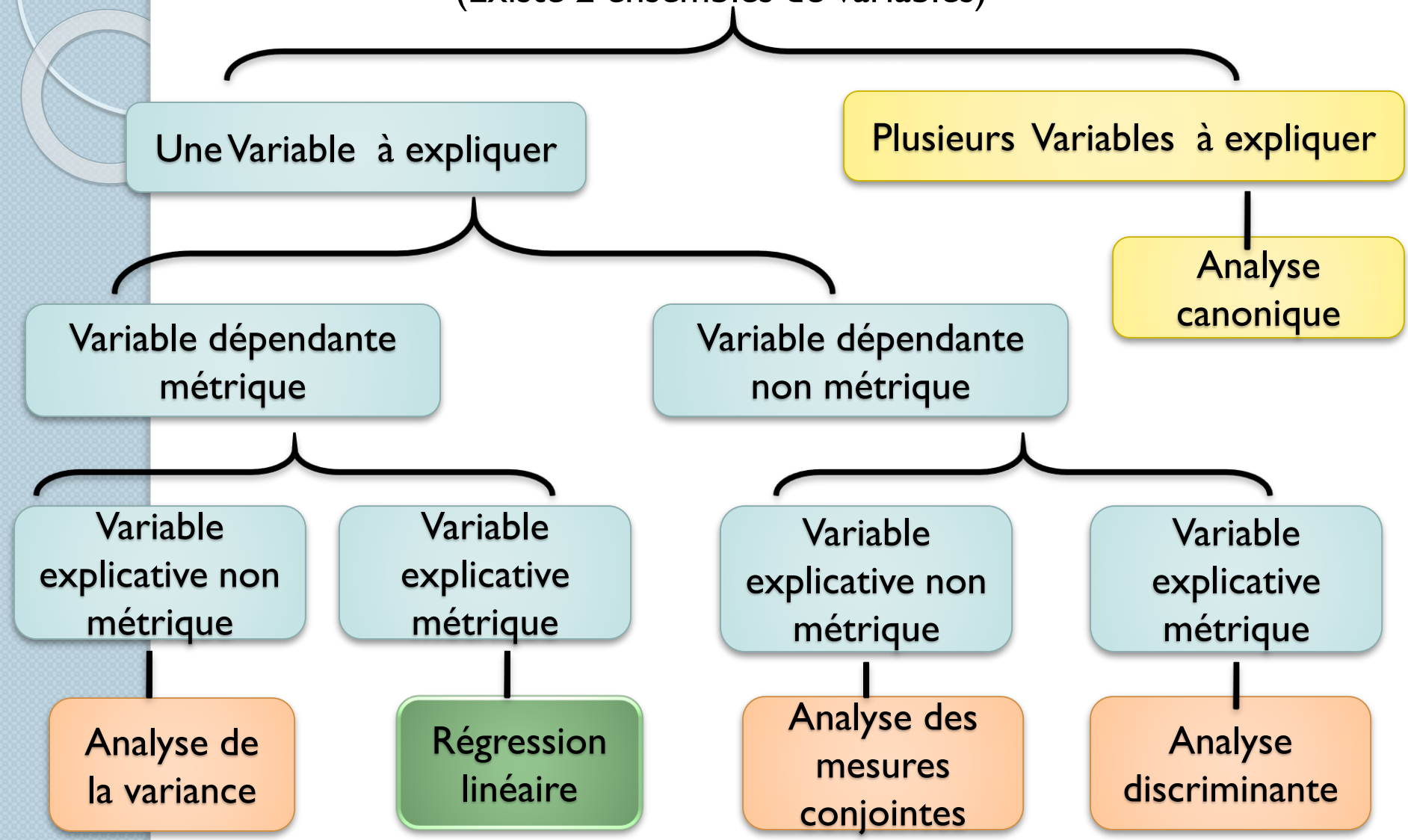
Méthodes descriptives

(N'existe pas 2 ensembles de variables)



Méthodes explicatives

(Existe 2 ensembles de variables)



Chapitre 1 : Analyse en Composantes Principales (ACP)

PLAN

Chapitre 1 : Analyse en Composantes Principales (ACP)

1-Définition

2- Principe

3- Démarche

4- Cas d'application

Chapitre 1 : Analyse en Composantes Principales

1-Définition

Analyse en Composantes Principales (ACP) est :

- Une méthode **descriptive** a pour objectif l'analyse des tableaux de données qui ne comportent pas des structure préalable (aucune distinction ni entre variable ni entre individu)
- Le but principale est de **résumer** l'information contenue dans un tableau composé d'un nombre élevé de ligne et de colonnes

➔ Un outil statistique de **synthèse** de l'information

➔ Un outil très important pour traiter les données **quantitatives**

2-Principe

Analyse en Composantes Principales (ACP) permet de :

- Résumer les informations contenant dans un tableau en **n** individus et **p** variables
- Remplacer les **p** variables avec **q** nouvelles variables avec **q < p**

2-Principe

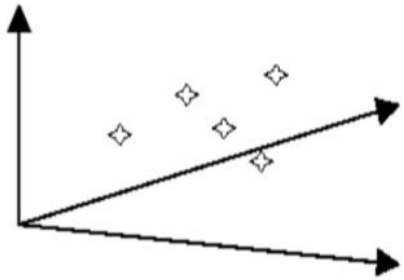
D'un point de vue géométrique

Le nuage de points représentant les données s'inscrit dans un espace de *P* dimensions, puisque chaque point représente un individu par rapport à $x_1, x_2, \ldots \ldots . x_n$

➔Il est difficile de visualiser les relations existant entre les variables dès que p>3

Si la dimension P= 2
Il est facile de présenter le nuage de points

Si la dimension P> 3
Il est difficile de présenter le nuage de points



3- Démarche

Soit : **n** individus caractérisés par **p** variables métriques

Ces données sont présentées dans un tableau appelé la Matrice des données de dimension n*p

➔ Les étapes pour déterminer la composante principale :

Centrage et réduction des données

↓

Déterminer les valeurs propres et les vecteurs propres sur la base de la matrice de corrélation entre les variables

↓

Déterminer les axes factoriels
Sélectionner les composantes principales

3- Démarche

➔ **Centrage et réduction des données:**

Soit : **n** individus caractérisés par **p** variables métriques

- Les **p** variables sont de nature différente, pour homogénéiser les unités, les **p** variables seront centrées et réduites,
- Les données sont centrées et réduites signifie que pour chaque variable la moyenne est nulle ($\overline{X}=0$) et la variance égale à 1 ($v=1$)

Matrice Centrée Réduite est obtenue par la formule suivante :

$$x_{ij} = \frac{x_{ij} - \overline{X}}{\sigma_i}$$

3- Démarche

➔ Centrage et réduction des données:

La Matrice des variances covariances permet de mesure la liaison linéaire qui peut exister entre un couple de variables statistiques

Var X1	Cov (X1,X2)	Cov (X1,X3)
Cov (X2,X1)	**Var X2**	Cov (X2,X3)
Cov (X3,X1)	Cov (X3,X2)	**Var X3**

Si Cov (X2,X1) =0 ➔ les variables X1 et X2 sont indépendantes
Si Cov (X2,X1) ≠ 0 ➔ les variables X1 et X2 sont dépendantes
(existe une relation linéaire entre les variable)

➡ Obtenue par la formule suivante :

$$V = \frac{1}{n} MC * MC^t$$

$MC : Matrice\ Centrée \Leftrightarrow MC : x_{ij} = (x_{ij} - \overline{X})$

$MC^t : Matrice\ Transposée$

Chapitre 1 : Analyse en Composantes Principales

3- Démarche

➔ **Centrage et réduction des données:**

Matrice des corrélations entre variables permet d'analyser les relations bilatérales entre les variables :

➡ Obtenue par la formule suivante :

$$\mathrm{U} = \frac{1}{n}\, CR^t * CR$$

Avec

CR : *Matrice Centrée Réduite*

CR^t : *Matrice Centrée Réduite Transposée*

Chapitre 1 : Analyse en Composantes Principales

Cas d'application

Une étude consiste à déterminer les facteurs de la localisation internationale d'une marque. Soit le tableau des données suivant:

	IDE	Taux croissance économique (%)	Taux d'inflation (%)
Pays A	300	2	6
Pays B	450	2	4
Pays C	950	8	2
Pays D	700	7	5

Chapitre 1 : Analyse en Composantes Principales

Travail à faire:

1. Calculer la moyenne et l'écart type des variables
2. Déterminer la Matrice Centrée Réduite
3. Déterminer la Matrice des variances covariance
4. Déterminer la Matrice des corrélations entre variables
5. Déterminer le polynôme caractéristique
6. Calculer les valeurs propres
7. Calculer et interpréter l'inertie des axes factoriels
8. Déterminer les vecteurs propres orthogonaux associés aux valeurs propres
9. Calculer et interpréter la corrélation des variables avec les composantes principales
10. Calculer et interpréter la contribution CONTR

Solution

1- Calculer la moyenne et l'écart type des variables

Définition :

La moyenne est un outil de calcul **permet de résumer** une liste de valeurs numériques en un seul nombre réel sans tenir compte de l'ordre de la liste.

$$\overline{X} = \frac{1}{n} \sum x_i$$

Solution

1- Calculer la moyenne et l'écart type des variables

Définition :
L'écart type est un outil de calcul permet de mesurer **la dispersion** des valeurs d'un échantillon. C'est la racine carrée de la variance :

$$\sigma = \sqrt{variance}$$

Avec la variance est la moyenne des carrées des écarts à la moyenne :

$$V = \frac{1}{n} \sum (x_i - \overline{X})^2$$

Avec

x_i : les valeurs de la variable

$\overline{X}$: la moyenne de la variable

Chapitre 1 : Analyse en Composantes Principales

Solution

1- Calculer la moyenne et l'écart type des variables

	IDE	Taux croissance économique (%)	Taux d'inflation (%)
Pays A	300	2	6
Pays B	450	2	4
Pays C	950	8	2
Pays D	700	7	5
Moyenne	**600**	**4,75**	**4,25**
Ecart type	**247,50**	**2,77**	**1,48**

Calcul: Pour la variable x1

$$\overline{X}_1 = \frac{300 + 450 + 950 + 700}{4} = 600$$

$$\sigma_1 = \sqrt{\frac{(300-600)^2 + (450-600)^2 + (950-600)^2 + (700-600)^2}{4}} = 247{,}5$$

Chapitre 1 : Analyse en Composantes Principales

2- Déterminer la Matrice Centrée Réduite (MCR)

MCR :

$$x_{ij} = \frac{x_{ij} - \overline{X}}{\sigma_i}$$

-1,21	-0,99	1,18
-0,80	-0,99	-0,16
1,41	1,17	-1,50
0,40	0,81	0,50

Calcul: Pour la variable x1

$$x_{11} = \frac{x_{ij} - \overline{X}}{\sigma_i} = \frac{300 - 600}{247,5} = -1,21$$

$$x_{21} = \frac{x_{ij} - \overline{X}}{\sigma_i} = \frac{450 - 600}{247,5} = -0,8$$

$$x_{31} = \frac{x_{ij} - \overline{X}}{\sigma_i} = \frac{950 - 600}{247,5} = 1,41$$

$$x_{41} = \frac{x_{ij} - \overline{X}}{\sigma_i} = \frac{700 - 600}{247,5} = 0,4$$

Chapitre 1 : Analyse en Composantes Principales

3- Déterminer la Matrice des variances covariances : $\mathrm{V} = \frac{1}{n} MC * MC^t$

Matrice centrée (MC):

$x_{ij} = (x_{ij} - \overline{\mathrm{X}})$

-300	-2,75	1,75
-150	-2,75	-0,25
350	3,25	-2,25
100	2,25	0,75

Calcul: Pour la variable x1

$$x_{11} = (x_{ij} - \overline{\mathrm{X}}) = 300 - 600 = -300$$
$$x_{21} = (x_{ij} - \overline{\mathrm{X}}) = 450 - 600 = -150$$
$$x_{31} = (x_{ij} - \overline{\mathrm{X}}) = 950 - 600 = -350$$
$$x_{41} = (x_{ij} - \overline{\mathrm{X}}) = 700 - 600 = -100$$

Chapitre 1 : Analyse en Composantes Principales

3- Déterminer la Matrice des variances covariances

$$V = \frac{1}{n} MC * MC^t$$

$$V = \frac{1}{4} \begin{vmatrix} -300 & -2,75 & 1,75 \\ -150 & -2,75 & -0,25 \\ 350 & 3,25 & -2,25 \\ 100 & 2,25 & 0,75 \end{vmatrix} * \begin{vmatrix} -300 & -150 & 350 & 100 \\ -2,75 & -2,75 & 3,25 & 2,25 \\ 1,75 & -0,25 & -2,25 & 0,75 \end{vmatrix}$$

$$V = \begin{vmatrix} 22503 & 45007 & -26253 & -7501 \\ 11252 & 5627 & -13127 & -3752 \\ 26253 & -13127 & 30629 & 8751 \\ -7501 & -3751 & 8751 & 2501 \end{vmatrix}$$

Chapitre 1 : Analyse en Composantes Principales

4- Déterminer la Matrice des corrélations entre variables

$$U = \frac{1}{n} CR^t * CR$$

$$U = \frac{1}{4} \begin{vmatrix} -1,21 & -0,80 & 1,41 & 0,40 \\ -0,99 & -0,99 & 1,17 & 0,81 \\ 1,18 & -0,16 & -1,50 & 0,50 \end{vmatrix} * \begin{vmatrix} -1,21 & -0,99 & 1,18 \\ -0,80 & -0,99 & -0,16 \\ 1,41 & 1,17 & -1,50 \\ 0,40 & 0,81 & 0,50 \end{vmatrix}$$

$$U = \begin{vmatrix} 1 & 0,99 & -0,8 \\ \mathbf{0,99} & 1 & -0,6 \\ \mathbf{-0,8} & -0,6 & 1 \end{vmatrix}$$

(X1; X2) =0,99 Forte corrélation positive entre IDE et Taux de Croissance

Taux de Croissance augmente ➔ IDE augmente

(X1; X3) = -0,8 Forte corrélation négative entre IDE et Taux d'inflation

Taux d'inflation augmente ➔ IDE diminue

Chapitre 1 : Analyse en Composantes Principales

5- Déterminer le polynôme caractéristique

$$\text{Det}\left| U - \lambda I \right| =$$

Avec

$$U = \begin{vmatrix} 1 & 0{,}99 & -0{,}8 \\ 0{,}99 & 1 & -0{,}6 \\ -0{,}8 & -0{,}6 & 1 \end{vmatrix} \qquad I = \begin{vmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{vmatrix}$$

$$\text{Det}\left| U - \lambda I \right| = \begin{vmatrix} 1-\lambda & 0{,}99 & -0{,}8 \\ 0{,}99 & 1-\lambda & -0{,}6 \\ -0{,}8 & -0{,}6 & 1-\lambda \end{vmatrix}$$

Chapitre 1 : Analyse en Composantes Principales

6- Calculer les valeurs propres de λ

$$\text{Det}\,|\,U-\lambda I\,|=0 \quad \text{Det}\begin{vmatrix} \mathbf{1-\lambda} & 0{,}99 & -0{,}8 \\ 0{,}99 & \mathbf{1-\lambda} & -0{,}6 \\ -0{,}8 & -0{,}6 & \mathbf{1-\lambda} \end{vmatrix} = \mathbf{0}$$

$$\text{Det}\ \mathbf{1-\lambda}\begin{vmatrix} 1-\lambda & -0{,}6 \\ -0{,}6 & 1-\lambda \end{vmatrix} \mathbf{-\ 0{,}99}\begin{vmatrix} 0{,}99 & -0{,}6 \\ -0{,}8 & 1-\lambda \end{vmatrix} \mathbf{-0{,}8}\begin{vmatrix} 0{,}99 & 1-\lambda \\ -0{,}8 & -0{,}6 \end{vmatrix}$$

$$\lambda = \frac{b \pm \sqrt{\Delta}}{2a}$$

$$\Delta = b^2 - 4ac$$

$$\lambda^2 - 3\lambda + 2{,}98 = 0$$

$$\lambda = 0$$

$$\lambda_1 = 2{,}35$$

$$\lambda_2 = 0{,}65$$

Chapitre 1 : Analyse en Composantes Principales

7- Calculer et interpréter l'inertie des axes factoriels

➔L'inertie de l'axe factoriel:

Le pourcentage (%) d'inertie exprimé par un axe factoriel permet d'évaluer la quantité d'information contenue dans cet axe :

$$Inertie\ d'un\ axe\ = \frac{Valeur\ propre\ correspondante}{somme\ des\ valeurs\ propres\ \ (Inertie\ totale)}$$

Chapitre 1 : Analyse en Composantes Principales

7- Calculer et interpréter l'inertie des axes factoriels

Axe 1 : $\lambda_1 = 2{,}35$

$$Inertie\ Axe\ 1 = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{2{,}35}{2{,}35 + 0{,}65} = 0{,}78$$

➔Cet axe contient 78% des informations

Axe 2: $\lambda_2 = 0{,}65$

$$Inertie\ Axe\ 2 = \frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{0{,}65}{2{,}35 + 0{,}65} = 0{,}22$$

➔ Cet axe contient 22% des informations

Chapitre 1 : Analyse en Composantes Principales

8- Déterminer les vecteurs propres orthogonaux associés aux valeurs propres

Axe 1 : $\lambda_1 = 2{,}35$ ➔ On cherche $\vec{a}\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ tel que :

$$U\vec{a} = \lambda_1 \vec{a}$$

$$\begin{vmatrix} 1 & 0{,}99 & -0{,}8 \\ 0{,}99 & 1 & -0{,}6 \\ -0{,}8 & -0{,}6 & 1 \end{vmatrix} * \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 2{,}35 \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

Chapitre 1 : Analyse en Composantes Principales

8- Déterminer les vecteurs propres orthogonaux associés aux valeurs propres

$$\begin{cases} 1x + 0{,}99y - 0{,}8z = 2{,}35x \\ 0{,}99x + 1y - 0{,}6z = 2{,}35y \\ -0{,}8x - 0{,}6y + 1z = 2{,}35z \end{cases}$$

$$\begin{cases} 1x + 0{,}99y - 0{,}8z - 2{,}35x = 0 \\ 0{,}99x + 1y - 0{,}6z - 2{,}35y = 0 \\ -0{,}8x - 0{,}6y + 1z - 2{,}35z = 0 \end{cases}$$

$$\begin{cases} \boldsymbol{x = 0,57} \\ \boldsymbol{y = 0,54} \\ \boldsymbol{z = 0,64} \end{cases} \Rightarrow U = \begin{pmatrix} \mathbf{0,57} \\ \mathbf{0,54} \\ \mathbf{0,64} \end{pmatrix}$$

Chapitre 1 : Analyse en Composantes Principales

8- Déterminer les vecteurs propres orthogonaux associés aux valeurs propres

Axe 2 : $\lambda_2 = 0{,}65$ ➔ On cherche $\vec{a}\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ tel que :

$$U\vec{a} = \lambda_2 \vec{a}$$

$$\begin{vmatrix} 1 & 0{,}99 & -0{,}8 \\ 0{,}99 & 1 & -0{,}6 \\ -0{,}8 & -0{,}6 & 1 \end{vmatrix} * \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 0{,}65 \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

Chapitre 1 : Analyse en Composantes Principales

8- Déterminer les vecteurs propres orthogonaux associés aux valeurs propres

$$\begin{cases} 1x + 0{,}99y - 0{,}8z = 0{,}65x \\ 0{,}99\,x + 1y - 0{,}6z = 0{,}65y \\ -0{,}8x - 0{,}6\,y + 1z = 0{,}65z \end{cases}$$

$$\begin{cases} 1x + 0{,}99y - 0{,}8z - 0{,}65x = 0 \\ 0{,}99\,x + 1y - 0{,}6z - 0{,}65y = 0 \\ -0{,}8x - 0{,}6y + 1z - 0{,}65z = 0 \end{cases}$$

$$\begin{cases} \boldsymbol{x = 0,53} \\ \boldsymbol{y = 0,99} \\ \boldsymbol{z = 0,34} \end{cases} \Rightarrow \boldsymbol{V = \begin{pmatrix} 0,53 \\ 0,99 \\ 0,34 \end{pmatrix}}$$

Chapitre 1 : Analyse en Composantes Principales

9- Calculer et interpréter la corrélation des variables avec les composantes principales

➔Consiste à la projection des individus sur les axes principaux

Avec :

MCR * U =➔ Projection sur axe 1

MCR *V =➔ Projection sur axe 2

MCR

-1,21	-0,99	1,18
-0,80	-0,99	-0,16
1,41	1,17	-1,50
0,40	0,81	0,50

$$U = \begin{pmatrix} 0,57 \\ 0,54 \\ 0,64 \end{pmatrix} \quad V = \begin{pmatrix} 0,53 \\ 0,99 \\ 0,34 \end{pmatrix}$$

Chapitre 1 : Analyse en Composantes Principales

9- Calculer et interpréter la corrélation des variables avec les composantes principales

MCR * U =➔ Projection sur **axe 1**

$$\begin{vmatrix} -1,21 & -0,99 & 1,18 \\ -0,80 & -0,99 & -0,16 \\ 1,41 & 1,17 & -1,50 \\ 0,40 & 0,81 & 0,50 \end{vmatrix} * \begin{pmatrix} 0,57 \\ 0,54 \\ 0,64 \end{pmatrix}$$

MCR * U =➔ Projection sur **axe 2**

$$\begin{vmatrix} -1,21 & -0,99 & 1,18 \\ -0,80 & -0,99 & -0,16 \\ 1,41 & 1,17 & -1,50 \\ 0,40 & 0,81 & 0,50 \end{vmatrix} * \begin{pmatrix} 0,53 \\ 0,99 \\ 0,34 \end{pmatrix}$$

Chapitre 1 : Analyse en Composantes Principales

9- Calculer et interpréter la corrélation des variables avec les composantes principales

Résultat de projection :

	AXE 1	AXE 2
I1	-0,46	-1,22
I2	-1,09	-1,46
I3	0,48	1,4
I4	0,99	1,2

Ce tableau permet de calculer la contribution des axes

➔ **La contribution des individus**

Pour calculer la contribution de chaque individu à l'inertie, on utilise la formule suivante :

$$CONTR = P_i \frac{\|x_i\|^2}{I_i}$$

x_i : Valeur CR

P_i : Poids $\frac{1}{n}$

I_i : La somme des valeurs propres $(\sum \lambda_i)$

Chapitre 1 : Analyse en Composantes Principales

9- Calculer et interpréter la corrélation des variables avec les composantes principales

	X1	X2	X3	**CONTR**
Pays A	0,12	0,08	0,17	**0,37**
Pays B	0,05	0,08	0,01	**0,14**
Pays C	0,17	0,11	0,19	**0,47**
Pays D	0,1	0,05	0,02	**0,17**
	0,44	**0,32**	**0,39**	

Calcul

$$\sum \lambda_i = 3$$

$$(x1\,;\ Pays\ A) \quad CONTR = \frac{1}{4} * \frac{\|1{,}21\|^2}{3} = 0{,}12$$

$$(x1\,;\ Pays\ B) \quad CONTR = \frac{1}{4} * \frac{\|0{,}8\|^2}{3} = 0{,}05$$

$$(x1\,;\ Pays\ C) \quad CONTR = \frac{1}{4} * \frac{\|1{,}41\|^2}{3} = 0{,}17$$

$$(x1\,;\ Pays\ D) \quad CONTR = \frac{1}{4} * \frac{\|0{,}4\|^2}{3} = 0{,}1$$

Chapitre 1 : Analyse en Composantes Principales

9- Calculer et interpréter la corrélation des variables avec les composantes principales

	X1	X2	X3	**CONTR**
Pays A	0,12	0,08	0,17	**0,37**
Pays B	0,05	0,08	0,01	**0,14**
Pays C	0,17	0,11	0,19	**0,47**
Pays D	0,1	0,05	0,02	**0,17**
	0,44	**0,32**	**0,39**	

➔ Le tableau permet de déterminer la contribution des individus dans l'analyse:

- Le pays **C** contribue de **0,47** (47%) pour expliquer le phénomène
- Le pays **A** contribue de **0,37** (37%) pour expliquer le phénomène

➔ Le tableau permet de déterminer la contribution des variables dans l'analyse:

- La variable X1 contribue de 44% dans l'analyse
- La variable X2 contribue de 32% dans l'analyse
- La variable X3 contribue de 39% dans l'analyse

Chapitre 1 : Analyse en Composantes Principales

10- Calculer et interpréter la contribution CONTR

➔**La contribution CONTR des axes est calculée par la formule suivante :**

Axe 1 : $\lambda_1 = 2{,}35$ ⇒ $CONTR = \mathrm{P}_i \dfrac{\|\mathrm{x}_i\|^2}{\lambda_1}$

Axe 2 : $\lambda_2 = 0{,}65$ ⇒ $CONTR = \mathrm{P}_i \dfrac{\|\mathrm{x}_i\|^2}{\lambda_2}$

Chapitre 1 : Analyse en Composantes Principales

10- Calculer et interpréter la contribution CONTR

➔**La contribution CONTR des axes**

	AXE 1	AXE 2	**CONTR AXE 1**	**CONTR AXE 2**
Pays A	-0,46	-1,22	**0,02**	**0,57**
Pays B	-1,09	-1,46	**0,12**	**0,82**
Pays C	0,48	1,4	**0,02**	**0,75**
Pays D	0,99	1,2	**0,1**	**0,55**

Calcul

$$(Axe\ 1\ ;\ Pays\ A) \quad CONTR = \frac{1}{4} * \frac{\|0{,}46\|^2}{2{,}35} = 0{,}02$$

$$(Axe\ 1\ ;\ Pays\ B) \quad CONTR = \frac{1}{4} * \frac{\|1{,}09\|^2}{2{,}35} = 0{,}12$$

$$(Axe\ 2\ ;\ Pays\ A) \quad CONTR = \frac{1}{4} * \frac{\|1{,}22\|^2}{0{,}65} = 0{,}57$$

$$(Axe\ 2\ ;\ Pays\ B) \quad CONTR = \frac{1}{4} * \frac{\|1{,}46\|^2}{0{,}65} = 0{,}82$$

Chapitre 1 : Analyse en Composantes Principales

10- Calculer et interpréter la contribution CONTR

➔**La contribution CONTR des axes**

	AXE 1	AXE 2	**CONTR AXE 1**	**CONTR AXE 2**
Pays A	-0,46	-1,22	**0,02**	**0,57**
Pays B	-1,09	-1,46	**0,12**	**0,82**
Pays C	0,48	1,4	**0,02**	**0,75**
Pays D	0,99	1,2	**0,1**	**0,55**

Interprétation

- Pour Axe 1, le point déterminant est 0,12 (12%)
- Pour Axe 2, le point déterminant est 0,82 (82%)

Chapitre 11 : Analyse Factorielle des Correspondances

PLAN

Chapitre II : Analyse Factorielle des Correspondances

1-Définition

2- Objectifs

3- Démarche

4- Cas d'application

Chapitre 11 : Analyse Factorielle des Correspondances

1- Définition

L'Analyse Factorielle des Correspondances (AFC), nommée également, Analyse des Correspondances Simples est :

- Une méthode exploratoire d'analyse des tableaux de contingence.
- Une analyse multidimensionnelle
- Développée par J.-P. Benzecri durant la période 1970-1990.

Chapitre 11 : Analyse Factorielle des Correspondances

1- Définition

➔ *Une analyse multidimensionnelle*

L'analyse factorielle traite des tableaux par :

- Remplacer un tableau de nombres difficile à analyser par une série de tableaux plus simples.
- Ces tableaux simples sont exprimables sous forme de graphiques

➔ *L'analyse de la relation entre variables:*

La variables numériques ➔ on analyse la Corrélation
La variables nominales ➔ on analyse la Correspondance

Chapitre 11 : Analyse Factorielle des Correspondances

2- Objectifs

L'Analyse Factorielle des Correspondances (AFC) :

➔Permet d'étudier la proximité qui caractérise les lignes (i) et les colonnes (j) sur la base des différences de poids affectés à (i) et (j)

➔Permet de répondre aux questions suivantes :

- Y-a-t-il un lien entre les deux caractères étudiés?
- Si OUI, comment se comporte un caractère par rapport à l'autre?

3- Démarche

A- Etude descriptive du tableau de contingence

Soit un tableau de données de lignes i et colonnes j. On fixe les notations suivantes :

$n_{i.j}$: effectif de la cellule (i, j)

$n_{i.}$: effectif total de la ligne i

$n_{.j}$: effectif total de la colonne j

$n_{..}$: effectif total

a-Tableau des fréquences

Les fréquences sont calculées par la formule suivante:

$$f_{ij} = \frac{\text{Effectif de la Cellule (i,j)}}{\text{Effectif Total}} = \frac{n_{i,j}}{N}$$

3- Démarche

b- Tableau des fréquences lignes

Les fréquences lignes (ou coordonnées des **profils lignes**) sont calculées par :

$$fL_{ij} = \frac{n_{i,j}}{n_{i.}} = \frac{f_{i,j}}{f_{i.}} = \frac{\text{Effectif de la Cellule (i, j)}}{\text{Effectif de la ligne i}}$$

$n_{i.}$: est la somme de la ligne i

➔ Les coordonnées du **profil ligne moyen** sont calculées par:

$$f_{.j} = \frac{n_{.j}}{n_{..}} = \frac{\text{Effectif de la colonne j}}{\text{Effectif Total}}$$

3- Démarche

c- Tableau des fréquences colonnes

Les fréquences colonnes (ou coordonnées des **profils colonnes**) sont calculées par :

$$fC_{ij} = \frac{n_{ij}}{n_{.j}} = \frac{f_{i,j}}{f_{.j}} = \frac{\text{Effectif de la Cellule (i, j)}}{\text{Effectif de la colonne J}}$$

➔ Les coordonnées du **profil colonnes moyen** sont calculées par:

$$f_{i.} = \frac{n_{i.}}{n_{..}} = \frac{\text{Effectif de la ligne i}}{\text{Effectif Total}}$$

3- Démarche

d- Distances entre profils

Chaque ligne du tableau des fréquences lignes peut être vue comme la liste des coordonnées d'un point dans un espace à q dimensions.
On obtient ainsi le nuage des individus-lignes.
On définit de même le nuage des individus-colonnes à partir du tableau des fréquences colonnes.

➔ Pour mesurer la distance entre 2 individus, on utilise la distance métrique.
On a la formule suivante :

$$d^2_{\emptyset^2}(L_i, L_{i'}) = \sum_j \frac{(fL_{i,j} - fL_{i',j})^2}{f_{.j}}$$

La distance $\emptyset^2$ s'exprime comme une pondération particulière de la distance euclidienne

3- Démarche

d- Distances entre profils

➔L'importance de la métrique $\emptyset^2$

- Avec la métrique $\emptyset^2$, la distance entre deux lignes ne dépend pas des poids respectifs des colonnes.

- La métrique $\emptyset^2$ possède la propriété *d'équivalence distributionnelle* : si on regroupe deux modalités lignes, les distances entre les profils-colonne, ou entre les autres profils-lignes restent inchangées.

3- Démarche

e- Taux de liaison

Les taux de liaison sont définis par la formule suivante:

$$T_{ij} = \frac{f_{i,j} - f_{i.} f_{.j}}{f_{i.} f_{.j}}$$

Le coefficient fi• f• j représente le "poids théorique" de chaque cellule dans le tableau. La somme de ces coefficients vaut 1.

Le taux de liaison permet de mesurer la variation du poids de la cellule (i,j) par rapport à la moyenne théorique;

A noter que, les valeurs prises par le taux de liaison sont :

- des nombres positifs quelconques (un score observé peut être 200% ou 300% supérieur au score théorique)
- des nombres négatifs compris entre -1 et 0 (le "déficit" le plus extrême d'un score observé est d'être 100% moins élevé que le score théorique).

Chapitre 11 : Analyse Factorielle des Correspondances

3- Démarche

B- L'analyse factorielle des correspondances

a- Calcul des matrices

b- Calcul des valeurs propres

Pour déterminer les valeurs propres, on utilise la même démarche que celle de ACP

➔ **AFC** est une méthode multidimensionnelle permettant d'étudier la dépendance entre deux variables qualitatives

AFC = 2 ACP

ACP	**AFC**
-Observations (lignes)	-Profils lignes
-Variables (colonnes)	-Profils colonnes
-Variables **quantitatives**	-Variables **qualitatives**

Chapitre 11 : Analyse Factorielle des Correspondances

Cas d'application

Une entreprise spécialisée dans les produits de Lux, elle cherche à choisir un nom pour un nouveau produit. Ce nom doit refléter une image d'un produit de qualité supérieure, prestigieuse, luxueuse, qui cible un public masculin, raffiné, distingué, de niveau socio-économique élevé.

L'entreprise propose 12 noms de marque, elle a mené une étudie pour sélectionner un nom de marque auprès d'un échantillon de clients potentiels.
L'entreprise a précisé 11 attributs, chaque interrogé doit évaluer les noms proposés selon les attributs.

Les résultats obtenus sont les suivants :

Chapitre 11 : Analyse Factorielle des Correspondances

Cas d'application

	Orly	Alezan	Corsaire	Directoire	Ducat	Fontenoy	Icare	Zodiaque	Pavois	Cocker	Escale	Hotesse	Total
Vieille	**1**	**2**	**14**	**38**	**18**	**10**	**9**	**5**	**9**	**4**	**0**	**1**	111
Riche	**20**	**9**	**1**	**11**	**10**	**9**	**1**	**1**	**20**	**9**	**7**	**12**	110
Elégant	**9**	**23**	**1**	**15**	**7**	**11**	**6**	**2**	**7**	**12**	**3**	**17**	113
Comique	**1**	**3**	**15**	**15**	**6**	**5**	**12**	**18**	**4**	**25**	**2**	**2**	108
Racé	**4**	**33**	**7**	**8**	**3**	**6**	**6**	**4**	**5**	**15**	**5**	**3**	99
Mièvre	**3**	**9**	**1**	**7**	**7**	**5**	**12**	**9**	**6**	**9**	**6**	**13**	87
Distingué	**11**	**9**	**1**	**17**	**4**	**21**	**6**	**1**	**5**	**4**	**5**	**27**	111
Vulgaire	**4**	**4**	**32**	**2**	**6**	**0**	**9**	**7**	**3**	**10**	**10**	**7**	94
Masculin	**9**	**12**	**23**	**4**	**7**	**13**	**5**	**5**	**10**	**5**	**13**	**0**	106
Féminin	**9**	**3**	**9**	**8**	**4**	**2**	**6**	**8**	**1**	**6**	**23**	**33**	112
Naturel	**7**	**5**	**2**	**7**	**11**	**2**	**6**	**11**	**9**	**24**	**10**	**0**	94
Total	78	112	106	132	83	84	78	71	79	123	84	115	**1145**

Travail à faire:

1. Déterminer le tableau des fréquences
2. Déterminer le tableau des fréquences lignes
3. Déterminer le tableau des fréquences colonnes
4. Calculer les distances entre profils

Chapitre 11 : Analyse Factorielle des Correspondances

1- Déterminer le tableau des fréquences

	Orly	Alezan	Corsaire	Directoire	Ducat	Fontenoy	Icare	Zodiaque	Pavois	Cocker	Escale	Hotesse	Total
Vieille	0,09%	0,17%	1,22%	3,32%	1,57%	0,87%	0,79%	0,44%	0,79%	0,35%	0,00%	0,09%	9,69%
Riche	1,75%	0,79%	0,09%	0,96%	0,87%	0,79%	0,09%	0,09%	1,75%	0,79%	0,61%	1,05%	9,61%
Elégant	0,79%	2,01%	0,09%	1,31%	0,61%	0,96%	0,52%	0,17%	0,61%	1,05%	0,26%	1,48%	9,87%
Comique	0,09%	0,26%	1,31%	1,31%	0,52%	0,44%	1,05%	1,57%	0,35%	2,18%	0,17%	0,17%	9,43%
Racé	0,35%	2,88%	0,61%	0,70%	0,26%	0,52%	0,52%	0,35%	0,44%	1,31%	0,44%	0,26%	8,65%
Mièvre	0,26%	0,79%	0,09%	0,61%	0,61%	0,44%	1,05%	0,79%	0,52%	0,79%	0,52%	1,14%	7,60%
Distingué	0,96%	0,79%	0,09%	1,48%	0,35%	1,83%	0,52%	0,09%	0,44%	0,35%	0,44%	2,36%	9,69%
Vulgaire	0,35%	0,35%	2,79%	0,17%	0,52%	0,00%	0,79%	0,61%	0,26%	0,87%	0,87%	0,61%	8,21%
Masculin	0,79%	1,05%	2,01%	0,35%	0,61%	1,14%	0,44%	0,44%	0,87%	0,44%	1,14%	0,00%	9,26%
Féminin	0,79%	0,26%	0,79%	0,70%	0,35%	0,17%	0,52%	0,70%	0,09%	0,52%	2,01%	2,88%	9,78%
Naturel	0,61%	0,44%	0,17%	0,61%	0,96%	0,17%	0,52%	0,96%	0,79%	2,10%	0,87%	0,00%	8,21%
Total	6,81%	9,78%	9,26%	11,53%	7,25%	7,34%	6,81%	6,20%	6,90%	10,74%	7,34%	10,04%	100,00%

Calcul :

$$f_{ij} = \frac{\text{Effectif de la Cellule (i,j)}}{\text{Effectif Total}} = \frac{n_{i,j}}{N}$$

Fréquence totale = 1145

Cellule (Orly ;Vieille) = 1 / 1145 = 0,09 %
Cellule (Orly ; Riche) = 20 / 1145 = 1,75%
Cellule (Orly ; Elégant) = 9 / 1145 = 0,79%

Chapitre 11 : Analyse Factorielle des Correspondances

2- Déterminer le tableau des fréquences lignes

	Orly	Alezan	Corsaire	Directoire	Ducat	Fontenoy	Icare	Zodiaque	Pavois	Cocker	Escale	Hotesse	Total
Vieille	0,90%	1,80%	12,61%	34,23%	16,22%	9,01%	8,11%	4,50%	8,11%	3,60%	0,00%	0,90%	100,00%
Riche	18,18%	8,18%	0,91%	10,00%	9,09%	8,18%	0,91%	0,91%	18,18%	8,18%	6,36%	10,91%	100,00%
Elégant	7,96%	20,35%	0,88%	13,27%	6,19%	9,73%	5,31%	1,77%	6,19%	10,62%	2,65%	15,04%	100,00%
Comique	0,93%	2,78%	13,89%	13,89%	5,56%	4,63%	11,11%	16,67%	3,70%	23,15%	1,85%	1,85%	100,00%
Racé	4,04%	33,33%	7,07%	8,08%	3,03%	6,06%	6,06%	4,04%	5,05%	15,15%	5,05%	3,03%	100,00%
Mièvre	3,45%	10,34%	1,15%	8,05%	8,05%	5,75%	13,79%	10,34%	6,90%	10,34%	6,90%	14,94%	100,00%
Distingué	9,91%	8,11%	0,90%	15,32%	3,60%	18,92%	5,41%	0,90%	4,50%	3,60%	4,50%	24,32%	100,00%
Vulgaire	4,26%	4,26%	34,04%	2,13%	6,38%	0,00%	9,57%	7,45%	3,19%	10,64%	10,64%	7,45%	100,00%
Masculin	8,49%	11,32%	21,70%	3,77%	6,60%	12,26%	4,72%	4,72%	9,43%	4,72%	12,26%	0,00%	100,00%
Féminin	8,04%	2,68%	8,04%	7,14%	3,57%	1,79%	5,36%	7,14%	0,89%	5,36%	20,54%	29,46%	100,00%
Naturel	7,45%	5,32%	2,13%	7,45%	11,70%	2,13%	6,38%	11,70%	9,57%	25,53%	10,64%	0,00%	100,00%
Total	6,81%	9,78%	9,26%	11,53%	7,25%	7,34%	6,81%	6,20%	6,90%	10,74%	7,34%	10,04%	100,00%

Calcul :

$$fL_{ij} = \frac{n_{i,j}}{n_{i.}} = \frac{f_{i,j}}{f_{i.}} = \frac{\text{Effectif de la Cellule (i, j)}}{\text{Effectif de la ligne i}}$$

Effectif de la ligne I = 111

Cellule (Orly ;Vieille) = 1 / 111 = 0,90 %
Cellule (Orly ; Riche) = 20 / 110 = 18,18%
Cellule (Orly ; Elégant) = 9 / 113 = 7,96%

Chapitre II : Analyse Factorielle des Correspondances

3- Déterminer le tableau des fréquences colonnes

	Orly	Alezan	Corsaire	Directoire	Ducat	Fontenoy	Icare	Zodiaque	Pavois	Cocker	Escale	Hotesse	Total
Vieille	1,28%	1,79%	13,21%	28,79%	21,69%	11,90%	11,54%	7,04%	11,39%	3,25%	0,00%	0,87%	9,69%
Riche	25,64%	8,04%	0,94%	8,33%	12,05%	10,71%	1,28%	1,41%	25,32%	7,32%	8,33%	10,43%	9,61%
Elégant	11,54%	20,54%	0,94%	11,36%	8,43%	13,10%	7,69%	2,82%	8,86%	9,76%	3,57%	14,78%	9,87%
Comique	1,28%	2,68%	14,15%	11,36%	7,23%	5,95%	15,38%	25,35%	5,06%	20,33%	2,38%	1,74%	9,43%
Racé	5,13%	29,46%	6,60%	6,06%	3,61%	7,14%	7,69%	5,63%	6,33%	12,20%	5,95%	2,61%	8,65%
Mièvre	3,85%	8,04%	0,94%	5,30%	8,43%	5,95%	15,38%	12,68%	7,59%	7,32%	7,14%	11,30%	7,60%
Distingué	14,10%	8,04%	0,94%	12,88%	4,82%	25,00%	7,69%	1,41%	6,33%	3,25%	5,95%	23,48%	9,69%
Vulgaire	5,13%	3,57%	30,19%	1,52%	7,23%	0,00%	11,54%	9,86%	3,80%	8,13%	11,90%	6,09%	8,21%
Masculin	11,54%	10,71%	21,70%	3,03%	8,43%	15,48%	6,41%	7,04%	12,66%	4,07%	15,48%	0,00%	9,26%
Féminin	11,54%	2,68%	8,49%	6,06%	4,82%	2,38%	7,69%	11,27%	1,27%	4,88%	27,38%	28,70%	9,78%
Naturel	8,97%	4,46%	1,89%	5,30%	13,25%	2,38%	7,69%	15,49%	11,39%	19,51%	11,90%	0,00%	8,21%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Calcul :

$$fC_{ij} = \frac{n_{ij}}{n_{.j}} = \frac{f_{i,j}}{f_{.j}} = \frac{\text{Effectif de la Cellule (i, j)}}{\text{Effectif de la colonne J}}$$

Effectif de la colonne I = 78

Cellule (Orly ;Vieille) = 1 / 78 = 1,28 %
Cellule (Orly ; Riche) = 20 / 78 = 25,64%
Cellule (Orly ; Elégant) = 9 / 78 = 11,54%

Chapitre 11 : Analyse Factorielle des Correspondances

4- Calculer les distances entre profils

➔ Les distances entre profils ligne

	Orly	Alezan	Corsaire	Directoire	Ducat	Fontenoy	Icare	Zodiaque	Pavois	Cocker	Escale	Hotesse	Total
Vieille	0,438	0,042	0,148	0,509	0,070	0,001	0,076	0,021	0,147	0,020	0,055	0,100	1,627
Riche	0,153	0,151	0,000	0,009	0,012	0,003	0,028	0,001	0,208	0,006	0,019	0,017	0,608
Elégant	0,073	0,316	0,183	0,000	0,001	0,036	0,049	0,358	0,009	0,146	0,001	0,173	1,344
Comique	0,014	0,954	0,050	0,029	0,009	0,003	0,037	0,257	0,003	0,060	0,014	0,001	1,432
Racé	0,001	0,540	0,038	0,000	0,035	0,000	0,088	0,064	0,005	0,022	0,005	0,141	0,938
Mièvre	0,061	0,005	0,000	0,046	0,027	0,236	0,103	0,144	0,008	0,042	0,008	0,088	0,769
Distingué	0,047	0,015	1,186	0,151	0,011	0,488	0,026	0,069	0,002	0,046	0,051	0,284	2,376
Vulgaire	0,026	0,051	0,165	0,002	0,000	0,205	0,035	0,012	0,056	0,033	0,004	0,055	0,644
Masculin	0,000	0,076	0,202	0,010	0,013	0,150	0,001	0,009	0,106	0,000	0,093	0,864	1,524
Féminin	0,001	0,007	0,038	0,000	0,091	0,000	0,002	0,034	0,109	0,379	0,134	0,864	1,658
Naturel	-	-	-	-	-	-	-	-	-	-	-	-	-

Calcul : on se base sur le tableau de fréquences lignes

$$d^2_{\emptyset^2}(L_i, L_{i'}) = \sum_j \frac{(fL_{i,j} - fL_{i',j})^2}{f_{.j}}$$

Chapitre II : Analyse Factorielle des Correspondances

4- Calculer les distances entre profils

	Orly	Alezan	Corsaire	Directoire	Ducat	Fontenoy	Icare	Zodiaque	Pavois	Cocker	Escale	Hotesse	Total
Vieille	0,438	0,042	0,148	0,509	0,070	0,001	0,076	0,021	0,147	0,020	0,055	0,100	1,627
Riche	0,153	0,151	0,000	0,009	0,012	0,003	0,028	0,001	0,208	0,006	0,019	0,017	0,608
Elégant	0,073	0,316	0,183	0,000	0,001	0,036	0,049	0,358	0,009	0,146	0,001	0,173	1,344
Comique	0,014	0,954	0,050	0,029	0,009	0,003	0,037	0,257	0,003	0,060	0,014	0,001	1,432
Racé	0,001	0,540	0,038	0,000	0,035	0,000	0,088	0,064	0,005	0,022	0,005	0,141	0,938
Mièvre	0,061	0,005	0,000	0,046	0,027	0,236	0,103	0,144	0,008	0,042	0,008	0,088	0,769
Distingué	0,047	0,015	1,186	0,151	0,011	0,488	0,026	0,069	0,002	0,046	0,051	0,284	2,376
Vulgaire	0,026	0,051	0,165	0,002	0,000	0,205	0,035	0,012	0,056	0,033	0,004	0,055	0,644
Masculin	0,000	0,076	0,202	0,010	0,013	0,150	0,001	0,009	0,106	0,000	0,093	0,864	1,524
Féminin	0,001	0,007	0,038	0,000	0,091	0,000	0,002	0,034	0,109	0,379	0,134	0,864	1,658
Naturel	-	-	-	-	-	-	-	-	-	-	-	-	-

Calcul :

$$d^2 \text{ (Vieille ; Riche ; Orly)} = (0{,}009-0{,}182)^2 / 0{,}068 = 0{,}438$$

$$d^2 \text{ (Riche ; Elégant ; Orly)} = (0{,}182-0{,}080)^2 / 0{,}068 = 0{,}153$$

Chapitre 11 : Analyse Factorielle des Correspondances

4- Calculer les distances entre profils

➔ **Distances entre profils colonnes**

	Orly	Alezan	Corsaire	Directoire	Ducat	Fontenoy	Icare	Zodiaque	Pavois	Cocker	Escale	Hotesse
Vieille	0,000	0,135	0,250	0,052	0,099	0,000	0,021	0,020	0,068	0,011	0,001	-
Riche	0,323	0,052	0,057	0,014	0,002	0,093	0,000	0,595	0,337	0,001	0,005	-
Elégant	0,082	0,389	0,110	0,009	0,022	0,030	0,024	0,037	0,001	0,039	0,127	-
Comique	0,002	0,140	0,008	0,018	0,002	0,094	0,105	0,436	0,247	0,341	0,000	-
Racé	0,685	0,604	0,000	0,007	0,014	0,000	0,005	0,001	0,040	0,045	0,013	-
Mièvre	0,023	0,066	0,025	0,013	0,008	0,117	0,010	0,034	0,000	0,000	0,023	-
Distingué	0,038	0,052	0,147	0,067	0,420	0,309	0,041	0,025	0,010	0,008	0,317	-
Vulgaire	0,003	0,863	1,001	0,040	0,064	0,162	0,003	0,045	0,023	0,017	0,041	-
Masculin	0,001	0,130	0,376	0,032	0,054	0,089	0,000	0,034	0,080	0,141	0,259	-
Féminin	0,080	0,035	0,006	0,002	0,006	0,029	0,013	0,102	0,013	0,518	0,002	-
Naturel	0,025	0,008	0,014	0,077	0,144	0,034	0,074	0,020	0,080	0,070	0,173	-
Total	1,262	2,474	1,996	0,330	0,834	0,957	0,297	1,349	0,899	1,191	0,960	-

Calcul : on se base sur le tableau de fréquences colonnes

$$d^2 \text{ (Vieille ; Alezan ; Corsaire)} = (0{,}018 - 0{,}132)^2 / 0{,}097 = 0{,}135$$

$$d^2 \text{ (Vieille ; Corsaire; Directoire)} = (0{,}132 - 0{,}288)^2 / 0{,}068 = 0{,}250$$

Chapitre III : Classification Ascendante Hiérarchique

PLAN

Chapitre III : Classification Ascendante Hiérarchique

1-Définition

2- La classification non hiérarchique

3- La classification hiérarchique

4- Cas d'application

Chapitre III : Classification Ascendante Hiérarchique

I- Définition

La classification a pour but de regrouper des individus en classes homogènes en fonction de l'étude de certaines caractéristiques des individus

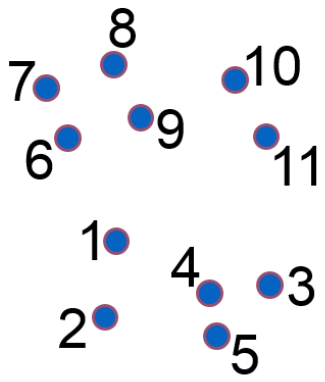
➔ Classes homogènes : consiste à regrouper les individus qui se ressemblent et séparer ceux qui sont éloignés.

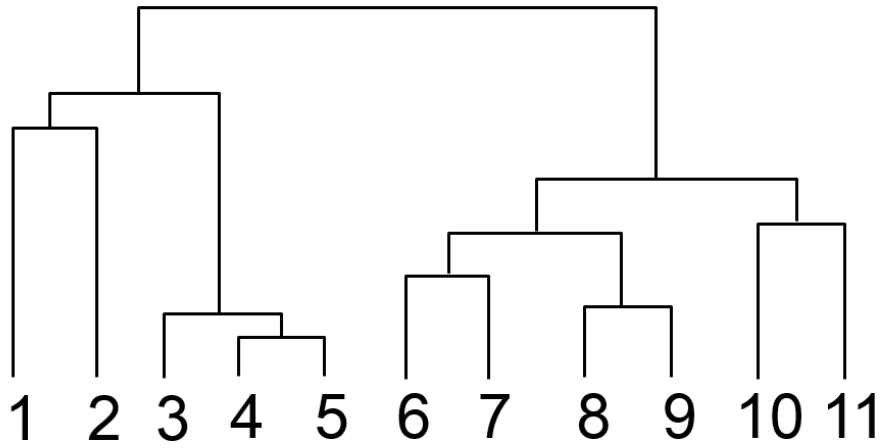
Principe
Les diverses techniques de classification visent à répartir *n individus, caractérisés par p variables X1, X2, ..., Xp en un* certain nombre *m* de sous-groupes aussi homogènes que possible.

Chapitre III : Classification Ascendante Hiérarchique

I- Définition

Exemple :





Chapitre III : Classification Ascendante Hiérarchique

I- Définition

On distingue deux types de techniques de classification :

➔ La classification non hiérarchique ou partitionnement:
la décomposition de l'ensemble de tous les individus en m ensembles disjoints ou classes d'équivalence mais *le nombre m de classes est fixé à l'avance*.

➔ La classification hiérarchique :
pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents.

Chapitre III : Classification Ascendante Hiérarchique

2- La classification non hiérarchique

Partitionner un ensemble d'observations (E) consiste à :

1- Regrouper les observations en *classes homogènes* (les sous-ensemble partagent des caractéristiques communes.

2- Regrouper les observations selon un critère:

- Critère de similarité (la proximité) pour inclure les observations qui se ressemblent dans une classe.
- Critère de différenciation (la distance) pour les exclure

2- La classification non hiérarchique

Notion de ***Partition***

Soit un ensemble E = { A.B.C.D }

Une partition de E est un ensemble de classe qui satisfait deux conditions:

Deux classes A et B : soient disjointes, soient confondues

➔A et B sont disjointes, Si

$$A \cap B = \emptyset$$

➔A et B sont confondues, si

$$A \cup B = A = B$$

➔L'Union de toutes les classe correspond à l'ensemble E

$$A \cup B \cup C \cup D = E$$

Chapitre III : Classification Ascendante Hiérarchique

3- La classification hiérarchique

Elle consiste à fournir un ensemble de partitions de E en classes de moins en moins fines obtenues par regroupements successifs de parties

➔ Pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents.

3- La classification hiérarchique

➔ Les mesures de l'homogénéité des classes

a- Notion de ***Distance***

Soit d est une mesure de distance sur E , si les trois axiomes sont satisfaites :

Axiome de séparation $d_{ih} = 0 \quad alors \quad i = h$

Axiome de symétrie $d_{ih} = d_{hi}$

Axiome d'inégalité triangulaire $d_{ih} \leq d_{ie} + d_{eh}$

Chapitre III : Classification Ascendante Hiérarchique

3- La classification hiérarchique

b- Les mesures de ***Distance***

On distingue de nombreuses mesures de la "distance" entre individus:

- Distance Euclidienne,
- Distance Euclidienne au carré,
- Distance du City-block (Manhattan),
- Distance de Tchebychev…

Le choix d'une (ou plusieurs) d'entre elles dépend des données étudiées.

➔ Distance Euclidienne : le type de distance le plus couramment utilisé :

$$d(I_i, I_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$

➔ Distance Euclidienne au carré :

$$d(I_i, I_j) = \sum_k (x_{ik} - x_{jk})^2$$

Chapitre III : Classification Ascendante Hiérarchique

3- La classification hiérarchique

c- Démarche

Centrage et réduction des données

⇩

Choix des critère de la classification (matrice de proximité)

⇩

Présentation des classes sous forme de dendrogramme

Chapitre III : Classification Ascendante Hiérarchique

4- Cas d'application

Soit les données d'une étude pour segmenter le marché d'une entreprise

	V1	V02	V03	V04
I1	200	2	39	418
I2	250	2	29	153
I3	100	1	40	309
I4	104	1	46	210
I5	420	3	47	117
I6	500	2	46	106
I7	820	3	52	198
I8	640	1	42	126

Travail à faire:

1- Présenter le tableau des données centrées réduites

2- Présenter la matrice de proximité par la distance euclidienne

3- Présenter le dendrogramme et interpréter les résultats des classes

Chapitre III : Classification Ascendante Hiérarchique

1- Présenter le tableau des données centrées réduites

	V1	V02	V03	V04
I1	200	2	39	418
I2	250	2	29	153
I3	100	1	40	309
I4	104	1	46	210
I5	420	3	47	117
I6	500	2	46	106
I7	820	3	52	198
I8	640	1	42	126
Total	**3034**	**15**	**341**	**1637**
Moyenne	**379,25**	**1,88**	**42,63**	**204,63**
Ecart type	**262,15**	**0,83**	**6,93**	**108,72**

Chapitre III : Classification Ascendante Hiérarchique

1- Présenter le tableau des données centrées réduites

	V1	V02	V03	V04
I1	-0,68	0,15	-0,52	1,96
I2	-0,49	0,15	-1,97	-0,47
I3	-1,07	-1,05	-0,38	0,96
I4	-1,05	-1,05	0,49	0,05
I5	0,16	1,35	0,63	-0,81
I6	0,46	0,15	0,49	-0,91
I7	1,68	1,35	1,35	-0,06
I8	0,99	-1,05	-0,09	-0,72
Total	0	0	0	0

Données centrées réduites : $x_{ij} = \frac{x_{ij} - \overline{X}}{\sigma_i}$

Chapitre III : Classification Ascendante Hiérarchique

2- Présenter la matrice de proximité par la distance euclidienne

➔ Distance Euclidienne :

$$d(I_i, I_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$

Chapitre III : Classification Ascendante Hiérarchique

2- Présenter la matrice de proximité par la distance euclidienne

	V01	V02	V03	V04	Distance
I1-I2	0,04	0,00	2,08	5,94	2,84
I1-I3	0,15	1,44	0,02	1,01	1,61
I1-I4	0,13	1,44	1,02	3,66	2,50
I1-I5	0,70	1,44	1,33	7,66	3,34
I1-I6	1,31	0,00	1,02	8,23	3,25
I1-I7	5,59	1,44	3,52	4,09	3,83
I1-I8	2,82	1,44	0,19	7,21	3,41
I2-I3	0,33	1,44	2,52	2,06	2,52
I2-I4	0,31	1,44	6,02	0,27	2,84
I2-I5	0,42	1,44	6,75	0,11	2,95
I2-I6	0,91	0,00	6,02	0,19	2,67
I2-I7	4,73	1,44	11,02	0,17	4,17
I2-I8	2,21	1,44	3,52	0,06	2,69
I3-I4	0,00	0,00	0,75	0,83	1,26
I3-I5	1,49	5,74	1,02	3,12	3,37
I3-I6	2,33	1,44	0,75	3,49	2,83
I3-I7	7,54	5,74	3,00	1,04	4,16
I3-I8	4,24	0,00	0,08	2,83	2,68
I4-I5	1,45	5,74	0,02	0,73	2,82
I4-I6	2,28	1,44	0,00	0,91	2,15
I4-I7	7,46	5,74	0,75	0,01	3,74
I4-I8	4,18	0,00	0,33	0,60	2,26
I5-I6	0,09	1,44	0,02	0,01	1,25
I5-I7	2,33	0,00	0,52	0,56	1,85
I5-I8	0,70	5,74	0,52	0,01	2,64
I6-I7	1,49	1,44	0,75	0,72	2,10
I6-I8	0,29	1,44	0,33	0,03	1,45
I7-I8	0,47	5,74	2,08	0,44	2,96

Chapitre III : Classification Ascendante Hiérarchique

2- Présenter la matrice de proximité par la distance euclidienne

	I1	I2	I3	I4	I5	I6	I7	I8
I1	0							
I2	2,84	0						
I3	1,61	2,52	0					
I4	2,50	2,84	1,26	0				
I5	3,34	2,95	3,37	2,82	0			
I6	3,25	2,67	2,83	2,15	1,25	0		
I7	3,83	4,17	4,16	3,74	1,85	2,10	0	
I8	3,41	2,69	2,68	2,26	2,64	1,45	2,96	0

La distance entre I6 et I8 est très réduite (1.45) ➔ alors les individus I6 et I8 se ressemblent.

La distance entre I2 et I7 est très importante (4.17) ➔ alors ces individus sont éloignés.

Chapitre III : Classification Ascendante Hiérarchique

3- Présenter le dendrogramme et interpréter les résultats des classes

	I1	I2	I3	I4	I5	I6	I7	I8
I1	0							
I2	2,84	0						
I3	1,61	2,52	0					
I4	2,50	2,84	1,26	0				
I5	3,34	2,95	3,37	2,82	0			
I6	3,25	2,67	2,83	2,15	1,25	0		
I7	3,83	4,17	4,16	3,74	1,85	2,10	0	
I8	3,41	2,69	2,68	2,26	2,64	1,45	2,96	0

Saut minimum ➔ $$D(A,B) = \min_{I \in A} \min_{J \in B} d(I,J)$$

	I1	I2	I3	I4	I5-I6	I7	I8
I1	0						
I2	2,84	0					
I3	1,61	2,52	0				
I4	2,50	2,84	1,26	0			
I5-I6	3,25	2,67	2,83	2,15	0,00		
I7	3,83	4,17	4,16	3,74	0,60	0	
I8	3,41	2,69	2,68	2,26	0,20	2,96	0

Les classes obtenues :
(I5-I6) = 1.25

Chapitre III : Classification Ascendante Hiérarchique

3- Présenter le dendrogramme et interpréter les résultats des classes

	I1	I2	I3	I4	I5-I6	I7	I8
I1	0						
I2	2,84	0					
I3	1,61	2,52	0				
I4	2,50	2,84	1,26	0			
I5-I6	3,25	2,67	2,83	2,15	0,00		
I7	3,83	4,17	4,16	3,74	0,60	0	
I8	3,41	2,69	2,68	2,26	0,20	2,96	0

	I1	I2	I3	I4	I5-I6-I8	I7
I1	0					
I2	2,84	0				
I3	1,61	2,52	0			
I4	2,50	2,84	1,26	0		
I5-I6-I8	3,25	2,67	2,68	2,15	0,00	
I7	3,83	4,17	4,16	3,74	0,60	0

Les classes obtenues :

(I5-I6) = 1.25

(I5-I6-I8) = 0.2

Chapitre III : Classification Ascendante Hiérarchique

3- Présenter le dendrogramme et interpréter les résultats des classes

	I1	I2	I3	I4	I5-I6-I8	I7
I1	0					
I2	2,84	0				
I3	1,61	2,52	0			
I4	2,50	2,84	1,26	0		
I5-I6-I8	3,25	2,67	2,68	2,15	0,00	
I7	3,83	4,17	4,16	3,74	0,60	0

	I1	I2	I3	I4	I5-I6-I8-I7
I1	0				
I2	2,84	0			
I3	1,61	2,52	0		
I4	2,50	2,84	1,26	0	
I5-I6-I8-I7	3,25	2,67	2,68	2,15	0,00

Les classes obtenues :

(I5-I6) = 1.25

(I5-I6-I8) = 0.2

(I5-I6-I8-I7)=0.6

Chapitre III : Classification Ascendante Hiérarchique

3- Présenter le dendrogramme et interpréter les résultats des classes

	I1	I2	I3	I4	I5-I6-I8-I7
I1	0				
I2	2,84	0			
I3	1,61	2,52	0		
I4	2,50	2,84	1,26	0	
I5-I6-I8-I7	3,25	2,67	2,68	2,15	0,00

	I1	I2	I3-I4	I5-I6-I8-I7
I1	0			
I2	2,84	0		
I3-I4	1,61	2,52	0,00	
I5-I6-I8-I7	3,25	2,67	2,15	0,00

	I3-I4-I1	I2	I5-I6-I8-I7
I3-I4-I1	0		
I2	1,23	0	
I5-I6-I8-I7	2,15	2,67	0,00

Les classes obtenues :

(I5-I6) = 1.25
(I5-I6-I8) = 0.2
(I5-I6-I8-I7)=0.6
(I3-I4) =1.26
(I3-I4-I1)=1.61

Chapitre III : Classification Ascendante Hiérarchique

3- Présenter le dendrogramme et interpréter les résultats des classes

	I3-I4-I1	I2	I5-I6-I8-I7
I3-I4-I1	0,00		
I2	1,23	0	
I5-I6-I8-I7	2,15	2,67	0,00

	I3-I4-I1-I2	I5-I6-I8-I7
I3-I4-I1-I2	0,00	
I5-I6-I8-I7	2,15	0,00

Les classes obtenues :

(I5-I6) = 1.25
(I5-I6-I8) = 0.2
(I5-I6-I8-I7)=0.6
(I3-I4) =1.26
(I3-I4-I1)=1.61
(I3-I4-I1-I2)=1.23
(I3-I4-I1-I2- I5-I6-I8-I7)=2.15