الجمهوريـــــــــــــــة الجزائريـــــــــة الديمقراطيـــــــة الشعبيـــــة
وزارة التعليـــــــم العـــــــالي و البحـــــــث العلـــــمي
جامعة وهران للعلوم والتكنولوجيا محمد بوضياف
كلية الرياضيات و الاعلام الالي

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur Et de la Recherche Scientifique

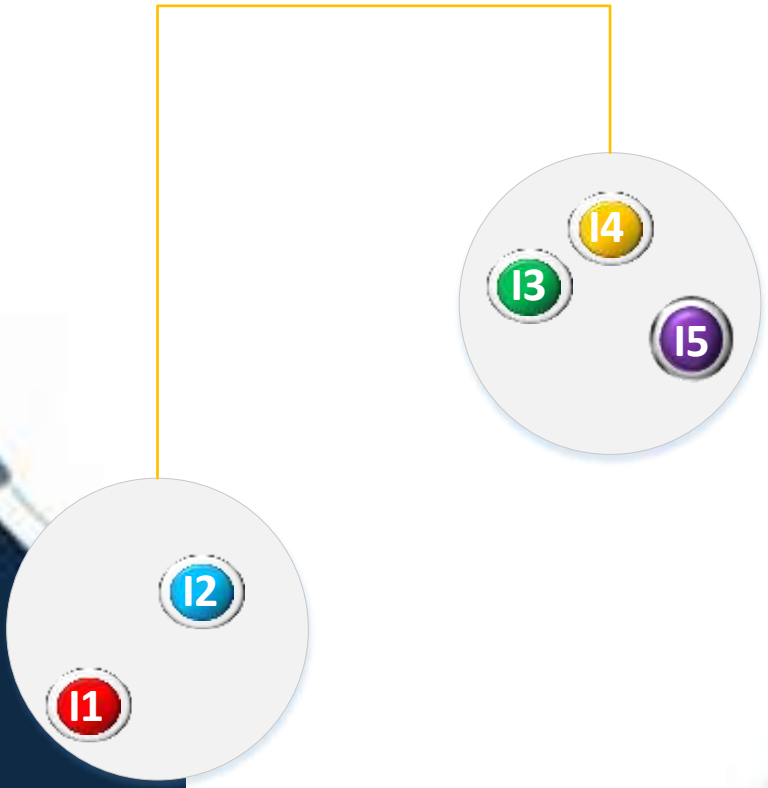
Université des Sciences et de la Technologie d'Oran Mohamed BOUDIAF

Faculté des Mathématiques et Informatique

Département d'Informatique

COURS D'ANALYSE DES DONNÉES

Destiné aux étudiants Master1-Informatique



AND

Dr Nabil NEGGAZ

Les techniques d'analyse des données ont connu un essor important surtout avec le développement de l'informatique et big data. Le volume important des données nécessite comme un prétraitement : la réduction des données, ce qui est l'objectif principal de l'analyse des données en premier lieu. Pour résoudre le problème de la dimensionalité, les méthodes multidimensionnelles telles que l'Analyse en Composantes Principales (ACP) et l'Analyse Factorielle des Correspondances (AFC) seront exploitées et expliquées en détail dans cet ouvrage. En second lieu, l'interprétation et la classification des données dans le domaine de la reconnaissance des formes, la fouille des données et l'intelligence artificielle font appel aux méthodes de classification plus particulièrement l'algorithme de la classification hiérarchique qui permet une représentation arborescente appelée dendrogramme et les méthodes de partitionnement « clustering » comme l'algorithme des centres mobiles qui est très utilisé dans l'apprentissage non-supervisé. En plus, les méthodes morphologiques à base des opérateurs de traitement d'image comme l'érosion, dilatation, ouverture et fermeture peuvent être utilisées dans le domaine de la classification.

La prévision dans le domaine d'économie et le domaine d'épidémiologie nécessitent des modèles statistiques puissants. Pour cela, nous avons introduit la méthode des moindres carrés et les séries chronologiques. Généralement, deux modèles sont exploités comme la prévision linéaire et la prévision exponentielle. Pour juger l'efficacité des modèles proposés, un coéfficient de corrélation doit être mesuré.

Cet ouvrage est organisé autour de cinq chapitres de la façon suivante :

- **Chapitre 1** décrit d'une façon détaillée la nature des variables, les mesures de similarité ainsi les différents types de tableaux de données. De plus, quelques ingrédients dédiés à l'analyse factorielle des correspondances (AFC) ont été proposés.
- **Chapitre 2** représente le noyau de cet ouvrage car il explique profondément l'algorithme d'ACP avec des exercices d'applications en variant la métrique selon des données homogènes/hétérogènes.
- **Chapitre 3** a un grand impact dans le clustering c.à.d. la classification non-supervisée. Ce chapitre explique 4 méthodes de classification : Classification Hiérarchique ascendante (CHA), Algorithme des Centres Mobiles (ACM), Maximum de Vraisemblance et les méthodes morphologiques
- **Chapitre 4** joue un rôle important dans la prévision en se basant sur des méthodes statistiques comme la méthode des moindres carrés.
- **Chapitre 5** : introduit les séries chronologiques qui entre dans le cadre de la prédiction en tenant compte de la composante temporelle.

Enfin, dans l'espoir que cet ouvrage constitue la première marche d'un long escalier et permet aux lecteurs d'acquérir des nouvelles connaissances en analyse des données.

SOMMAIRE

DESCRIPTION DES TABLEAUX DE DONNÉES

1.1 Introduction

Ce chapitre est dédié à l'introduction de quelques notions de bases de statistique multi-dimensionnelle comme la nature des variables et leurs codages adéquats, les tableaux de données, les mesures de similarités/dissimilarités, les tableaux de fréquences, les profils lignes et profiles colonnes. Il est important de connaitre les différentes métriques utilisées pour le calcul de la distance. Généralement, deux métriques sont exploitées comme la métrique d'identité et la métrique $\aleph 2$ utilisée pour l'analyse factorielle des correspondances (AFC). La réduction des tableaux de données reste un vrai challenge qui sera introduit dans ce chapitre en définissant le tableau de BURT et la réduction par regroupement.

1.2 Mesures de similarité

- **Distance**

Notons E l'ensemble des N objets à classer, une distance est une application de $E \times E$ dans $\mathbb{R}^+$ telle que :

$$\begin{cases} d(i,j) = d(j,i) \\ d(i,j) \geq 0 \\ d(i,i) = 0 \Leftrightarrow i = j \\ d(i,j) \leq d(i,k) + d(k,j) \end{cases}$$

Généralement, on trouve la distance euclidienne qui est utilisée pour des variables mesurables (quantitative). La distance euclidienne est calculée par :

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

D'une façon générale, la distance est munie d'une métrique M et elle est définie par :

$$d^2(X,Y) = (X-Y)^t M (X-Y)$$

$$d^2(X,Y) = \|X-Y\|_M^2 = (X-Y)^t .M. (X-Y)$$

$$d_M^2 = \sum_{i=1}^{n} (x_i - y_i)^2 \; si \; M = I \Rightarrow d_E$$

- **Mesures de similarité**

On parle de dissimilarité si s est une application telle que : $\begin{cases} s(i,j) = s(j,i) \\ s(i,j) \geq 0 \\ s(i,i) \geq s(i,j) \end{cases}$

Les mesures de similarités sont utilisées dans le cas où les individus sont décrits par la présence ou l'absence de p caractéristiques. Plusieurs indices de similarité ont été

proposées qui combinent de diverse manières les quatre nombres suivant associés à un couple d'individus :

↶	1	0
1	a	b
0	c	d

- **a** représente le nombre de fois où $x_{ij}=1$ et $x_{i'j'}=1$.
- **b** représente le nombre de fois où $x_{ij}=0$ et $x_{i'j'}=1$.
- **c** représente le nombre de fois où $x_{ij}=1$ et $x_{i'j'}=0$.
- **d** représente le nombre de fois où $x_{ij}=0$ et $x_{i'j'}=0$

Les indices suivants compris entre 0 et 1 et qui sont facilement transformables en dissimilarité par complémentation à 1 [1]:

1) L'indice de Jaccard : $S(I,J)=\dfrac{a}{a+b+c}$
2) L'indice de Dice : $S(I,J)=\dfrac{2a}{2a+b+c}$
3) L'indice de Russsel et Rao : $S(I,J)=\dfrac{a}{a+d+b+c}$
4) L'indice de Rogers et Tanimoto : $S(I,J)=\dfrac{a+d}{a+d+2(b+c)}$
5) L'indice de Jaccard : $S(I,J)=\dfrac{2a}{2a+b+c}$
6) L'indice de Sokal et Sneath : $S(I,J)=\dfrac{a}{a+2(c+d)}$
7) L'indice de Kulzinsky : $S(I,J)=\dfrac{a}{c+d}$

La notion fondamentale en statistique est celle de groupe ou d'ensemble d'objets équivalents à une population. Ces objets sont appelés des **individus**. Alors un individu est décrit par un ensemble de caractéristiques appelées **variables**.

On distingue principalement deux types de variables [2]:

- **Variables quantitatives** par exemple Age, poids et taille, s'expriment par des nombres réels sur lesquels on peut appliquer des opérations arithmétiques (moyenne) ont un sens. Certaines sont discrètes (ensemble dénombrable des modalités) comme le nombre d'articles lus par un chercheur quotidiennement, le nombre d'enfants, etc , par ailleurs d'autres sont continues si toutes les valeurs d'un intervalle de $\mathbb{R}$ sont acceptables.
- **Variables qualitatives** par exemple couleur des yeux, La mention obtenue en bac, etc s'exprimant à l'appartenance à une modalité d'un ensemble fini. Deux types se trouvent dans la littérature : les variables qualitatives nominale par exemple la catégorie socio-professionnelle d'un travailleur (Cadre, employé, ouvrier…), tandis que les variables

qualitatives ordinales requièrent une relation d'ordre entre les modalités ; par exemple : les modalités obtenues en bac sont : passable, assez bien, bien et très bien.

1.2.1 Codage des variables qualitatives nominales (VQN)

Soit la fonction $\mathcal{N}$: VQN $\rightarrow \{0,1\}^M$; M est le nombre maximum que peut prendre la variable nominale $\mathcal{N}(x_{ij}) = (0,0,0\cdots,1,0,0\cdots 0)$: $\begin{cases} 1 & pour\ la\ coordonnée\ de\ rang\ x_{ij} \\ 0 & ailleur \end{cases}$

1.2.2 Codage des variables qualitatives ordinales ($\mathcal{O}$)

$\mathcal{O}(x_{ij}) = (1,1,1,1,0,0\cdots 0)$: $\begin{cases} 1 & pour\ toute\ les\ coordonnées\ \ jusqu'au\ rang\ x_{ij} \\ 0 & ailleur \end{cases}$

1.3 Construction de tableau de données

- **Tableau individus/variables** : ce type de tableau contient n lignes et n p colonnes de telle sorte les lignes représentent les individus, tandis que les colonnes représentent les variables.

  Exemple : le tableau suivant contient 10 individus caractérisés par 3 variables : couleur, poids et forme.

Produit	Couleur	Poids	Forme
P1	Bleu	Lourd	Rond
P2	Blanc	Léger	Carré
P3	Vert	Léger	Rond
P4	Bleu	Très lourd	Carré
P5	Blanc	Léger	Rond
P6	Vert	Très lourd	Carré
P7	Bleu	Léger	Rond
P8	Blanc	Très lourd	Carré
P9	Vert	Très lourd	Rond
P10	Bleu	Lourd	Carré

- **Tableau de contingence** : est un tableau d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur une même population de n individus. On peut dire aussi que c'est un tableau de type variable/variable avec $K = \sum_i \sum_j Kij$.

Exemple :

- **Tableaux de contingences : (Couleur/ Poids) ;** $K = \sum_i \sum_j Kij = 10$

Couleur \ Poids	Léger	Lourd	Très Lourd
Bleu	**1**	**2**	**1**
Blanc	**2**	**0**	**1**
Vert	**1**	**0**	**2**

- **Tableaux de contingences : (Couleur/ Formes) ;** $K = \sum_i \sum_j Kij = 10$

Couleur \ Forme	Rond	Carré
Bleu	**2**	**2**
Blanc	**1**	**2**
Vert	**2**	**1**

- **Tableaux de contingences : (Poids/Formes) ;** $K = \sum_i \sum_j Kij = 10$

Poids \ Forme	Rond	Carré
Léger	**3**	**1**
Lourd	**1**	**1**
Très Lourd	**1**	**3**

- **Tableau de codage et tableau de codage disjonctif complet** : lorsque les variables sont toutes qualitatives, le tableau où x_i^j indique le numéro de la modalité de la variable X^J à laquelle appartient l'individu i est le tableau de codage. Les numéros de des modalités étant arbitraire, on lui associera le tableau disjonctif à $m_1 + m_2 + \cdots + m_p$ colonnes constitué de la façon suivante :

  A toute variable à m_j modalités on substitue un ensemble de m_j variables valant 0 ou 1.

  Pour meix comprendre le principe de codage, nous traitons l'exemple suivant :

  Exemple :

Produit	**Couleur**	**Poids**	**Forme**
P1	Bleu	Lourd	Rond
P2	Blanc	Léger	Carré
P3	Vert	Léger	Rond
P4	Bleu	Très lourd	Carré
P5	Blanc	Léger	Rond
P6	Vert	Très lourd	Carré
P7	Bleu	Léger	Rond
P8	Blanc	Très lourd	Carré
P9	Vert	Très lourd	Rond
P10	Bleu	Lourd	Carré

- Couleur (bleu, blanc, vert) : variable qualitative nominale qui a 3 modalités alors :

N (bleu)=(1,0,0) ,

N (blanc)=(0,1,0) et N (vert)=(0,0,1) ;

- **Poids (**Léger, Lourd, Très lourd) possède 3 modalités avec Léger <Lourd <Très lourd alors Poids représente une variable qualitative ordinale.
  O (Léger)= (1, 0, 0) ; **O** (Lourd)= (1, 1, 0) ; **O (**Très lourd)= (1, 1, 1)
  Forme (Rond, carré) : possdè deux modalités et elle est qualitative nominale
  N (Rond) =(1, 0) ; **N** (Carré) =(0, 1)

Donc le tableau **de codage disjonctif** contient 10 lignes et 8 colonnes .

	Couleur			Poids			Forme	
	V1	V2	V3	V4	V5	V6	V7	V8
P1	1	0	0	1	1	0	1	0
P2	0	1	0	1	0	0	0	1
P3	0	0	1	1	0	0	1	0
P4	1	0	0	1	1	1	0	1
P5	0	1	0	1	0	0	1	0
P6	0	0	1	1	1	1	0	1
P7	1	0	0	1	0	0	1	0
P8	0	1	0	1	1	1	0	1
P9	0	0	1	1	1	1	1	0
P10	1	0	0	1	1	0	0	1

- **Tableau de distance** : ce tableau est symétrique tels que les lignes et les colonnes représentent les individus.

Exemple : Pour déterminer le tableau de distance, on a besoin de définir une mesure de dissemblance/ressemblance.

Par exemple on a les deux individus suivants :

$I1:1010$

$I2:1001$; La mesure de dissemblance : $d(I,J)=\dfrac{b+c}{a+b+c+d}$

- **a** représente le nombre de fois où $x_{ij}=1$ et $x_{i'j'}=1$.
- **b** représente le nombre de fois où $x_{ij}=0$ et $x_{i'j'}=1$.
- **c** représente le nombre de fois où $x_{ij}=1$ et $x_{i'j'}=0$.
- **d** représente le nombre de fois où $x_{ij}=0$ et $x_{i'j'}=0$

↶	1	0
1	a	b
0	c	d

La distance entre I1 et I2 : $\Rightarrow a=1;b=1;c=1;d=1.\Rightarrow d(I1,I2)=2/4=1/2$

On suppose qu'on a le tableau de codage suivant :

	Réponse 1		Réponse 2	
	V1	V2	V3	V4
I1	1	0	1	0
I2	1	0	0	1
I3	0	1	1	0
I4	0	1	0	1
I5	1	0	1	0
I6	1	0	0	1
I7	0	1	1	0

Alors le tableau de distance correspondant en utilisant La mesure de dissemblance : $d(I,J)=\frac{b+c}{a+b+c+d}$

Est le suivant :

D	I1	I2	I3	I4	I5	I6	I7
I1	**0**	**1/2**	**1/2**	**1**	**0**	**1/2**	**1/2**
I2		**0**	**1**	**½**	**1/2**	**0**	**1**
I3			**0**	**½**	**1/2**	**1**	**0**
I4				**0**	**1**	**1/2**	**1/2**
I5					**0**	**1/2**	**1/2**
I6						**0**	**1/2**
I7							**0**

- **Tableau de BURT** : sert à réduire le tableau de codage disjonctif complet en calculant le produit matriciel $X'X$.

Remarque :

Dans le cas où les variables possèdent un codage qualitatif ordinal, on doit ignorer la notion d'ordre au niveau des variables qualitatives ordinales c.à.d. on doit construire un tableau de codage disjonctif complet et puis on calcule le tableau de Burt par $X'X$.

1.4 Statistique à deux variables

Dans la partie statistique à deux variable, on trouve l'algorithme d'AFC qui a été introduit par Bezécri sous le nom d'analyse des correspondance [3]. Cette dernière traite le tableau de contingence (variable/variable), appelé aussi tableau croisé qui contient n lignes et p colonnes relatifs aux modalités des deux variables x et Y, respectivement.

Y / X	y_1	y_2		y_j		y_p
x_1						
x_2						
x_i				K_{ij}		
x_n						

1.4.1 Les ingrédients nécessaires pour une analyse factorielle des correspondances (AFC)

- **Tableau de fréquences** : ce tableau est obtenu après normalisation du tableau de contingence en divisant par K (le nombre total des individus) c.à.d. $f_{ij} = \frac{K_{ij}}{K}$ avec $K = \sum_i \sum_j Kij$. Les $f_{i\bullet}$ et les $f_{\bullet j}$ s'appellent respectivement les fréquences marginales lignes et les fréquences marginales colonnes telles que :

$$f_{i.} = \sum_{j=1}^{p} f_{ij} \text{ et } f_{.j} = \sum_{i=1}^{n} f_{ij}$$

Exemple : soit le **tableaux de contingences : (Couleur/ Formes) ;** $K = \sum_i \sum_j Kij = 10$

Couleur / Forme	Rond	Carré
Bleu	**2**	**2**
Blanc	**1**	**2**
Vert	**2**	**1**

Alors le **tableau de fréquence :** $f_{ij} = \frac{K_{ij}}{K}$ est :

f_{ij}	f_{i1}	f_{i2}	$f_{i.}$
f_{1j}	**2/10**	**2/10**	**4/10**
f_{2j}	**1/10**	**2/10**	**3/10**
f_{3j}	**2/10**	**1/10**	**3/10**
$f_{.j}$	**5/10**	**5/10**	**1**

- **Tableau des profils lignes** F_J^i : est un tableau des fréquences conditionnelles $F_J^i = \frac{f_{ij}}{f_{i\bullet}}$, cette valeur représente la probabilité d'avoir la modalité j de la variable Y sachant que la modalité de la variable X est i. Le nuage de point est un couple composé de cordonnées des individus F_J^i associés à leurs poids $f_{i\bullet}$, noté par : $N(I) = \left\{\left(F_J^i, f_{i\bullet}\right), i = 1.....n\right\}$.

Exemple :

- **Tableau des profils lignes :** F_J^i : $F_J^i = \frac{f_{ij}}{f_{i.}}$

F_J^i	F_1^i	F_2^i	$f_{i.}$
F_J^1	**2/4**	**2/4**	**4/10**
F_J^2	**1/3**	**2/3**	**3/10**
F_J^3	**2/3**	**1/3**	**3/10**

$$N(I)=\left\{\left(F_J^i, f_{i.}\right), i=1.....n\right\}=\left\{\left[(2/4;2/4),4/10\right] \;\; ;\left[(1/3;2/3),3/10\right] ;\left[(2/3;1/3),3/10\right]\right\}$$

Un raisonnement similaire peut être réalisé pour les colonnes, alors on parle de s profiles colonnes.

- **Tableau des profils colonnes** : F_I^j : est un tableau des fréquences conditionnelles $F_I^j=\frac{f_{ij}}{f_{\bullet j}}$, cette valeur représente la probabilité d'avoir la modalité i de la variable X sachant que la modalité de la variable Y est j. Le nuage de point est un couple composé de cordonnées des individus F_I^j associés à leurs poids $f_{\bullet j}$, noté par : $N(J)=\left\{\left(F_I^j, f_{\bullet j}\right), j=1.....p\right\}$.

Exemple :

- **Tableau des profils colonnes :** F_I^j : $F_I^j=\frac{f_{ij}}{f_{.j}}$

F_I^j	F_I^1	F_I^2
F_1^j	**2/5**	**2/5**
F_2^j	**1/5**	**2/5**
F_3^j	**2/5**	**1/5**
$f_{.j}$	**5/10**	**5/10**

$$N(J)=\left\{\left(F_I^j, f_{.j}\right), j=1..... p\right\}=\left\{\left[(2/5;1/5;2/5),5/10\right] \;\; ;\left[(2/5;2/5;1/5),5/10\right] ;\right\}$$

1.4.2 La ressemblance entre profils :

Pour déterminer la ressemblance entre deux individus de profils lignes ou profiles colonnes, on calcule la distance munie d'une métrique $\aleph^2$ qui est définie par :

de la façon suivante :

- La matrice diagonale $D_{\frac{1}{f_{\bullet j}}}$ définie la métrique pour le nuage $N(I)$, tandis que la matrice diagonale $D_{\frac{1}{f_{i\bullet}}}$ définie la métrique pour le nuage $N(J)$. Alors la distance entre deux individus profils lignes est donnée par : $d^2\left(F_J^i, F_j^{i'}\right)=\left\|F_J^i-F_j^{i'}\right\|^2_{D_{\frac{1}{f.j}}}=\sum_{j=1}^{p}\left(F_J^i-F_j^{i'}\right)^2/f_{.j}$, tandis que la distance entre deux individus profils colonnes est donnée par : $d^2\left(F_I^j, F_i^{j'}\right)=\left\|F_I^j-F_i^{j'}\right\|^2_{D_{\frac{1}{fi.}}}=\sum_{i=1}^{n}\left(F_I^j-F_I^{j'}\right)^2/f_{i.}$

Exemple :

$$d^2\left(F_J^1, F_J^2\right) = \left\|F_J^1 - F_J^2\right\|^2_{D_{\frac{1}{f_{.j}}}}$$

- **Le tableau de distance pour le nuage** $N(I)$ **:**

$$= \sum_{j=1}^{p} \left(F_J^1 - F_J^2\right)^2 / f_{.j}$$

Soit le Tableau des profils lignes : $F_J^i \ : \ F_J^i = \dfrac{f_{ij}}{f_{i.}}$

F_J^i	F_1^i	F_2^i	$f_{i.}$
F_J^1	**2/4**	**2/4**	**4/10**
F_J^2	**1/3**	**2/3**	**3/10**
F_J^3	**2/3**	**1/3**	**3/10**

$$N(I) = \left\{\left(F_J^i, f_{i.}\right), i = 1.....n\right\} = \left\{[(2/4;2/4),4/10] \quad ;[(1/3;2/3),3/10] \ ;[(2/3;1/3),3/10]\right\}$$

$f_{.j} = \begin{pmatrix} 5/10 \\ 5/10 \end{pmatrix}$ alors :

$$d^2\left(F_J^1, F_J^2\right) = \left\|F_J^1 - F_J^2\right\|^2_{D_{\frac{1}{f_{.j}}}} = \sum_{j=1}^{p} \left(F_J^1 - F_J^2\right)^2 / f_{.j} = \frac{(2/4 - 1/3)^2}{5/10} + \frac{(2/4 - 2/3)^2}{5/10} = \frac{1}{9}$$

$$d^2\left(F_J^1, F_J^3\right) = \left\|F_J^1 - F_J^3\right\|^2_{D_{\frac{1}{f_{.j}}}} = \sum_{j=1}^{p} \left(F_J^1 - F_J^3\right)^2 / f_{.j} = \frac{(2/4 - 2/3)^2}{5/10} + \frac{(2/4 - 1/3)^2}{5/10} = \frac{1}{9}$$

$$d^2\left(F_J^2, F_J^3\right) = \left\|F_J^2 - F_J^3\right\|^2_{D_{\frac{1}{f_{.j}}}} = \sum_{j=1}^{p} \left(F_J^2 - F_J^3\right)^2 / f_{.j} = \frac{(1/3 - 2/3)^2}{5/10} + \frac{(2/3 - 1/3)^2}{5/10} = \frac{4}{9}$$

$d^2\left(F_J^i, F_J^{i'}\right)$	F_J^1	F_J^2	F_J^3
F_J^1	**0**	**1/9**	**1/9**
F_J^2	**1/9**	**0**	**4/9**
F_J^3	**1/9**	**4/9**	**0**

- **Centre de gravité pour chaque nuage** : généralement, le centre de gravité est donné par la formule suivante : $g = \frac{\sum_{i=1}^{n} P_i * x_i}{\sum_{i=1}^{n} P_i}$ alors le Centre de gravité pour le nuage $N(I)$ est défini par :

$$g_{N(I)} = \frac{\sum_{i=1}^{n} f_{i.} * F_J^i}{\sum_{i=1}^{n} f_{i.}} =$$

$$\frac{\sum_{i=1}^{n} f_{.j} * \frac{f_{ij}}{f_{.j}}}{\sum_{i=1}^{n} f_{i.}} = \frac{\sum_{i=1}^{n} f_{ij}}{\sum_{i=1}^{n} f_{i.}}$$

$$= \frac{f_{.j}}{\sum_{i=1}^{n}\sum_{j=1}^{p} f_{ij}} = \frac{f_{.j}}{\sum_{i=1}^{n}\sum_{j=1}^{p} \frac{K_{ij}}{K}} = \frac{f_{.j}}{\frac{1}{K}\sum_{i=1}^{n}\sum_{j=1}^{p} K_{ij}} = \frac{f_{.j}}{\frac{1}{K} * K}$$

$$= f_{.j}$$

Concernant le centre de gravité pour le nuage $N(J)$ est défini par :

$$g_{N(J)} = \frac{\sum_{j=1}^{p} f_{.j} * F_I^j}{\sum_{j=1}^{p} f_{.j}}$$

$$= \frac{\sum_{j=1}^{p} f_{.j} * \frac{f_{ij}}{f_{.j}}}{\sum_{j=1}^{p} f_{.j}} = \frac{\sum_{j=1}^{p} f_{ij}}{\sum_{j=1}^{p} f_{.j}}$$

$$= \frac{f_{i.}}{\sum_{j=1}^{p}\sum_{i=1}^{n} f_{ij}} = \frac{f_{i.}}{\frac{1}{K}\sum_{j=1}^{p}\sum_{i=1}^{n} K_{ij}} = \frac{f_{i.}}{\frac{1}{K} * K}$$

$$= f_{i.}$$

Remarque :

Le centre de gravité pour le nuage $N(I)$ corresponds aux fréquences marginales colonnes, tandis que e centre de gravité pour le nuage $N(J)$ corresponds aux fréquences marginales lignes.

Exemple :

- **Le centre de gravité du nuage** $N(I)$

$$g_{N(I)} = \frac{\sum_{i=1}^{n} p_i.x_i}{\sum_{i=1}^{n} p_i} = \frac{\sum_{i=1}^{n} f_{i.} * F_J^i}{\sum_{i=1}^{n} f_{i.}} = \frac{4}{10}\begin{pmatrix}2/4\\2/4\end{pmatrix} + \frac{3}{10}\begin{pmatrix}1/3\\2/3\end{pmatrix} + \frac{3}{10}\begin{pmatrix}2/3\\1/3\end{pmatrix} = \begin{pmatrix}5/10\\5/10\end{pmatrix} = f_{.j}$$

- **Le centre de gravité du nuage** $N(J)$

$$g_{N(J)} = \frac{\sum_{j=1}^{p} f_{.j} * F_I^j}{\sum_{j=1}^{p} f_{.j}} = \frac{5}{10}\begin{pmatrix}2/5\\1/5\\2/5\end{pmatrix} + \frac{5}{10}\begin{pmatrix}2/5\\2/5\\1/5\end{pmatrix} = \begin{pmatrix}4/10\\3/10\\3/10\end{pmatrix} = f_{i.}$$

1.4.3 Calcul d'inertie :

L'inertie représente la variance et qui est calculée par : $I = \sum_{i=1}^{n} P_i d^2(x_i, g)$

- L'inertie pour le nuage $N(I)$ est calculée par : $I_{N(I)} = \sum_{i=1}^{n} f_{i.} d^2\left(F_J^i, f_{.j}\right)$ avec

$$d^2\left(F_J^i, f_{.j}\right) = \left\|F_J^i - f_{.j}\right\|^2_{D_{\frac{1}{f.j}}} = \sum_{j=1}^{p}\left(F_J^i - f_{.j}\right)^2 / f_{.j}$$

Donc : $I_{N(I)} = \sum_{i=1}^{n} f_{i.} \sum_{j=1}^{p}\left(F_J^i - f_{.j}\right)^2 / f_{.j}$

- L'inertie pour le nuage $N(J)$ est déterminée par : $I_{N(J)} = \sum_{j=1}^{p} f_{.j} d^2\left(F_I^j, f_{i.}\right)$ avec

$$d^2\left(F_I^j, f_{i.}\right) = \left\|F_I^j - f_{i.}\right\|^2_{D_{\frac{1}{fi.}}} = \sum_{i=1}^{n}\left(F_I^j - f_{i.}\right)^2 / f_{i.}$$

Donc : $I_{N(J)} = \sum_{j=1}^{p} f_{.j} \sum_{i=1}^{n}\left(F_I^j - f_{i.}\right)^2 / f_{i.}$

Exemple :

- **L'inertie du nuage** $N(I)$

$$N(I) = \left\{\left(F_J^i, f_{i.}\right), i = 1.....n\right\} = \left\{[(2/4;2/4),4/10] \;\; ;[(1/3;2/3),3/10] \;; [(2/3;1/3),3/10]\right\}$$

$$f_{.j} = \begin{pmatrix}5/10\\5/10\end{pmatrix}$$

$$I_{N(J)} = \sum_{j=1}^{p} f_{.j} d^2\left(F_I^j, f_{i.}\right)$$

$$d^2\left(F_I^j, f_{i.}\right) = \left\|F_I^j - f_{i.}\right\|^2_{D_{\frac{1}{fi.}}} = \sum_{i=1}^{n}\left(F_I^j - f_{i.}\right)^2 / f_{i.}$$

-

$$I_{N(I)} = \sum_{i=1}^{n} f_{i.} d^2\left(F_J^i, f_{.j}\right) \quad et \quad d^2\left(F_J^i, f_{.j}\right) = \left\| F_J^i - f_{.j} \right\|^2_{D_{\frac{1}{f_{.j}}}} = \sum_{j=1}^{p} \left(F_J^i - f_{.j}\right)^2 / f_{.j}$$

$$I_{N(I)} = f_{1.}.d^2\left(F_J^1, f_{.j}\right) + f_{2.}.d^2\left(F_J^2, f_{.j}\right) + f_{3.}.d^2\left(F_J^3, f_{.j}\right) = 4/10.\left[\frac{(2/4-5/10)^2}{5/10} + \frac{(2/4-5/10)^2}{5/10}\right]$$

$$+3/10.\left[\frac{(1/3-5/10)^2}{5/10} + \frac{(2/3-5/10)^2}{5/10}\right] + 3/10.\left[\frac{(2/3-5/10)^2}{5/10} + \frac{(1/3-5/10)^2}{5/10}\right] =$$

- **Calcul d'inertie du nuage $N(J)$:**

$$N(J) = \left\{\left(F_I^j, f_{.j}\right), j = 1 \ldots\ldots p\right\} = \left\{[(2/5;1/5;2/5),5/10] \quad ;[(2/5;2/5;1/5),5/10] \ ;\right\} \ f_{i.} = \begin{pmatrix} 4/10 \\ 3/10 \\ 3/10 \end{pmatrix}$$

$$I_{N(J)} = \sum_{j=1}^{p} f_{.j} d^2\left(F_I^j, f_{i.}\right) \ et \quad d^2\left(F_I^j, f_{i.}\right) = \left\| F_I^j - f_{i.} \right\|^2_{D_{\frac{1}{f_{i.}}}} = \sum_{i=1}^{n} \left(F_I^j - f_{i.}\right)^2 / f_{i.}$$

$$I_{N(J)} = f_{.1}.d^2\left(F_I^1, f_{i.}\right) + f_{.2}.d^2\left(F_I^2, f_{i.}\right) = 5/10.\left[\frac{(2/5-4/10)^2}{4/10} + \frac{(1/5-3/10)^2}{3/10} + \frac{(2/5-3/10)^2}{3/10}\right]$$

$$+5/10.\left[\frac{(2/5-4/10)^2}{4/10} + \frac{(2/5-3/10)^2}{3/10} + \frac{(1/5-3/10)^2}{3/10}\right] =$$

Exercice 1 :

Soit un ensemble de 12 personnes à qui nous avons posé ces deux questions :

Q1 : Lisez- vous le journal « le quotidien d'Oran» ?

Q2 : Quelle est votre fréquence de lecture ?

Leurs réponses étaient comme suivantes :

P1=(Oui, Souvent) ; P2=(Non, Jamais) ; P3=(Oui, Rarement) ; P4=(Oui, Toujours) ;

P5=(Non, Jamais) ; P6=(Oui, Souvent) ; P7=(Oui, Souvent) ; P8=(Oui, Toujours) ;

P9=(Non, Jamais) ; P10=(Oui, Souvent) ; P11=(Oui, Rarement) ; P12=(Oui, Souvent) ;

En supposant que les fréquences de lecture obéissent à la relation d'ordre suivante : Jamais<rarement<Souvent<Toujours

1. Donner le tableau de codage correspondant.
2. Expliquer brièvement comment peut-on réduire ce tableau ?
3. Réduire le tableau de codage.
4. Une étude parallèle s'intéresse à une analyse factorielle des correspondances
   - Donner le tableau de contingence correspondant.
   - Etablir les nuages de points à étudier et montrer la relation entre ces deux nuages.

1. Le tableau de codage correspondant.

Réponse1 (Oui , Non) : Oui(1, 0) ; Non (0, 1) ;
Réponse 2(Jamais, Rarement, Souvent, Toujours) : Jamais<rarement<Souvent<Toujours

Jamais (1, 0, 0,0) ; Rarement (1, 1, 0,0) ; Souvent(1, 1, 1,0) ; Toujours (1,1,1,1)

	Réponse1		**Réponse2**			
	V1	**V2**	**V3**	**V4**	**V5**	**V6**
P1	**1**	**0**	**1**	**1**	**1**	**0**
P2	**0**	**1**	**1**	**0**	**0**	**0**
P3	**1**	**0**	**1**	**1**	**0**	**0**
P4	**1**	**0**	**1**	**1**	**1**	**1**
P5	**0**	**1**	**1**	**0**	**0**	**0**
P6	**1**	**0**	**1**	**1**	**1**	**0**
P7	**1**	**0**	**1**	**1**	**1**	**0**
P8	**1**	**0**	**1**	**1**	**1**	**1**
P9	**0**	**1**	**1**	**0**	**0**	**0**
P10	**1**	**0**	**1**	**1**	**1**	**0**
P11	**1**	**0**	**1**	**1**	**0**	**0**
P12	**1**	**0**	**1**	**1**	**1**	**0**

2. Pour réduire le tableau de codage, il faut tout d'abord ignorer la relation d'ordre de la deuxième variable (Fréquence de lecture) ensuite, on calcule le tableau de Burt.

- Tableau de codage disjonctif complet

	Réponse1		**Réponse2**			
	V1	**V2**	**V3**	**V4**	**V5**	**V6**
P1	**1**	**0**	**0**	**0**	**1**	**0**
P2	**0**	**1**	**1**	**0**	**0**	**0**
P3	**1**	**0**	**0**	**1**	**0**	**0**
P4	**1**	**0**	**0**	**0**	**0**	**1**
P5	**0**	**1**	**1**	**0**	**0**	**0**
P6	**1**	**0**	**0**	**0**	**1**	**0**
P7	**1**	**0**	**0**	**0**	**1**	**0**
P8	**1**	**0**	**0**	**0**	**0**	**1**
P9	**0**	**1**	**1**	**0**	**0**	**0**
P10	**1**	**0**	**0**	**0**	**1**	**0**
P11	**1**	**0**	**0**	**1**	**0**	**0**
P12	**1**	**0**	**0**	**0**	**1**	**0**

- Tableau de Burt

B=X^t.X=

9	**0**	**0**	**2**	**5**	**2**
0	3	3	0	0	0
0	3	3	0	0	0
2	0	0	2	0	0
5	0	0	0	5	0
2	0	0	0	0	2

3. Analyse factorielle des correspondances

- Le tableau de contingence correspondant. $K = \sum_i \sum_j K_{ij} = 12$

R1 \ R2	Jamais	Rarement	Souvent	Toujours
Oui	0	2	5	2
Non	3	0	0	0

- Les nuages de points :

Tableau de fréquence : $f_{ij} = \frac{K_{ij}}{K}$

f_{ij}	f_{i1}	f_{i2}	f_{i3}	f_{i4}	$f_{i.}$
f_{1j}	**0**	**2/12**	**5 /12**	**2/12**	**9/12**
f_{2j}	**3/12**	**0**	**0**	**0**	**3/12**
$f_{.j}$	**3/12**	**2/12**	**5/12**	**2/12**	

Tableau des profils lignes : $F_J^i \ : \ F_J^i = \frac{f_{ij}}{f_{i.}}$

F_J^i	F_1^i	F_2^i	F_3^i	F_4^i	$f_{i.}$
F_J^1	**0**	**2/9**	**5/9**	**2/9**	**9/12**
F_J^2	**1**	**0**	**0**	**0**	**3/12**

$$N(I) = \{(F_J^i, f_{i.}), i = 1.....n\} = \{[(0;2/9;5/9;2/9), 9/12]; [(1;0;0;0), 3/12]\}$$

Tableau des profils colonnes : $F_I^j \ : \ F_I^j = \frac{f_{ij}}{f_{.j}}$

F_I^j	F_I^1	F_I^2	F_I^3	F_I^4
F_1^j	**0**	**1**	**1**	**1**
F_2^j	**1**	**0**	**0**	**0**
$f_{.j}$	**3/12**	**2/12**	**5/12**	**2/12**

$$N(J) = \{(F_I^j, f_{.j}), j = 1..... p\} = \{[(0;1), 3/12]; [(1;0), 2/12]; [(1;0), 5/12]; [(1;0), 2/12]\}$$

- **La relation entre le nuage $N(I)$ et $N(J)$**
- **Calcul d'inertie du nuage $N(I)$:**

$$N(I) = \{(F_J^i, f_{i.}), i = 1.....n\} = \{[(0;2/9;5/9;2/9), 9/12]; [(1;0;0;0), 3/12]\}$$

$$f_{.j} = \begin{pmatrix} 3/12 \\ 2/12 \\ 5/12 \\ 2/12 \end{pmatrix}$$

$$I_{N(I)} = \sum_{i=1}^{n} f_{i.} d^2\left(F_J^i, f_{.j}\right) \quad et \quad d^2\left(F_J^i, f_{.j}\right) = \left\| F_J^i - f_{.j} \right\|^2_{D_{\frac{1}{f.j}}} = \sum_{j=1}^{p} \left(F_J^i - f_{.j}\right)^2 / f_{.j}$$

$$I_{N(I)} = f_{1.} . d^2\left(F_J^1, f_{.j}\right) + f_{2.} . d^2\left(F_J^2, f_{.j}\right) = 9/12 . \left[\frac{(0-3/12)^2}{3/12} + \frac{(2/9-2/12)^2}{2/12} + \frac{(5/9-5/12)^2}{5/12} + \frac{(2/9-2/12)^2}{2/12} \right]$$

$$+ 3/12 .. \left[\frac{(1-3/12)^2}{3/12} + \frac{(0-2/12)^2}{2/12} + \frac{(0-5/12)^2}{5/12} + \frac{(0-2/12)^2}{2/12} \right] =$$

- **Calcul d'inertie du nuage $N(J)$:**

$$N(J) = \left\{ \left(F_I^j, f_{.j}\right), j = 1 ..... p \right\} = \left\{ [(0;1), 3/12]; [(1;0), 2/12]; [(1;0), 5/12]; [(1;0), 2/12] \right\} \; f_{i.} = \begin{pmatrix} 9/12 \\ 3/12 \end{pmatrix}$$

$$I_{N(J)} = \sum_{j=1}^{p} f_{.j} d^2\left(F_I^j, f_{i.}\right)$$ et

$$d^2\left(F_I^j, f_{i.}\right) = \left\| F_I^j - f_{i.} \right\|^2_{D_{\frac{1}{fi.}}} = \sum_{i=1}^{n} \left(F_I^j - f_{i.}\right)^2 / f_{i.}$$

$$I_{N(J)} = f_{.1} . d^2\left(F_I^1, f_{i.}\right) + f_{.2} . d^2\left(F_I^2, f_{i.}\right) + f_{.3} . d^2\left(F_I^3, f_{i.}\right) + f_{.4} . d^2\left(F_I^4, f_{i.}\right) =$$

$$3/12 . \left[\frac{(0-9/12)^2}{9/12} + \frac{(1-3/12)^2}{3/12} \right] + 2/12 . \left[\frac{(1-9/12)^2}{9/12} + \frac{(0-3/12)^2}{3/12} \right] +$$

$$5/12 . \left[\frac{(1-9/12)^2}{9/12} + \frac{(0-3/12)^2}{3/12} \right] + 2/12 . \left[\frac{(1-9/12)^2}{9/12} + \frac{(0-3/12)^2}{3/12} \right] =$$

-

La relation entre les deux nuage réside dans l'égalité d'inertie

Exercice 2 :

Soit un ensemble de sept individus ayant répondu à deux questions comme suivant :

I1=(Oui, Oui) ; I2=(Oui, Non) ; I3=(Non, OUI) ; I4=(Non, Non) ; I5=(Oui, Oui) ; I6=(Oui, Non) ; I7=(Non, Oui)

1. Déterminer le tableau de codage disjonctif complet.
2. En utilisant la mesure $d(I,J)=\frac{b+c}{a+b+c+d}$, donner le tableau de distance correspondant.
3. En déduire une classification en 4 classes, 3 classes.
4. Réduire le tableau de codage.

- **Solution :**

1. **Tableau de codage disjonctif :**

Réponse 1(Oui, Non) : N(Oui)=(1, 0) ; N(Non)=(0, 1)

Réponse 2(Oui, Non) : N(Oui)=(1, 0) ; N(Non)=(0, 1)

	Réponse 1		Réponse 2	
	V1	**V2**	**V3**	**V4**
I1	**1**	**0**	**1**	**0**
I2	**1**	**0**	**0**	**1**
I3	**0**	**1**	**1**	**0**
I4	**0**	**1**	**0**	**1**
I5	**1**	**0**	**1**	**0**
I6	**1**	**0**	**0**	**1**
I7	**0**	**1**	**1**	**0**

- **a** représente le nombre de fois où $x_{ij}=1$ et $x_{i'j'}=1$.
- **b** représente le nombre de fois où $x_{ij}=0$ et $x_{i'j'}=1$.
- **c** représente le nombre de fois où $x_{ij}=1$ et $x_{i'j'}=0$.
- **d** représente le nombre de fois où $x_{ij}=0$ et $x_{i'j'}=0$

Calcul de la distance entre I1 et I2 :

$I1:1010$

$I2:1001$; La mesure de dissemblance : $d(I,J)=\frac{b+c}{a+b+c+d}$

$\Rightarrow a=1;b=1;c=1;d=1. \Rightarrow d(I1,I2)=2/4=1/2$

2. Tableau de distance

D	I1	I2	I3	I4	I5	I6	I7
I1	**0**	**1/2**	**1/2**	**1**	**0**	**1/2**	**1/2**
I2		**0**	**1**	**1/2**	**1/2**	**0**	**1**
I3			**0**	**1/2**	**1/2**	**1**	**0**
I4				**0**	**1**	**1/2**	**1/2**
I5					**0**	**1/2**	**1/2**
I6						**0**	**1/2**
I7							**0**

3. Classification en 4 classes :

$C_1 = \{I_1, I_5\}$; $C_2 = \{I_2, I_6\}$; $C_3 = \{I_3, I_7\}$ **et** $C_4 = \{I_4\}$

- **Classification en 3 classes :**

  $C_1 = \{I_1, I_5\}$; $C_2 = \{I_2, I_6\}$; $C_3 = \{I_3, I_4, I_7\}$ **ou bien** $C_1 = \{I_1, I_5\}$; $C_2 = \{I_2, I_6\}$; $C_3 = \{I_3, I_7\}$

4. Réduction du tableau de codage :

- **Tableau de Burt :**

B=X^t.X=

	V1	V2	V3	V4
V1	**4**	**0**	**2**	**2**
V2	**0**	**3**	**1**	**2**
V3	**2**	**1**	**3**	**0**
V4	**2**	**2**	**0**	**4**

- **Réduction par regroupement :**

	V1	V2	V3	V4
C1	**2**	**0**	**2**	**0**
C2	**2**	**0**	**0**	**2**
C3	**2**	**1**	**3**	**0**
C4	**0**	**2**	**2**	**0**

1.5 Conclusion

Dans ce chapitre, nous avons introduit les notions de base en statistique et en analyse des données en décrivant les différents tableaux de données ainsi les mesures de similarité. Après, nous avons défini les ingrédients nécessaires d'une analyse factorielle des correspondances en introduisant les tableaux des profils lignes et colonnes, le centre de gravité de chaque nuage et les métriques utilisées. L'analyse factorielle sera détaillée dans le chapitre suivant.

ANALYSE FACTORIELLE

2.1 Introduction

L'analyse en composantes principales (ACP) est une des premières analyses factorielles et qui attire l'attention des scientifiques jusqu'à présent sachant que l'ACP a été conçue par Karl Pearson en 1901 [4].

Plusieurs applications font appel à l'intégration de l'ACP comme une méthode de sélection des attributs. Souvent, l'ACP est appliqué comme une méthode de prétraitement dans le domaine d'intelligence artificielle [5]. Plusieurs variantes de l'ACP figurent dans la littérature comme l'ACP non-normée (données homogènes) qui utilise un nuage de point centré, ou bien l'ACP centrée réduite (données hétérogènes) qui utilise un nuage de point centré et réduit. L'algorithme d'ACP permet de traiter un tableau de type individu/variable <n, p>.

Nous trouvons aussi d'autres variantes telles que l'analyse en composantes curviligne pour remédier la linéarité des projections [6], ou encore l'analyse en composantes indépendantes pour la séparation de source [7]. L'analyse factorielle des correspondances (AFC) permet une représentation simultanée des individus et des variables et qui peut être traitée comme double ACP sur un tableau croisé. [8].

2.2 Les données

Les données pour l'ACP sont généralement présentées sous la forme d'un tableau où les lignes indiquent les individus, tandis que les colonnes représentent les variables. Elle traite des données quantitatives. Nous notons x_{ij}, la valeur de la variable j de l'individu i. N désigne le nombre total des individus et P indique le nombre total des variables.

2.3 Les objectifs

Deux objectifs sont envisageables par l'ACP :

- Le premier cherche à représenter graphiquement les individus en calculant les composantes principales, qui représentent la projection des individus dans l'espace réduit.
- Le deuxième cherche à représenter les variables en calculant les coefficients de corrélation entre les variables et les composantes principales.

Elle peut être représentée selon deux points de vue :

- La recherche d'un ensemble réduit de variables non corrélées (combinaison linéaire des variables initiales).
- La recherche de sous espace représentant au mieux le nuage initial.

2.4 Les types d'inertie

➢ **L'inertie d'un point autour de son centre de gravité g :**

L'inertie est une notion fondamentale en ACP, puisqu'elle est une mesure de dispersion du nuage de points autour de son centre de gravité.

$$I_{x_i} = P_i.d^2(x_i, g)$$

➢ **L'inertie totale du nuage de points autour de son centre de gravité g :**

$$I = \sum_{i=1}^{n} P_i.d^2(x_i, g)$$

➢ **L'inertie d'un point par rapport à un axe**

$$I_{X_i/\Delta} = P_i d^2_{i/\Delta}$$

➢ **L'inertie du nuage de points par rapport à un axe**

- **Principe de l'ACP**

Pour visualiser le nuage des individus, il est nécessaire de réduire la dimension de l'espace qui le porte. L'ACP réduit cette dimension par une projection orthogonale sur un sous espace. Donc l'inertie I du nuage autour du sous espace linéaire Δ est donnée par :

$$I_{/\Delta} = \sum_{i=1}^{n} P_i d^2_{i/\Delta}$$

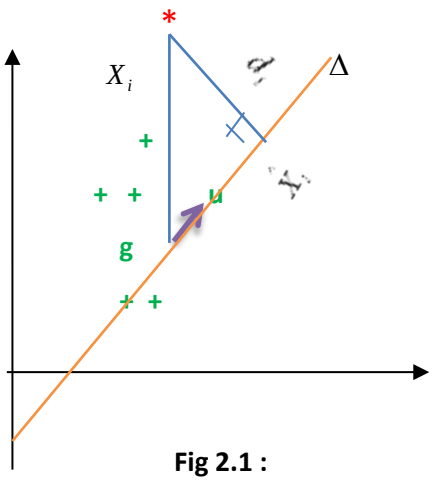


Fig 2.1 :

D'après le théorème de Pythagore : $\|X_i\|^2_M = \left\|\hat{X}_i\right\|_M + \|d_{i/\Delta}\|^2 \Rightarrow \|d_{i/\Delta}\|^2 = \|X_i\|^2_M - \left\|\hat{X}_i\right\|^2_M$

$$I_{/\Delta}=\sum_{i=1}^{n}P_i d^2_{i/\Delta} et \left\|d_{i/\Delta}\right\|^2=\left\|X_i\right\|^2_M-\left\|\hat{X}_i\right\|^2_M \Rightarrow I_{/\Delta}=\sum_{i=1}^{n}P_i.\left(\left\|X_i\right\|^2_M-\left\|\hat{X}_i\right\|^2_M\right)$$

$$\Rightarrow I_{/\Delta}=\sum_{i=1}^{n}P_i.\left\|X_i\right\|^2_M-\sum_{i=1}^{n}P_i.\left\|\hat{X}_i\right\|^2_M \rightarrow M: Métrique$$

$$\Rightarrow I_{/\Delta}=\sum_{i=1}^{n}P_i.\left\|X_i\right\|^2_M-\sum_{i=1}^{n}P_i.\left[< X_i,U >_M\right]^2$$

$$\Rightarrow I_{/\Delta}=\sum_{i=1}^{n}P_i.\left\|X_i\right\|^2_M-\sum_{i=1}^{n}P_i.< X_i,U >^t_M .< X_i,U >_M$$

$$\Rightarrow I_{/\Delta}=\sum_{i=1}^{n}P_i.\left\|X_i\right\|^2_M-\sum_{i=1}^{n}P_i.\left(X_i.M.U\right)^t.\left(X_i.M.U\right)$$

$$\Rightarrow I_{/\Delta}=\sum_{i=1}^{n}P_i.\left\|X_i\right\|^2_M-\sum_{i=1}^{n}P_i.U^t.M^t.X_i^t.X_i.M.U \rightarrow M^t=M$$

$$\Rightarrow I_{/\Delta}=\sum_{i=1}^{n}P_i.\left\|X_i\right\|^2_M-\sum_{i=1}^{n}P_i.U^t.M.X_i^t.X_i.M.U$$

$$\Rightarrow I_{/\Delta}=\sum_{i=1}^{n}P_i.\left\|X_i\right\|^2_M-U^t.M\sum_{i=1}^{n}P_i.X_i^t.X_i.M.U \rightarrow V=\sum_{i=1}^{n}P_i.X_i^t.X_i$$

$$\Rightarrow I_{/\Delta}=\sum_{i=1}^{n}P_i.\left\|X_i\right\|^2_M-U^t.M.V.M.U$$

$$Min \quad I_{/\Delta} \Rightarrow Max\left[U^t.M.V.M.U\right]$$

$V.M.U_k=\lambda_k.U_k$ Tels que U_k : Vecteurs propres de la matrice $V.M$; λ_k : Valeurs propres de la matrice $V.M$

La métrique $M=\begin{cases} I & \text{: Variables homogènes} \\ D_{\frac{1}{\sigma_j^2}} & \text{: Variables hétérogènes} \end{cases}$

Remarque : Les vecteurs propres constituent une base orthonormée c-à-d :

$$< U_i,U_j >_M = U_i^t M U_j = 0 \ \forall i \neq j \ \& \ \left\|U_i\right\|^2 =< U_i,U_i >_M = U_i^t M U_i = 1$$

- **Les composantes principales** : $C_k^i =< X_i,U_k >_M = X_i^t M U_k$ et $C_k = XMU_k$.

2.5 Les propriétés de la composante principale :

1) $Moy(C_k)=0 \Rightarrow \dfrac{\sum_{i=1}^{N} P_i.C_k^i}{\sum_{i=1}^{n} P_i}=0 \Rightarrow \dfrac{1}{N}\sum_{i=1}^{N} C_k^i=0 \Rightarrow \sum_{i=1}^{N} C_k^i=0$

2) $Var(C_k)=\lambda_k \Rightarrow Var(C_k)=\sum_{i=1}^{N} P_i.(C_k^i)^2=\dfrac{1}{N}C_k^t * C_k$ 3) 3

3) $Cor(C_i,C_j)=\dfrac{Cov(C_i,C_j)}{\sigma_{C_i}.\sigma_{C_j}}=0 \Rightarrow Cov(C_i,C_j)=0 \Rightarrow \dfrac{1}{N}C_i^t * C_j=0$

2.6 Algorithme d'ACP:

1. Centrer le tableau ($X_{<n,p>}$) : $X^{'}=X-g$ et $g^j=\dfrac{\sum_{i=1}^{N} p_i.x_{ij}}{\sum_{i=1}^{N} p_i}; j=1...p; i=1...N \;\&\; p_i=\dfrac{1}{N}$
2. Calculer la matrice variance –covariance : $V=\dfrac{1}{N}X^t.X$
3. Déterminer la métrique $M=\begin{cases} I \\ D_{\frac{1}{\sigma_j^2}} \end{cases}$ Données homogènes/ hétérogènes.
4. Recherche des axes principaux U_k de la matrice (VM)
   - Calculer les valeurs propres : $dét(VM-\lambda I)=0$
   - Trier les valeurs propres par ordre décroissant : $\lambda_1 > \lambda_2 > ..... > \lambda_p$.
5. Calculer la qualité de représentation : $Q_j=\dfrac{\sum_{i=1}^{J}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \geq 80\%$.
6. Calculer les vecteurs propres U_k de la matrice (VM) en utilisant la formule : $VMU_k=\lambda_k U_k$.
7. Calculer les composantes principales : $C_k^i =< X_i, U_k >_M = X_i^t M U_k$ et $C_k = XMU_k$.
8. Représenter graphiquement les individus dans l'espace réduit en utilisant les composantes principales.
9. Les contributions aux inerties :
   - Part d'inertie de X_i prise en compte par l'axe U_k : $\cos^2(\theta_{ik})=\dfrac{(C_k^i)^2}{\|X_i\|_M^2}$.
   - Contribution relative de l'individu X_i à l'inertie expliquée de l'axe U_k :

$$\rho_{ik}=\frac{P_i.(C_k^i)^2}{\sum_{i=1}^{n} P_i.(C_k^i)^2}=\frac{P_i.(C_k^i)^2}{Var(C_k)}=\frac{P_i.(C_k^i)^2}{\lambda_k}$$

10. Représentation des variables à l'aide du coefficient de corrélation :

$$Cor(X^j,C_K)=\frac{Cov(X^j,C_k)}{\sigma_{X^j}.\sigma_{C_k}}=\frac{\sum_{i=1}^{N} P_i.X_i^j.C_k^i}{\sigma_{X^j}.\sqrt{\lambda_k}}=\frac{\frac{1}{N}(X^j)^t.C_k}{\sigma_{X^j}.\sqrt{\lambda_k}}$$

2.7 Algorithme d'AFC :

1. Tableau [Variable/Variable] $\rightarrow$ deux tableaux de profils [Individus/ Variables]
2. Application de deux ACP $\rightarrow N(I)$
   $\downarrow N(J)$
3. Les valeurs propres significatives du nuage $\lambda_k \in]0,1[$
   $N(I) \mapsto \lambda_k, U_k$
   $N(J) \mapsto \lambda_k, V_k$
4. Calculer les composantes principales : $N(I) \mapsto C_k = F_J^I . D_{1/f.j} U_k \ \& N(J) \mapsto d_k = F_I^J . D_{1/fi.} V_k$
5. Les formules de transitions : $C_k^i = \frac{1}{\sqrt{\lambda_K}} \sum_{j=1}^{P} F_J^I . d_k^j \ \& d_k^j = \frac{1}{\sqrt{\lambda_K}} \sum_{i=1}^{n} F_I^J . C_k^i$

Exemple sur l'analyse en composantes principales dans le cas des données homogènes :

Soit un ensemble de six individus caractérisés par trois notes chacun.

	N1	N2	N3
I1	8	1	0
I2	4	6	5
I3	6	8	7
I4	10	4	7
I5	8	2	5
I6	0	3	6

- Appliquer l'analyse en composante principale ($\lambda_1 = 12$).

1. Le tableau centré $X_{\prec n,p \succ}$:

   - **Calcul du centre de gravité :** $g = \frac{\sum_{i=1}^{n} p_i . x_i}{\sum_{i=1}^{n} p_i}$

$P_i = \frac{1}{N}$ où N représente le nombre d'individus. Cela implique que $P_i = \frac{1}{6}$; $i = 1,\ldots\ldots,n;1\ldots6$; $j = 1,\ldots\ldots p;1\ldots3$

$$g=\begin{pmatrix}\frac{1}{6}(8+4+6+10+8+0)\\ \frac{1}{6}(1+6+8+4+2+3)\\ \frac{1}{6}(0+5+7+7+5+6)\end{pmatrix}=\begin{pmatrix}6\\4\\5\end{pmatrix}$$

	N1	N2	N3
I1	8	1	0
I2	4	6	5
I3	6	8	7
I4	10	4	7
I5	8	2	5
I6	0	3	6
g	**6**	**4**	**5**

Tableau centré →

$$X=\begin{pmatrix}8-6 & 1-4 & 0-5\\ 4-6 & 6-4 & 5-5\\ 6-6 & 8-4 & 7-5\\ 10-6 & 4-4 & 7-5\\ 8-6 & 2-4 & 5-5\\ 0-6 & 3-4 & 6-5\end{pmatrix}=\begin{pmatrix}+2 & -3 & -5\\ -2 & +2 & 0\\ 0 & +4 & +2\\ +4 & 0 & +2\\ +2 & -2 & 0\\ -6 & -1 & +1\end{pmatrix}$$

2. Calcul de la matrice variance –covariance :

$$V=\frac{1}{N}X^{t}.X\Rightarrow V=\frac{1}{6}\begin{pmatrix}64 & -8 & -8\\ -8 & 34 & 22\\ -8 & 22 & 34\end{pmatrix}.$$

3. Détermination de la métrique :

Les données sont homogènes car elles possèdent le même type de mesure $\Rightarrow M=Id=\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}$

4. Recherche des axes principaux U_k de la matrice (VM) :

- Calcul des valeurs propres λ_k de la matrice (VM) en utilisant la formule suivante :

$$dét(VM-\lambda I)=0$$

- $$dét(VM-\lambda I)=\frac{1}{6}\begin{vmatrix}64-6\lambda & -8 & -8\\ -8 & 34-6\lambda & 22\\ -8 & 22 & 34-6\lambda\end{vmatrix}=0\Rightarrow \lambda^3-22\lambda^2+136\lambda-192=0$$

$$\begin{array}{l|l}\lambda^3-22\lambda^2+136\lambda-192 & \lambda-12\\ \underline{-\lambda^3+12\lambda^2} & \hline \lambda^2-10\lambda+16\\ -10\lambda^2+136\lambda-192 & \\ \underline{+10\lambda^2-120\lambda} & \\ 16\lambda-192 & \\ \underline{-16\lambda+192} & \\ 0 & \end{array}$$

$$dét(VM-\lambda I)=0\Rightarrow \lambda^3-22\lambda^2+136\lambda-192=0\Rightarrow(\lambda-12).(\lambda^2-10\lambda+16)=0$$

$$\Rightarrow\begin{cases}\lambda-12=0\\ \lambda^2-10\lambda+16=0\end{cases}\Rightarrow\begin{cases}\lambda_1=12\\ \Delta=b^2-4ac=100-4.1.16=36\rightarrow\sqrt{\Delta}=6\Rightarrow\begin{cases}\lambda_2=\dfrac{-b+\sqrt{\Delta}}{2a}=\dfrac{10+6}{2}=8\\ \lambda_3=\dfrac{-b-\sqrt{\Delta}}{2a}=\dfrac{10-6}{2}=2\end{cases}\end{cases}$$

$$\lambda_1=12\succ\lambda_2=8\succ\lambda_3=2$$

5. Calcul de la qualité de représentation

$$Q_j=\frac{\sum_{i=1}^{j}\lambda_i}{\sum_{i=1}^{p}\lambda_i}\geq 80\%\Rightarrow Q_1=\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3}=\frac{12}{12+8+2}=0.54=54\%\prec 80\%$$

$$Q_2=\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3}=\frac{12+8}{12+8+2}=0.9=90\%\geq 80\%$$ → **Il y'a deux axes principaux U_1,U_2 relatifs aux valeurs propres λ_1,λ_2**

6. Calcul des vecteurs propres U_1,U_2 de la matrice (VM) : $VMU_k=\lambda_kU_k$

$$VMU_1=\lambda_1U_1\Rightarrow\frac{1}{6}\begin{pmatrix}64&-8&-8\\-8&34&22\\-8&22&34\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=12\begin{pmatrix}x\\y\\z\end{pmatrix}\Rightarrow\begin{cases}64x-8y-8z=72x\\-8x+34y+22z=72y\\-8x+22y+34z=72z\end{cases}\Rightarrow\begin{cases}-8x-8y-8z=0\rightarrow(1)\\-8x-38y+22z=0\rightarrow(2)\\-8x+22y-38z=0\rightarrow(3)\end{cases}$$

$$\begin{matrix}Eq(1)-Eq(2):30y-30z=0\Rightarrow y=z\rightarrow(4)\\ Eq(4)dansEq(3):-8x+22y-38y=0\Rightarrow x=-2y\end{matrix}\Rightarrow U_1^*=\begin{pmatrix}-2\\1\\1\end{pmatrix}$$

$$\left\|U_1^*\right\|_M^2=U_1^{*t}.M.U_1^*=\begin{pmatrix}-2&1&1\end{pmatrix}.Id.\begin{pmatrix}-2\\1\\1\end{pmatrix}=6\Rightarrow\left\|U_1^*\right\|_M=\sqrt{6}\Rightarrow U_1=\frac{U_1^*}{\left\|U_1^*\right\|_M}=\frac{1}{\sqrt{6}}\begin{pmatrix}-2\\1\\1\end{pmatrix}$$

$$VMU_2=\lambda_2U_2\Rightarrow\frac{1}{6}\begin{pmatrix}64&-8&-8\\-8&34&22\\-8&22&34\end{pmatrix}\begin{pmatrix}x'\\y'\\z'\end{pmatrix}=8\begin{pmatrix}x'\\y'\\z'\end{pmatrix}\Rightarrow\begin{cases}64x'-8y'-8z'=48x'\\-8x'+34y'+22z'=48y'\\-8x'+22y'+34z'=48z'\end{cases}\Rightarrow\begin{cases}16x'-8y'-8z'=0\rightarrow(a)\\-8x'-14y'-8z'=0\rightarrow(b)\\-8x'+22y'-14z'=0\rightarrow(c)\end{cases}$$

$$\begin{matrix}Eq(b)-Eq(c):-36y'+36z'=0\Rightarrow y=z\rightarrow(d)\\ Eq(d)dansEq(a):16x'-8y'-8y'=0\Rightarrow x=y\end{matrix}\Rightarrow U_2^*=\begin{pmatrix}1\\1\\1\end{pmatrix}$$

$$\left\|U_2^*\right\|_M^2=U_2^{*t}.M.U_2^*=\begin{pmatrix}1&1&1\end{pmatrix}.Id.\begin{pmatrix}1\\1\\1\end{pmatrix}=3\Rightarrow\left\|U_2^*\right\|_M=\sqrt{3}\Rightarrow U_2=\frac{U_2^*}{\left\|U_2^*\right\|_M}=\frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}$$

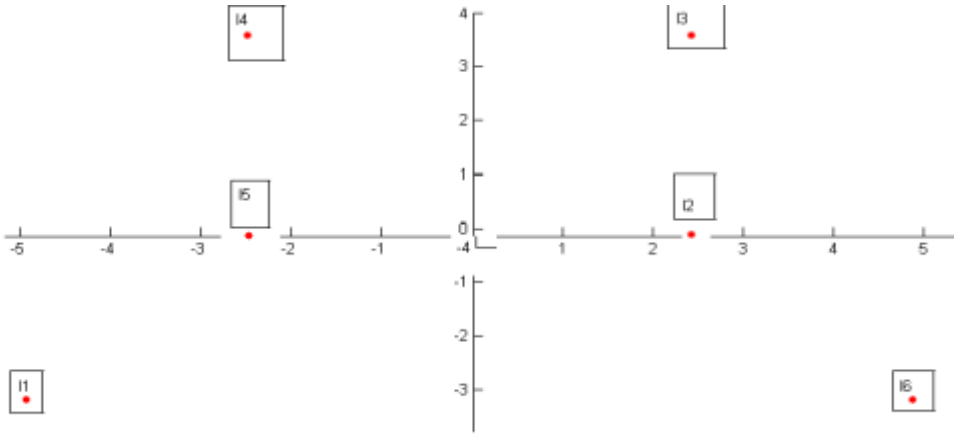
7. Calcul des composantes principales $C_k = X.M.U_k$

$$C_1 = X.M.U_1 = X.Id.U_1 = X.U_1 = \begin{pmatrix} +2 & -3 & -5 \\ -2 & +2 & 0 \\ 0 & +4 & +2 \\ +4 & 0 & +2 \\ +2 & -2 & 0 \\ -6 & -1 & +1 \end{pmatrix} . \frac{1}{\sqrt{6}} \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{6}} \begin{pmatrix} -12 \\ +6 \\ +6 \\ -6 \\ -6 \\ +12 \end{pmatrix} = \sqrt{6} \begin{pmatrix} -2 \\ +1 \\ +1 \\ -1 \\ -1 \\ +2 \end{pmatrix}$$

$$C_2 = X.M.U_2 = X.Id.U_2 = X.U_2 = \begin{pmatrix} +2 & -3 & -5 \\ -2 & +2 & 0 \\ 0 & +4 & +2 \\ +4 & 0 & +2 \\ +2 & -2 & 0 \\ -6 & -1 & +1 \end{pmatrix} . \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} -6 \\ 0 \\ +6 \\ +6 \\ 0 \\ -6 \end{pmatrix} = \sqrt{3} \begin{pmatrix} -2 \\ 0 \\ +2 \\ +2 \\ 0 \\ -2 \end{pmatrix}$$

8. Représentation graphique des individus :

$$I_1 = \left(-2\sqrt{6}, -2\sqrt{3}\right); I_2 = \left(\sqrt{6}, 0\right); I_3 = \left(\sqrt{6}, 2\sqrt{3}\right); I_4 = \left(-\sqrt{6}, 2\sqrt{3}\right); I_5 = \left(-\sqrt{6}, 0\right); I_6 = \left(2\sqrt{6}, -2\sqrt{3}\right)$$



9. Calcul des contributions :

- **Contribution relative de l'individu X_i à l'inertie expliquée de l'axe U_k :**

$$\rho_{ik} = \frac{P_i.\left(C_k^i\right)^2}{\sum_{i=1}^{n} P_i.\left(C_k^i\right)^2} = \frac{P_i.\left(C_k^i\right)^2}{Var\left(C_k\right)} = \frac{P_i.\left(C_k^i\right)^2}{\lambda_k}$$

- **Par rapport à l'axe U_1 : i=1…6**

$$\rho_{11} = \frac{P_1.\left(C_1^1\right)^2}{\lambda_1} = \frac{\frac{1}{6}.\left(-2\sqrt{6}\right)^2}{12} = \frac{1}{3} = 0.33$$

$$\rho_{21} = 0.08; \rho_{31} = 0.08; \rho_{41} = 0.08; \rho_{51} = 0.08; \rho_{61} = 0.33.$$

- **Par rapport à l'axe U_2 : i=1…6**

$$\rho_{12} = \frac{P_1.\left(C_2^1\right)^2}{\lambda_2} = \frac{\frac{1}{6}.\left(-2\sqrt{3}\right)^2}{8} = \frac{1}{4} = 0.25;$$

$$\rho_{22} = 0; \rho_{32} = 0.25; \rho_{42} = 0.25; \rho_{52} = 0; \rho_{62} = 0.25.$$

Remarque : Si $\rho_{ik} \approx 1$ alors le i ème individu doit être retiré du tableau initial et dans ce cas, il faut refaire l'ACP.

- **Part d'inertie de X_i prise en compte par l'axe U_k :** $\cos^2(\theta_{ik}) = \frac{\left(C_k^i\right)^2}{\|X_i\|_M^2}$.

- **Par rapport à l'axe U_1 : i=1…6**

$$\cos^2(\theta_{11}) = \frac{\left(C_1^1\right)^2}{\|X_1\|_M^2} = \frac{\left(-2\sqrt{6}\right)^2}{\begin{pmatrix}2 & -3 & -5\end{pmatrix}.Id.\begin{pmatrix}2\\-3\\-5\end{pmatrix}} = \frac{24}{4+9+25} = \frac{24}{38} = 0.63$$

$$\cos^2(\theta_{21}) = \frac{\left(C_1^2\right)^2}{\|X_2\|_M^2} = \frac{\left(\sqrt{6}\right)^2}{4+4+0} = \frac{6}{8} = 0.75;\ \cos^2(\theta_{31}) = 0.3;\ \cos^2(\theta_{41}) = 0.3;\ \cos^2(\theta_{51}) = 0.75;$$

$$\cos^2(\theta_{61}) = 0.63;$$

- **Par rapport à l'axe U_2 : i=1…6**

$$\cos^2(\theta_{12}) = \frac{\left(C_2^1\right)^2}{\|X_1\|_M^2} = \frac{\left(-2\sqrt{3}\right)^2}{\begin{pmatrix}2 & -3 & -5\end{pmatrix}.Id.\begin{pmatrix}2\\-3\\-5\end{pmatrix}} = \frac{12}{4+9+25} = \frac{12}{38} = 0.32$$

$$\cos^2(\theta_{22}) = \frac{\left(C_2^2\right)^2}{\|X_2\|_M^2} = \frac{(0)^2}{4+4+0} = 0;\ \cos^2(\theta_{32}) = 0.6;\ \cos^2(\theta_{42}) = 0.6;\ \cos^2(\theta_{52}) = 0;$$

$$\cos^2(\theta_{62}) = 0.32;$$

10. Représentation des variables à l'aide du coefficient de corrélation

$$Cor\left(X^{j},C_{K}\right)=\frac{Cov\left(X^{j},C_{k}\right)}{\sigma_{X^{j}}.\sigma_{C_{k}}}=\frac{\sum_{i=1}^{n}P_{i}.X_{i}^{j}.C_{k}^{i}}{\sigma_{X^{j}}.\sqrt{\lambda_{k}}}=\frac{\frac{1}{N}\left(X^{j}\right)^{t}.C_{k}}{\sigma_{X^{j}}.\sqrt{\lambda_{k}}}$$

$$Cor\left(X^{1}:N_{1},C_{1}\right)=\frac{Cov\left(X^{1},C_{1}\right)}{\sigma_{X^{1}}.\sigma_{C_{1}}}=\frac{\sum_{i=1}^{n}P_{i}.X_{i}^{1}.C_{1}^{i}}{\sigma_{X^{1}}.\sqrt{\lambda_{1}}}=\frac{\frac{1}{N}\left(X^{1}\right)^{t}.C_{1}}{\sigma_{X^{1}}.\sqrt{\lambda_{1}}}=\frac{\frac{1}{6}\begin{pmatrix}2&-2&0&4&2&-6\end{pmatrix}*\sqrt{6}\begin{pmatrix}-2\\+1\\+1\\-1\\-1\\+2\end{pmatrix}}{\sqrt{64/6}*\sqrt{12}}=-\frac{\sqrt{3}}{2}$$

$$Cor\left(X^{1}:N_{1},C_{2}\right)=\frac{Cov\left(X^{1},C_{2}\right)}{\sigma_{X^{1}}.\sigma_{C_{2}}}=\frac{\sum_{i=1}^{n}P_{i}.X_{i}^{1}.C_{2}^{i}}{\sigma_{X^{1}}.\sqrt{\lambda_{2}}}=\frac{\frac{1}{N}\left(X^{1}\right)^{t}.C_{2}}{\sigma_{X^{1}}.\sqrt{\lambda_{2}}}=\frac{\frac{1}{6}\begin{pmatrix}2&-2&0&4&2&-6\end{pmatrix}*\sqrt{3}\begin{pmatrix}-2\\0\\+2\\+2\\0\\-2\end{pmatrix}}{\sqrt{64/6}*\sqrt{8}}=\frac{1}{2}$$

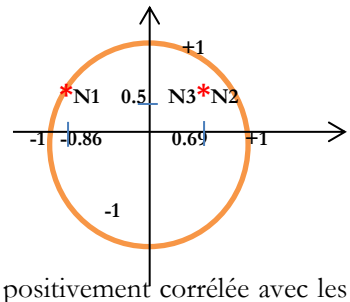
$$Cor\left(X^{2}:N_{2},C_{1}\right)=\frac{Cov\left(X^{2},C_{1}\right)}{\sigma_{X^{2}}.\sigma_{C_{1}}}=\frac{\sum_{i=1}^{n}P_{i}.X_{i}^{2}.C_{1}^{i}}{\sigma_{X^{2}}.\sqrt{\lambda_{1}}}=\frac{\frac{1}{N}\left(X^{2}\right)^{t}.C_{1}}{\sigma_{X^{2}}.\sqrt{\lambda_{1}}}=\frac{\frac{1}{6}\begin{pmatrix}-3&2&4&0&-2&-1\end{pmatrix}*\sqrt{6}\begin{pmatrix}-2\\+1\\+1\\-1\\-1\\+2\end{pmatrix}}{\sqrt{34/6}*\sqrt{12}}=\sqrt{6/17}=0.59$$

$$Cor\left(X^{2}:N_{2},C_{2}\right)=\frac{Cov\left(X^{2},C_{2}\right)}{\sigma_{X^{2}}.\sigma_{C_{2}}}=\frac{\sum_{i=1}^{n}P_{i}.X_{i}^{2}.C_{2}^{i}}{\sigma_{X^{2}}.\sqrt{\lambda_{2}}}=\frac{\frac{1}{N}\left(X^{2}\right)^{t}.C_{2}}{\sigma_{X^{2}}.\sqrt{\lambda_{2}}}=\frac{\frac{1}{6}\begin{pmatrix}-3&2&4&0&-2&-1\end{pmatrix}*\sqrt{3}\begin{pmatrix}-2\\0\\+2\\+2\\0\\-2\end{pmatrix}}{\sqrt{34/6}*\sqrt{8}}=\frac{4}{\sqrt{34}}=0.68$$

$$Cor\left(X^3 : N_3, C_1\right) = \frac{Cov\left(X^3, C_1\right)}{\sigma_{X^3}.\sigma_{C_1}} = \frac{\sum_{i=1}^{n} P_i.X_i^3.C_1^i}{\sigma_{X^3}.\sqrt{\lambda_1}} = \frac{\frac{1}{N}\left(X^3\right)^t.C_1}{\sigma_{X^3}.\sqrt{\lambda_1}} = \frac{\frac{1}{6}\left(-5 \quad 0 \quad 2 \quad 2 \quad 0 \quad 1\right) * \sqrt{6} \begin{pmatrix} -2 \\ +1 \\ +1 \\ -1 \\ -1 \\ +2 \end{pmatrix}}{\sqrt{34/6} * \sqrt{12}} = \frac{1}{2}$$

$$Cor\left(X^3 : N_3, C_2\right) = \frac{Cov\left(X^3, C_2\right)}{\sigma_{X^3}.\sigma_{C_2}} = \frac{\sum_{i=1}^{n} P_i.X_i^3.C_2^i}{\sigma_{X^3}.\sqrt{\lambda_2}} = \frac{\frac{1}{N}\left(X^3\right)^t.C_2}{\sigma_{X^3}.\sqrt{\lambda_2}} = \frac{\frac{1}{6}\left(-5 \quad 0 \quad 2 \quad 2 \quad 0 \quad 1\right) * \sqrt{3} \begin{pmatrix} -2 \\ 0 \\ +2 \\ +2 \\ 0 \\ -2 \end{pmatrix}}{\sqrt{34/6} * \sqrt{8}} = \frac{4}{\sqrt{34}} = 0.68$$

Fig. 2.2 : Cercle de corrélation.



- La deuxième composante C_2 est positivement corrélée avec les trois variables (N1, N2 et N3). C_2 nous informe sur le résultat général de l'étudiant ($C_2 \approx$ moyenne).
- C_1 est positivement corrélée avec N2, N3 et négativement corrélée avec N1. C1 nous informe sur la différence entre les résultats des deux derniers examens et le 1ier examen.

Exemple sur l'analyse en composantes principales dans le cas des données hétérogènes :

Soit un ensemble de six individus dont on a mesuré le poids et la taille.

	Poids	Taille
I1	20	15
I2	5	2
I3	12	21
I4	21	13
I5	2	7
I6	12	20

1. Appliquer une Analyse en composante principale.
2. Déduire toutes les partitions possibles, en définissant le sens physique des classes.

- Le tableau centré $X_{<n,p>}$:
  - Calcul du centre de gravité : $g = \dfrac{\sum_{i=1}^{n} p_i . x_i}{\sum_{i=1}^{n} p_i}$

$P_i = \dfrac{1}{N}$ où N représente le nombre d'individus. Cela implique que $P_i = \dfrac{1}{6}$; $i = 1,\ldots\ldots, n;1\ldots 6$; $j = 1,\ldots. p;1\ldots 2$

$$g = \begin{pmatrix} \frac{1}{6}(20+5+12+21+2+12) \\ \frac{1}{6}(15+2+21+13+7+20) \end{pmatrix} = \begin{pmatrix} 12 \\ 13 \end{pmatrix}$$

	Poids	Taille
I1	20	15
I2	5	2
I3	12	21
I4	21	13
I5	2	7
I6	12	20
g	**12**	**13**

Tableau centré $\longrightarrow$ $X = \begin{pmatrix} 20-12 & 15-13 \\ 5-12 & 2-13 \\ 12-12 & 21-13 \\ 21-12 & 13-13 \\ 2-12 & 7-13 \\ 12-12 & 20-13 \end{pmatrix} = \begin{pmatrix} 8 & 2 \\ -7 & -11 \\ 0 & 8 \\ 9 & 0 \\ -10 & -6 \\ 0 & 7 \end{pmatrix}$

- Calcul de la matrice variance –covariance :

$$V = \frac{1}{N} X^t . X \Rightarrow V = \frac{1}{6}\begin{pmatrix} 294 & 153 \\ 153 & 274 \end{pmatrix}.$$

- Détermination de la métrique :

Les données sont hétérogènes $\Rightarrow M = D_{\frac{1}{\sigma_j^2}}\ j = 1.....P$

$$M = \begin{pmatrix} 6/294 & 0 \\ 0 & 6/274 \end{pmatrix} \Rightarrow VM = \begin{pmatrix} 1 & 153/274 \\ 153/294 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.56 \\ 0.52 & 1 \end{pmatrix}$$

- Recherche des axes principaux U_k de la matrice (VM) :
  - Calcul des valeurs propres λ_k de la matrice (VM) en utilisant la formule suivante : $dét(VM-\lambda I)=0$
  - $dét(VM-\lambda I)=\frac{1}{6}\begin{vmatrix}1-\lambda & 0.56\\ 0.52 & 1-\lambda\end{vmatrix}=0\Rightarrow \lambda^2-2\lambda+0.71=0$

$$\Delta=b^2-4ac=(-2)^2-4.(1)(0.71)=1.16\rightarrow\sqrt{\Delta}=6\Rightarrow\begin{cases}\lambda_2=\frac{-b+\sqrt{\Delta}}{2a}=\frac{2-1.08}{2}=0.46\\ \lambda_3=\frac{-b-\sqrt{\Delta}}{2a}=\frac{2+1.08}{2}=1.54\end{cases}$$

$$\lambda_1=1.54\succ\lambda_2=0.46$$

- **Calcul de la qualité de représentation**

$$Q_j=\frac{\sum_{i=1}^{j}\lambda_i}{\sum_{i=1}^{p}\lambda_i}\geq 80\%\Rightarrow Q_1=\frac{\lambda_1}{\lambda_1+\lambda_2}=\frac{1.54}{1.54+0.46}=0.77=77\%\prec 80\%$$

$$Q_2=\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2}=\frac{1.54+0.46}{1.54+0.46}=1=100\%\geq 80\%\rightarrow$$ **Il y'a deux axes principaux U_1,U_2 relatifs aux valeurs propres λ_1,λ_2**

- **Calcul des vecteurs propres U_1,U_2 de la matrice (VM) :** $VMU_k=\lambda_k U_k$

$$VMU_1=\lambda_1U_1\Rightarrow\begin{pmatrix}1 & 0.56\\ 0.52 & 1\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}=1.54\begin{pmatrix}x\\ y\end{pmatrix}\Rightarrow\begin{cases}x+0.56y=1.54x\\ 0.52x+y=1.54y\end{cases}\Rightarrow\{-0.54x+0.5y=0\rightarrow x=1.04y$$

$$\Rightarrow U_1^*=\begin{pmatrix}1.04\\ 1\end{pmatrix}$$

$$\left\|U_1^*\right\|_M^2=U_1^{*t}.M.U_1^*=U_1^{*t}.D_{\frac{1}{\sigma_j^2}}.U_1^*=\begin{pmatrix}1.04 & 1\end{pmatrix}\begin{pmatrix}0.02 & 0\\ 0 & 0.02\end{pmatrix}\begin{pmatrix}1.04\\ 1\end{pmatrix}=0.04\Rightarrow\left\|U_1^*\right\|_M=0.2\Rightarrow U_1=\frac{U_1^*}{\left\|U_1^*\right\|_M}=\begin{pmatrix}5.2\\ 5\end{pmatrix}$$

$$VMU_2=\lambda_2U_2\Rightarrow\begin{pmatrix}1 & 0.56\\ 0.52 & 1\end{pmatrix}\begin{pmatrix}x'\\ y'\end{pmatrix}=0.46\begin{pmatrix}x'\\ y'\end{pmatrix}\Rightarrow\begin{cases}x'+0.56y'=0.46x'\\ 0.52x'+y'=0.46y'\end{cases}\Rightarrow\{0.54x'+0.56y'=0\rightarrow x'=-1.04y'$$

$$\Rightarrow U_2^*=\begin{pmatrix}-1.04\\ 1\end{pmatrix}$$

$$\left\|U_2^*\right\|_M^2=U_2^{*t}.M.U_2^*=U_2^{*t}.D_{\frac{1}{\sigma_j^2}}.U_2^*=\begin{pmatrix}-1.04 & 1\end{pmatrix}\begin{pmatrix}0.02 & 0\\ 0 & 0.02\end{pmatrix}\begin{pmatrix}-1.04\\ 1\end{pmatrix}=0.04\Rightarrow\left\|U_2^*\right\|_M=0.2\Rightarrow U_1=\frac{U_2^*}{\left\|U_2^*\right\|_M}=\begin{pmatrix}-5.2\\ 5\end{pmatrix}$$

- **Calcul des composantes principales** $C_k = X.M.U_k$

$$C_1 = X.M.U_1 = X.D_{\frac{1}{\sigma_j^2}}.U_1 = \begin{pmatrix} 8 & 2 \\ -7 & -11 \\ 0 & 8 \\ 9 & 0 \\ -10 & -6 \\ 0 & 7 \end{pmatrix} \cdot \begin{pmatrix} 0.02 & 0 \\ 0 & 0.02 \end{pmatrix} \begin{pmatrix} 5.2 \\ 5 \end{pmatrix} = \begin{pmatrix} 1.03 \\ -1.83 \\ 0.8 \\ 0.9 \\ -1.64 \\ 0.7 \end{pmatrix}$$

$$C_2 = X.M.U_2 = X.D_{\frac{1}{\sigma_j^2}}.U_2 = \begin{pmatrix} 8 & 2 \\ -7 & -11 \\ 0 & 8 \\ 9 & 0 \\ -10 & -6 \\ 0 & 7 \end{pmatrix} \cdot \begin{pmatrix} 0.02 & 0 \\ 0 & 0.02 \end{pmatrix} \begin{pmatrix} -5.2 \\ 5 \end{pmatrix} = \begin{pmatrix} -0.63 \\ -0.37 \\ 0.8 \\ -0.9 \\ 0.44 \\ 0.7 \end{pmatrix}$$

- **Représentation graphique des individus :**

$I_1 = (1.03,-0.63); I_2 = (-1.83,-0.37); I_3 = (0.8,0.8); I_4 = (0.9,-0.9); I_5 = (-1.64,0.44); I_6 = (0.7,0.7)$

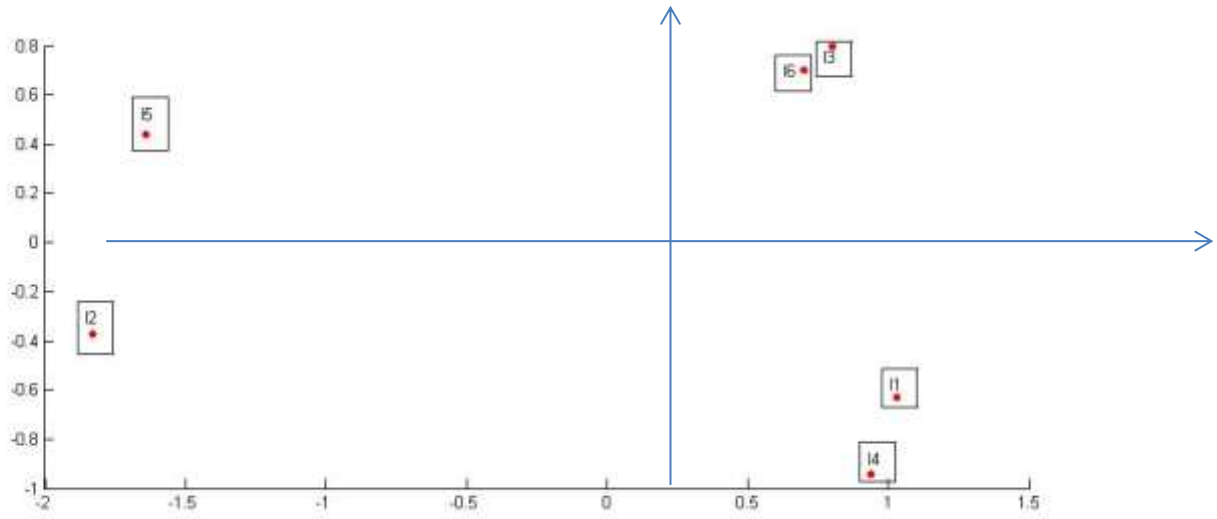


Fig.2.3 : Représentation graphique des individus

- **Calcul des contributions :**

$$C_1 = \begin{pmatrix} 1.03 \\ -1.83 \\ 0.8 \\ 0.9 \\ -1.64 \\ 0.7 \end{pmatrix} \quad C_2 = \begin{pmatrix} -0.63 \\ -0.37 \\ 0.8 \\ -0.9 \\ 0.44 \\ 0.7 \end{pmatrix}$$

- **Contribution relative de l'individu X_i à l'inertie expliquée de l'axe U_k :**

$$\rho_{ik} = \frac{P_i.(C_k^i)^2}{\sum_{i=1}^{n} P_i.(C_k^i)^2} = \frac{P_i.(C_k^i)^2}{Var(C_k)} = \frac{P_i.(C_k^i)^2}{\lambda_k}$$

- **Par rapport à l'axe U_1 : i=1…6**

$$\rho_{11} = \frac{P_1.(C_1^1)^2}{\lambda_1} = \frac{\frac{1}{6}.(1.03)^2}{1.54} = 0.11$$

$$\rho_{21} = 0.36; \rho_{31} = 0.07; \rho_{41} = 0.09; \rho_{51} = 0.29; \rho_{61} = 0.05.$$

- **Par rapport à l'axe U_2 : i=1…6**

$$\rho_{12} = \frac{P_1.(C_2^1)^2}{\lambda_2} = \frac{\frac{1}{6}.(-0.63)^2}{8} = 0.14;$$

$$\rho_{22} = 0.05; \rho_{32} = 0.23; \rho_{42} = 0.29; \rho_{52} = 0.07; \rho_{62} = 0.18.$$

- **Part d'inertie de X_i prise en compte par l'axe U_k :** $\cos^2(\theta_{ik}) = \frac{(C_k^i)^2}{\|X_i\|_M^2}$.

- **Par rapport à l'axe U_1 : i=1…6**

$$\cos^2(\theta_{11}) = \frac{(C_1^1)^2}{\|X_1\|_M^2} = \frac{(1.03)^2}{\begin{pmatrix}8 & 2\end{pmatrix}.\begin{pmatrix}0.02 & 0\\ 0 & 0.02\end{pmatrix}.\begin{pmatrix}8\\ 2\end{pmatrix}} = 0.78$$

$$\cos^2(\theta_{21}) = \frac{(C_1^2)^2}{\|X_2\|_M^2} = \frac{(-1.83)^2}{\begin{pmatrix}-7 & -11\end{pmatrix}.\begin{pmatrix}0.02 & 0\\ 0 & 0.02\end{pmatrix}.\begin{pmatrix}-7\\ -11\end{pmatrix}} = 0.98$$

$$\cos^2(\theta_{31}) = 0.5;\ \cos^2(\theta_{41}) = 0.54;\ \cos^2(\theta_{51}) = 0.99;\ \cos^2(\theta_{61}) = 0.5;$$

- **Par rapport à l'axe U_2 : i=1…6**

$$\cos^2(\theta_{12}) = \frac{(C_2^1)^2}{\|X_1\|_M^2} = \frac{(-0.63)^2}{\begin{pmatrix}8 & 2\end{pmatrix}.\begin{pmatrix}0.02 & 0\\ 0 & 0.02\end{pmatrix}.\begin{pmatrix}8\\ 2\end{pmatrix}} = 0.29$$

$$\cos^2(\theta_{22}) = \frac{(C_2^2)^2}{\|X_2\|_M^2} = \frac{(-0.37)^2}{\begin{pmatrix}-7 & -11\end{pmatrix}.\begin{pmatrix}0.02 & 0\\ 0 & 0.02\end{pmatrix}.\begin{pmatrix}-7\\ -11\end{pmatrix}} = 0.04;$$

$$\cos^2(\theta_{32})=0.5;\ \cos^2(\theta_{42})=0.54;\ \cos^2(\theta_{52})=0.07;\ \cos^2(\theta_{62})=0.5;$$

- **Représentation des variables à l'aide du coefficient de corrélation**

$$Cor(X^j,C_K)=\frac{Cov(X^j,C_k)}{\sigma_{X^j}.\sigma_{C_k}}=\frac{\sum_{i=1}^{n}P_i.X_i^j.C_k^i}{\sigma_{X^j}.\sqrt{\lambda_k}}=\frac{\frac{1}{N}(X^j)^t.C_k}{\sigma_{X^j}.\sqrt{\lambda_k}}$$

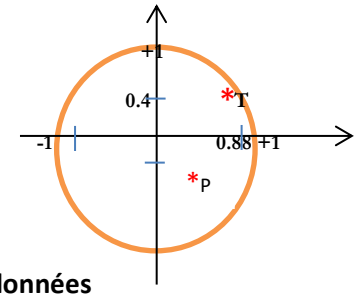
$$Cor(X^1:Poids,C_1)=\frac{Cov(X^1,C_1)}{\sigma_{X^1}.\sigma_{C_1}}=\frac{\sum_{i=1}^{n}P_i.X_i^1.C_1^i}{\sigma_{X^1}.\sqrt{\lambda_1}}=\frac{\frac{1}{N}(X^1)^t.C_1}{\sigma_{X^1}.\sqrt{\lambda_1}}=\frac{\frac{1}{6}\begin{pmatrix}8&-7&0&9&-10&0\end{pmatrix}*\begin{pmatrix}1.03\\-1.83\\0.8\\0.9\\-1.64\\0.7\end{pmatrix}}{\sqrt{294/6}*\sqrt{1.54}}=0.88$$

$$Cor(X^1:Poids,C_2)=\frac{Cov(X^1,C_2)}{\sigma_{X^1}.\sigma_{C_2}}=\frac{\sum_{i=1}^{n}P_i.X_i^1.C_2^i}{\sigma_{X^1}.\sqrt{\lambda_2}}=\frac{\frac{1}{N}(X^1)^t.C_2}{\sigma_{X^1}.\sqrt{\lambda_2}}=\frac{\frac{1}{6}\begin{pmatrix}8&-7&0&9&-10&0\end{pmatrix}*\begin{pmatrix}-0.63\\-0.37\\0.8\\-0.9\\0.44\\0.7\end{pmatrix}}{\sqrt{294/6}*\sqrt{0.46}}=-0.54$$

$$Cor(X^2:Taille,C_1)=\frac{Cov(X^2,C_1)}{\sigma_{X^2}.\sigma_{C_1}}=\frac{\sum_{i=1}^{n}P_i.X_i^2.C_1^i}{\sigma_{X^2}.\sqrt{\lambda_1}}=\frac{\frac{1}{N}(X^2)^t.C_1}{\sigma_{X^2}.\sqrt{\lambda_1}}=\frac{\frac{1}{6}\begin{pmatrix}2&-11&8&0&-6&7\end{pmatrix}*\begin{pmatrix}1.03\\-1.83\\0.8\\0.9\\-1.64\\0.7\end{pmatrix}}{\sqrt{274/6}*\sqrt{1.54}}=0.86$$

$$Cor(X^2:Taille,C_2)=\frac{Cov(X^2,C_2)}{\sigma_{X^2}.\sigma_{C_2}}=\frac{\sum_{i=1}^{n}P_i.X_i^2.C_2^i}{\sigma_{X^2}.\sqrt{\lambda_2}}=\frac{\frac{1}{N}(X^2)^t.C_2}{\sigma_{X^2}.\sqrt{\lambda_2}}=\frac{\frac{1}{6}\begin{pmatrix}2&-11&8&0&-6&7\end{pmatrix}*\begin{pmatrix}-0.63\\-0.37\\0.8\\-0.9\\0.44\\0.7\end{pmatrix}}{\sqrt{274/6}*\sqrt{0.46}}=0.42$$

Cercle de corrélation :



Les partitions possibles

$P_1 = \{\{I_1, I_4\}, \{I_3, I_6\}, \{I_2\}, \{I_5\}\}$

$C_1 \quad C_2 \quad C_3 \quad C_4$

C_1 : Poids et taille importants / Poids >taille.

C_2 : Poids et taille importants / Poids <taille.

C_3 : Poids et taille petits / Poids >taille.

C_4 : Poids et taille petits / Poids <taille.

$P_2 = \{\{I_1, I_2, I_4\}, \{I_3, I_5, I_6\}\}$

$C_1 \quad C_2$

C_1 : Poids >taille.

C_2 : Poids <taille.

$P_3 = \{\{I_1, I_3, I_4, I_6\}, \{I_2, I_5\}\}$

$C_1 \quad C_2$

C_1 : Poids et taille importants.

C_2 : Poids et taille importants.

$P_4 = \{\{I_1, I_4\}, \{I_3, I_6\}, \{I_2, I_5\}\}$

$C_1 \quad C_2 \quad C_3$

C_1 : Poids et taille importants / Poids >taille.

C_2 : Poids et taille importants / Poids <taille.

C_3 : Poids et taille petits.

$P_5 = \{\{I_1, I_3, I_4, I_6\}, \{I_2\}, \{I_5\}\}$

$C_1 \quad C_2 \quad C_3$

C_1 : Poids et taille importants.

C_2 : Poids et taille petits / Poids >taille.

C_3 : Poids et taille petits/ Poids< taille.

Exercice :

On a relevé dans trois magasins (M1, M2 et M3) d'un même quartier appartenant à des chaines différentes. Les prix affichés pour quatre produits vendus sous quatre marques différentes (A, B, C et D).

	M1	M2	M3
A	16	20	12
B	20	12	22
C	16	24	26
D	28	24	20

1. On veut faire effectuer l'analyse en composantes principales de ce tableau (Données homogènes).
   - Vérifier que $U_1^t = 1/\sqrt{3}(1 \quad 1 \quad 1)$ et $U_2^t = 1/\sqrt{6}(1 \quad 1 \quad -2)$ sont des vecteurs propres de cette ACP
2. Représenter le nuage des points produits dans le plan principal.
3. Représenter le produit supplémentaire E ayant les prix suivants : $E^t = (16 \quad 8 \quad 12)$
4. Représenter graphiquement les trois variables (M1, M2 et M3).

Solution :

1. U_1, U_2 **sont des vecteurs propres de VM** $\Rightarrow \exists \lambda_1 \, tq \quad VMU_1 = \lambda_1 U_1 ; \exists \lambda_2 \, tq \quad VMU_2 = \lambda_2 U_2$

- Le tableau centré $X_{<n,p>}$:
  - Calcul du centre de gravité : $g = \dfrac{\sum_{i=1}^{n} p_i . x_i}{\sum_{i=1}^{n} p_i}$

$P_i = \dfrac{1}{N}$ où N représente le nombre d'individus. Cela implique que $P_i = \dfrac{1}{4}$; $i = 1,\ldots\ldots, n; i = 1\ldots'$; $j = 1,\ldots\ldots p; 1\ldots 3$

$$g = \begin{pmatrix} \frac{1}{4}(16+20+16+28) \\ \frac{1}{4}(20+12+24+24) \\ \frac{1}{4}(12+22+26+20) \end{pmatrix} = \begin{pmatrix} 20 \\ 20 \\ 20 \end{pmatrix}$$

	M1	M2	M3
A	16	20	12
B	20	12	22
C	16	24	26
D	28	24	20
g	**20**	**20**	**20**

Tableau centré $\longrightarrow$ $X = \begin{pmatrix} 16-20 & 20-20 & 12-20 \\ 20-20 & 12-20 & 22-20 \\ 16-20 & 24-20 & 26-20 \\ 28-20 & 24-20 & 20-20 \end{pmatrix} = \begin{pmatrix} -4 & 0 & -8 \\ 0 & -8 & 2 \\ -4 & 4 & 6 \\ 8 & 4 & 0 \end{pmatrix}$

- Calcul de la matrice variance –covariance :

$$V = \frac{1}{N} X^t . X \Rightarrow V = \begin{pmatrix} 24 & 4 & 2 \\ 4 & 24 & 2 \\ 2 & 2 & 26 \end{pmatrix}.$$

- Détermination de la métrique :

Les données sont homogènes car elles possèdent le même type de mesure $\Rightarrow M = Id = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

- U_1, U_2 **sont des vecteurs propres de VM** $\Rightarrow \exists \lambda_1 \; tq \quad VMU_1 = \lambda_1 U_1 ; \exists \lambda_2 \; tq \quad VMU_2 = \lambda_2 U_2$

$$\Rightarrow \exists \lambda_1 \; tq \quad VMU_1 = \lambda_1 U_1 \Rightarrow \begin{pmatrix} 24 & 4 & 2 \\ 4 & 24 & 2 \\ 2 & 2 & 26 \end{pmatrix} . \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \lambda_1 \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \Rightarrow \frac{30}{\sqrt{3}} = \frac{\lambda_1}{\sqrt{3}} \Rightarrow \lambda_1 = 30$$

$$\Rightarrow \exists \lambda_2 \; tq \quad VMU_2 = \lambda_2 U_2 \Rightarrow \begin{pmatrix} 24 & 4 & 2 \\ 4 & 24 & 2 \\ 2 & 2 & 26 \end{pmatrix} . \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} = \lambda_2 \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} \Rightarrow \frac{24}{\sqrt{6}} = \frac{\lambda_1}{\sqrt{6}} \Rightarrow \lambda_2 = 24$$

Le plan principal : On cherche à projeter les données dans un espace à deux dimensions U_1, U_2

- $Tr(VM) = \sum_{j=1}^{P} \lambda_j \Rightarrow 74 = 30 + 24 + \lambda_3 \Rightarrow \lambda_3 = 20 \; (\lambda_1 = 30 \succ \lambda_2 = 24 \succ \lambda_3 = 20)$

2. Coordonnées des produits (Individus) : $C_k = X.M.U_k$

$$C_1 = X.M.U_1 = X.Id.U_1 = X.U_1 = \begin{pmatrix} -4 & 0 & -8 \\ 0 & -8 & 2 \\ -4 & 4 & 6 \\ 8 & 4 & 0 \end{pmatrix} . \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} -12 \\ -6 \\ +6 \\ +12 \end{pmatrix} = \sqrt{3} \begin{pmatrix} -4 \\ -2 \\ +2 \\ +4 \end{pmatrix}$$

$$C_2 = X.M.U_2 = X.Id.U_2 = X.U_2 = \begin{pmatrix} -4 & 0 & -8 \\ 0 & -8 & 2 \\ -4 & 4 & 6 \\ 8 & 4 & 0 \end{pmatrix} . \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} = \frac{1}{\sqrt{6}} \begin{pmatrix} +12 \\ -12 \\ -12 \\ +12 \end{pmatrix} = \sqrt{6} \begin{pmatrix} +2 \\ -2 \\ -2 \\ +2 \end{pmatrix}$$

- **Représentation graphique des individus :**

$A = (-4\sqrt{3}, +2\sqrt{6}); B = (-2\sqrt{3}, -2\sqrt{6}); C = (2\sqrt{3}, -2\sqrt{6}); D = (4\sqrt{3}, 2\sqrt{6});$

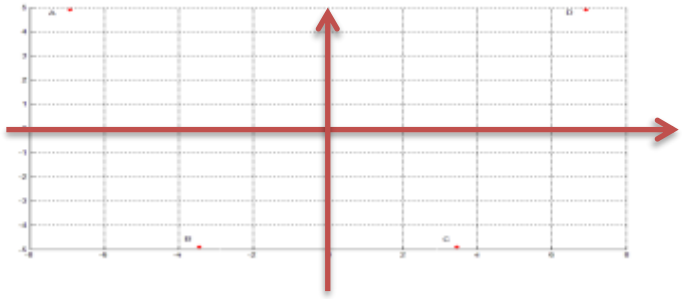


Fig. 2.4 : Représentation des individus dans l'espace réduit.

3. Coordonnées de E=(16 8 12)^t

$$C_1 = X.M.U_1 = X.Id.U_1 = X.U_1 = (16-20 \quad 8-20 \quad 12-20).\frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix} = (-4 \quad -12 \quad -8)\frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix} = -8\sqrt{3}$$

$$C_2 = X.M.U_2 = X.Id.U_2 = X.U_2 = (-4 \quad -12 \quad -8).\frac{1}{\sqrt{6}}\begin{pmatrix}1\\1\\-2\end{pmatrix} = 0$$

$$E = (-8\sqrt{3}, 0);$$

4. Représentation des variables à l'aide du coefficient de corrélation

$$Cor(X^j, C_K) = \frac{Cov(X^j, C_k)}{\sigma_{X^j}.\sigma_{C_k}} = \frac{\sum_{i=1}^{n} P_i.X_i^j.C_k^i}{\sigma_{X^j}.\sqrt{\lambda_k}} = \frac{\frac{1}{N}(X^j)^t.C_k}{\sigma_{X^j}.\sqrt{\lambda_k}}$$

$$Cor(X^1 : M_1, C_1) = \frac{Cov(X^1, C_1)}{\sigma_{X^1}.\sigma_{C_1}} = \frac{\sum_{i=1}^{n} P_i.X_i^1.C_1^i}{\sigma_{X^1}.\sqrt{\lambda_1}} = \frac{\frac{1}{N}(X^1)^t.C_1}{\sigma_{X^1}.\sqrt{\lambda_1}} = \frac{\frac{1}{4}(-4 \quad 0 \quad -4 \quad -8)*\sqrt{3}\begin{pmatrix}-4\\-2\\+2\\+4\end{pmatrix}}{\sqrt{24}*\sqrt{30}} = -\frac{\sqrt{3}}{2\sqrt{5}} = 0.39$$

$$Cor(X^1 : M_1, C_2) = \frac{Cov(X^1, C_2)}{\sigma_{X^1}.\sigma_{C_2}} = \frac{\sum_{i=1}^{n} P_i.X_i^1.C_2^i}{\sigma_{X^1}.\sqrt{\lambda_2}} = \frac{\frac{1}{N}(X^1)^t.C_2}{\sigma_{X^1}.\sqrt{\lambda_2}} = \frac{\frac{1}{4}(-4 \quad 0 \quad -4 \quad -8)*\sqrt{6}\begin{pmatrix}+2\\-2\\-2\\+2\end{pmatrix}}{\sqrt{24}*\sqrt{24}} = \frac{-1}{\sqrt{6}} = -0.41$$

$$Cor(X^2 : M_2, C_1) = \frac{Cov(X^2, C_1)}{\sigma_{X^2}.\sigma_{C_1}} = \frac{\sum_{i=1}^{n} P_i.X_i^2.C_1^i}{\sigma_{X^2}.\sqrt{\lambda_1}} = \frac{\frac{1}{N}(X^2)^t.C_1}{\sigma_{X^2}.\sqrt{\lambda_1}} = \frac{\frac{1}{4}(0 \quad -8 \quad 4 \quad 4)*\sqrt{3}\begin{pmatrix}-4\\-2\\+2\\+4\end{pmatrix}}{\sqrt{24}*\sqrt{30}} = \frac{\sqrt{5}}{2\sqrt{3}} = 0.64$$

$$Cor(X^2 : N_2, C_2) = \frac{Cov(X^2, C_2)}{\sigma_{X^2}.\sigma_{C_2}} = \frac{\sum_{i=1}^{n} P_i.X_i^2.C_2^i}{\sigma_{X^2}.\sqrt{\lambda_2}} = \frac{\frac{1}{N}(X^2)^t.C_2}{\sigma_{X^2}.\sqrt{\lambda_2}} = \frac{\frac{1}{4}(0 \quad -8 \quad 4 \quad 4)*\sqrt{6}\begin{pmatrix}+2\\-2\\-2\\+2\end{pmatrix}}{\sqrt{24}*\sqrt{24}} = \frac{1}{\sqrt{6}} = 0.41$$

$$Cor\left(X^3 : N_3, C_1\right) = \frac{Cov\left(X^3, C_1\right)}{\sigma_{X^3}.\sigma_{C_1}} = \frac{\sum_{i=1}^{n} P_i.X_i^3.C_1^i}{\sigma_{X^3}.\sqrt{\lambda_1}} = \frac{\frac{1}{N}\left(X^3\right)^t.C_1}{\sigma_{X^3}.\sqrt{\lambda_1}} = \frac{\frac{1}{4}\begin{pmatrix}-8 & 2 & 6 & 0\end{pmatrix} * \sqrt{3}\begin{pmatrix}-4\\-2\\+2\\+4\end{pmatrix}}{\sqrt{26} * \sqrt{30}} = \frac{\sqrt{5}}{\sqrt{13}} = 0.62$$

$$Cor\left(X^3 : N_3, C_2\right) = \frac{Cov\left(X^3, C_2\right)}{\sigma_{X^3}.\sigma_{C_2}} = \frac{\sum_{i=1}^{n} P_i.X_i^3.C_2^i}{\sigma_{X^3}.\sqrt{\lambda_2}} = \frac{\frac{1}{N}\left(X^3\right)^t.C_2}{\sigma_{X^3}.\sqrt{\lambda_2}} = \frac{\frac{1}{4}\begin{pmatrix}-8 & 2 & 6 & 0\end{pmatrix} * \sqrt{6}\begin{pmatrix}+2\\-2\\-2\\+2\end{pmatrix}}{\sqrt{26} * \sqrt{24}} = \frac{-4}{\sqrt{26}}$$

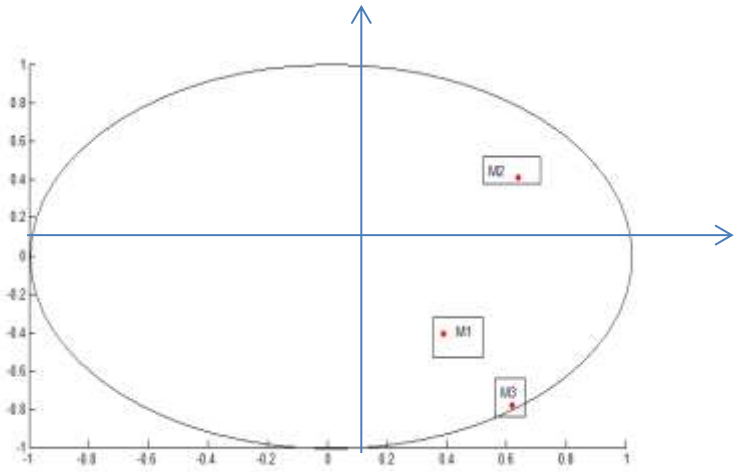


Fig. 2.5 Cercle de corrélation.

La première composante C_1 est positivement corrélée avec les trois variables (M1, M2 et M3). C_1 nous informe sur la quantité moyenne stockée dans les trois magasins ($C_1 \approx$moyenne).

C_2 est positivement corrélée avec M2 et négativement corrélée avec M1, M3. C1 nous informe sur la différence entre la quantité stockée dans le magasin M2 et les deux autres magasins M1, M3.

On peut aussi déduire que la Qte-moy(D)>Qte-moy(C)>Qte-moy(B)>Qte-moy(A).

5. Cordonnée de E :

$$C_1 = X.M.U_1 = X.Id.U_1 = X.U_1 = (16-20 \quad 8-20 \quad 12-20).\frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix} = (-4 \quad -12 \quad -8)\frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix} = -8\sqrt{3}$$

$$C_2 = X.M.U_2 = X.Id.U_2 = X.U_2 = (-4 \quad -12 \quad -8).\frac{1}{\sqrt{6}}\begin{pmatrix}1\\1\\-2\end{pmatrix} = 0 \Rightarrow E = (-8\sqrt{3},0);$$

2.8 Conclusion

L'ACP est une méthode statistique puissantes qui permet de synthétiser de vastes populations décrites par plusieurs variables quantitatives. Elle permet aussi de classifier les individus et de réaliser un bilan des liaisons entre les variables. L'ACP permet une double visualisation graphique i.e. une représentation graphique pour les individus, tandis que l'autre visualisation pour les variables à l'aide de cercle de corrélation.

MÉTHODES DE CLASSIFICATION

3.1 Les méthodes de classification

Le but des méthodes de classification est de construire une partition, ou une suite de partitions emboitées, d'un ensemble d'objets dont on connait les distances deux à deux. Les classes formées doivent être le plus homogène possible [9].

3.2 Classification hiérarchiques

3.2.1 Définition

Elles constituent en un ensemble de partitions de Ω en classes de moins en moins fines obtenues par regroupement successifs de parties. Une classification hiérarchique se représente par un dendrogramme ou arbre de classification (voir figure 3.1).

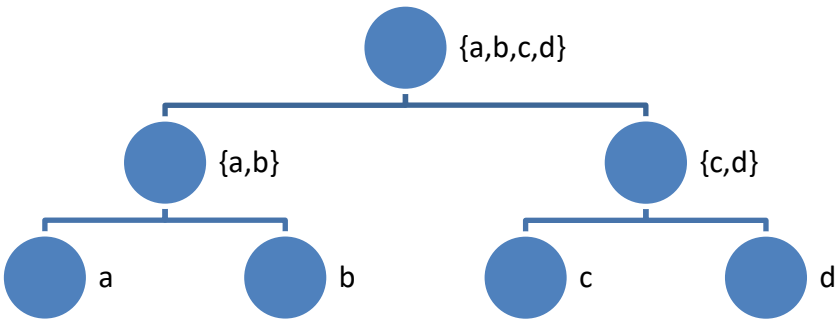


Figure 3.1 : Dendrogramme (Arbre hiérarchique).

Cet arbre est obtenu de manière ascendante en regroupant tout d'abord les deux individus les plus proches qui forment un nœud, il ne reste plus que n-1 objets et on itère le processus jusqu'à regroupement complet.

- **Aspect formel**

Soit Ω un ensemble fini, H un ensemble de parties non vides de Ω est une hiérarchie si :

$$\begin{cases} \Omega \in H \\ \forall I_i \in \Omega \Rightarrow \{I_i\} \in H \\ \forall h, h' \in H \Rightarrow h \cap h' = \varnothing \ ou \ h \subset h' \ ou \ h' \subset h \end{cases}$$

Exemple : $H = \{\varnothing, \{a\}, \{b\}, \{c\}, \{d\}, \{a,b\}, \{c,d\}, \{a,b,c,d\}\}$

Pour déterminer les partitions possibles, nous traçons une ligne horizontale entre chaque niveau d'agrégation et en recueillant les morceaux.

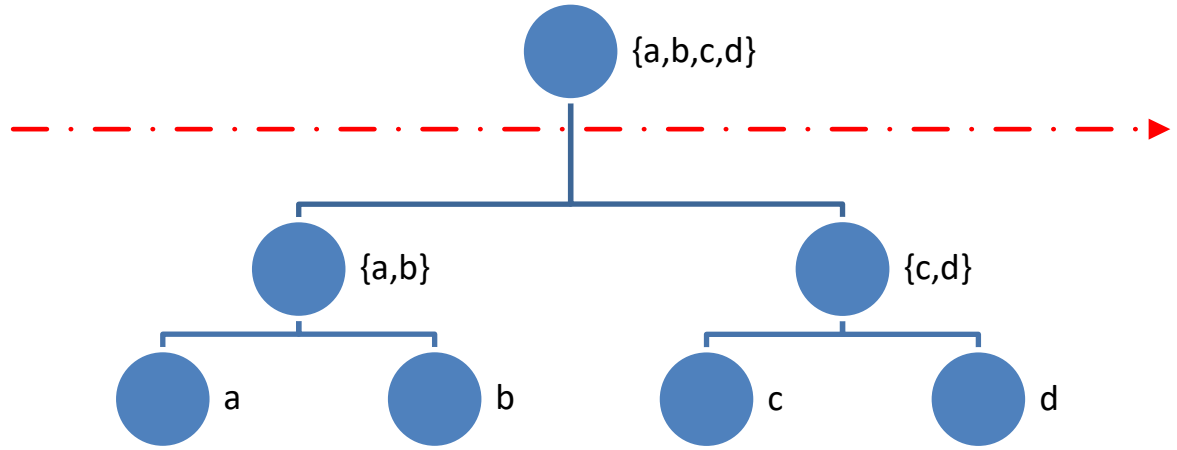


Figure 3.2 : La ligne de découpage.

3.2.2 Algorithme

La table 3.1 montre les étapes de l'algorithme CHA.

Algorithme de la classification hiérarchique ascendante (CHA)
Input : le critère d'agrégation **1-** Initialiser les n singletons et calculer la matrice de distance deux à deux. **2-** Regrouper les deux éléments les plus proches au sens de la distance entre groupes choisis **3-** Mettre à jour le tableau de distance en remplaçant les deux classes regroupées par la nouvelle et en calculant sa distance avec chacune des autres classes. **4-** Répéter les étapes 2 et 3 jusqu'à l'agrégation en une seule classe. **Output** : Les partitions et le dendrogramme

3.2.3 Définition d'une hiérarchie indicée

Une hiérarchie indicée est un couple (H,ν) tels que : H : hiérarchie et ν est une application dans $\mathbb{R}^+$ avec :

$$\begin{cases} \nu(\{I_i\}) = 0; \forall I_i \in \Omega \\ \forall h, h' \in H, si\ h \subset h' \Rightarrow \nu(h) \prec \nu(h') \end{cases}$$

3.2.4 Les critères d'agrégation

Plusieurs critères sont développés pour la classification hiérarchique ascendante. Nous citons [10]:

- Critère du saut minimum

On prend ici comme distance entre parties la plus petite distance avec :

$$\delta^h\left(t, s_h \cup s'_h\right) = Min\left(\delta^{h-1}\left(t, s_h\right), \delta^{h-1}\left(t, s'_h\right)\right).$$

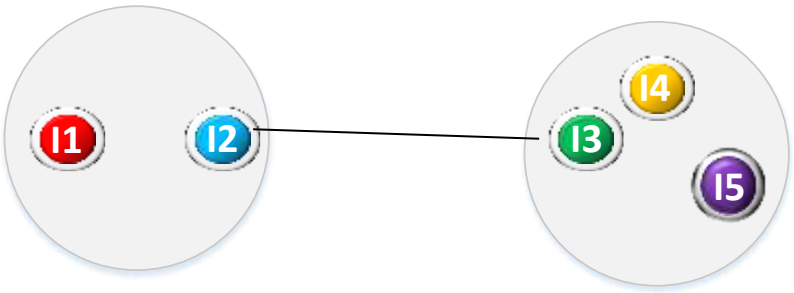


Figure 3.3 : Principe du saut minimum.

- Critère du saut maximum

On prend ici comme distance entre parties la plus grande distance avec :

$$\delta^h\left(t, s_h \cup s'_h\right) = Max\left(\delta^{h-1}\left(t, s_h\right), \delta^{h-1}\left(t, s'_h\right)\right).$$

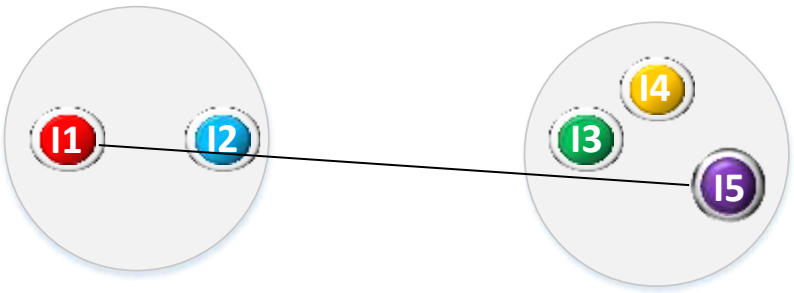


Figure 3.4 : Principe du saut minimum.

- Critère de la minimisation d'inertie de la réunion de deux classes

$$\delta^0\left(\{i\}, \{i'\}\right) = \frac{m_i * m_{i'}}{m_i + m_{i'}} \left\| i - i' \right\|^2$$

$$\delta^h\left(t, s_h \cup s'_h\right) = \frac{\left(m_t + m_{s_h}\right)\delta^{h-1}\left(t, s_h\right) + \left(m_t + m_{s'_h}\right)\delta^{h-1}\left(t, s'_h\right) \cup + \left(m_{s_h} + m_{s'_h}\right)\delta^{h-1}\left(s_h, s'_h\right) - \left(m_t v(t) - m_{s_h} v(s_h) - m_{s'_h} v(s'_h)\right)}{m_t + m_{s_h} + m_{s'_h}}$$

- Critère de Ward

$$\delta^h(A,B)=\frac{P_A P_B}{P_A+P_B}d^2(g_A,g_B)$$

A chaque itération, on agrège de manière à avoir un gain minimum d'inertie intra-classes

$$\delta^h(A,B\cup C)=\frac{(P_A+P_B)\delta^h(A,B)+(P_A+P_C)\delta^h(A,C)-P_C\delta^h(B,C)}{P_A+P_B+P_C}$$

3.2.5 Exemple

Soit l'ensemble des six individus suivants (poids d'un individu= 1) :

	I1	I2	I3	I4	I5	I6
X	1	2	6	2	3	6
Y	1	2	2	6	6	4

1. Appliquer une classification hiérarchique avec le critère de la minimisation de l'inertie de la réunion de deux classes.

2. Donner toutes les partitions trouvées.

3. Déduire l'inertie intra classe et inter classe pour chaque partition.

4. Quelle est la classe la plus excentrique de la partition en trois classes ?

5. Quelle est la contribution de chacun des deux axes à son excentricité ?

Solution :

1. Application de l'algorithme CHA :

$$\delta^0(\{i\},\{i'\})=\frac{m_i * m_{i'}}{m_i+m_{i'}}\|i-i'\|^2$$

$$\delta^h(t,s_h\cup s_h')=\frac{(m_t+m_{s_h})\delta^{h-1}(t,s_h)+(m_t+m_{s_h'})\delta^{h-1}(t,s_h')\cup+(m_{s_h}+m_{s_h'})\delta^{h-1}(s_h,s_h')-(m_t\nu(t)-m_{s_h}\nu(s_h)-m_{s_h'}\nu(s_h'))}{m_t+m_{s_h}+m_{s_h'}}$$

Etape 1 :

δ^0	$\{I_1\}$	$\{I_2\}$	$\{I_3\}$	$\{I_4\}$	$\{I_5\}$	$\{I_6\}$
$\{I_1\}$	**0**					
$\{I_2\}$	**1**	**0**				
$\{I_3\}$	**13**	**8**	**0**			
$\{I_4\}$	**13**	**8**	**16**	**0**		
$\{I_5\}$	**29/2**	**17/2**	**25/2**	**1/2**	**0**	
$\{I_6\}$	**17**	**10**	**2**	**10**	**13/2**	**0**

$$\delta^0(\{I_1\},\{I_2\})=\frac{1*1}{1+1}\left[(1-2)^2+(1-2)^2\right]^2$$

$$\delta^0(\{I_1\},\{I_2\})=1 \textbf{ et } \nu(\{I_7\})=1/2$$

$$\{I_7\}=\{I_4,I_5\}$$

Etape 2 :

δ^1	$\{I_7\}=\{I_4,I_5\}$	$\{I_1\}$	$\{I_2\}$	$\{I_3\}$	$\{I_6\}$
$\{I_7\}$	**0**				
$\{I_1\}$	**56/3**	**0**			
$\{I_2\}$	**34/3**	**1**	**0**		
$\{I_3\}$	**58/3**	**13**	**8**	**0**	
$\{I_6\}$	**34/3**	**17**	**10**	**2**	**0**

$\nu(\{I_8\})=1$

$\{I_8\}=\{I_1,I_2\}$

$$\delta^1(\{I_1\},\{I_4\}\cup\{I_5\})=\frac{(m_{\{I_1\}}+m_{\{I_4\}})\delta^0(\{I_1\},\{I_4\})+(m_{\{I_1\}}+m_{\{I_5\}})\delta^0(\{I_1\},\{I_5\})+(m_{\{I_4\}}+m_{\{I_5\}})\delta^0(\{I_4\},\{I_5\})-m_{\{I_1\}}\nu(\{I_1\})-m_{\{I_4\}}\nu(\{I_4\})-m_{\{I_5\}}\nu(\{I_5\})}{m_{\{I_1\}}+m_{\{I_4\}}+m_{\{I_5\}}}$$

$$\delta^1(\{I_1\},\{I_7\})=\delta^1(\{I_1\},\{I_4\}\cup\{I_5\})=\frac{(1+1)*13+(1+1)*\frac{29}{2}+(1+1)*\frac{1}{2}}{1+1+1}=\frac{56}{3}$$

$$\delta^1(\{I_2\},\{I_7\})=\delta^1(\{I_2\},\{I_4\}\cup\{I_5\})=\frac{(1+1)*8+(1+1)*\frac{17}{2}+(1+1)*\frac{1}{2}}{1+1+1}=\frac{34}{3}$$

$$\delta^1(\{I_3\},\{I_7\})=\delta^1(\{I_3\},\{I_4\}\cup\{I_5\})=\frac{(1+1)*16+(1+1)*\frac{25}{2}+(1+1)*\frac{1}{2}}{1+1+1}=\frac{58}{3}$$

$$\delta^1(\{I_6\},\{I_7\})=\delta^1(\{I_6\},\{I_4\}\cup\{I_5\})=\frac{(1+1)*10+(1+1)*\frac{13}{2}+(1+1)*\frac{1}{2}}{1+1+1}=\frac{34}{3}$$

Etape 3 :

δ^2	$\{I_8\}=\{I_1,I_2\}$	$\{I_7\}$	$\{I_3\}$	$\{I_6\}$
$\{I_8\}$	**0**			
$\{I_7\}$	**91/3**	**0**		
$\{I_3\}$	**44/3**	**58/3**	**0**	
$\{I_6\}$	**56/3**	**34/3**	**2**	**0**

$\nu(\{I_9\})=2$

$\{I_9\}=\{I_3,I_6\}$

$$\delta^2(\{I_7\},\{I_1\}\cup\{I_2\})=\frac{(m_{\{I_7\}}+m_{\{I_1\}})\delta^1(\{I_7\},\{I_1\})+(m_{\{I_7\}}+m_{\{I_2\}})\delta^1(\{I_7\},\{I_2\})+(m_{\{I_1\}}+m_{\{I_2\}})\delta^1(\{I_1\},\{I_2\})-m_{\{I_7\}}\nu(\{I_7\})-m_{\{I_1\}}\nu(\{I_1\})-m_{\{I_2\}}\nu(\{I_2\})}{m_{\{I_7\}}+m_{\{I_1\}}+m_{\{I_2\}}}$$

$$\delta^1(\{I_7\},\{I_8\})=\delta^1(\{I_7\},\{I_1\}\cup\{I_2\})=\frac{(2+1)*\frac{56}{3}+(2+1)*\frac{34}{3}+(1+1)*1-2*\frac{1}{2}}{2+1+1}=\frac{91}{4}$$

$$\delta^1(\{I_3\},\{I_8\})=\delta^1(\{I_3\},\{I_1\}\cup\{I_2\})=\frac{(1+1)*13+(1+1)*8+(1+1)*1}{1+1+1}=\frac{44}{4}$$

$$\delta^1(\{I_6\},\{I_8\})=\delta^1(\{I_6\},\{I_1\}\cup\{I_2\})=\frac{(1+1)*17+(1+1)*10+(1+1)*1}{1+1+1}=\frac{56}{3}$$

Etape 4 :

δ^3	$\{I_9\}=\{I_3,I_6\}$	$\{I_8\}=\{I_1,I_2\}$	$\{I_7\}$
$\{I_9\}$	**0**		
$\{I_8\}$	**102/4**	**0**	
$\{I_7\}$	**95/4**	**91/4**	**0**

$\nu(\{I_{10}\})=\frac{91}{4}$

$\{I_{10}\}=\{I_7,I_8\}$

$$\delta^3(\{I_8\},\{I_3\}\cup\{I_6\})=\frac{(m_{\{I_8\}}+m_{\{I_3\}})\delta^2(\{I_8\},\{I_3\})+(m_{\{I_8\}}+m_{\{I_6\}})\delta^2(\{I_8\},\{I_6\})+(m_{\{I_3\}}+m_{\{I_6\}})\delta^2(\{I_3\},\{I_6\})-m_{\{I_8\}}\nu(\{I_8\})-m_{\{I_3\}}\nu(\{I_3\})-m_{\{I_6\}}\nu(\{I_6\})}{m_{\{I_8\}}+m_{\{I_3\}}+m_{\{I_6\}}}$$

$$\delta^3(\{I_8\},\{I_9\})=\delta^3(\{I_8\},\{I_3\}\cup\{I_6\})=\frac{(2+1)*\frac{44}{3}+(1+1)*\frac{56}{3}+(1+1)*2-2*1}{2+1+1}=\frac{102}{4}$$

$$\delta^3(\{I_7\},\{I_9\})=\delta^3(\{I_7\},\{I_3\}\cup\{I_6\})=\frac{(2+1)*\frac{58}{3}+(1+1)*\frac{34}{3}+(1+1)*2-2*\frac{1}{2}}{2+1+1+1}=\frac{95}{4}$$

Etape 5:

δ^4	$\{I_{10}\}=\{I_7,I_8\}$	$\{I_9\}$
$\{I_{10}\}$	**0**	
$\{I_9\}$	**281/6**	**0**

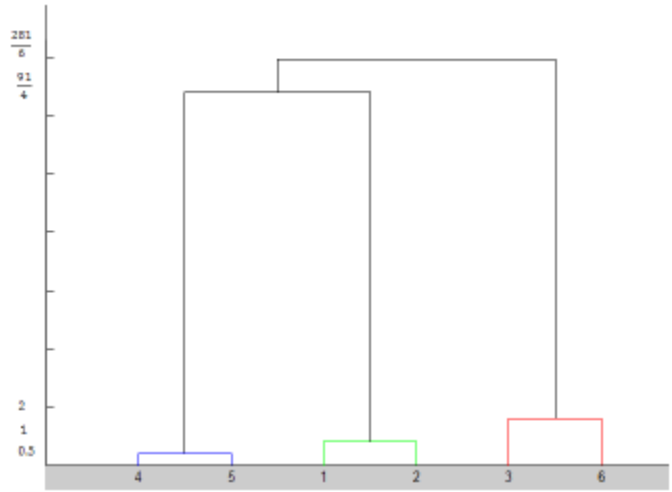
$\nu(\{I_{10}\})=\frac{281}{6}$

$\{I_{11}\}=\{I_9,I_{10}\}$

$$\delta^4(\{I_9\},\{I_7\}\cup\{I_8\})=\frac{(m_{\{I_9\}}+m_{\{I_7\}})\delta^3(\{I_9\},\{I_7\})+(m_{\{I_9\}}+m_{\{I_8\}})\delta^3(\{I_9\},\{I_8\})+(m_{\{I_7\}}+m_{\{I_8\}})\delta^3(\{I_7\},\{I_8\})-m_{\{I_9\}}\nu(\{I_9\})-m_{\{I_7\}}\nu(\{I_7\})-m_{\{I_8\}}\nu(\{I_8\})}{m_{\{I_9\}}+m_{\{I_7\}}+m_{\{I_8\}}}$$

$$\delta^4(\{I_9\},\{I_{10}\})=\delta^3(\{I_9\},\{I_7\}\cup\{I_8\})=\frac{(2+2)*\frac{95}{4}+(2+2)*\frac{102}{4}+(2+2)*\frac{91}{4}-2*2-2*\frac{1}{2}-2*1}{2+2+2}=\frac{281}{6}$$

- **L'arbre de la classification**



2. Les partitions possibles

$$P_0 = \{\{I_1\},\{I_2\},\{I_3\},\{I_4\},\{I_5\}\{I_6\}\}$$

$$P_1 = \{\{I_1\},\{I_2\},\{I_3\},\{I_6\},\{I_4,I_5\}\}$$

$$P_2 = \{\{I_1,I_2\},\{I_3\},\{I_4,I_5\},\{I_6\}\}$$

$$P_3 = \{\{I_1,I_2\},\{I_4,I_5\},\{I_3,I_6\}\}$$

$$P_4 = \{\{I_1,I_2,I_4,I_5\},\{I_3,I_6\}\}$$

$$P_5 = \{\{I_1,I_2,I_3,I_4,I_5,I_6\}\}$$

3. Calcul des inerties inter-classes et intra-classes pour chaque partition

$$P_0 = \{\{I_1\},\{I_2\},\{I_3\},\{I_4\},\{I_5\}\{I_6\}\}$$

$$I_{intra}(P_0) = v(\{I_1\}) + v(\{I_2\}) + v(\{I_3\}) + v(\{I_4\}) + v(\{I_5\}) + v(\{I_6\})$$

$$I_{intra}(P_0) = 0+0+0+0+0+0 \Rightarrow I_{intra}(P_0) = 0$$

$$I_{totale}(P_0) = I_{intra}(P_0) + I_{inter}(P_0) \Rightarrow I_{inter}(P_0) = I_{totale}(P_0) - I_{intra}(P_0)$$

$$I_{inter}(P_0) = \frac{281}{6} - 0 \Rightarrow I_{inter}(P_0) = \frac{281}{6} = 46.83$$

$$P_1 = \{\{I_1\},\{I_2\},\{I_3\},\{I_6\},\{I_4,I_5\}\}$$

$$I_{intra}(P_1) = v(\{I_1\}) + v(\{I_2\}) + v(\{I_3\}) + v(\{I_6\}) + v(\{I_4,I_5\})$$

$$I_{intra}(P_1)=0+0+0+0+\frac{1}{2}\Rightarrow I_{intra}(P_1)=\frac{1}{2}$$

$$I_{totale}(P_1)=I_{intra}(P_1)+I_{inter}(P_1)\Rightarrow I_{inter}(P_1)=I_{totale}(P_1)-I_{intra}(P_1)$$

$$I_{inter}(P_1)=\frac{281}{6}-\frac{1}{2}=\frac{281-3}{6}\Rightarrow I_{inter}(P_1)=\frac{279}{6}=46.33$$

$$P_2=\{\{I_1,I_2\},\{I_3\},\{I_4,I_5\},\{I_6\}\}$$

$$I_{intra}(P_2)=v(\{I_1,I_2\})+v(\{I_3\})+v(\{I_4,I_5\})+v(\{I_6\})$$

$$I_{intra}(P_2)=1+0+\frac{1}{2}+0\Rightarrow I_{intra}(P_2)=\frac{3}{2}$$

$$I_{totale}(P_2)=I_{intra}(P_2)+I_{inter}(P_2)\Rightarrow I_{inter}(P_2)=I_{totale}(P_2)-I_{intra}(P_2)$$

$$I_{inter}(P_2)=\frac{281}{6}-\frac{3}{2}=\frac{281-9}{6}\Rightarrow I_{inter}(P_2)=\frac{272}{6}=45.33$$

$$P_3=\{\{I_1,I_2\},\{I_4,I_5\},\{I_3,I_6\}\}\rightarrow I_{intra}(P_3)=v(\{I_1,I_2\})+v(\{I_4,I_5\})+v(\{I_3\},\{I_6\})$$

$$I_{intra}(P_3)=1+\frac{1}{2}+2\Rightarrow I_{intra}(P_3)=\frac{7}{2}$$

$$I_{totale}(P_3)=I_{intra}(P_3)+I_{inter}(P_3)\Rightarrow I_{inter}(P_3)=I_{totale}(P_3)-I_{intra}(P_3)$$

$$I_{inter}(P_3)=\frac{281}{6}-\frac{7}{2}=\frac{281-21}{6}\Rightarrow I_{inter}(P_3)=\frac{260}{6}=43.33$$

$$P_4=\{\{I_1,I_2,I_4,I_5\},\{I_3,I_6\}\}$$

$$I_{intra}(P_4)=v(\{I_1,I_2,I_4,I_5\})+v(\{I_3\},\{I_6\})\rightarrow I_{intra}(P_4)=\frac{91}{4}+2=\frac{99}{4}$$

$$I_{totale}(P_4)=I_{intra}(P_4)+I_{inter}(P_4)\Rightarrow I_{inter}(P_4)=I_{totale}(P_4)-I_{intra}(P_4)$$

$$I_{inter}(P_4)=\frac{281}{6}-\frac{99}{4}=\frac{2*281-3*99}{12}\Rightarrow I_{inter}(P_4)=\frac{265}{12}=22.08$$

$$P_5=\{\{I_1,I_2,I_3,I_4,I_5,I_6\}\}\rightarrow I_{intra}(P_5)=\frac{281}{6}\Rightarrow I_{inter}(P_5)=0$$

4. La classe la plus excentrique de la partition en trois classes

$$P_3=\{\{I_1,I_2\},\{I_4,I_5\},\{I_3,I_6\}\};g=\begin{pmatrix}\frac{1+2+6+2+3+6}{6}\\ \frac{1+2+2+6+6+6+4}{6}\end{pmatrix}\Rightarrow g=\begin{pmatrix}\frac{10}{3}\\ \frac{7}{2}\end{pmatrix}$$

$$C_1=\{I_1,I_2\}\Rightarrow g_{C_1}=\begin{pmatrix}\frac{1+2}{2}\\ \frac{1+2}{2}\end{pmatrix}\Rightarrow g_{C_1}=\begin{pmatrix}\frac{3}{2}\\ \frac{3}{2}\end{pmatrix}$$

- **L'excentricité de la classe** C_1

$$\rho^2(C_1)=d^2(g_{C_1},g)=\left(\frac{3}{2}-\frac{10}{3}\right)^2+\left(\frac{3}{2}-\frac{7}{2}\right)^2\Rightarrow\rho^2(C_1)=\frac{121}{36}+4=\frac{265}{36}=7.36$$

$$C_1=\{I_1,I_2\}\Rightarrow g_{C_1}=\begin{pmatrix}\frac{1+2}{2}\\ \frac{1+2}{2}\end{pmatrix}\Rightarrow g_{C_1}=\begin{pmatrix}\frac{3}{2}\\ \frac{3}{2}\end{pmatrix}$$

- **L'excentricité de la classe** C_2

$$C_2=\{I_4,I_5\}\Rightarrow g_{C_2}=\begin{pmatrix}\frac{2+3}{2}\\ \frac{6+6}{2}\end{pmatrix}\Rightarrow g_{C_2}=\begin{pmatrix}\frac{5}{2}\\ 6\end{pmatrix}$$

$$\rho^2(C_2)=d^2(g_{C_2},g)=\left(\frac{5}{2}-\frac{10}{3}\right)^2+\left(6-\frac{7}{2}\right)^2\Rightarrow\rho^2(C_2)=\frac{25}{36}+\frac{25}{4}=\frac{250}{36}=6.94$$

- **L'excentricité de la classe** C_3

$$C_3=\{I_3,I_6\}\Rightarrow g_{C_2}=\begin{pmatrix}\frac{6+6}{2}\\ \frac{2+4}{2}\end{pmatrix}\Rightarrow g_{C_2}=\begin{pmatrix}6\\ 3\end{pmatrix}$$

$$\rho^2(C_3)=d^2(g_{C_3},g)=\left(6-\frac{10}{3}\right)^2+\left(3-\frac{7}{2}\right)^2\Rightarrow\rho^2(C_3)=\frac{64}{9}+\frac{1}{4}=\frac{265}{36}=7.36$$

$\left(\rho^2(C_1)=\rho^2(C_3)\right)\succ\rho^2(C_2)\Rightarrow C_3$ ou C_1 est la classe la plus excentrique.

- **Contribution de chacun des deux axes à son excentricité**

$$Cont(X)=\left(x_{g_{C_3}}-x_g\right)^2/\rho^2(C_3)=\left(6-\frac{10}{3}\right)^2\Big/\left(6-\frac{10}{3}\right)^2+\left(3-\frac{7}{2}\right)^2=\frac{256}{265}\Rightarrow Cont(X)\cong 0.97$$

$$Cont(Y)=\left(y_{g_{C_3}}-y_g\right)^2/\rho^2(C_3)=\left(3-\frac{7}{2}\right)^2\Big/\left(6-\frac{10}{3}\right)^2+\left(3-\frac{7}{2}\right)^2=\frac{9}{265}\Rightarrow Cont(Y)\cong 0.03$$

- **Remarque : Dans** le cas où nous utilisons le critère de la minimisation d'inertie de la réunion de deux classes, la valeur du dernier sommet indique l'inertie totale tandis que les indices d'agrégation de chaque nœuds représentent l'inertie intra-classes.

3.3 Définition d'une classification

Une classification est un partitionnement de N individus en K classes. Ce processus est déterminé par la minimisation d'inertie intra-classes [11]. Les algorithmes de classification cherchent à répartir les données en K clusters $(C_1, C_2, \ldots, C_K)$ avec :

$$\begin{cases} \cup_{C_k} = E \\ \forall i \neq j, C_k \cap C_k = \varnothing \end{cases}$$

3.4 La notion d'inertie intra-classes et inter-classes

Etant donné une partition en n K classes d'un nuage de N points , on définira les quantités suivantes : g_{c_k} centre de gravité de la classe C_k, g représente le centre de gravité du nuage de points et I_{tot} représente l'inertie totale qui est égale à la somme des carrés des distances des individus aux centres de gravité g, décrite par l'équation suivante [12] :

$$I_{tot} = \sum_{x_i \in c_k} P_i.d^2(x_i, g)$$

- **L'inertie inter-classes :** cette mesure représente la somme des carrés des distances des centres g_{c_k} de classe C_k au centre de gravité global g, L'inertie inter-classes est égale à :

  $I_{inter} = \sum_{k=1}^{K} P_{c_k}.d^2(g_{c_k}, g)$ avec g représente le centre de gravité du nuage de points.

- **L'inertie intra-classes :** cette mesure représente la somme des carrés des distances des individus au centre de gravité g_{c_k} de chaque classe C_k, L'inertie intra-classes est égale à :

  $I_{intra} = \sum_{k=1}^{K} \sum_{x_i \in c_k} P_i.d^2(x_i, g_{c_k})$ avec g_{c_k} représente le centre de gravité de la classe C_k.

L'inertie totale I_{tot} des N points autour du centre de gravité global g est alors égal à la somme des deux termes suivants : inertie intra-classes et inertie inter-classes. I_{tot} est déterminée à l'aide de théorème de König-Huyghens et qui peut être formulée par : $I_{tot} = I_{intra} + I_{inter}$

Un critère usuel de la classification consiste à chercher la partition telle que l'inertie intra-classes I_{intra} soit minimale pour avoir des classes bien homogènes, ce qui revient à chercher L'inertie inter-classes maximale.

Remarque :

$$I_{tot} = I_{intra} \searrow + I_{inter} \nearrow \Rightarrow \quad I_{tot} = \sum_{x_i \in c_k} P_i.d^2(x_i, g) = \sum_{k=1}^{K} \sum_{x_i \in c_k} P_i.d^2(x_i, g_{c_k}) \searrow + \sum_{k=1}^{K} P_{c_k}.d^2(g_{c_k}, g) \nearrow$$

Exemple : On suppose qu'on a obtenu les trois partitions suivantes : $P_1 = \left\{ \overbrace{\{I_1, I_2\}}^{C_1}, \overbrace{\{I_3, I_4\}}^{C_2}, \overbrace{\{I_5\}}^{C_3} \right\}$,

$P_2 = \left\{ \overbrace{\{I_1, I_5\}}^{C_1}, \overbrace{\{I_2, I_4\}}^{C_2}, \overbrace{\{I_3\}}^{C_3} \right\}$ et $P_3 = \left\{ \overbrace{\{I_1, I_3\}}^{C_1}, \overbrace{\{I_2, I_4\}}^{C_2}, \overbrace{\{I_5\}}^{C_3} \right\}$ à partir du tableau de données suivant :

	X	Y
I_1	0	0
I_2	1	0
I_3	5	5
I_4	4	5
I_5	10	10

Calcul d'inertie intra-classes de chaque partition :

Partition 1 :

$$g_{C_1} = \begin{pmatrix} \frac{0+1}{2} \\ \frac{0+0}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}, g_{C_2} = \begin{pmatrix} \frac{5+4}{2} \\ \frac{5+5}{2} \end{pmatrix} = \begin{pmatrix} \frac{9}{2} \\ 5 \end{pmatrix}, g_{C_3} = \begin{pmatrix} 10 \\ 10 \end{pmatrix}$$

$$d^2(I, g_{c_k}) = \left[\sqrt{(x_I - x_{g_{c_k}})^2 + (y_I - y_{g_{c_k}})^2} \right]^2 = (x_I - x_{g_{c_k}})^2 + (y_I - y_{g_{c_k}})^2$$

$$I_{intra}(P_1) = \sum_{I \in c_k} P_i . d^2(I, g_{c_k})$$

$$= P_1 . d^2(I_1, g_{c_1}) + P_2 . d^2(I_2, g_{c_1}) + P_3 . d^2(I_3, g_{c_2}) + P_4 . d^2(I_4, g_{c_2}) + P_4 . d^2(I_4, g_{c_2}) + P_5 . d^2\left(I_5, g_{c_3}\right)$$

$$= 1$$

$$I_{intra}(P_1) = 1 \ (1^{ère} \ partition \ P_1)$$

Partition 2 : $P_2 = \left\{ \overbrace{\{I_1, I_5\}}^{C_1}, \overbrace{\{I_2, I_4\}}^{C_2}, \overbrace{\{I_3\}}^{C_3} \right\}$ alors $g_{C_1} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, g_{C_2} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{2} \end{pmatrix}, g_{C_3} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$

$$I_{intra}(P_2) = \sum_{I \in c_k} P_i . d^2(I, g_{c_k})$$

$$= P_1 * d^2(I_1, g_{c_1}) + P_5 * d^2(I_5, g_{c_1}) + P_2 * d^2(I_2, g_{c_2}) + P_4 * d^2(I_4, g_{c_2}) + P_3 * d^2\left(I_3, g_{c_3}\right)$$

$$= 117$$

$$I_{intra}(P_2) = 117 \ (P_1 \ est \ mieulleure \ que \ P_2, car \ I_{intra}(P_1) < I_{intra}(P_2))$$

Partition 3 : $P_3 = \left\{ \overbrace{\{I_1, I_3\}}^{C_1}, \overbrace{\{I_2, I_4\}}^{C_2}, \overbrace{\{I_5\}}^{C_3} \right\}$ P3 $g_{C_1} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{2} \end{pmatrix}, g_{C_2} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{2} \end{pmatrix}, g_{C_3} = \begin{pmatrix} 10 \\ 10 \end{pmatrix}$

$$I_{intra}(P_3) = \sum_{I \in c_k} P_i . d^2(I, g_{c_k})$$

$$= P_1 * d^2(I_1, g_{c_1}) + P_5 * d^2(I_3, g_{c_1}) + P_2 * d^2(I_2, g_{c_2}) + P_4 * d^2(I_4, g_{c_2}) + P_5 * d^2\left(I_5, g_{c_3}\right)$$

$$= 42$$

$I_{intra}(P_1) < I_{intra}(P_3) < I_{intra}(P_2)$ *alors la meilleur partition est* P_1

Calcul d'inertie inter-classes de chaque partition :

Partition 1 :

$$g_{C_1} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}, g_{C_2} = \begin{pmatrix} \frac{9}{2} \\ 5 \end{pmatrix}, g_{C_3} = \begin{pmatrix} 10 \\ 10 \end{pmatrix}, g\begin{pmatrix} 4 \\ 4 \end{pmatrix} \textit{centre de gravité du nuage}$$

$$I_{inter}(P_1) = \sum_{k=1}^{K=3} P_{C_k} d^2(g_{C_k}, g)$$

$$= P_{C_1} . d^2(g_{C_1}, g) + P_{C_2} . d^2(g_{C_2}, g) + P_{C_3} . d^2(g_{C_3}, g)$$

$$= 2\left[\frac{49}{4} + 16\right] + 2\left[\frac{1}{4} + 1\right] + 1[36 + 36]$$

$$= 131$$

Partition 2 :

$$g_{C_1} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, g_{C_2} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{2} \end{pmatrix}, g_{C_3} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, g\begin{pmatrix} 4 \\ 4 \end{pmatrix} \textit{centre de gravité du nuage}$$

$$I_{inter}(P_2) = \sum_{k=1}^{K=3} P_{C_k} d^2(g_{C_k}, g)$$

$$= P_{C_1} . d^2(g_{C_1}, g) + P_{C_2} . d^2(g_{C_2}, g) + P_{C_3} . d^2(g_{C_3}, g)$$

$$= 2[1+1] + 2\left[\frac{9}{4} + \frac{9}{4}\right] + 1[1+1]$$

$$= 15$$

Partition 3 :

$$g_{C_1} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{2} \end{pmatrix}, g_{C_2} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{2} \end{pmatrix}, g_{C_3} = \begin{pmatrix} 10 \\ 10 \end{pmatrix}, g \begin{pmatrix} 4 \\ 4 \end{pmatrix} centre \text{ de } g ravité \text{ du } nuage$$

$$I_{inter}(P_3) = \sum_{k=1}^{K=3} P_{C_k} d^2(g_{C_k}, g)$$

$$= P_{C_1}.d^2(g_{C_1}, g) + P_{C_2}.d^2(g_{C_2}, g) + P_{C_3}.d^2(g_{C_3}, g)$$

$$= 2\left[\frac{9}{4} + \frac{9}{4}\right] + 2\left[\frac{9}{4} + \frac{9}{4}\right] + 1[36 + 36]$$

$$= 90$$

$$I_{inter}(P_1) > I_{inter}(P_3) > I_{inter}(P_2) \text{ alors la meilleur partition est } P_1$$

$$I_{tot} = I_{intra} + I_{inter} \Rightarrow I_{tot} = 1 + 131 = 117 + 15 = 42 + 90 = 132$$

Remarque : On peut calculer l'inertie totale en utilisant la formule suivante :

$$I_{tot} = \sum_{i=1}^{N} P_i.d^2(I_i, g) = d^2(I_1, g) + d^2(I_2, g) + d^2(I_3, g) + d^2(I_4, g) + d^2(I_5, g)$$

$$\Rightarrow = (16 + 16) + (9 + 16) + (1 + 1) + (0 + 1) + (72) = 132$$

3.5Classification par partitionnement

3.5.1 Algorithme des centres mobiles (ACM)

- Principe :

Il s'agit de déterminer une partition de l'ensemble I des individus en K classes C_k.

K étant fixé a priori. Cette tache nécessite de calculer la distance euclidienne entre les individus et les centroides (noyaux) de chaque classe, puis, on affecte l'individu à la classe la plus proche. Dans ACM, deux fonctions sont appliquées [13]:

- La fonction d'affectation : chaque individu i est affecté à la classe C_k dont il est le plus proche au sens de la distance. Cette fonction est formulée mathématiquement par : $F(L_k) = C_k = \{x \in \Omega / d(x, L_i) \leq d(x, L_j), \forall i \neq j\}$
- La fonction de représentation : elle s'agit de déterminer l'ensemble L des K noyaux optimisant le critère W. Pour cela, pour toute classe C_k, il suffit de chercher le noyau L_k qui minimise la quantité $\sum_{i \in c_k} d^2(x_i, L_k)$, donc $L_k = g(C_k)$ avec L_k *représente centre de gravité de la classe* C_k

Algorithme ACM est un algorithme itératif de type non-supervisé qui cherche à trouver une partition $P = \{C_1, C_2, \cdots, C_K\}$ de l'ensemble I et un ensemble de K noyaux : $L = \{L_1, L_2, \cdots, L_K\}$ en minimisant le critère d'inertie intra-classes : $W = \sum_{k=1}^{K} D(C_k, L_k) = \sum_{k=1}^{K} \sum_{x \in C_i} m_x d^2(x, L_k)$

Les étapes de l'algorithme ACM sont décrites dans la table 3.2.

Algorithme des centres mobiles (ACM)
Input : les centres de départ
5- Définir aléatoirement les noyaux ou tirer au hasard certains centres de l'espace initial des individus $(L_1, L_2, \cdots, L_k)$
6- Appliquer la fonction d'affectation : $F(L_k) = C_k = \{x \in \Omega / d(x, L_k) \leq d(x, L_j), \forall k \neq j\}$
7- Appliquer la fonction de représentation $L_k = g(C_k)$; L_k: *centre de gravité de la classe* C_k
8- Répéter les étapes 2 et 3 tant que l'inertie intra-classes diminue
Output : La partition

Table 3.2 : Algorithmes des centres mobiles

3.5.2 Avantages et inconvénients de l'algorithme ACM

Comme avantage principal, cet algorithme est facile à implémenter et qui nécessite uniquement un seul paramètre d'entrée (k : le nombre de classes). On note aussi, que l'ACM est très utilisé dans l'apprentissage non-supervisé dans le cadre de traitement d'image satellitaire. Néanmoins, certains inconvénients sont soulignés comme la convergence vers des minimas locaux et cela est due à l'initialisation aléatoire des centres.

Un autre inconvénient majeur est sa limite vis-à-vis les classes sphériques. Pour résoudre ce problème, il est nécessaire d'utiliser des métriques adaptatives au niveau de la fonction d'affectation.

- **Exemple1 :**

Nous voulons classer les six points suivants en trois classes : (poids d'un individu= 1) :

	I1	I2	I3	I4	I5	I6
X	1	2	3	4	6	8
Y	2	5	2	7	7	2

Nous utiliserons l'algorithme des centres mobiles.

1. Donner la partition (P_1) obtenue pour les points de départ :

$$L_1 = \binom{2}{2}, \quad L_2 = \binom{5}{6}, \quad L_3 = \binom{10}{2}$$

2. Donner la partition (P_2) obtenue pour les points de départ :

$$L_1 = \binom{1}{2}, \quad L_2 = \binom{4}{4}, \quad L_3 = \binom{7}{2}$$

3. Laquelle des deux partitions est la meilleure ? Justifiez votre réponse.

Solution :

- **La partition $P1$:**

$$L_1 = \binom{2}{2}, \quad L_2 = \binom{5}{6}, \quad L_3 = \binom{10}{2}$$

- **Fonction d'affectation :** $F(L_k) = C_k = \{x \in \Omega / d(x, L_k) \leq d(x, L_j), \forall k \neq j\}$

d^2	L_1	L_2	L_3
I1	**1**	32	81
I2	**9**	10	73
I3	**1**	20	49
I4	29	**2**	61
I5	41	**2**	41
I6	64	25	**4**

$$C_1 = \{I1, I2, I3\}$$
$$C_2 = \{I4, I5\}$$
$$C_3 = \{I6\}$$

$W = \sum_{k=1}^{K} D(C_k, L_k) = \sum_{k=1}^{K} \sum_{x \in C_k} m_x d^2(x, L_k) = 1 + 9 + 1 + 2 + 2 + 1 = 19$

- **Fonction de représentation :** $L_k = g(C_k)$; L_k: *centre de gravité de la classe* C_k .

$$L_1 = \begin{pmatrix} \frac{1+2+3}{3} \\ \frac{2+5+2}{3} \end{pmatrix} ; L_2 = \begin{pmatrix} \frac{4+6}{2} \\ \frac{7+7}{2} \end{pmatrix} ; L_3 = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$$

$$L_1 = \begin{pmatrix} 2 \\ 3 \end{pmatrix} ; L_2 = \begin{pmatrix} 5 \\ 7 \end{pmatrix} ; L_3 = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$$

- **Fonction d'affectation :** $F(L_k) = C_k = \{x \in \Omega / d(x, L_k) \leq d(x, L_j), \forall k \neq j\}$

d^2	L_1	L_2	L_3
I1	**2**	41	49
I2	**4**	13	45
I3	**2**	29	25
I4	20	**1**	41
I5	32	**1**	29
I6	37	34	**0**

$$C_1 = \{I1, I2, I3\}$$

$$C_2 = \{I4, I5\}$$

$$C_3 = \{I6\}$$

$W = \sum_{k=1}^{K} D(C_k, L_k) = \sum_{i=1}^{K} \sum_{x \in C_k} m_x d^2(x, L_k) = 2 + 4 + 2 + 1 + 1 + 0 = 10$

Stabilité des résultats implique arrêt

La partition obtenue est : $P_1 = \{\{I_1, I_2, I_3\}, \{I_4, I_5\}, \{I_6\}\}$

- Partition **P2 :**

$$L_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix} ; L_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix} ; L_3 = \begin{pmatrix} 7 \\ 2 \end{pmatrix}$$

- **Fonction d'affectation :** $F(L_k) = C_k = \{x \in \Omega / d(x, L_k) \leq d(x, L_j), \forall k \neq j\}$

d^2	L_1	L_2	L_3
I1	**0**	13	36
I2	10	**5**	34
I3	4	5	16
I4	34	**9**	34
I5	50	**13**	26
I6	49	20	**1**

$$C_1 = \{I1, I3\}$$

$$C_2 = \{I2, I4, I5\}$$

$$C_3 = \{I6\}$$

$W = \sum_{k=1}^{K} D(C_k, L_k) = \sum_{k=1}^{K} \sum_{x \in C_k} m_x d^2(x, L_k) = 0 + 4 + 5 + 9 + 13 + 1 = 32.$

- **Fonction de représentation :** $L_k = g(C_k)$; L_i: *centre de gravité de la classe* C_k .

$$L_1 = \begin{pmatrix} \frac{1+3}{2} \\ \frac{2+2}{2} \end{pmatrix} ; L_2 = \begin{pmatrix} \frac{2+4+6}{3} \\ \frac{5+7+7}{3} \end{pmatrix} ; L_3 = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$$

$$L_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} ; L_2 = \begin{pmatrix} 5 \\ \frac{19}{3} \end{pmatrix} ; L_3 = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$$

d^2	L_1	L_2	L_3
I1	**1**	27,77	49
I2	9	**5,77**	45
I3	**1**	19,77	25
I4	29	**0,44**	41
I5	41	**4,44**	29
I6	36	34,77	**0**

$$C_1 = \{I1, I3\}$$
$$C_2 = \{I2, I4, I5\}$$
$$C_3 = \{I6\}$$

$W = \sum_{k=1}^{K} D(C_k, L_k) = \sum_{k=1}^{K} \sum_{x \in C_k} m_x d^2(x, L_k) = 1 + 1 + 5{,}77 + 0{,}44 + 4{,}44 + 0 = 12{,}65$

Stabilité des résultats implique arrêt

La partition obtenue est : $P_2 = \{\{I_1, I_3\}, \{I_2, I_4, I_5\}, \{I_6\}\}$

La meilleure partition est P1 car l'inertie intra-classe est minimale (W1<W2).

La forme la plus forte est la forme $C_3 = \{I6\}$.

- **Exemple n°2 :**

Soit l'ensemble des six individus suivants (poids d'un individu= 1) :

	I1	I2	I3	I4	I5	I6
X	0	1	6	1	2	6
Y	0	1	5	5	5	6

- Trouver une classification en trois classes, en utilisant l'algorithme des centres mobiles dans le cadre des nuées dynamiques avec les noyaux de départ suivant : $L_1 = \binom{0}{0}$, $L_2 = \binom{1}{5}$, $L_3 = \binom{20}{20}$

Solution :

- **Classification en trois classes en utilisant l'algorithme des centres mobiles**

$L_1 = \binom{0}{0}$, $L_2 = \binom{1}{5}$, $L_3 = \binom{20}{20}$

➢ **Fonction d'affectation :** $F(L_k) = C_k = \{x \in \Omega / d(x, L_k) \leq d(x, L_j), \forall k \neq j\}$

d^2	L_1	L_2	L_3
I1	**0**	26	800
I2	**2**	16	722
I3	61	**25**	421
I4	26	**0**	586
I5	29	**1**	549
I6	72	26	**392**

$$C_1 = \{I1, I2\}$$
$$C_2 = \{I3, I4, I5\}$$
$$C_3 = \{I6\}$$

$W = \sum_{k=1}^{K} D(C_k, L_k) = \sum_{k=1}^{K} \sum_{x \in C_k} m_x d^2(x, L_k) = 0 + 2 + 25 + 0 + 1 + 392 = 420$

➢ **Fonction de représentation :** $L_k = g(C_k)$; L_i: *centre de gravité de la classe* C_k .

➢ $L_1 = \begin{pmatrix} \frac{0+1}{2} \\ \frac{0+1}{2} \end{pmatrix}$; $L_2 = \begin{pmatrix} \frac{6+1+2}{3} \\ \frac{5+5+5}{3} \end{pmatrix}$; $L_3 = \begin{pmatrix} 6 \\ 6 \end{pmatrix}$

$L_1 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$; $L_2 = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$; $L_3 = \begin{pmatrix} 6 \\ 6 \end{pmatrix}$

➢ **Fonction d'affectation :** $F(L_k) = C_k = \{x \in \Omega / d(x, L_k) \leq d(x, L_j), \forall k \neq j\}$

d^2	L_1	L_2	L_3
I1	**0.5**	34	72
I2	**0.5**	20	50
I3	101/2	9	**1**
I4	82/2	**4**	26
I5	90/4	**1**	17
I6	121/2	10	**0**

$$C_1 = \{I1, I2\}$$
$$C_2 = \{I4, I5\}$$
$$C_3 = \{I3, I6\}$$

$$W = \sum_{k=1}^{K} D(C_k, L_k) = \sum_{k=1}^{K} \sum_{x \in C_k} m_x d^2(x, L_k) = 0.5 + 0.5 + 1 + 4 + 1 + 0 = 7$$

➢ **Fonction de représentation :** $L_k = g(C_k)$; L_i: *centre de gravité de la classe* C_k .

➢ $L_1 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$; $L_2 = \begin{pmatrix} 3/2 \\ 5 \end{pmatrix}$; $L_3 = \begin{pmatrix} 6 \\ 5.5 \end{pmatrix}$

➢ **Fonction d'affectation :** $F(L_k) = C_k = \{x \in \Omega / d(x, L_k) \leq d(x, L_j), \forall k \neq j\}$

d^2	L_1	L_2	L_3
I1	**0.5**	9/4+25	36+121/2
I2	**0.5**	¼+16	25+81/4
I3	101/2	81/6	**¼**
I4	82/2	**1/4**	25+1/4
I5	90/4	**¼**	16+1/4
I6	121/2	81/4	**¼**

$$C_1 = \{I1, I2\}$$

$$C_2 = \{I4, I5\} \quad C_3 = \{I3, I6\}$$

$$W = \sum_{k=1}^{K} D(C_k, L_k) = \sum_{k=1}^{K} \sum_{x \in C_k} m_x d^2(x, L_k) = 0.5 + 0.5 + 1$$

$$/4 + 1/4 + 1/4 + 1/4 = 2$$

Stabilité des résultats →arret

3.6 Méthode des nuées dynamiques

Algorithme des nuées dynamique est **une généralisation** de l'algorithme des centres mobiles où le noyau peut être représenté par une droite ou par une fonction de densité de probabilité qui permet de produire un nouveau algorithme appelé Maximum de vraisemblance [14].

3.6.1 Méthode du Maximum de Vraisemblance

3.6.1.1 Définition : soit $P(C_k)$ la probabilité de l'évènement : une classe C_k existe.

$P(C_k/X)$: Probabilité conditionnelle ; C_k contient X

$P(X/C_k)$: Probabilité conditionnelle ; $X \in C_k$

X : probabilité de l'évènement : X existe

f : fonction d'affectation $f(L)=P$.

g : fonction de représentation $g(P)=L$.

3.6.1.2 Règle de Bayes

$$P(X/C_k)=\frac{P(X\cap C_k)}{P(C_k)} \Rightarrow P(X\cap C_k)=P(X/C_k)*P(C_k) \rightarrow (1)$$

$$P(C_k/X)=\frac{P(C_k\cap X)}{P(X)} \Rightarrow P(C_k\cap X)=P(C_k/X)*P(X) \rightarrow (2)$$

De $(1)\&(2) \Rightarrow P(X/C_k)*P(C_k)=P(C_k/X)*P(X)$

$$\Rightarrow P(C_k/X)=\frac{P(X/C_k)*P(C_k)}{P(X)}$$

On peut définir une fonction d'affectation f : $X \in C_k$ si $\forall j \quad P(C_k/X) \succ P(C_j/X)$

$$\Rightarrow \frac{P(X/C_k)*P(C_k)}{P(X)} \succ \frac{P(X/C_j)*P(C_j)}{P(X)} \Rightarrow P(X/C_k)*P(C_k) \succ P(X/C_j)*P(C_j)$$

Remarque : Il est difficile d'estimer des probabilités à priori d'occurrence des classes en pratique on adopte l'hypothèse de classes équiprobables c.-à-d. : $\forall i \neq j \quad P(C_i)=P(C_j) \Rightarrow X \in C_k$ si $\forall j$ $P(X/C_k) \succ P(X/C_j)$

- **Cas d'une distribution Normale**

$$P(X/C_k)=\frac{1}{(2\pi)^{1/2}.(\det(V_k))^{1/2}}\exp\left[-\frac{1}{2}(X-\mu_k)^t.V_k^{-1}.(X-\mu_k)\right]$$

V_k : Matrice de variance- covariance

μ_k : Moyenne de la classe C_k

- **Simplification**

$$\ln(P(X/C_k)) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\det(V_k)) - \frac{1}{2}(X-\mu_k)^t.V_k^{-1}.(X-\mu_k)$$

$$\Rightarrow X \in C_k \text{ si } \forall j \quad P(X/C_k) \succ P(X/C_j) \Rightarrow \ln(P(X/C_k)) \succ \ln(P(X/C_j))$$

$$-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\det(V_k)) - \frac{1}{2}(X-\mu_k)^t.V_k^{-1}.(X-\mu_k) \succ -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\det(V_j)) - \frac{1}{2}(X-\mu_j)^t.V_j^{-1}.(X-\mu_j)$$

- **Fonction de représentation**

$$g(P) = L = (\mu_1, V_1; \mu_2, V_2; \mu_3, V_3; .....; \mu_i, V_i; ..... ; \mu_k, V_k)$$

3.6 Méthode non paramétrique

3.7.1 Classification par morphologie mathématique

3.7.1.1 Discrétisation

Soit $\{y_1, y_2, \ldots\ldots\ldots, y_n\}$ un échantillon de n n observation multidimensionnelle. $y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{ip} \end{pmatrix}$

1) L'origine est translatée au point : $O = \begin{pmatrix} \min y_{i1} \\ \min y_{i2} \\ \vdots \\ \min y_{id} \\ \vdots \\ \min y_{ip} \end{pmatrix}$

2) La transformation diagonale : $y'_{ij} = \dfrac{y_{ij} - \min\ y_{ij}}{Max\ y_{ij} - \min\ y_{ij}}.R, \quad avec\ R : resolution$

Cette transformation permet de situer les observations dans un hyper cube de coté R .

3) chaque Axe du nouvel espace est découpé en R intervalles adjacents égaux de longueur unité.

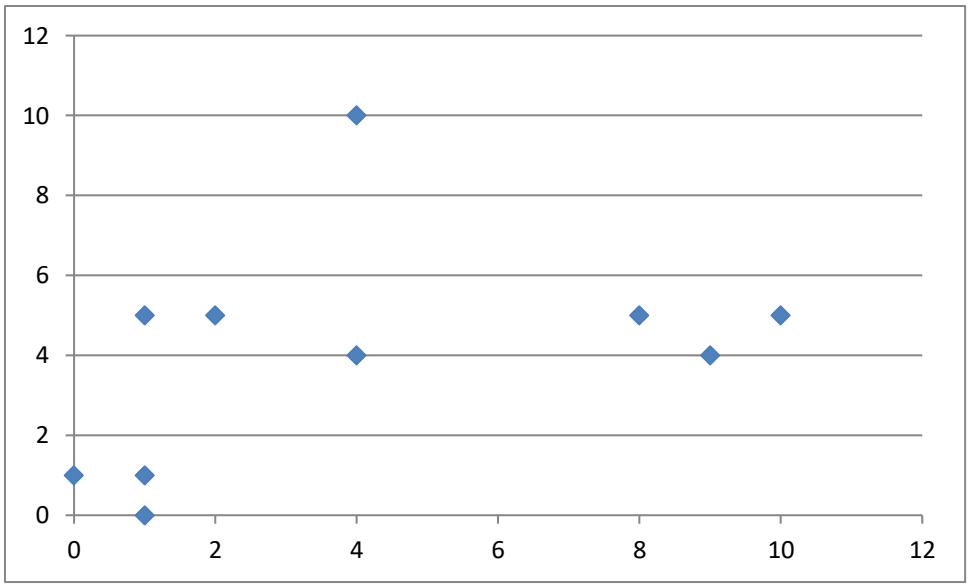
Cette discrétisation définit un ensemble d'hyper cube H de R de côté unité dont chacun est repérer par les p parties entiers des coordonnées de son centre.

Individu y_i est situer dans l'hyper cube de coordonnées.

$$H_i \begin{pmatrix} \text{int}(y'_{i1}) \\ \text{int}(y'_{i2}) \\ \vdots \\ \text{int}(y'_{id}) \\ \vdots \\ \text{int}(y'_{ip}) \end{pmatrix}, \quad \text{int} : partie\ entiere$$

Exemple : méthode morphologiques

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}
X	0	1	1	2	4	4	10	8	9	1
Y	1	5	1	5	4	10	5	5	4	0



$$\begin{cases} Min\ x = 0 \\ Max\ x = 10 \end{cases}, \qquad \begin{cases} Min\ y = 0 \\ Max\ y = 10 \end{cases}$$

R=10

$$I_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\begin{cases} X'_{I_1} = \dfrac{0-0}{10-0}.10 = 0 \\ Y'_{I_1} = \dfrac{1-0}{10-0}.10 = 1 \end{cases}, \quad \begin{cases} X'_i = \dfrac{X_i - \min X}{MaxX - MinX}.R \\ Y'_i = \dfrac{Y_i - \min Y}{MaxY - MinY}.R \end{cases}$$

Donc pour toute les autres transformations on aura Y'=Y X'=X

On trace l'ensemble d'hyper cube

Discrétisation : R=10

10						1					
9											
8											
7											
6											
5	0	1	1	0	0	0	0	1	0	1	0
4	0	0	0	0	1	0	0	0	1	0	0
3	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0
	0	1	2	3	4	5	6	7	8	9	10

$$X = \{(1,0),(1,1),(0,1),(4,4),(9,4),(1,5),(2,5),(8,5),(10,5),(4,10)\}$$

$$\text{R=5}\begin{cases} Min\ x = 0 \\ Max\ x = 10 \end{cases},\begin{cases} Min\ y = 0 \\ Max\ y = 0 \end{cases},\begin{cases} X'_i = \dfrac{X_i - \min X}{MaxX - MinX}.R = \dfrac{X-0}{10-0}.5 = \dfrac{X}{2} \\ Y'_i = \dfrac{Y_i - \min Y}{MaxY - MinY}.R = \dfrac{Y-0}{10-0}.5 = \dfrac{Y}{2} \end{cases}$$

Donc le tableau change :

X	0	0 .5	0.5	1	2	2	5	4	4.5	0.5
Y	0.5	2.5	0.5	2.5	2	5	2.5	2.5	2	0

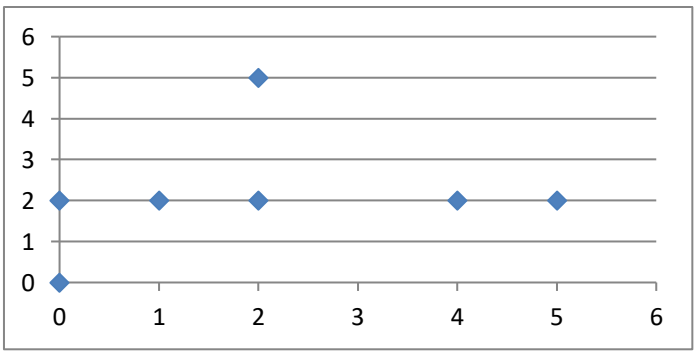
$$I_1\begin{pmatrix} 0 \\ 0.5 \end{pmatrix} = H\begin{pmatrix} \text{int}(0) \\ \text{int}(0.5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$I_4\begin{pmatrix} 1 \\ 2.5 \end{pmatrix} = H\begin{pmatrix} \text{int}(1) \\ \text{int}(2.5 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

X	0	0	0	1	2	2	5	4	4	0
Y	0	2	0	2	2	5	2	2	2	0

$$X = \{(0,0),(0,2),(1,2),(2,2),(4,2),(5,2),(2,5)\}$$

5						
4						
3						
2	**1**	**1**	**1**		**1**	**1**
1						
0	**1**					
	0	1	2	3	4	5



Remarque :

La valeur du paramètre R est décisif pour les résultats de la discrétisations en effet si R est trop grand Le nuage va être trop dispersé et la détection des classes devient difficile par contre si R est trop petit une confusion entre les différentes classes va

* le choix de la valeur la plus approprié de R est fait comme suivant :

1- choix d'un intervalle de variation de R

2- Application de la discrétisation et la détection des classes pour toutes les valeurs de R

3- traçage graphique nombre de classe détecté en fonction de R.

Finalement la valeur de R est choisie au centre de plus grand intervalle de stabilité du graphe.

3.7.1.2 Elément structurant :

Les transformations morphologiques consistent à comparer l'ensemble à analyser à un élément structurant à fin d'extraire ces caractéristiques structurale et morphologique l'élément structurant est un ensemble discret (image binaire) généralement plus petit que l'ensemble à analyser, il est défini par sa structure et un point de référence appelé l'origine.

Exemple :

0	1	0
1	1	1
0	1	0

L'origine

$$S = \{\overbrace{(0,0)}^{S_1}, \overbrace{(1,0)}^{S_2}, \overbrace{(0,1)}^{S_3}, \overbrace{(-1,0)}^{S_4}, \overbrace{(0,-1)}^{S_4}\}$$

3.8 Dilatation : $X \oplus S = \cup_{s \in S} (X)_s$

$X = \{(1,0),(1,1),(0,1),(4,4),(9,4),(1,5),(2,5),(8,5),(10,5),(4,10)\} \rightarrow E$

$(X)_{S_1} = X$: *ça change pas car* $S_1 = (0,0)$

$(X)_{S_2} = \{(2,0),(2,1),(1,1),(5,4),(10,4),(2,5),(3,5),(9,5),(11,5),(5,10)\}$

$(X)_{S_3} = \{(1,1),(1,2),(0,2),(4,5),(9,5),(1,6),(2,6),(8,6),(10,6),(4,11)\}$

$(X)_{S_4} = \{(0,0),(0,1),(-1,1),(3,4),(8,4),(0,5),(1,5),(7,5),(9,5),(3,10)\}$

$(X)_{S_5} = \{(1,-1),(1,0),(0,0),(4,3),(9,3),(1,4),(2,4),(8,4),(10,4),(4,9)\}$

$X \oplus S = \{X \cup (2,0),(2,1),(5,4),(10,4),(3,5),(9,5),\cancel{(11,5)},(5,10),$

$(0,2),(1,2),(4,5),(1,6),(2,6),(8,6),(10,6),\cancel{(4,11)},(0,0),\cancel{(-1,1)},$

$(3,4),(8,4),(0,5),(7,5),(3,10),(1,-1),(4,3),(9,3),(1,4),(2,4),(4,9)\}$

Sur l'expression ***E*** on ajoute $X \oplus S$ on marque les autres points pas encore marque

$(2,0),(2,1),(5,4)....(4,0)$ }

$(11,5),(4,11),(-1,1),(1,-1)$: On ne les met pas car ça dépasse le graphe.

- **Propriétés :**

1- Opération locale.

2- Opération croissante : $X \subset Y \Rightarrow X \oplus S \subset Y \oplus S$

3- Opération extensive : $X \subset X \oplus S$

4- Distributive : $(X \cup Y) \oplus S = (X \oplus S) \cup (Y \oplus S)$

5- Itérative : $(X \oplus S) \oplus S' = X \oplus (S \oplus S')$

3.9 Erosion : $X \Theta S = \cap_{s \in S} (X)_{-S}, ou = \cap_{s \in S} (X)_S$

Exemple précédent (déclaration)

10						1					
9											
8											
7											
6											
5	0	1	1	0	0	0	0	1	0	1	0
4	0	0	0	0	1	0	0	0	1	0	0
3	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0
	0	1	2	3	4	5	6	7	8	9	10

0	1	0
1	1	1
0	1	0

Élément structurant

$$X=\{(0,0),(0,2),(1,2),(2,2),(4,2),(5,2),(2,5)\} \quad S=\left\{\overset{S_1}{(0,0)},\overset{S_2}{(1,0)},\overset{S_3}{(0,1)},\overbrace{(-1,0)}^{S_4},\overbrace{(0,-1)}^{S_4}\right\}$$

$(X)_{S_1}=X$: *ça change pas car* $S_1=(0,0)$

$(X)_{S_2}=\{(0,0),(0,1),(-1,1),(0,5),(1,5),(3,4),(7,5),(8,4),(9,5),(10,3)\}$

$(X)_{S_3}=\{(1,-1),(1,0),(0,0),(1,4),(2,4),(4,3),(8,4),(9,3),(10,4),(4,9)\}$

$(X)_{S_4}=\{(2,6),(2,1),(1,1),(2,5),(3,5),(5,4),(9,5),(10,4),(11,5),(5,10)\}$

$(X)_{S_5}=\{(1,1),(1,2),(0,2),(1,6),(2,6),(4,5),(8,6),(9,5),(10,6),(9,11)\}$

$X\Theta S=\{\ \}=\varnothing$

Dilatation permet de grandir l'ensemble par contre l'érosion réduit

- **Dilatation :** On met les valeurs de tableau à la structure donnée

- **Erosion :** on cherche la structure sur le tableau

3.10 Propriété d'érosion

-opération locale croissante : $X\subset Y\Rightarrow(X\Theta S)\subset(Y\Theta S)$

-opération anti-extensive : $X\Theta S\subset X$

-relation de distributivité : $(X\cap Y)\Theta S=(X\Theta S)\cap(Y\Theta S)$

-itération : $(X\Theta S)\Theta S_2=X\Theta(S\Theta S_2)$

Remarque :

1- la solution consiste à copié la structure de l'élément structurant dans l'ensemble discret X

2- l'érosion consiste à rechercher la structure de l'élément structurant dans X Ceci signifie plus l'élément structurant est grand plus l'effet de filtrage est important

3.11 Ouverture : $X\circ S=(X\Theta S)\oplus S$

3.12 Fermeture : $X\bullet S=(X\oplus S)\Theta S$

Idempotente (change pas après plusieurs application)

$X\circ S=(X\circ S)\circ S$ *et* $X\bullet S=(X\bullet S)\bullet S$

Remarque : une 1ère méthode de classification (détection de région modale) par la morphologie mathématique fut l'opération ouverture

- le nombre de région connexe du résultat est le nombre de classe recherché

10											
9											
8											
7											
6											
5	0	1	1	0	0	1	0	0	0	0	0
4	0	0	0	0	0	0	0	0	**1**	0	0
3	0	0	0	0	0	0	0	0	0	**1**	0
2	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
	0	1	2	3	4	5	6	7	8	9	10

On garde que les 1 qui vérifient la structure de l'élément structurant

0	1	0
1	1	1
0	1	0

D'après le tableau initial on a

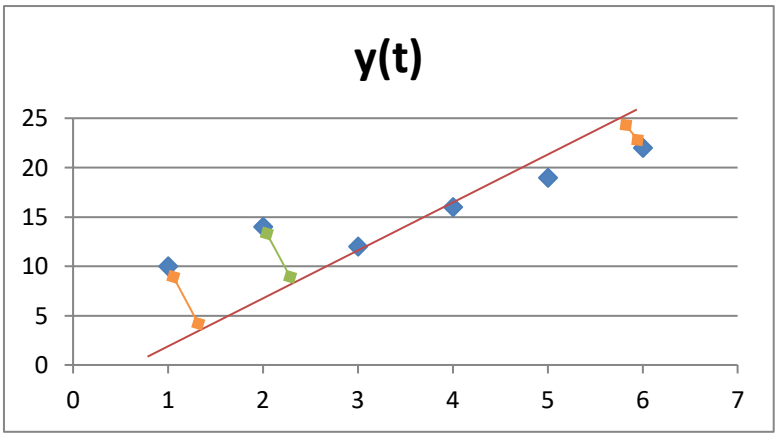
$C = \{I_8, I_9\}, \{I_5\}, \{I_2, I_4\}, \{I_3\}, \{I_1, I_6, I_7, I_{10}\}$

$\Rightarrow$ *représente des classes*

Les 1 montrés en gras (8,5),(9,4) sont $\{I_8, I_9\}$ d'après le tableau.

RÉGRESSION & CORRÉLATION

4.1 Méthode des moindres carrés



4.1.1 Le but de la méthode : la méthode des moindres carrés sert à ajuster les points c.à.d. trouver la courbe qui représente mieux les données [15]. D'une façon mathématique, on cherche à minimiser le critère suivant : $\sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (y_{cal} - y_i)^2 \rightarrow Min$

Dans notre cours, on se limitera à une représentation polynomiale tel que $y = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m$.

On considère que X , Y sont mesurés donc les inconnus sont les paramètres $a_0, a_1, a_2, ..., a_m$

4.1.2 Principe :

$$\begin{cases} a_0 + a_1 x_1 + a_2 x_1^2 + ... + a_m x_1^M - y_1 = d_1 \\ a_0 + a_1 x_2 + a_2 x_2^2 + ... + a_m x_2^M - y_2 = d_2 \\ \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + ... + a_m x_n^M - y_n = d_n \end{cases}$$

$$\sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (y_{cal} - y_i)^2 = \sum_{i=1}^{N} \left(a_0 + a_1 x_i + a_2 x_i^2 + ... + a_m x_i^M - y_i\right)^2 = F(a_0, a_1, a_2, ..., a_m) \rightarrow Min$$

$$\begin{cases} \frac{\partial F}{\partial a_0} = 2\sum_{i=1}^{N} \left(a_0 + a_1 x_i + a_2 x_i^2 + ... + a_m x_i^M - y_i\right) = 0 \\ \frac{\partial F}{\partial a_1} = 2\sum_{i=1}^{N} \left(a_0 + a_1 x_i + a_2 x_i^2 + ... + a_m x_i^M - y_i\right) x_i = 0 \\ \frac{\partial F}{\partial a_2} = 2\sum_{i=1}^{N} \left(a_0 + a_1 x_i + a_2 x_i^2 + ... + a_m x_i^M - y_i\right) x_i^2 = 0 \\ \frac{\partial F}{\partial a_m} = 2\sum_{i=1}^{N} \left(a_0 + a_1 x_i + a_2 x_i^2 + ... + a_m x_i^M - y_i\right) x_i^M = 0 \end{cases} \Rightarrow \begin{cases} n.a_0 + a_1 \sum_{i=1}^{N} x_i + a_2 \sum_{i=1}^{N} x_i^2 + ... + a_m \sum_{i=1}^{N} x_i^M = \sum_{i=1}^{N} y_i \\ a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 + a_2 \sum_{i=1}^{N} x_i^3 + ... + a_m \sum_{i=1}^{N} x_i^{M+1} = \sum_{i=1}^{N} x_i . y_i \\ a_0 \sum_{i=1}^{N} x_i^2 + a_1 \sum_{i=1}^{N} x_i^3 + a_2 \sum_{i=1}^{N} x_i^4 + ... + a_m \sum_{i=1}^{N} x_i^{M+2} = \sum_{i=1}^{N} x_i^2 . y_i \\ a_0 \sum_{i=1}^{N} x_i^M + a_1 \sum_{i=1}^{N} x_i^{M+1} + a_2 \sum_{i=1}^{N} x_i^{M+2} + ... + a_m \sum_{i=1}^{N} x_i^{2M} = \sum_{i=1}^{N} x_i^M . y_i \end{cases}$$

Dans le cas d'une droite : $y = a_0 + a_1 x$

Le système devient : $\begin{cases} n.a_0 + a_1 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \\ a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i.y_i \end{cases} \Rightarrow \begin{cases} a_0 = \bar{y} - a_1 \bar{x} \\ a_1 = \dfrac{Cov(x,y)}{Var(x)} \end{cases}$

4.1.3 Démonstration :

$$\begin{cases} n.a_0 + a_1 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \\ a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i.y_i \end{cases} \Rightarrow Eq(1): a_0 = \frac{\sum_{i=1}^{N} y_i}{n} - a_1 \frac{\sum_{i=1}^{N} x_i}{n} \Rightarrow a_0 = \bar{y} - a_1 \bar{x}$$

$$(\bar{y} - a_1 \bar{x}) \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i.y_i \Rightarrow \frac{(\bar{y} - a_1 \bar{x}) \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2}{n} = \frac{\sum_{i=1}^{N} x_i.y_i}{n} \Rightarrow (\bar{y} - a_1 \bar{x}) \bar{x} + a_1 \overline{x^2} = \overline{xy}$$

$$(\bar{y} - a_1 \bar{x}) \bar{x} + a_1 \overline{x^2} = \overline{xy} \Rightarrow \bar{y}.\bar{x} - a_1 . \left[(\bar{x})^2 - \overline{x^2} \right] = \overline{xy} \Rightarrow a_1 = \frac{\overline{xy} - \bar{x}.\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

$$Cov(x,y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N} \sum_{i=1}^{N} x_i.y_i - x_i.\bar{y} - \bar{x}.y_i + \bar{x}.\bar{y} = \frac{1}{N} \sum_{i=1}^{N} x_i.y_i - \frac{1}{N} \sum_{i=1}^{N} x_i.\bar{y} - \frac{1}{N} \sum_{i=1}^{N} \bar{x}.y_i + \frac{1}{N} \sum_{i=1}^{N} \bar{x}.\bar{y}$$

$$Cov(x,y) = \overline{xy} - \bar{y}.\bar{x} - \bar{x}.\bar{y} + \bar{x}.\bar{y} = \overline{xy} - \bar{y}.\bar{x}$$

$$Var(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - x_i.\bar{x} - \bar{x}.x_i + (\bar{x})^2 = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \frac{1}{N} \sum_{i=1}^{N} x_i.\bar{x} - \frac{1}{N} \sum_{i=1}^{N} \bar{x}.x_i + \frac{1}{N} \sum_{i=1}^{N} (\bar{x})^2$$

$$Var(x) = \overline{x^2} - (\bar{x})^2 - (\bar{x})^2 + (\bar{x})^2 = \overline{x^2} - (\bar{x})^2$$

4.2 Le coefficient de corrélation :

$-1 \le r \le 1$: nous informe s'il y a ou il n'y a pas de relation fonctionnelle entre x et y .

$$r = \frac{Cov(x,y)}{\sigma(x) * \sigma(y)} = \frac{\frac{1}{N} \sum_{i=1}^{N} (x - \bar{x})(y - \bar{y})}{\sqrt{\frac{1}{N} (x - \bar{x})^2} . \sqrt{\frac{1}{N} (y - \bar{y})^2}} = \frac{\overline{xy} - \bar{x}.\bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} * \sqrt{\overline{y^2} - (\bar{y})^2}} \rightarrow droite: y = ax + b$$

4.3 Le coefficient de corrélation généralisé:

$$Si: m \geq 2 \Rightarrow r^2 = \frac{\sum_{i=1}^{N}\left(y_{cal} - \bar{y}\right)}{\sum_{i=1}^{N}\left(y - \bar{y}\right)}$$

4.4 Test de corrélation linéaire : nous appliquons les statistiques.

Le critère suivant : $C = \frac{r^2.(n-2)}{1-r^2}$ qui suit la loi de Fisher : F avec $\nu_1 = 1$ et $\nu_2 = n-2$ degré de liberté.

4.5 Règle de décision :

Si $C \prec F_{\nu_1,\nu_2}$ alors on ne rejette pas l'hypothèse d'indépendance entre Y et X .

- Exemple :

On considère un échantillon dont le nombre de points $n = 4$ et le coefficient de corrélation $r = 0{,}98$.

$$n = 4 \text{ et } r = 0{,}98 \Rightarrow C = \frac{(0{,}98)^2.(4-2)}{1-(0{,}98)^2} = 48{,}5$$

$\nu_1 = 1$ et $\nu_2 = 2 : F_{\nu_1,\nu_2} = 98{,}49$ avec $\alpha = 0{,}01$

$C \prec F_{\nu_1,\nu_2}$ alors on ne rejette pas l'hypothèse d'indépendance entre Y et X

4.6 Echantillonnage de la régression et la corrélation :

a) Régression : b représente le coefficient de régression de l'échantillon et β représente le coefficient de régression de la population.

Test : Hypothèse : « β valeur approchée »

On utilise la statistique : $t = \frac{\beta - b}{S_{yx}/\sigma_x}\sqrt{n-2}$ et $S_{yx} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_{cal}\right)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} d_i^{\,2}}{n}}$;

$$\sigma_x = \sqrt{\frac{1}{N}\left(x - \bar{x}\right)^2} = \sqrt{\overline{x^2} - \left(\bar{x}\right)^2}$$

t : suit la distribution de Student à $n-2$ degré de liberté

b) Corrélation : Régression : r représente le coefficient de régression de l'échantillon et ρ représente le coefficient de régression de la population.

H_0 : Hypothèse « $\rho = 0$ Pas de dépendance »

$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$:Student à $n-2$ degré de liberté.

- Exemple :

On considère un échantillon dont le nombre de points $n=5$ et le coefficient de régression de l'échantillon $b=-0{,}9586$. Tester l'hypothèse que $\beta=0{,}6$ sachant que $S_{yx}=1{,}4656\&$ $\sigma_x=0{,}7699$

$$t=\frac{\beta-b}{S_{yx}/\sigma_x}\sqrt{n-2}\Rightarrow t=\frac{0{,}6+0{,}9586}{1{,}4656/0{,}7699}\sqrt{5-2}$$

$$\Rightarrow t=1{,}2875$$

$$\nu=n-2=5-2=3\rightarrow t_{0{,}95}=2{,}35\&t_{0{,}99}=4{,}54$$

$t\prec t_{0{,}95}\Rightarrow$ On ne rejette pas l'hypothèse que $\beta=0{,}6$

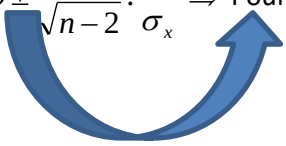
Exercice :

Calculer les limites de confiance à 95% pour le coefficient de régression de l'échantillon $b=0{,}96$ Sachant que $S_{yx}=1{,}4656, n=5\ \&\ \sigma_x=0{,}7699$

Solution :

$$t=\frac{\beta-b}{S_{yx}/\sigma_x}\sqrt{n-2}\Rightarrow \beta=b+\frac{t}{\sqrt{n-2}}.\frac{S_{yx}}{\sigma_x}$$

$\beta=b\pm\frac{t}{\sqrt{n-2}}.\frac{S_{yx}}{\sigma_x}\Rightarrow$ Pour avoir un intervalle symétrique (distribution symétrique)



Pour $\alpha=5\%$, intervalle symétrique $\Rightarrow \alpha=\frac{0{,}05}{2}=0{,}025; pour: \nu=5-2=3\Rightarrow t_{0{,}975}=3{,}18.$

$$\Rightarrow \beta=0{,}96\pm\frac{3{,}18}{\sqrt{3}}.\frac{1{,}4656}{0{,}7699}=0{,}96\pm1{,}9036$$

Exercice :

$r=0{,}32$ pour un échantillon de $n=18$. Peut-on en conclure que le coefficient de régression de la population ρ est significativement à zéro ?

Solution :

H_0 : Hypothèse « $\rho=0$ Pas de dépendance »

$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$: Student à $n-2$ degré de liberté

$$t=\frac{0{,}32\sqrt{18-2}}{\sqrt{1-(0{,}32)^2}}=1{,}35$$

$\nu=16: t_{0{,}95}=1{,}75\ \&\ t_{0{,}99}=2{,}58$

$\left.\begin{matrix} t\prec 1{,}75 \\ t\prec 2{,}58 \end{matrix}\right\}\rightarrow$ On ne rejette pas l'hypothèse H_0

Exercice :

$r = 0{,}32$, $\alpha = 0{,}05$. Quelle est la taille minimale de l'échantillon pour que le coefficient de régression de la population ρ soit supérieur à zéro ?

H_0 : Hypothèse « $\rho = 0$ Pas de dépendance »

$t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$:Student à $n-2$ degré de liberté

$t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$: $\rho \succ 0$, il suffit de rejeter l'hypothèse H_0

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \Rightarrow t^2 = \frac{r^2.(n-2)}{1-r^2} \Rightarrow t^2.(1-r^2) = r^2.(n-2) \Rightarrow n = 2 + \frac{t^2.(1-r^2)}{r^2}$$

$$n = 2 + \frac{t^2.(1-r^2)}{r^2} \Rightarrow n = 2 + \frac{t^2.(1-(0{,}32)^2)}{(0{,}32)^2} \rightarrow (*)$$

D'après la table de distribution de Student, $\nu = \infty, t_{0,95} = 1{,}64$; on remplace dans la formule $(*)$

$$(*) \Rightarrow n = 2 + \frac{(1{,}64)^2.(1-(0{,}32)^2)}{(0{,}32)^2} = 25{,}6 \cong 26 ;$$

$$n = 26 \Rightarrow t_{cal} = \frac{0{,}32\sqrt{26-2}}{\sqrt{1-(0{,}32)^2}} = \frac{1{,}57}{0{,}95} = 1{,}65$$

$n = 26 \mapsto \nu = n - 2 = 24, t_{0,95} = 1{,}71$;

$\Rightarrow t_{cal} = 1{,}65 \rightarrow t \prec t_{0,95}$: On ne rejette pas l'hypothèse H_0

$$n = 27 \Rightarrow t_{cal} = \frac{0{,}32\sqrt{27-2}}{\sqrt{1-(0{,}32)^2}} = 1{,}69 \Rightarrow t_{cal} = 1{,}65 \rightarrow t \prec t_{0,95}$$: On ne rejette pas l'hypothèse H_0

$$n = 28 \Rightarrow t_{cal} = \frac{0{,}32\sqrt{28-2}}{\sqrt{1-(0{,}32)^2}} = 1{,}72 \Rightarrow t_{cal} = 1{,}72 \rightarrow t \succ t_{0,95}$$: On rejette l'hypothèse H_0 :« $\rho = 0$ »

Exercice :

Pour $n = 12$,On a obtenu la droite de régression $y = 35{,}82 + 0{,}476x$. Evaluer les limites de confiances à 95%, pour les valeurs de y_p pour $x = 65$.

Remarque : $x = 65$ e valeur théorique (calculée) d'interpolation.

Solution :

$$t=\frac{(y_0-y_p)\sqrt{n-2}}{S_{yx}.\sqrt{n+1+\left[n(x_0-\bar{x})^2/S_x^2\right]}};$$

$$S_{yx}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-y_{cal})^2}{n}}=\sqrt{\frac{\sum_{i=1}^{n}d_i^{\,2}}{n}};$$

$$\sigma_x=\sqrt{\frac{1}{N}(x-\bar{x})^2}=\sqrt{\overline{x^2}-(\bar{x})^2}$$

$$y_p=y_0\pm\frac{t.S_{yx}.\sqrt{n+1+\left[n(x_0-\bar{x})^2/S_x^2\right]}}{\sqrt{n-2}}$$

Pour avoir un intervalle symétrique

(distribution symétrique)

Pour $\alpha=5\%, \nu=12-2=10$, intervalle symétrique $\Rightarrow \alpha=\frac{0,05}{2}=0,025\Rightarrow t_{0,975}=2,23.$

$$y_0=35,82+0,476*65=66,67$$

$$y_p=66,76\pm\frac{2,23.128.\sqrt{12+1+(12*12,78/7,08)}}{3,16}$$

$$\Rightarrow y_p=66,76\pm3,8$$

$$\Rightarrow y_p\in[62,96-70,56]$$

Exercice :

$r=0,75$ pour un échantillon de $n=24$. Peut-on rejeter l'hypothèse que a) $\rho=0,6$, b) $\rho=0,5$?

$\alpha=0,05$

Solution :

$$Z=1,1513*\log\left(\frac{1+r}{1-r}\right)$$

$$\mu_Z=1,1513*\log\left(\frac{1+\rho}{1-\rho}\right)$$

$$\sigma_Z=\frac{1}{\sqrt{n-3}}$$

$$\Rightarrow Z=1,1513*\log\left(\frac{1+0,75}{1-0,75}\right)=0,9730$$

$$\Rightarrow \mu_Z=1,1513*\log\left(\frac{1+0,6}{1-0,6}\right)=0,6932$$

$$\Rightarrow \sigma_Z=\frac{1}{\sqrt{24-3}}=0,2182$$

$$z=\frac{Z-\mu_Z}{\sigma_Z}=\frac{0,9730-0,6932}{0,2182}=1,28$$

$\phi(z)=0,95\Rightarrow z=1,64\mapsto z\prec1,64$: on ne peut pas rejeter l'hypothèse que a) $\rho=0,6$

Même méthode pour $\rho = 0{,}5$ $z \succ 1{,}64$: on rejette l'hypothèse que b) $\rho = 0{,}5$

Exercice n°1 : Soit le tableau de mesures suivant :

X	0	1	2	3	4
Y	1	1	3	7	13

Ajuster aux données une courbe de régression.

Courbe de régression

$$y = a_0 + a_1 x \rightarrow \begin{cases} a_0 = \bar{y} - a_1 \bar{x} \\ a_1 = \dfrac{Cov(x,y)}{Var(x)} \end{cases} \rightarrow \begin{cases} Cov(x,y) = \overline{xy} - \bar{x}.\bar{y} \\ Var(x) = \overline{x^2} - (\bar{x})^2 \\ \overline{xy} = \dfrac{\sum_{i=1}^{N} x_i.y_i}{N} \\ \overline{x^2} = \dfrac{\sum_{i=1}^{N} x_i^2}{N} \end{cases} \quad \begin{cases} \bar{x} = \dfrac{\sum_{i=1}^{N} x_i}{N} = \dfrac{10}{5} = 2 \\ \bar{y} = \dfrac{\sum_{i=1}^{N} y_i}{N} = \dfrac{25}{5} = 5 \end{cases}$$

x_i	0	1	2	3	4	$\sum_{i=1}^{5} x_i = 10$
y_i	1	1	3	7	13	$\sum_{i=1}^{5} y_i = 25$
$x_i.y_i$	0	1	6	21	52	$\sum_{i=1}^{5} x_i.y_i = 80$
x_i^2	0	1	4	9	16	$\sum_{i=1}^{5} x_i^2 = 30$
x_i^3	0	1	8	27	64	$\sum_{i=1}^{5} x_i^3 = 100$
x_i^4	0	1	16	81	256	$\sum_{i=1}^{5} x_i^4 = 354$
$x_i^2.y_i$	0	1	12	63	208	$\sum_{i=1}^{5} x_i^2.y_i = 284$

$$\begin{cases} \overline{xy} = \dfrac{\sum_{i=1}^{N} x_i.y_i}{N} = \dfrac{80}{5} = 16 \\ \overline{x^2} = \dfrac{\sum_{i=1}^{N} x_i^2}{N} = \dfrac{30}{5} = 6 \end{cases}$$

$$\rightarrow \begin{cases} Cov(x,y) = \overline{xy} - \bar{x}.\bar{y} = 16 - 2.5 = 6 \\ Var(x,y) = \overline{x^2} - (\bar{x})^2 = 6 - 2^2 = 2 \end{cases}$$

$$\begin{cases} a_0 = \bar{y} - a_1 \bar{x} = 5 - 3.2 = -1 \\ a_1 = \dfrac{Cov(x,y)}{Var(x)} = \dfrac{6}{2} = 3 \end{cases} \Rightarrow y = -1 + 3x$$

Remarque :

On peut résoudre le système $\begin{cases} n.a_0 + a_1 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \\ a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i.y_i \end{cases}$

pour trouver les deux coefficients de la droite $y = a_0 + a_1 x$

$$\begin{cases} n.a_0 + a_1 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \\ a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i.y_i \end{cases} \Rightarrow \begin{cases} 5a_0 + 10a_1 = 25 \rightarrow (1) \\ 10a_0 + 30a_1 = 80 \rightarrow (2) \end{cases} \Rightarrow \begin{cases} Eq1 \rightarrow a_0 = 5 - 2a_1 \\ a_0 + 3a_1 = 8 \rightarrow Eq2/10 \end{cases} \Rightarrow \begin{cases} a_0 = 5 - 2a_1 \\ 5 - 2a_1 + 3a_1 = 8 \end{cases} \Rightarrow (a_0, a_1) = (-1,3)$$

Le critère à minimiser

$$\sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (y_{cal} - y_i)^2 = (-1-1)^2 + (2-1)^2 + (5-3)^2 + (8-7)^2 + (11-13)^2 = 14 \neq 0 \Rightarrow y = a_0 + a_1 x + a_2 x^2$$

$$\begin{cases} n.a_0 + a_1 \sum_{i=1}^{N} x_i + a_2 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i \\ a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 + a_2 \sum_{i=1}^{N} x_i^3 = \sum_{i=1}^{N} x_i.y_i \\ a_0 \sum_{i=1}^{N} x_i^2 + a_1 \sum_{i=1}^{N} x_i^3 + a_2 \sum_{i=1}^{N} x_i^4 = \sum_{i=1}^{N} x_i^2.y_i \end{cases} \rightarrow \begin{cases} n.a_0 + a_1 \sum_{i=1}^{N} x_i + a_2 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i \\ a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 + a_2 \sum_{i=1}^{N} x_i^3 = \sum_{i=1}^{N} x_i.y_i \\ a_0 \sum_{i=1}^{N} x_i^2 + a_1 \sum_{i=1}^{N} x_i^3 + a_2 \sum_{i=1}^{N} x_i^4 = \sum_{i=1}^{N} x_i^2.y_i \end{cases} \rightarrow \begin{cases} 5a_0 + 10a_1 + 30a_2 = 25 \\ 10a_0 + 30a_1 + 100a_2 = 80 \\ 30a_0 + 100a_1 + 354a_2 = 284 \end{cases}$$

$$\Rightarrow y = 1 - x + x^2 \Rightarrow \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (y_{cal} - y_i)^2 = (1-1)^2 + (1-1)^2 + (3-3)^2 + (7-7)^2 + (13-13)^2 = 0$$

Exercice n°2 :

Une société a mis au point un produit. Une étude préalable a montré une relation entre le prix X proposé pour ce produit et le nombre de clients Y disposé à l'acheter à ce prix. Le chiffre d'affaire potentiel Z, correspondant au choix du prix X est donné par Z=X.Y. L'enquête menée auxprès de 500 personnes a donné le tableau suivant :

X :Prix DA	**40**	**35**	**32**	**28**	**24**	**20**	**16**	**12**	**10**	**8**
Y : Client	60	80	130	200	240	350	390	420	440	500

- Etablir la droite de régression : $y = a_0 + a_1 x$.
- Déterminer le chiffre d'affaire maximal.

$$y = a_0 + a_1 x \rightarrow$$

$$\begin{cases} a_0 = \bar{y} - a_1 \bar{x} \\ a_1 = \dfrac{Cov(x,y)}{Var(x)} \end{cases} \rightarrow \begin{cases} Cov(x,y) = \overline{xy} - \bar{x}.\bar{y} \\ Var(x) = \overline{x^2} - \left(\bar{x}\right)^2 \\ \overline{xy} = \dfrac{\sum_{i=1}^{N} x_i.y_i}{N} \\ \overline{x^2} = \dfrac{\sum_{i=1}^{N} x_i^2}{N} \end{cases} \Rightarrow \begin{cases} \bar{x} = \dfrac{\sum_{i=1}^{N} x_i}{N} = \dfrac{40+35+32+24+20+16+12+10+8}{10} = 22,5 \\ \bar{y} = \dfrac{\sum_{i=1}^{N} y_i}{N} = \dfrac{60+80+130+200+240+350+390+420+440+500}{10} = 281 \end{cases}$$

$$\sum_{i=1}^{10} x_i.y_i = 47400 ; \sum_{i=1}^{10} x_i^2 = 6173$$

$$\begin{cases} \overline{xy} = \dfrac{\sum_{i=1}^{10} x_i.y_i}{10} = 4740 \\ \overline{x^2} = \dfrac{\sum_{i=1}^{N} x_i^2}{10} = 617,3 \end{cases} \rightarrow \begin{cases} Cov(x,y) = \overline{xy} - \bar{x}.\bar{y} = 4740 - 22,5*281 = -1582,5 \\ Var(x,y) = \overline{x^2} - \left(\bar{x}\right)^2 = 617,3 - 22,5^2 = 111,05 \end{cases}$$

$$\begin{cases} a_0 = \bar{y} - a_1 \bar{x} = 281 + 14,25*22,5 = 601,63 \\ a_1 = \dfrac{Cov(x,y)}{Var(x)} = -14,25 \end{cases} \Rightarrow y = 601,63 - 14,25x$$

Le chiffre d'affaire maximal :

$$z = x*y \Rightarrow z = x*(601,63 - 14,25x) \Rightarrow z = 601,63x - 14,25x^2.$$

$$\Rightarrow z_{Max} = 601,63*21,11 - 14,25*(21,11)^2 = 6350,15DA.$$

4.7 Lissage exponentiel

$$\Rightarrow y = a^t.b \Rightarrow \ln(y) = (\ln a).t + \ln b \rightarrow Y' = At + B \Rightarrow \begin{cases} A = \ln a \Rightarrow a = e^A \\ B = \ln b \Rightarrow b = e^B \end{cases}$$

$$y = At + B \rightarrow \begin{cases} B = \overline{\ln y} - A\bar{t} \\ A = \dfrac{Cov(t, \ln y)}{Var(t)} \end{cases} \rightarrow \begin{cases} Cov(t, \ln y) = \overline{t \ln(y)} - \bar{t}.\overline{\ln(y)} \\ Var(t) = \overline{t^2} - \left(\bar{t}\right)^2 \\ \overline{t \ln(y)} = \dfrac{\sum_{i=1}^{N} t_i.\ln(y_i)}{N} \\ \bar{t} = \dfrac{\sum_{i=1}^{N} t_i}{N} \text{ et } \overline{\ln(y)} = \dfrac{\sum_{i=1}^{N} \ln(y_i)}{N} \end{cases}$$

SÉRIES CHRONOLOGIQUES

5.1 Introduction

Les séries chronologiques ont connu un essor important dans plusieurs domaines comme l'économie, finance, la biologie, la météorologie et pollution. Le but principal réside dans :

- La compréhension du passé c.à.d. analyser et expliquer les valeurs observées ;
- La prédiction du future c.à.d. bâtir des prévisions pour les valeurs non encore observées ;
- L'étude du lien avec d'autres séries chronologiques.

5.2 Définition d'une série chronologique

On appelle série chronologique une suite finie de données quantitatives indexée par le temps [16]. L'indice de temps peut être selon le cas, la seconde, la minute, l'heure, le jour, le mois, le trimestre, le semestre, le quadrimestre, l'année ,……

La série chronologique $\{y_t\}_{t\in T}$ avec $T=\{t_1,t_2\cdots,t_n\}$ n'est rien d'autre que la série statistique double $\left(t_j,y_{t_j}\right)_{1\le j\le n}$, où :

- La première composante est le temps
- La deuxième composante est une variable numérique y prenant ses valeurs aux instants t.

Une série chronologique est composée de partie essentielle qui sont la tendance $X(t)$ et la composante saisonnière telle que $Y(t)=f(X(t),S(t))$

Composantes saisonnières

Evolution d'une variable dans le temps

Tendance

5.3 Représentation graphique

On représente graphiquement la série chronologique $\{y_t\}_{t\in T}$

1- En dessinant le nuage formé par les points $\left(t_j,y_{t_j}\right)_{1\le j\le n}$;

2- En reliant les points entre eux par des segments de droite, pour indiquer la chronologie.

5.4 Analyse de la tendance

La tendance représente l'évolution générale de la série chronologique et la composante saisonnière représente l'effet de différentes saisons sur la série chronologique.

Le modèle s'écrit comme suit : $Y(t)=f(X(t),S(t))$ $\begin{cases} X(t)+S(t) \\ X(t)*S(t) \end{cases}$

Telle que $f = \begin{cases} + & Modéleadditif \\ * & ModèleMultiplicatif \end{cases}$

Il y a deux types de la tendance (linéaire, exponentielle).

5.4.1 Tendance linéaire

La tendance $X(t)$ d'une série chronologique est linéaire si les coefficients $X(t)-X(t-1)=Cst$

$$X(t)-X(t-1)=a \Rightarrow X(t)=a+X(t-1)=$$
$$a+a+X(t-2)=2a+X(t-2)=$$
$$3a+X(t-3)=at+X(t-t)$$
$$\Rightarrow X(t)=at+X(0)$$
$$\Rightarrow X(t)=a.t+b$$

$$y=at+b \rightarrow \begin{cases} b=\bar{y}-a\bar{t} \\ a=\dfrac{Cov(t,y)}{Var(t)} \end{cases} \rightarrow \begin{cases} Cov(t,y)=\overline{ty}-\bar{t}.\bar{y} \\ Var(t)=\dfrac{n^2-1}{12} \\ \overline{ty}=\dfrac{\sum_{i=1}^{N} t_i.y_i}{N} \\ \bar{t}=\dfrac{\sum_{i=1}^{N} t_i}{N} et \bar{y}=\dfrac{\sum_{i=1}^{N} y_i}{N} \end{cases}$$

5.4.2 Tendance exponentielle

La tendance $X(t)$ d'une série chronologique est exponentielle si les coefficients

$$X(t)/X(t-1)\approx Cst$$

$$X(t)/X(t-1)=a \Rightarrow X(t)=a*X(t-1)=a*(a*X(t-2))=a^2*X(t-2)=a^3*X(t-3)=a^t*X(0)$$
$$\Rightarrow X(t)=a^t*b$$

Pour étudier la tendance exponentielle, il faut transformer ce modèle à une tendance linéaire.

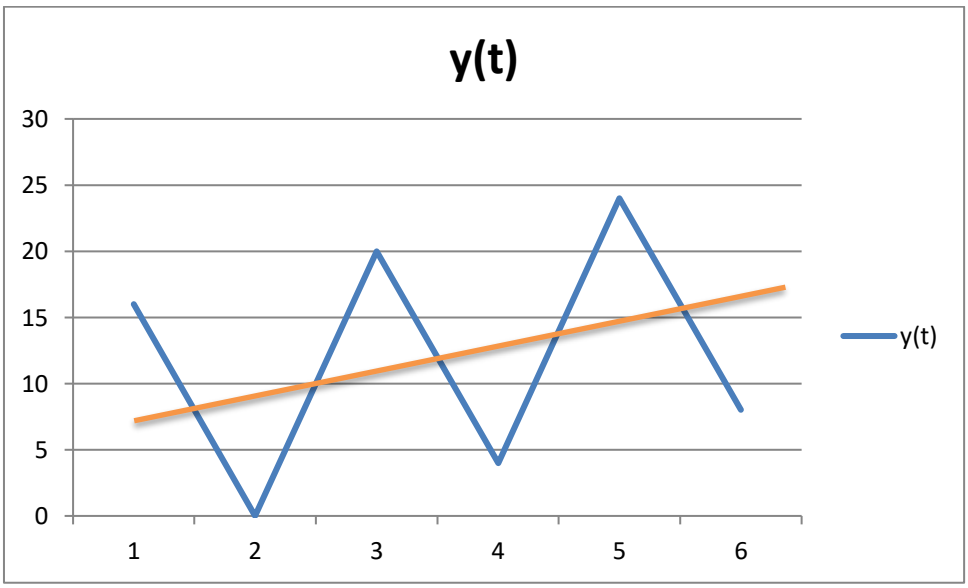
$$\Rightarrow X(t)=a^t.b \Rightarrow \ln(X(t))=(\ln a).t+\ln b \rightarrow X'(t)=At+B$$

$$\Rightarrow \begin{cases} A=\ln a \Rightarrow a=e^A \\ B=\ln b \Rightarrow b=e^B \end{cases}$$

$$X'(t)=At+B\rightarrow\begin{cases}B=\overline{\ln y}-A\bar{t}\\ A=\dfrac{Cov(t,\ln y)}{Var(t)}\end{cases}\rightarrow\begin{cases}Cov(t,\ln y)=\overline{t\ln(y)}-\bar{t}.\overline{\ln(y)}\\ Var(t)=\dfrac{n^2-1}{12}\\ \overline{t\ln(y)}=\dfrac{\sum_{i=1}^{N}t_i.\ln(y_i)}{N}\\ \bar{t}=\dfrac{\sum_{i=1}^{N}t_i}{N}\ et\ \overline{\ln(y)}=\dfrac{\sum_{i=1}^{N}\ln(y_i)}{N}\end{cases}\rightarrow$$

Remarque :

La tendance de la série chronologique est déterminée après élimination des effets de la composante saisonnière en effectuant un lissage de la série. Pour cela, il faut calculer **la moyenne mobile pour mieux apparaître l'allure de la tendance**.



5.5 La moyenne mobile

- Si le nombre de saison est impair :

Exemple : si l'ordre de la moyenne mobile =3. Alors $Z(t)=\frac{1}{3}[y(t-1)+y(t)+y(t+1)]$

si l'ordre de la moyenne mobile =5. Alors $Z(t)=\frac{1}{5}[y(t-2)+y(t-1)+y(t)+y(t+1)+y(t+2)]$.

- Si le nombre de saison est pair :

Exemple : si l'ordre de la moyenne mobile =2. Alors $Z(t)=\frac{1}{2}\left[\frac{1}{2}y(t-1)+y(t)+\frac{1}{2}y(t+1)\right]$

si l'ordre de la moyenne mobile =4. Alors $Z(t)=\frac{1}{4}\left[\frac{1}{2}y(t-2)+y(t-1)+y(t)+y(t+1)+\frac{1}{2}y(t+2)\right]$

Remarque :

La tendance de la série chronologique est déterminée donc par la moyenne mobile $Z(t)$. On évalue donc les quantités suivantes : $Z(t)-Z(t-1)$ & $Z(t)/Z(t-1)$

5.6 Composantes saisonnières

La modélisation des composantes saisonnières se fait selon les deux cas suivants :

a) **Modèle additif :** il combine une tendance et une saisonnalité de période p (nombre de saison) de la manière suivante : $Y(t)=X(t)+S(t)\Rightarrow S(t)=Y(t)-X(t)=\Delta(t)$ & $\sum_{t=1}^{nbde\ saison} S(t)=0$

b) **Modèle Multiplicatif :** ce modèle combine une tendance et une saisonnalité de période p(nombre de saison) de la manière suivante : $Y(t)=X(t)*S(t)\Rightarrow S(t)=Y(t)/X(t)$ & $\sum_{t=1}^{nbde\ saison} S(t)=$ nombre de saison

5.7 Exemple d'une prévision par lissage linéaire

Le tableau suivant représente la quantité (en tonnes) de papier consommé dans une administration

Année	**2017**		**2018**		**2019**
Semestre	1er semestre	2ème semestre	1er semestre	2ème semestre	1er semestre
Qté	6	18	10	22	14

1. Donner le modèle de la série chronologique avec composante saisonnière (Modèle additif).
2. Selon le modèle, quelle est la quantité de papier à consommer à la fin de l'année 2019 ?

L'administration devra faire une demande de budget supplémentaire si la quantité de papier consommée est supérieure à 100 tonnes (Qté>100). Quand est-ce que ceci risque de se produire ?

- **Solution :**

1. Le modèle de la série chronologique avec composante saisonnière (Modèle additif).

t	y(t)	Z(t)	Z(t)-Z(t-1)	Z(t)/Z(t-1)
1	6	-	-	-
2	18	13	-	-
3	10	15	2	1.15
4	22	17	2	1.13
5	14	-	-	-

Le nombre de saison =2

La moyenne mobile :

$$Z(t)=\frac{1}{2}\left[\frac{1}{2}y(t-1)+y(t)+\frac{1}{2}y(t+1)\right]$$

$Z(t)-Z(t-1)=C^{ste}\Rightarrow$ **tendance linéaire** $\Rightarrow X(t)=a.t+b$

$$y=at+b\rightarrow\begin{cases}b=\bar{y}-a\bar{t}\\ a=\dfrac{Cov(t,y)}{Var(t)}\end{cases}\rightarrow\begin{cases}Cov(t,y)=\overline{ty}-\bar{t}.\bar{y}\\ Var(t)=\dfrac{n^2-1}{12}\\ \overline{ty}=\dfrac{\sum_{i=1}^{N}t_i.y_i}{N}\\ \bar{t}=\dfrac{\sum_{i=1}^{N}t_i}{N}\ et\ \bar{y}=\dfrac{\sum_{i=1}^{N}y_i}{N}\end{cases}\rightarrow\begin{cases}Var(t)=\dfrac{n^2-1}{12}=\dfrac{5^2-1}{12}=2\\ \overline{ty}=\dfrac{1.6+2.18+3.10+4.22+5.14}{5}=46\\ \bar{t}=\dfrac{1+2+3+4+5}{5}=3\\ \bar{y}=\dfrac{6+18+10+22+14}{5}=14\end{cases}$$

$$\begin{cases}b=\bar{y}-a\bar{t}=14-2.3=8\\ a=\dfrac{Cov(t,y)}{Var(t)}=\dfrac{46-3.14}{2}=2\end{cases}\Rightarrow y=2t+8\Rightarrow X(t)=2t+8$$

Composante saisonnière (modèle additif) : $Y(t)=X(t)+S(t)\Rightarrow S(t)=Y(t)-X(t)=\Delta(t)$

t	$Y(t)$	$X(t)$	$\Delta(t)$
1	6	10	-4
2	18	12	+6
3	10	14	-4
4	22	16	+6
5	14	18	-4

$$Moy(S_1)=\frac{-4-4-4}{3}=-4;Moy(S_2)=\frac{6+6}{2}=6$$

$$\sum_i S_i = -4+6 = 2 \neq 0 \Rightarrow C = \frac{Moy(S_1)+Moy(S_2)}{2} = \frac{-4+6}{2} = 1$$

$$Moy(S_1) \leftarrow Moy(S_1) - C \leftarrow -5; Moy(S_2) \leftarrow Moy(S_2) - C \leftarrow 5$$

$$\Rightarrow \sum_i S_i = -5+5 = 0$$

$$Le\ modèle\ : \begin{cases} Y(t) = X(t) + S(t) \\ X(t) = 2t+8 \\ S(1) = -5; S(2) = 5 \\ S(t+2) = S(t) \end{cases}$$

2. la quantité de papier à consommer à la fin de l'année 2019:

$$t = 6 \Rightarrow Y(6) = X(6) + S(6) = 2.6 + 8 + 5 = 25$$

3. $2t+8+S(t) \succ 100 \Rightarrow \begin{cases} 2t+8-5 \succ 100 \\ 2t+8+5 \succ 100 \end{cases} \Rightarrow \begin{cases} 2t+8-5 \succ 100 \\ 2t+8+5 \succ 100 \end{cases} \Rightarrow \begin{cases} 2t \succ 97 \\ 2t \succ 87 \end{cases} \Rightarrow \begin{cases} t \succ 48.5 \\ t \succ 43.5 \end{cases} \Rightarrow t \cong 44$

Le risque peut se produire en : $2017 + \frac{44}{2} - 1 = 2038$

5.8 Exemple d'une prévision par lissage exponentiel

Le tableau suivant représente le nombre de personnes atteintes d'une maladie contagieuse dans une ville donnée.

2017		**2018**		**2019**
1er semestre	2ème semestre	1er semestre	2ème semestre	1er semestre
11	7	29	43	101

1. Donner le modèle de la série chronologique avec composante saisonnière (Modèle multiplicatif). Selon le modèle, quelle est l'estimation du nombre de personnes atteintes à la fin de l'année 2019?

1. Le modèle de la série chronologique avec composante saisonnière (Modèle Multiplicatif).

t	y(t)	Ln(y(t))	Z(t)	Z(t)-Z(t-1)	Z(t)/Z(t-1)
1	11	2.398	-	-	-
2	7	1.946	13.5	-	-
3	29	3.367	27	13.5	2
4	43	3.761	54	27	2
5	101	4.615	-		-

Le nombre de saison =2

La moyenne mobile : $Z(t) = \frac{1}{2}\left[\frac{1}{2}y(t-1) + y(t) + \frac{1}{2}y(t+1)\right]$

$Z(t)/Z(t-1)=C^{ste} \Rightarrow$ **tendance** **exponentielle**

$$\Rightarrow X(t)=a^t.b \Rightarrow \ln(X(t))=(\ln a).t+\ln b \rightarrow X'(t)=At+B \Rightarrow \begin{cases} A=\ln a \Rightarrow a=e^A \\ B=\ln b \Rightarrow b=e^B \end{cases}$$

$$X'(t)=At+B \rightarrow \begin{cases} B=\overline{\ln y}-A\bar{t} \\ A=\dfrac{Cov(t,\ln y)}{Var(t)} \end{cases} \rightarrow \begin{cases} Cov(t,\ln y)=\overline{t\ln(y)}-\bar{t}.\overline{\ln(y)} \\ Var(t)=\dfrac{n^2-1}{12} \\ \overline{t\ln(y)}=\dfrac{\sum_{i=1}^{N} t_i.\ln(y_i)}{N} \\ \bar{t}=\dfrac{\sum_{i=1}^{N} t_i}{N} \text{ et } \overline{\ln(y)}=\dfrac{\sum_{i=1}^{N} \ln(y_i)}{N} \end{cases}$$

$$\rightarrow \begin{cases} Var(t)=\dfrac{n^2-1}{12}=\dfrac{5^2-1}{12}=2 \\ \overline{t\ln(y)}=\dfrac{2.398+3.892+10.101+15.044+23.075}{5}=10.902 \\ \bar{t}=\dfrac{1+2+3+4+5}{5}=3 \\ \overline{\ln(y)}=\dfrac{2.398+1.946+3.367+3.761+4.615}{5}=3.217 \end{cases}$$

$$\begin{cases} B=\overline{\ln y}-A\bar{t}=3.217-0.626*3=1.339 \\ A=\dfrac{Cov(t,y)}{Var(t)}=\dfrac{10.902-(3*3.217)}{2}=0.626 \end{cases} \Rightarrow X'(t)=0.626.t+1.339$$

$$\Rightarrow \begin{cases} a=e^{0.626}=1.870 \\ b=e^{1.339}=3.815 \end{cases} \rightarrow X(t)=(1.870)^t*3.815$$

Composante saisonnière : $Y(t)=X(t); S(t) \Rightarrow S(t)=Y(t)/X(t)=Q(t)$

t	$Y(t)$	$X(t)$	$Q(t)$
1	**11**	**7.134**	**1.542**
2	**7**	**13.341**	**0.525**
3	**29**	**24.947**	**1.162**
4	**43**	**46.651**	**0.922**
5	**101**	**87.237**	**1.158**

$$Moy(S_1)=\frac{1.542+1.162+1.158}{3}=1.287; Moy(S_2)=\frac{0.525+0.922}{2}=0.724$$

$$\sum_i S_i=1.287+0.724=2.011\neq 2 \Rightarrow C=\frac{2.011-2}{2}=0.006$$

$$Moy(S_1) \leftarrow Moy(S_1) - C \leftarrow 1.281;$$
$$Moy(S_2) \leftarrow Moy(S_2) - C \leftarrow 0.718 \Rightarrow \sum_i S_i = 1.999 \begin{cases} Y(t) = X(t) * S(t) \\ X(t) = (1.87)^t * 3.815 \\ S(1) = 1.281; S(2) = 0.718 \\ S(t+2) = S(t) \end{cases}$$

2. *Le nombre de personnes atteintes à la fin de l'année* ***2019 :***

$t = 6 \Rightarrow Y(6) = X(6) * S(6) = 117.130 \approx 117$

5.9 Conclusion

Dans ce chapitre, nous avons montré plusieurs concepts en commençant par la définition d'une série chronologique, la modélisation mathématique à base d'un lissage linaire et exponentiel. Après, nous avons traité deux exemples réels pour mieux comprendre la prédiction par les séries chronologiques. Les séries chronologiques jouent un rôle important dans notre vie quotidienne où nous pouvons dans plusieurs domaines comme la prédiction des ventes ou la propagation de corona virus qui est devenu le sujet d'actualité.

[1] M. JAMBU. (1999). Méthode de base de l'analyse de données.

[2] T. Yves. Cours de statistique descriptive.

[3] G. SAPORTA (2006). Probabilités analyse des données et statistique. EditionTECHNIP, paris. France.

[4] J.P.BENZECRI (1980). L'analyse de données (Tome1) la taxonomie. Dunod

[5] J. DUDA (2002). Pattern recognition. MIT.

[6] P.DEMARTINES et J. HeRAULT (1997). Curvilinear component analysis : a self organizing neural network for non linear of mappinf of data set. IEEE transactions on neural networks, 8(1) :148-54

[7] L. LEBART , A. MORINEAU et M. PIRON (1995). Statistique exploratoire multidimensionnele. Dunod

[8] J. ZHANG (2012). Kernel principal compenent analysis. Expert systems with application. Vol10 :11-25.

[9] F. CHEVALIER et J. LE BALLAC (2013). Rapport sur la classification. Université de RENNES1.

[10] G. CELEUX, E. DIDAY, G. GOVAERT (1989). Classification automatique des données. Dunod.

[11] G. BROSSIER (2003). Les éléments fondamentaux de la classification. Hermes Sciences publication.

[12] N. WICKER (2001). Cours d'analyse de données. North-western European Journal of mathematics

[13] F.DAZY et J-F LE BARZIC (1996). L'analyse des données évolutives « méthodes et applications ». Edition TECHNIP.

[14] C. E. Lawrence and A. A. Reilly (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences," *Proteins: Structure, Function, and Bioinformatics,* vol. 7, pp. 41-51.

[15] J. De Lagarde (1995). Initiation à l'analyse des données. Dunod.

[16] A. LAGNOUX (2018). Série chronologique. Université de TOULOUS.

Les techniques d'analyse des données ont connu un essor important surtout avec le développement de l'informatique et big data. Le volume important des données nécessite comme un prétraitement : la réduction des données, ce qui est l'objectif principal de l'analyse des données en premier lieu. Pour résoudre le problème de la dimensionalité, les méthodes multidimensionnelles telles que l'Analyse en Composantes Principales (ACP) et l'Analyse Factorielle des Correspondances (AFC) seront exploitées et expliquées en détail dans cet ouvrage.

En second lieu, l'interprétation et la classification des données dans le domaine de la reconnaissance des formes, la fouille des données et l'intelligence artificielle font appel aux méthodes de classification plus particulièrement l'algorithme de la classification hiérarchique qui permet une représentation arborescente appelée dendrogramme et les méthodes de partitionnement « clustering » comme l'algorithme des centres mobiles qui est très utilisé dans l'apprentissage non-supervisé. En plus, les méthodes morphologiques à base des opérateurs de traitement d'image comme l'érosion, dilatation, ouverture et fermeture peuvent être utilisées dans le domaine de la classification.

La prévision dans le domaine d'économie et le domaine d'épidémiologie nécessitent des modèles statistiques puissants. Pour cela, nous avons introduit la méthode des moindres carrés et les séries chronologiques. Généralement, deux modèles sont exploités comme la prévision linéaire et la prévision exponentielle. Pour juger l'efficacité des modèles proposés, un coéfficient de corrélation doit être mesuré.

Enfin, dans l'espoir que cet ouvrage constitue la première marche d'un long escalier et permet aux lecteurs d'acquérir des nouvelles connaissances en analyse des données.