République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

FACULTÉ DES SCIENCES EXACTES ET SCIENCES DE LA NATURE ET LA VIE
DÉPARTEMENT DE MATHÉMATIQUES

Mémoire Présenté En Vue De L'obtention Du

DIPLÔME De Master en Mathématiques

Option : Statistique

Par

Bassa Radia

Titre

Classification Automatique

Membres du Comité d'Examen

		UMKB	Président
Dr.	MERAGHNI Djamel	UMKB	Rapporteur
		UMKB	Examinateur

Juin 2012

Dédicace

D'abord,je remercie mon dieu,pour la force et le courage qu'il m'a donné pour élaborer ce travail.
A celui qui mai indique la bonne voie en me rappelant que la volonté fait toujours les grands hommes...

...mon pére

A celle qui a attendu avec patience les fruits de sa bonne éducation ,...

...ma mére

A mes fréres ;Saddam,Tarek,Mohamed Taher,Faris ,Abd elrahman et le bébé Abd ella
A mes soeurs ;Sabrina ,Samiha,Selsabil,Sirin,Meryam,Salima
A mes oncles et Ames tantes
A toute la famille Bassa et Tennech

Remerciements

Au nom de DIEU Le Plus Clément et Le Plus Miséricordieux.

Louange à Allah, Dieu de l'univers et de tous les hommes, que sa grâce, son salut, son pardon et sa bénédiction soient accordés au meilleur de ses créatures notre prophète Mohamed ainsi qu'aux membres purs de sa famille et à tous ses compagnons.

Je tiens à remercier en premier Dieu le tout puissant qui nous a accordé la volonté et le courage pour réaliser ce projet.

Tout d'abords je tiens à exprimer ma profonde gratitude à Monsieur Meraghni Djamel,M.A.C.C à l'Université de Biskra, d'avoir accepter la charge de m'encadrer,a eu confiance en moi et m'a permis de travai ller sur un sujet de mémoire qui m'était difficile car sans précédent, mais très intéressant. Il a su me donner une grande liberté d'initiative tout en restant toujours présent pour discuter des problèmes rencontrés, des résultats obtenus et des orientations à suivre. Son enthousiasme et son dynamisme m'ont à chaque fois permis de rebondir dans les moments difficiles. Je le remercie vivement pour son aide précieuse et tous les conseils qu'il a pu me fournir durant la préparation de ce mémoire.

Bien évidemment, je remercie D.Yahia qui m'a toujours aidé. J'ai rarement vu quelqu'un d'aussi gentille et serviable.

Un merci sans limites à ma très chère et vertueuse maman, qui a sû comment me tisser et m'illuminer le chemin de la réussite. Une mention spéciale aux êtres qui me sont très chers : mes sœurs et mon petit frère. Je leur exprime toute ma gratitude pour le soutien et l'encouragement qu'ils m'ont apportée pour mener à bien mon travail.

Table des matières

Table des figures

Liste des tableaux

Introduction

Statistique, branche des mathématiques ayant pour objet la collecte, le traitemen-
tet l'analyse des données numériques relatives à un ensemble d'objets, d'indivi-
dusou d'éléments. La statistique constitue un outil précieux pour l'expérimentation de projets, la gestion des entreprises ou encore l'aide à la décision.
Les méthodes d'analyse des données (ACP, AFc„ ACM,CAH,...) font partie dela statistique déscriptive. Dans ce mémoire, composé de trois chapitres :

Chapitre1 :Données Multivariées

Dans ce chapitre aplique les données multivariées on passe en revue les diférents résultats de l'ACP. Ces derniers constituent la base de toutes les autres méthodes d'analyse des données.

Chapitre2 :classification

Dans ce chapitre, on explique les différentes approches utilisées en classification : hiérarchique (ascendantes et descendantes) et non hiérarchique. La méthode de classification hiérarchique ascendante a divers algorithmes :saut minimum, diamètre, distance moyenne et Ward. Les différentes méthodes de classification non hiérarchique sont : centres mobiles, nuées dynamiques et k means.

Chapitre3 :Application

Dans cette chapitre application des méthode classification.

Chapitre 1

Données multivariées

La plupart du temps les données se presentent sous la forme suivante :on a relevé sur n uniteapplées ('individus')p variables numeriques lorsque n et p grands on cherche à synthetiser cette masse d'informations sous une forme exploitable et compréhensible :une premiere etape consiste ontà décrire séparement les resultats obtenus pour chaque variable.
On considerera donc ici
S'interesse qu' a une variable X ,appelée encore caractere,dont on possede n valeurs $x_1, ..., x_n$.

1.1 Formes des données

1.1.1 Tableau des données

L'analyse en composents principales(A.C.P)s'applique à des tableaux à deux dimensions croisant des individus et des variable quantitives appelés de façon concise tableaux individus X variables quantitatives selon un usage bien établi.
Les lignes du tableau représentent les individus et les colonnes représentent les variables.AL'intersection de la ligne i et de la colonne$j$ se trouve la valeur de la variable j pour l'individu i :

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

On définit :

$x_j = (x_{1j}; x_{2j}, ...; x_{nj})^t \in \mathbb{R}^n$,la variable $j$de la liste des n valeurs qu'elle prend sur les n individus .$e_i = (x_{i1}; x_{i2}; ....; x_{ip})^t \in \mathbb{R}^P$,l'individu i vecteur à p composantes

1.1.2 Individu_variables

Soit (Ω) un forme par n individus note $(\omega_1; \omega_2; ..\omega_n)$ ou $(e_1; ...; e_n), \Omega = \{e_1; ....; e_n\}$
On définit sur (Ω)
p variable statistique$(x_1, ..., x_p)$.(quantitative en generale) pour $(j = 1, ..., p)$:

$$\begin{aligned} x_j : \Omega &\to \mathbb{R} \\ \varrho_{i\to} &x_{ij} \\ e_i &\to x_j(e_i) = x_{ij}. \end{aligned}$$

Les termes d'individus et de variable recouvrent des notions différentes ,les questions qui l'on se pose sur les individus et sur les variable ne sont pas le mème nature.

1.1.3 Standardisation des données

Avant de faire l'A.C.P.,il est utile de transformer le tableau initial
$X = (x_{ij})_{1\leq i\leq n;\ 1\leq j\leq p}$en un tableau standard Z de terme général :

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{S_j}$$

Dans le but d'uniformiser l'unité de mesure des variables.

$$X = (x_{ij}) \to Y = (y_{ij}) = (x_{ij} - \overline{x_i}) \to Z = (z_{ij})$$

standardisation
Telle que :
Y est un tableau centré de terme général y_{ij},
Si on désigne par $D_{\frac{1}{S}}$la matrice diagonale des inverses des écarts types, alors on peut écrire Z sous la forme matricielle,

$$Z = Y\ D_{\frac{1}{S}}$$

1.2 Nuage des individus

S'interesser aux variables revient à envisager le tableau x(ou$z$) entant que juxtaposition de colonnes ,Achaque individu est associée une suite de n nombres selon ce point de vue ,un individu peut être présentée comme un point de l'espace vectoriel $\mathbb{R}^p$ dont les dimensions correspond à une variable.
Donc,l'ensemble de individus constitue dans $\mathbb{R}^p$, un nuage de points dont le centre de gravité est g.

1.2.1 Resemblance

Pour les individus on essaye d'evaluer leurs ressemblance ,on dira que deux individus se ressemblance qunad ils possédant des valeurs proches pour l'ensemble des variables. On doit alors définir entre individus ,pour cela et puisque les dimension de $\mathbb{R}^p$(les variables) ne sont pas de même nature.

1.2.2 Matrice de poids

Chaque individu e_i,est affécté d'un poids $p_i$positive telque :

$$\sum_{i=1}^{n} p_i = 1.$$

Les poids sont regroupés dans une matrice diagonale D appelée matrice des poids :

$$D = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & 0 & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & p_n \end{bmatrix}$$

Dans le cas usuel ou tous les individus ont la même importance ;on a

$$p_i = \frac{1}{n}; i = 1, 2..., n$$

Et donc :

$$D = \frac{1}{n} 1_n$$

1.2.3 Centre de gravité

On appelle centre de gravité le vecteur des moyennes arithmétiques des variable on le note par :$g = (\bar{x}_1, ..., \bar{x}_p)^t$où :

$$\bar{x}_j = \sum_{i=1}^{n} p_i x_{ij}; j = 1, 2, ..., p.$$

On a :

$$g = X^t D 1_n$$

Ou 1désigne le vecteur de $\mathbb{R}^p$dont toutes les composantes sont égales à 1.
Le tableau centre s'ecrit :

$$Y = X - 1_n g^t = \left(1_n - 1_n 1_n^t D\right) X.$$

1.2.4 Metrique

On doit utiliser une metrique M ,qui est une matrice carrée d'ordre p définie positive .la distance entre deux individus e_i et $e_{i'}$,est alors :

$$d_M(e_i, e_{i'}) := \sqrt{(e_i - e_{i'})^t M (e_i - e_{i'})}, i; i^{'} = 1, ..., n.$$

La métrique la plus utilisée en analyse des données est la matrice diagonale des inverses des variances notée par $D_{\frac{1}{s^2}}$. cela revient à diviser chaque variable par son écart -type ce qui permet d'eliminer l'influence des unités de mesure .l'utilisation de la métrique $D_{\frac{1}{s^2}}$ sur le tableau initial$X$ (ou sur Y) est equivalente àl'utilisation de la métrique usuelle (canonique,triviale) I_p sur le tableau standard Zdont les variables sont sans untiés ,cela d'où l'importance de procéder àla standardisation des données avant de commencer l'analyse.

1.2.5 Inertie

On apple inertie totale d'un nuage de points ,la moyenne (podérée) des carrées des distance par rapport au centre de gravité ,on la note par I_g :

$$\begin{aligned} I_g &:= \sum_{i=1}^{n} p_i d_M^2(e_i, g) \\ &= \sum_{i=1}^{n} p_i \langle e_i - g; e_i - g \rangle_M \\ &= \sum_{i=1}^{n} p_i (e_i - g)^t M (e_i - g) \end{aligned}$$

L'inertie d'un nuage en un point quelconque $a \in \mathbb{R}^p$est :

$$\begin{aligned} I_a &= \sum_{i=1}^{n} p_i d_M^2(e_i, a) \\ &= \sum_{i=1}^{n} p_i (e_i - a)^t M (e_i - a) \end{aligned}$$

La relation suivante ,dite `relation de huyghens.`lie les deux inertie :

$$I_a := I_g + d_M^2(a, g)$$

Démontration

$\cdot 1^{ere}$ cas :
Si $g = 0_{\mathbb{R}^p}$ (cas des tableau Y et Z)

$$I_a = I_0 + d_M^2(a) \Leftrightarrow \sum_{i=1}^{n} p_i d_M^2(e_i, a) = \sum_{i=1}^{n} p_i \|e_i\|^2 + \|a\|^2$$

On a :

$$\begin{aligned} d^2(e_i, a) &= \|e_i - a\|^2 \\ \|e_i - a\|^2 &= \langle e_i - a; e_i - a\rangle \\ &= \langle e_i, e_i\rangle - \langle e_i, a\rangle - \langle a, e_i\rangle + \langle a, a\rangle \\ &= \|e_i\|^2 + \|a\|^2 - 2\langle a, e_i\rangle \end{aligned}$$

Donc

$$\begin{aligned} I_a &= \sum_{i=1}^{n} p_i \|e_i - a\|^2 = \sum_{i=1}^{n} p_i \|e_i\|^2 + \sum_{i=1}^{n} p_i \|a\|^2 - 2\sum_{i=1}^{n} p_i \langle e_i, a\rangle \\ I_a &= I_0 + \|a\|^2 - 2\sum_{i=1}^{n} p_i \langle e_i, a\rangle \\ e_i &= [x_{i1}\ x_{i2}\ \cdots x_{ip}]^t \text{ et } a = [a_1\ a_2 \cdots a_p]^t \\ \sum_{i=1}^{n} p_i \langle e_i, a\rangle &= \sum_{i=1}^{n} p_i \sum_{i=1}^{n} x_{ij} a_j = \sum_{i=1}^{n} a_j \sum_{i=1}^{n} p_i x_{ij} \\ \sum_{i=1}^{n} x_{ij} a_j &= 0 \text{ mais} g = (x_1, x_2, \cdots, x_p)^t = (0, 0, \cdots, 0)^t \end{aligned}$$

Donc

$$I_a = I_0 + \|a\|_M^2$$

$\cdot 2^{éme}$cas :
Si $g \neq 0_{\mathbb{R}^p}$

$$\begin{aligned} \|e_i - a\|^2 &= \langle e_i - a, e_i - a\rangle \\ &= \langle e_i - g + g - a, e_i - g + g - a\rangle \\ &= \langle e_i - g; e_i - g\rangle + \langle g - a; g - a\rangle + \langle e_i - g; g - a\rangle + \langle g - a; e_i - g\rangle \\ I_a &= \sum_{i=1}^{n} p_i \langle e_i - a; e_i - a\rangle = \sum_{i=1}^{n} p_i \langle e_i - g; e_i - g\rangle + \sum_{i=1}^{n} p_i \langle g - a; g - a\rangle + 2\sum_{i=1}^{n} p_i \langle e_i - g; g - a\rangle \\ I_a &= I_g + \|g - a\|^2 + 2\sum_{i=1}^{n} p_i \langle e_i - g; g - a\rangle \end{aligned}$$

Telle que :

$$\sum_{i=1}^{n} p_i \langle e_i - g; g - a \rangle = 0$$

Car

$$\langle e_i - g; g - a \rangle = \langle e_i - g; -(a - g) \rangle = -\langle e_i - g; a - g \rangle$$

$$\begin{aligned}
\sum_{i=1}^{n} p_i \langle e_i - g; g - a \rangle_M &= \sum_{i=1}^{n} p_i \, (g - a)^t \, M \, (e_i - g) \\
&= (g - a)^t \, M \left(\sum_{i=1}^{n} p_i (e_i - g \right) \\
&= (g - a)^t \, M \left(\sum_{i=1}^{n} p_i e_i - \sum_{i=1}^{n} p_i g \right) \\
&= (g - a)^t \, M \, (g - g) \\
&= 0
\end{aligned}$$

L'inertie s'exprime en termes de la métrique et la matrice de covariance :

$$I_g = tr\,(MV) = tr(VM)$$

Démonstration

$\cdot 1^{ier}$cas :
Si X est centré.

$$\begin{aligned}
I_g &= \sum_{i=1}^{n} p_i e_i^t M e_i \in \mathbb{R} \\
&= tr \left(\sum_{i=1}^{n} p_i e_i^t M e_i \right) = tr \left(M \sum_{i=1}^{n} p_i e_i^t e_i \right) \\
&= tr \left(M \underbrace{X^t D X} \right) = tr\,(MV)
\end{aligned}$$

$\cdot 2^{éme}$cas :
SiX n'est pas centré :c.à.d.$g \neq 0_{\mathbb{R}^p}$

$$\begin{aligned} I_g &= \sum_{i=1}^{n} p_i \left(e_i - g\right)^t M \left(e_i - g\right) \in \mathbb{R} \\ &= tr\left(\sum_{i=1}^{n} p_i \left(e_i - g\right)^t M \left(e_i - g\right)\right) \\ &= \sum_{i=1}^{n} tr\left(p_i \left(e_i - g\right)^t M \left(e_i - g\right)\right) \\ &= \sum_{i=1}^{n} tr\left(M p_i \left(e_i - g\right)^t \left(e_i - g\right)\right) \\ &= tr\left[M \sum_{i=1}^{n} p_i \left(e_i - g\right)^t \left(e_i - g\right)\right] \\ &= tr\left[MV\right] \end{aligned}$$

·Si $M = I$ alors l'inertie est égale à la somme des variances des p variables.
·Si $M = D_{\frac{1}{s^2}}$ alors l'inertie est égale au nombre de variables,Elle ne depend pas de leurs valeurs.

1.3 Nuage des variables

S'interesser aux variables revient à envisager le tableau X (ou$Z$) entant que juxtaposition de colonnes ,Achaque variable est assoicée une suite de n nombres . Selon ce point de vue , une variable peut être présentée comme un vecteur de l'espace vectoriel $\mathbb{R}^n$ dont les dimension sont représentées par les individus.
L'ensemble des p variables fprment un nuage de points dans $\mathbb{R}^n$ appelé nuage de variables.

1.3.1 Liaison

Pour deux variables x_j et $x_{j'}$,on s'interesse au degré de leur liaison qu' on mesure par le coefficient de corrélation linéaire $r_{jj'}$ pour étudier la proximité entre deux variables ,on mumit $\mathbb{R}^n$ de la métrique des poids D, le produit scalaire est définie comme suit :

$$\langle x_j, x_{j'} \rangle := x_j^t D x_{j'}, j; j^{'} = 1, ..., p$$

Cela correspond (dans le cas ou' les variables sont centrées) àla covariance entre x_j et $x_{j'}$,la norme d'une variable centrée x_j est donc égale à son ecart type :

$$\|x_j\|_D = s_j; j = 1, ..., p.$$

1.3.2 Matrice de covariance

La matrice de covariance est une matrice carrée symétrique d'ordre p notée par V :

$$V = \begin{bmatrix} varx_1 & cov\,(x_1, x_2) & \cdots & cov\,(x_1, x_p) \\ cov\,(x_2, x_1) & varx_2 & \cdots & cov\,(x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ cov\,(x_p, x_1) & cov\,(x_p, x_2) & \cdots & varx_p \end{bmatrix}$$

Son terme général dans le cas de poids égaux est :

$$V_{jj'} = cov\left(x_j, x_{j'}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{ij'} - \bar{x}_{j'}\right) = \frac{1}{n}\sum_{i=1}^{n} y_{ij}y_{ij'}/j.j^{'} = 1, 2, ..., p.$$

D'ou' la formule matricielle de V :

$$V = \frac{1}{n}y^t y = \frac{1}{n}x^t x - gg^t.$$

·Dans le cas des poids inégaux on a :

$$V = Y^t DY = X^t DX - gg^t.$$

1.3.3 Matrice de corrélation

La matrice de corrélation est un matrice carrée symétrique d'ordre p notée par R :

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \cdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Ou' :

$$r_{jj'} = r_{j'j} = \frac{s_{jj'}}{s_j s_{j'}}$$

Si on note par $D_{\frac{1}{s}}$ la matrice diagonale des inverses des écarts -types ,

$$D_{\frac{1}{s}} = \begin{bmatrix} \frac{1}{s_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{s_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{s_p} \end{bmatrix}$$

On remarque que :

$$(i) : Z = Y.D_{\frac{1}{s}} et R = D_{\frac{1}{s}} V D_{\frac{1}{s}}.$$
$$(ii) : R = Z^t DZ \text{car} R = D_{\frac{1}{s}} \left(Y^t DY\right)_{D_{\frac{1}{s}}} = \left(YD_{\frac{1}{s}}\right)^t D \left(YD_{\frac{1}{s}}\right) = Z^t DZ.$$

R est la matrice de correlation de x, c'est ausi est la matrice de covariance du tableau standard Z, car

$$s_{z_j} = 1, j = 1, ..., p.$$

1.3.4 Variable engendreé

Soit Δ un axe ,de l'espace $\mathbb{R}^p$,engendré par un vecteur unitaire a ,on projette tous les individus $e_i, i = 1, ..., n$.par cet axe ,la liste des corrdonnées $c_i = \langle a, e_i \rangle_M$ des projections des $e_i$sur Δ forment une nouvelle variables $c = (c_1, .., c_n)^t$ dite variable artificielle qui s'ecrit :

$$c = XU$$

Avec

$$U := Ma \in \mathbb{R}^p$$

Ainsi,la variable c'est une combinaison lineaire des variables initiales x_j, qui sont les colonnes de x

$$c = \sum_{j=1}^{p} u_j x_j$$

Ou u_j est la $j^{éme}$coordonnée de u.

1.4 Analyse en composantes principales

Dans la plupart des situation,on dispose de plusieurs observations sur chaque individu constituant la population d'etude .On a donc à prendre en compte p variables par individu ,p étant strictement supérieur à 2.C'est le rôle de la statistitque multyificatorielle que d'analyse des données dans leur ensemble . L'ACP est une méthode centrale dans l'analyse des données multidmensionnelles , lorsque ces déniéres sont de type numérique (de préférence dans la même unité). On trouve deux approches différentes à l'ACP :
·Elle peut être présentée comme la recherche d'un ensemble réduit de variables non corrélées ,combinasons linéaires des variables initiales résumant avec précision les données (approche anglo -saxonne).
·Une autre interprétation repose sur la représentation des données initiales à l'aide de nuage de points dans un espace géométrique . l'objectif est alors de trouver des sous espaces de dimensions plus faibles (droite ,plan ,...)qui représentent au mieux le nuage initial . C'est cettedernriére appproche que nous aborderons par la suite .

1.4.1 Principe et caracteéristiques

Le principe de l'ACP est d'obtenir une representation approchée du nuage des individus (ou des variables) dans un sous -espace de dimension faible . ceci se fait par projection des individus sur un sous -espace F_k de $\mathbb{R}^p$, de dimension $k \prec p$,choisi de telle sorte que les distances entre les points projrtés ressemblent le plus possible aux distance entre les points du nuage initial ,ce qui ,en d'autre termes revient à deformer le moins possible les distance en projection .

Construction de F_k

Puisque l'operation de projection a pour effet de réduire les distances ,alors F_k doit être telque la moyenne des carré des distance entre les points projetés.
Soit la plus grande possible ,c'est à dire l'inertie du nuage projeté sur F_k doit être maximale ,on désigne par P_k l'opérateur de projection $M-$orthogonale sur F_k ,$P_k$est indemptent et $M-$Symétrique :

$$P_k^2 = P_k \text{et} P_k^t M = M P_k$$

La $m-$orthogonalité des droites $\Delta_1, \Delta_2, ...$ est garantie par le fait que la matrice MV et $M-$symétrique et donc ayont des vecteurs proprs $M-$orthogonaux deux à deux .

Définition 1.4.1 *(axes principaux)*

On appelle un axes principal a_j d'inertie les p vecteurs propres $M-$normés de la matrice VM :

$$VMa_j = \lambda_j a_j \text{et} \|a_j\| = 1, j = 1, ..., p.$$

Définition 1.4.2 *(facteurs principaux)*

Si a_j est un axe principal alors le vecteur

$$u_j := Ma_j$$

Est appelé facteur principal associé.
Les facteurs principaux sont aussi les vecteurs propres $M^{-1}-$normés de la matrice MV :

$$MVu_j = \lambda_j u_j \text{et} \|u_j\|_{M^{-1}} = 1, j = 1, ..., p.$$

Définition 1.4.3 *(composantes principales)*

Si u_j est un facteur principal alors le vecteur

$$c_j := Xu_j$$

Est appelé composent principale associé
Les composantes principales c_j sont des combinaisons linéaires des variables initiales $x_1, ..., x_p$:

$$c_j = \sum_{k=1}^{p} u_{kj} x_k, j = 1, ..., p,$$

Ou'u_{kj} est la $k^{éme}$ coordonnée du facteur principal u_j.
Les composantes principales sont aussi les vecteurs propres de matrice XMX^tD :

$$XMX^tDc_j = \lambda_j c_j, j = 1, ..., p.$$

·c_j est le vecteur de $\mathbb{R}^n$ foemé par les coordonnées c_{ij} des projections $M-$orthogonales des individus e_i sur l'axe $\Delta_j$de vecteur directeur l'axe principal a_j.
·Les composantes principales sont non corrélées entre elle .leurs variances sont égales aux valeurs propres

$$cov\left(c_j, c_{j'}\right) = \lambda_j \delta_{jj'}, j, j = 1, ..., p.$$

Où

$$\delta_{jj'} = 1, j = j^{'} \text{et } 0 \text{ sinon}$$

Est le symbole de `kronecker.`
Preuve Soit X un tableau au centré $(g = 0), c_j$ et$c_{j'}$ $\left(j \neq j^{'} = 1, \cdots, p\right)$ deux composentes principales. ■
1·La moyenne de $c_j$est :

$$\begin{aligned} \overline{c_j} &= c_j^t D 1_n \\ &= \left(X u_j\right)^t D 1_n \\ &= u_j^t \underbrace{X^t D 1_n}_{g} \\ &= 0 \end{aligned}$$

2·La variance de $c_j$est :

$$\begin{aligned} varc_j &= c_j^t D c_j \\ &= \left(X u_j\right)^t D \left(X u_j\right) \\ &= u_j^t X^t D X u_j \\ &= u_j^t S u_j \end{aligned}$$

On à :

$$Su_j = \lambda_j M^{-1} u_j$$

D'ou :

$$\begin{aligned} varc_j &= \lambda_j u_j^t M^{-1} u_j \\ &= \lambda_j \|u_j\|^2_{M-1} \\ &= \lambda_j \end{aligned}$$

3·La covariance entré $c_j$et$c_{j'}$ est :

$$\begin{aligned} cov\left(c_j, c_{j'}\right) &= c_j^t D c_j \\ &= \left(X u_j\right)^t D \left(X u_j\right) \\ &= u_j^t X^t D X u_{j'} \\ &= \langle u_j, u_{j'} \rangle_S \\ &= 0 \end{aligned}$$

1.4.2 ACP sur données centrées reduites

L'utilisation de la métrique $D_{\frac{1}{s^2}}$ sur le tableau X (centré) revient à réduire les variables initiales du tableau X éliminant ainsi l'influence des unités de mesure lorsque les variables sont hétérogénes .En patique ,on travaille sur le tableau standard Z en utilisant la métrique I_p .la matrice de covariance de Z n'est autre que la matrice de corrélation R . Dans ce cas ,les axes principaux et les facteurs principaux sont confondus . Ce sont les vecteurs propre de R rangés selon l'ordre décroissant des valeurs propres

$$Ru_j = \lambda_j u_j \text{ et } c_j = Zu_j$$

Avec

$$\|u_j\| = 1, j = 1, ..., p.$$

1.4.3 Reconstitution des données

Reconstitution du tableau X

Le tableau X est obtenu à partir des facteurs principaux et composantes principales par la formule suivante :

$$X = \sum_{j=1}^{p} c_j u_j^t M^{-1}.$$

Si on se contente de k facteurs ($k \prec p$), on obtient la formule de reconstitution approchée suivante :

$$X \simeq \sum_{j=1}^{k} c_j u_j^t M^{-1}.$$

Reconstitution des matrices MV et VM Les matrice MV et VM sont repectivement obtenues à partir des facteurs et axes principaux comme suit :

$$MV = \sum_{j=1}^{p} \lambda_j u_j u_j^t M^{-1}.\text{et} VM = \sum_{j=1}^{p} \lambda_j a_j a_j^t M^{-1}.$$

Lorsque $M = I_p$,on obtient la matrice de covariance :

$$V = \sum_{j=1}^{p} \lambda_j u_j u_j^t = \sum_{j=1}^{p} \lambda_j a_j a_j^t.$$

1.4.4 Interprétation des résultats de l'ACP

Le but de l'ACP est d'obtenir une représentation des individus (ou variables) dans un sous -espace de dimension k ($k \prec p$) en construisant de nouvelles variables (artificielles) appélées composantes principales . Ces derniéres fournissent une image approchée du nuage des individus .IL y a donc une perte d'information due à la réduction de la dimension de l'espace des individus . Notons que la meilleure représentation plane est obtenue sur le plan ,engendré par les deux premiers axes principaux ,appelé premier plan principal . La question qui se pose est d'apprécier la validité de la représentation de l'ACP en mesurant la qualité de l'approximation pour chacun des points du nuage ainsi que pour l'ensemble du nuage.

Qualité de représentation du nuage

La qualité de représentation $QLT(F_k)$ du nuage projeté sur le sous -espace principal F_k est habituellement mesurée à l'aide de ce que l'on appelle critére du pourcentage d'inertie totale expliquée :

$$QLT(F_k) = \frac{I(F_k)}{I_g} = \frac{\lambda_1 + ... + \lambda_K}{\lambda_1 + ... + \lambda_p}, k \prec p.$$

Plus cette quantité est grande ,plus la représentation est satisfaisante.

Qualité de représentation d'un i individu sur un axe l Notée par $QLT_l(i)$, elle est définie par le rapport :

$$QLT_l(i) := \frac{\text{inertie de la projection de } i \text{ sur } l}{\text{inertie totale de } i}, i = 1, ..., p, \text{et} l = 1, 2, ...$$

C'est le carré du cosinus de l'angle fprmé par le vecteur liant le centre de gravité g à l'individu e_i et l'axe l.
La qualité de représentation $QLT_{ll'}(i)$ d'un individu e_i sur le plan engendré par deux axes l et $l^{'}$ est egale à la somme des qualités de représentation sur chacun des deux axes

$$QLT_{ll'}(i) := QLT_l(i) + QLT_{l'}(i), i = 1, ..., n \text{ et} l \neq l^{'} = 1, 2, ...$$

Contribution d'un individu i à la contruction d'un axe l C'est une quantité comprise entre 0 et 1,donnée par :

$$CTR_l(i) = \frac{p_i c_{il}^2}{\sum_{1=1}^{n} p_i c_{il}^2} = \frac{p_i c_{il}^2}{\lambda_l}, i = 1, ..., n \text{ et} l = 1, 2, ...,$$

Où c_{il} est la valeur de la composante principale l pour l'individu i.
La contribubution d'un individu e_i est considérée importante lorsqu' elle dépasse son poids p_i .La contribution d'un groupe d'individus à la construction d'un axe est égale à la somme des contributions individuelles.

Corrélation des variables L'interprétation des résultats relatifs aux variables s'effectue par le biais des corrélations entre ces variables (variables initiales) et les composantes principales (variables artificielles). Pour chaque variable x_j,on calcule le coefficient $r(x_j, c_l)$ de sa corrélation linéaire avec chaque composants principale c_l . on obtient :

$$r(x_j, c_l) = \sqrt{\lambda_l} u_{jl}, j = 1, ..., p \text{ et } l = 1, 2, ...$$

Où u_{jl} est la $j^{iéme}$ coordonnée du facteur principal u_l assoicé à c_l et λ_l la valeur propre correspondante.
Les corrélation d'une variable $x_j$avec les deux premiéres composantes principales c_1 et c_2 sont exprimées sur une figure appélée `cercle des corrélation.`
C'est un cercle de rayon 1 dont les points ont pour abscisse $r(x_j, c_2)$.

Contribution d'une variable j à la construction d'un axe l

$$CTR_l(x_j) = \frac{r^2(x_j, c_l)}{\sum_{j=1}^{p} r^2(x_j, c_l)} = u_{jl}^2, j = 1, ..., p \text{ et } l = 1, 2, ...$$

La contribuction d'un groupe de variables à un axe est égale à la somme des contribution individuelles .

1.4.5 Individus et variables supplémentaires

IL arive souvent que l'on désire placer des individus ou des variables n'ayant pas participer à l'ACP (appelée individus et variables supplémentaires),sur les grahiques obtenue à la la fin de l'analyse.

Individu supplémentaire Soit $d = (d_1, ..., d_p)^t \in \mathbb{R}^p$ un individu supplémentaire que l'on veut positionner sur un grahique par rapport aux axes principaux . Au moyen des facteur principaux $u_1, ..., u_k$, on calcule les combinisons linéaires $d^t u_1, ..., d^t u_k$ qui sont les coordonnées de ω dans le systéme des axes principaux.

Variable supplémentaire Soit t une variable supplémentaire que l'on veut plcer sur le cercle des corrélations par exemple . Deux cas se présentent ;on bien on connait ses corrdonnées ou bien on connait les coefficients de ses corrélations avec les variables initiales.
(i) si on connait les n coordonnées de la variable t, on détermine ses deux coordonnées dans le cercle par :

$$r\left(t, c_l\right) = \frac{t^t D c_l}{\sqrt{\lambda_l}}, l = 1, 2,$$

(i) si on ne connait pas les coordonnées de t ,alors qu' on connait les coefficients de ses corrélation avec $x_1, ..., x_p$, on détermine ses deux coordonnées dans le cercle par :

$$r\left(t, c_l\right) = \frac{\omega^t u_l}{\sqrt{\lambda_l}}, l = 1, 2,$$

Où ω est vecteur des covariances de t avec les variables x_j.

Chapitre 2

Classification

Comme les autre methode de l'analyse des données,dontelle fait partie,laclassification a pourbut d'obtenir une représentation schématique simple d'un tableau rectangulaire de données dont les colonnes,suivant l'usage,sont des descipteurs de l'ensemble des objects,placées en lignes.
l'objectif le plus simple d'une classification est de répartir l'echantillion en groupes d'ellements homogénes,chaque groupe étant bien differencié des autre le plus souvent,cependant,cet objectif est plus raffiné ;on veut en général obtenir des sections à l'interieur des groupes principaux ,puis des subdivisions plus petites de ces sections ,et ainsi de suite ,En bref ,on desire avoir une hiérarchie,c'est a dire une suite de partitions "emboîteés",de plus en plus finies ,sur l'ensemble d'objectsinitiaux.
On distingue deux grandes familles de techniques de classification :

·**La classification hiérarchique** Pour un niveau de précision donnée,deux individus peuvent être confondus dans un même groupe,alors qu'àun niveau de précision plus elevé,ils séront distinguéset appartiendront à déux sous-groupes différents.

·**La classification non hiérarchique(ou partitionnement)** oboutissant à la décomposition de l'ensemble de tous les individus en k ensembles disjoints ou classes d'equivalence ,le nombre k de classes étant fixé d'avance.
-La classification est étape essentielle du processus de traitement,car elle ensemble les élements similaires en groupes homgénes,correspondant à des régions de l'espace vectoriel contenant le nuage des individus.
La partition en classes des données permet de réduire de taille et la dimension de l'ensemble initial,tout en faisant apparaît une organisation significative.
On classifie les individus ou les variables,mais pas les deux.
Ces groupement sont progressifs,hiérarchisés et sont effectués de maniére progressive par agglomération (classification ascendante),oudégressive par subdivision(classification descendante).
L'un des plus grands classificateurs asans aucun douté,été le savant suédoiscarl linné qui ,au disc-huitiéme siécle,a établi une classification du monde vivant en général et du régne

végétal en particulier,classification encore en vigueur aujoud'hui chez les spécialistes des sciences naturelles.

2.1 Classification hiérarchique

La classification hiérarchique consite à reparttir les n éléments d'un ensemble E en groupes,c'est -à-dire établir une partition de cet ensemble.Différentes contraintes sont bien sûr imposées.Chaque groupe devant être le plus homogéne possible et les groupe devant être les plus distancts possible entre eux.
De plus,on ne se contente pas d'une partition,mais on cherche une hiérarchie de parties qui constituent un arbre binaire appellé `dendogramme` ou `arbre de classification` dont le nombre de noeuds est égal à $n-1$.
On obtient une partition de E en coupant cet arbre selon une ligne horizontale et recueillant les "`morceaux`".Notons que l'arbre de classification peut avoir une forme horizontale.

2.1.1 Tableau des données

La classification peut être réalisée sur :
-Un tableau de valeurs numériques de$n$observations(individus)décrites par$p$caracteres(variable),de la forme suivante :

$$X = (x_{ij})_{\substack{1\leq i\leq n \\ 1\leq j\leq p}}$$

-Un tableau de ‹‹presence -absence››.
-Un tableau carré symétrique de similarités ou de dissimilarités (distances par exemple).

2.1.2 Distance

Une distance d une application definie sur$E \times E$àvaleur dans$\mathbb{R}^+$vérifaiant les proprités de reflexivité,symétrie et l'inégalité triangulaire.pour$(x, y, z) \in E$.

$\cdot d(x, x) = 0$
$\cdot d(x, y) = d(y, x)$ (symétrie)
$\cdot d(x, y) \leq d(x, z) + d(y, z)$ (l'inégalité triangulaire)

Lorsqu' on a seulement$d(x, y) \geq 0, d(x, y) = d(y, x)$,et $d(x, x) = 0$ on dite que d est une dissimilarité.

$$d(x, y) \equiv x = y \Rightarrow d(x, x) = 0$$

-Une distance est une dissimilarité n'est pas forcément une distance.

-Une distance est dite euclidienne quand elle engendrée par produit scalaire :

$$\begin{aligned} d_M^2(x,y) &= \langle x-y, x-y\rangle_M \\ &= (x-y)^t M(x-y) \\ &= \sum_{k=1}^{p} (x-y)^2 \end{aligned}$$

-Lorsque les données se présentent sous la forme d'un tableau X est p caractéres numériques,on utilise souvent une des distances liées aux métriques suivante :
·La distance euclidienne avec$M = D_{\frac{1}{s^2}}$est une diagonale des inverses des variances des p caractéres.
·La distance `Mahalanobis`$M = V^{-1}$est la matrice de covariance du tableau X.

2.1.3 Similarité (Indice de similarité)

Un indice de similarité*s* (ou similarité*s*)définie sur E est une application du produit cartésien$E \times E$dans $\mathbb{R}^+$véifiant pour tout$x, y \in E$:

$$\begin{aligned} &\cdot s(x,y) \geq 0 \\ &\cdot s(x,y) = s(y,x) \text{ et} s(x,x) \geq s(x,y) \end{aligned}$$

2.1.4 Indice de dissimilarité

On définit un indice de dissimilarité sur E comme une application δ de$E \times E$ dans $\mathbb{R}^+$verifiant pour tout x, y dans$E$:

$$\delta(x,y) = \delta(y,x) \text{ et} \delta(x,x) \prec \delta(x,y)$$

Souvent on suppose que

$$s(x,x) = 1$$

Dans ce cas ,on définit la dissimilaritépar :

$$\delta(x,y) := 1 - s(x,y)$$

·Une distance est une dissimilarité ,mais une dissimilarité n'est pas forcément une distance.

Similarité entre objets décrits par des variables binaires

Ce cas trés fréquent conserne des données du type suivant :
Si n (individus) objets sont décrits par la presence ou l'absence de p caractéristiques ,alors plusieurs indices de similarité sont proposés en fonction des quatre nombres suivants associés à un couple (i, j) d'individus .

·a :nombre de caractéristiques communesà i et à j.

·b :nombre de caractéristiques possédées par i et pas par j.

·c :nombre de caractéristiques possédées par$j$etpas par$i$.

·d :nombre de caractéristiques non possédées ni par$i$,ni par j.

Jaccard : $\frac{a}{a+b+c}$

Sokal & **Michener** : $\frac{a+b}{n}$

Sokal &**Sneath** : $\frac{a}{a+2(b+c)}$

Rogers &**Tanimoto** : $\frac{a+d}{a+2(b+c)+d}$

Sorensen : $\frac{2a}{2a+b+c}$

Gower&**Legendre** : $\frac{a-b-c+d}{n}$

Ochiai : $\frac{a}{\sqrt{(a+b)(a+c)}}$

Sokal &**Sneath** : $\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$

Phi de Pearson : $\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$

Les indices sont compris entre zéro et un (0et1) et se transforment en indice de dissimilarité par complémentation à un c'est -a- dire :

$$d(x, y) = 1 - s(x, y)$$

2.2 Hiérarchie :

Définition 2.2.1 *:*

une partition est un sous-ensemble de parties deux disjointes dont la reunion fait l'ensemble tout entier.

$\{E_1, E_2, ..., E_k\}$ partition de$E$.

$$i \neq j \Rightarrow E_i \cap E_j = \varnothing, \cup_{k-1}^{k} E_k = E$$

Une hiérarchie H est une famille de partie de E c'est -à-dire :

$$H \subset P(E) \text{ ou} P(A) = \{A/A \subset E\}.$$

On dit que H est une hiérarchie si :

·E et tous les singlentons appartiennent à H.

$$\forall A, B \in H / A \cap B = \{\varnothing, A, B\}$$

En d'autre termes deux classes sont soit disjointes,soit contenues l'une dans l'autres tous hiérarchie H correspond à un arbre de classification au dendogramme.

$H = \{\varnothing, a, b, c, d, e, cd, cde, bcde, abcde\}$voir figure :

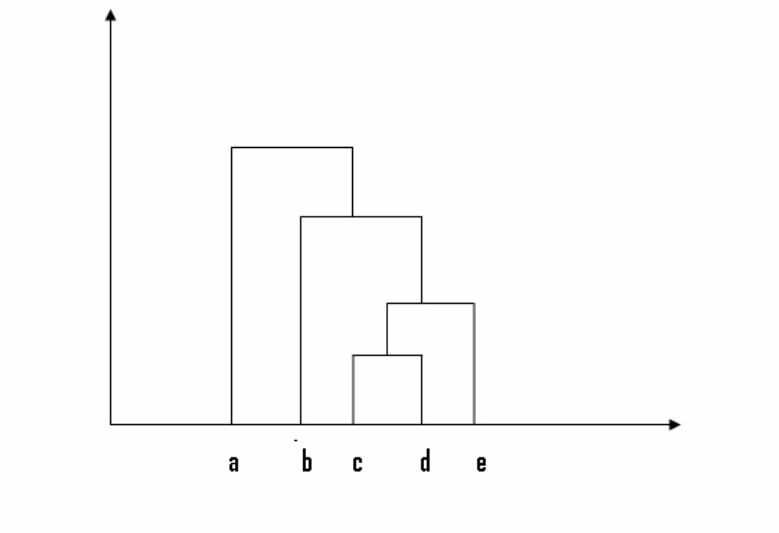


FIG. 2.1 – Dendogramme de l'ensemble H

Cet arbre est obtenu dans la plupart des méthodes de maniére ascendante :on regroupe d'abord les deux individus les plus proches qui forment un "sommet" il ne reste plus que$n-1$objets et onitére le processus jusqu' a regroupement complet.

Une partition de$E$est dite compatible avec une hiérarchie$H$si les classes de$E$sont des l'element de$H$.Graphiquement c'est une partition obtenue en coupant l'arbre de classification de$H$selon une horizontale et en reueillant les morceaux comme l'indice que la figure :

Les différentes partitions de l'ensemble$\{a, b, c, d, e\}$représentées dans la figure sont :

- $p_0 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}\}$correspond à la distance$d = 0$;
- $p_1 = \{\{a\}, \{b\}, \{c, d\}, \{e\}\}$correspond à la distance$d = 1$;
- $p_2 = \{\{a\}, \{b\}, \{c, d, e\}\}$correspond à la distance$d = 2$;
- $p_3 = \{\{a\}, \{b, c, d, e\}\}$correspond à la distance$d = 3$;
- $p_4 = \{\{a, b, c, d, e\}\}$correspond à la distance$d = 4$;

Lorsqu' on peut dire qu' un élement ou une partie A est reliée à B avant quene soit reliée à D,on dit qu' on a affaire a une hierarchie `stratifiée.`

Une hiérarchie est dite indiceé s il existe une partition i de H dans $\mathbb{R}^+$croissante ;telleque :

$$A \subset B \Rightarrow i(A) \leq i(B)$$

L'indice$i(A)$est aussi appelé `niveau d'agrégation de`A,c'est le niveau auquel on trouve agrégé pour la premiére fois tous les constituants de$A$,Ainsi dans la figure :

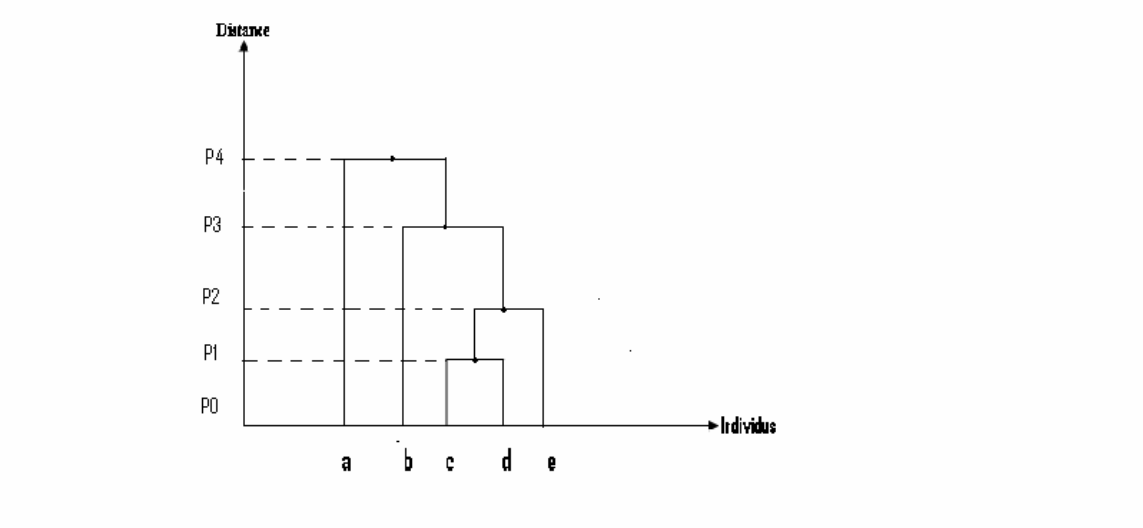


FIG. 2.2 – Partition de l'ensemble H

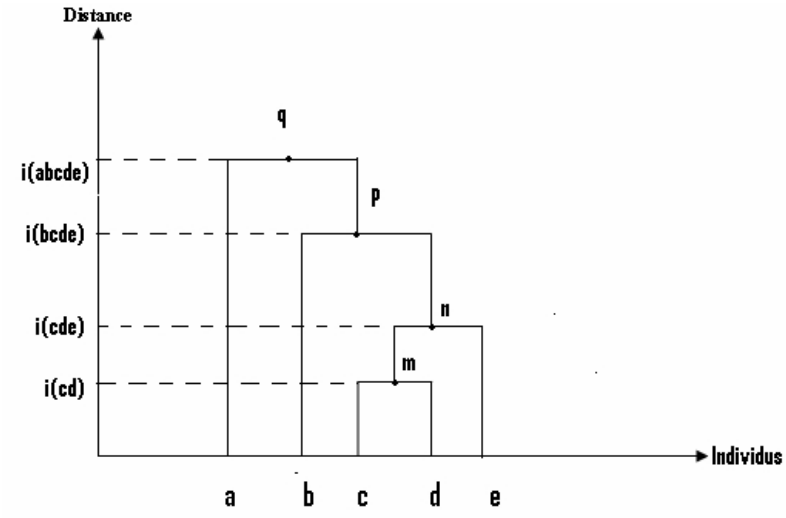


FIG. 2.3 – Indices de dissimilarite de H

Les noeuds(m, n, p, q)symbolisent les diverses subdivision de l'echantillion,les élements de ces subdivisions étant les objets(a, b, c, d, e),placés à l'extremite inférieure des branches chez qui leur sont reliées.
Les niveaux d'agrégation sont en général égaux à l'indice de dissimilarité des deux parties constituant la réunion

$$i\left(cde\right) = \delta\left(\left(cd\right), e\right)$$

2.3 Méthode de classification ascendantes

2.3.1 Principe général des constructions ascendantes

Ces méthodes sont les plus utilisées dans la plupart du temps ,elle partent de la partition triviale de E ou ces éléments(singletons)pour agglutisier(regrouper)les élements puis les classés jusqu'à l'obtention de la partition triviale globale égale à E tout entier :
On regroupe tout d'abord les deux élements,les plus proches qui forment un sommet,il ne reste plus que$(n-1)$objets à classifier.
On i tére(relait)les processus jusqu' à regroupement complet.on peut résumer les étapes de la classification ascendante de n objects comme suit :

·$Etape1$:Commencer par n classés(contrnant chacune un seul objet)et une matrice symétrique de dissimilarité (distance)$D = d_{ij} \in M\left(n\right)$.

·$Etape2$:Regrouper les deux classes les plus proches c'est -à-dire celles correspondants au terme minimum de la matrice$D$,soita, b ses deux classes et $d\left(a, b\right)$leur distance (dissimilarité) on noté par ab la nouvelle classe obtenue aprés regroupement de a et b.

·$Etape3$:Modifier la matrice$D$ en eliminant les lignes etles colonnes des classes a et$b$ et en ajoutant une ligne et une colonne pour les distances entre ab et les classes restantes (autre classes).

·$Etape4$:R épéter les étapes 2 et 3 un nombre de fois égal à $n-1$ pour enfin aboutir à la classe globale E.

2.4 différents algorithmes

IL existe plusieurs algorithmes pour les méthodes de classification ascendantes,ils différent les unes des autre par la maniére de definir la dissimilarité entre deux classes.

2.4.1 Méthode du saut minimum

Cette méthode aussi applée `sous -dominante` (single linkage) consiste à écrire que :

$$d\left(ab, c\right) = \inf\{d\left(a, c\right), d\left(b, c\right)\}\text{ou}d\left(A, B\right) = \min_{i\in A, j\in B} d\left(i, j\right)$$

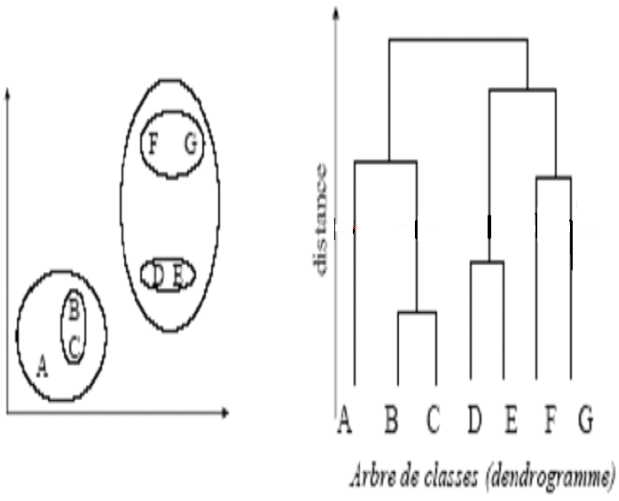


FIG. 2.4 – Méthode classification hiérarchique ascendante

La distance entre deux classes est donc la plus petite distance (voir figure) entre les élements de ces deux classes

2.4.2 Méthode du diamétre (complete linkage)

On prend ici comme distance entre les parties la plus grand distance existant entre les objets de ces classes .c'est -à-dire les voisins les plus éloignés

$$d(A,B) = \max_{i\in A, j\in B} d(i,j) \text{ ou} d(ab,c) = \sup\{d(a,c), d(b,c)\}$$

Elle est aussi appeleé méthode de la `distance maximale`(ou "complet linkage "en anglais)

2.4.3 Méthode de la distance moyenne (average linkage)

La distance entre deux classes$A$et$B$est calculée comme la moyenne des distances entre tous les objets pris dans l'une et l'autre des deux classes.D'ou' l'appellation `"average linkage " en englais`

$$d(A,B) = \frac{1}{n_A n_B} \sum_{i\in A} \sum_{j\in B} d(i,j)$$

Où $n_A$et $n_B$désignent les cardinaux respectifs de$A$et$B$ $(n_A = cardA; n_B = cardB)$.

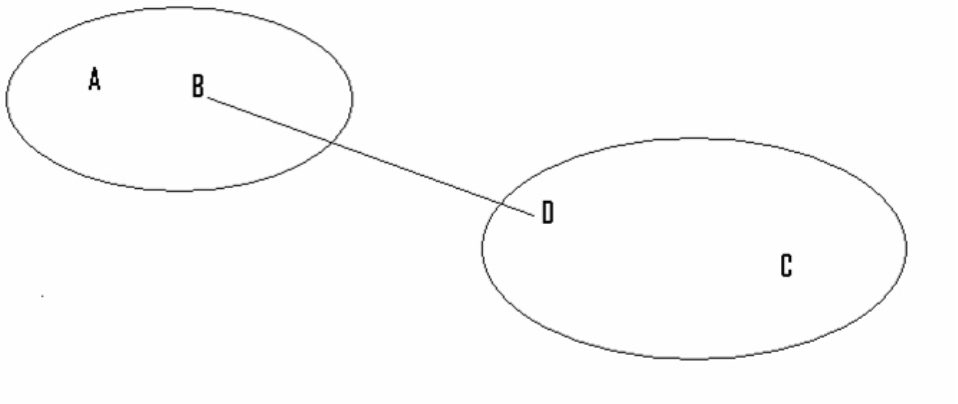


FIG. 2.5 – Méthode de saut minimum

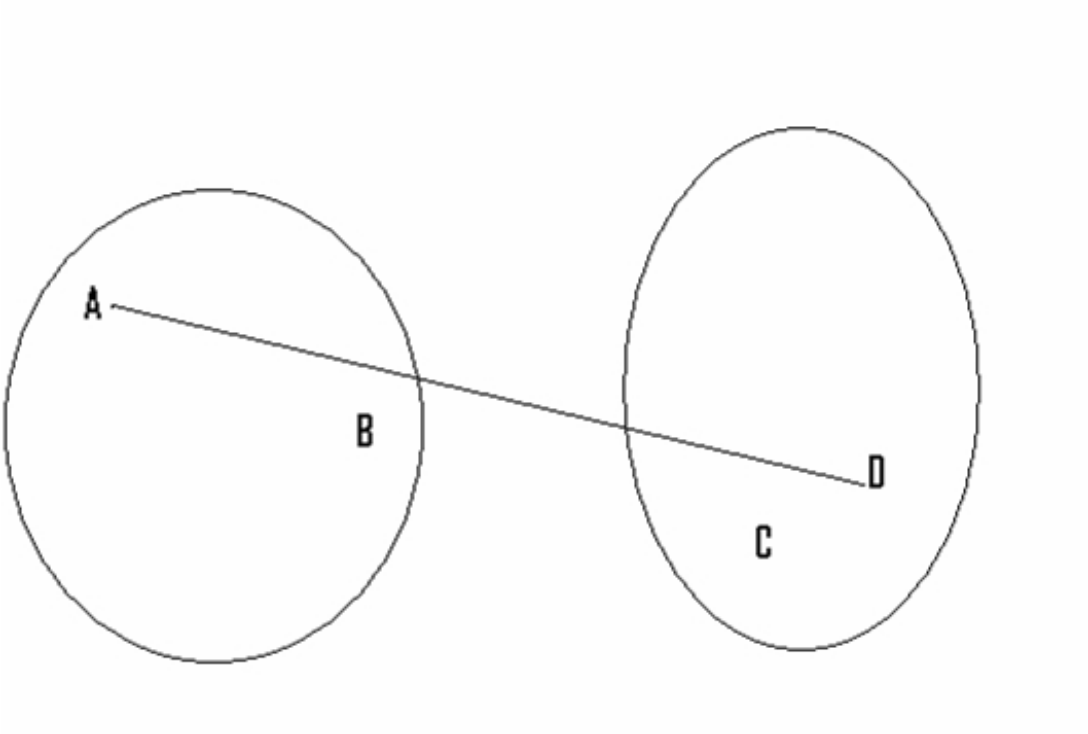


FIG. 2.6 – Méthode de diamétre

2.4.4 méthode de ward

La léthode de ward,égaleement appélée méthode de variance minimale,regroupe à chaque étape les deux classes qui minimisent l'inertié intraclasse.
Ce qui correspond en même temps à un maximun de l'inertié interclasse du fait du théoréme de huyghense qui dit que l'inertié totale est égale à la somme des inerties intraclasses et interclasse.

$$\boldsymbol{Inertiétotale = Inertiéinerterclasse + Inertiéintraclasse}$$

$$I = I_B + I_W$$

·On obtient ainsi une variablité faible à l'intérieur de chaque classe et une variabilité élevée d'une classe à l'autre. Dans cette méthode,l'indice de dissimilarité entre deux classes A et$B$,noté par $\delta(A, B)$,est alors égale à la perte d'inertie interclasse resultant de leur regroupement,cette perte ou variation doit être maximale pour que l'inertie interclasse le soit et par conséquence l'inertié intraclasse soit minimale,dans ce cas le niveau d'agrégation est égale à la perte d'inertié.

Calculons cette perte d'inertié Soit g_A et g_B les centres de gravité des classes A et B respectivement et g_{AB} le centre de gravité de leur réuniou
$(A \cup B)$on a :

$$g_{AB} = \frac{P_A g_A + P_B g_B}{P_A + P_B}$$

Où$P_A$et $P_B$sont les poids des deux classes(voir figure)

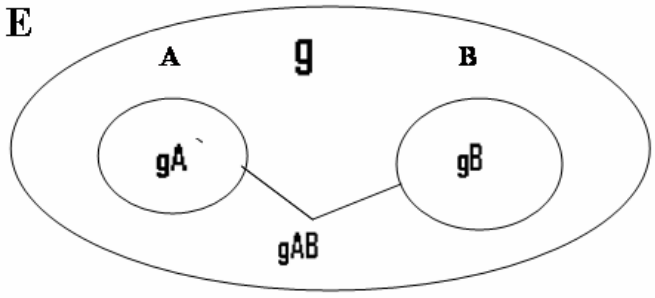


FIG. 2.7 – Méthode de ward

SoitI_1l'inertié avant le regroupement de$A$ et$B$ etI_2l'inertié aprés leur reproupement.
La perte d'inertié est égale à$I_1 - I_2$telleque :

$$I_1 = P_1 d^2 (g_A, g) + P_B d^2 (g_B, g)$$

$$I_2 = P_{A\cup B} d^2 (g_{AB}, g) \, avec P_{A\cup B} = P_A + P_B$$

Et g est le centre de gravité de E alors :

$$I_1 - I_2 = P_A d^2 (g_A, g) + P_B d^2 (g_B, g) - P_{A\cup B} d^2 (g_{AB}, g)$$

·Un calcul élementaire montre que cette variation vaut :

$$\frac{P_A P_B}{P_A + P_B} d^2 (g_A, g_B)$$

(qui est positive).
·Le niveau d'agrégation est égal à la perte d'inertié.
Si on pose :

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

Démontration $I_1 - I_2 = P_A d^2 (g_A, g) + P_B d^2 (g_B, g) - (P_A + P_B) \, d^2 g_{AB}, g$
$= P_A (g_A - g)^2 + P_B (g_B - g)^2 - (P_A + P_B) (g_{AB} - g)^2$
$= P_A (g_A^2 + g^2 - g g_A) + P_B (g_B^2 + g^2 - g g_B) - (P_A + P_B) \left(\frac{P_A g_A + P_B g_B}{P_A + P_B} - g \right)^2$
$= P_A g_A^2 + P_A g^2 - 2 P_A g_A g + P_B g_B^2 + P_B g^2 - 2 P_B g_B g - (P_A + P_B) \left(\frac{P_A g_A + P_B g_B}{P_A + P_B} \right)^2 +$
$2 (P_A + P_B) \frac{P_A g_A + P_B g_B}{P_A + P_B} g - (P_A + P_B)^2$
$= P_A g_A^2 + P_A g^2 - 2 P_A g_A g + P_B g_B^2 + P_B g^2 - 2 P_B g_B g - \frac{(P_A g_A + P_A g_B)^2}{P_A + P_B} + 2 P_A g_A g +$
$2 P_A g_B g - P_A g^2 - P_B g^2$
$= P_A g_A^2 + P_A g_B^2 - \frac{(P_A g_A + P_B g_B)^2}{P_A + P_B}$
$= P_A g_A^2 + P_B g_B^2 - \frac{P_A^2 g_A^2 + P_B^2 g_B^2 + 2 P_A P_B g_A g_B}{P_A + P_B}$
$= \frac{(P_A + P_B)\left(P_A g_A^2 + P_B g_B^2 \right) - P_A^2 g_A^2 - P_B^2 g_B^2 - 2 P_A P_B g_A g_B}{P_A + PB}$
$= \frac{\left(P_A^2 g_A^2 + P_A P_B g_A^2 + P_A P_B g_A^2 + P_B^2 g_B^2 \right) - P_A^2 g_A^2 - P_B^2 g_B^2 - 2 P_A P_B g_A g_B}{P_A + P_B}$
$= \frac{P_A P_B g_A^2 + P_A P_B g_B^2 - 2 P_A P_B g_A g_B}{P_A + P_B}$
$= \frac{P_A P_B (g_A - g_B)^2}{P_A + P_B} = \frac{P_A P_B}{P_A + P_B} d^2 (g_A, g_B)$

$$I_1 - I_2 = \frac{P_A P_B}{P_A + P_B} d^2 (g_A, g_B)$$

On agrége les classe A, B pour les quel$\delta (A, B)$est minimale.
1−Si$A = \{a\}$et$B = \{b\}$alors :

$$\delta(a,b) = \frac{P_a P_b}{P_a + P_b} d^2(a,b).$$

2−Si$A = \{a,b\}$et$B = \{c\}$alors :

$$\delta(ab,c) = \frac{(P_a + P_c)\,\delta(a,c) + (P_b + P_c)\,\delta(b,c) - P_c \delta(a,b)}{P_a + P_b + P_c}$$

S'appelle la formule de lance et wilioms.

3−Le poids d'une classe c est égale àlasomme des poids de ses élements :

$$Pc = \sum_{e_i \in c} Pe_i$$

4−La somme des niveaux d'agrégation des différents noeuds de l'arbre de classification dont le nombre est égale à $n-1$doit être égale à l'inertié totale du nuage E puisque la somme des pertes d'inertié est égale à l'inertié totale.

$$\sum_{i=1}^{n-1} \delta_i = Inertiétotale\ dunuageE.$$

5−La méthode de ward s'applique dans le cas de distance euclidienne.

2.4.5 Methode de van den driessche

Elle est basée sur la miniùisation de la dispersion d'une classe.
La dispersion d'une classe A est égale à la moyenne des distances internes de la classe :

$$D(A) = \frac{\sum_{i=1}^{L} \sum_{j=1}^{L} d(i,j)}{\frac{L(L-1)}{2}}, L = cardA$$

2.5 Méthode de classification descendantes

La classification hiérarchique descendantes procéde par divisions successives de l'ensemble des objets ,Elle considére l'ensemble de donnnées comme un gros cluster unique,et le scinde en deux cluster "descendants".
La scission s'opére de façon à ce que les distance entre les deux descendantes soient la plus grande possible ,de façon à créer deux cluster bien séparées.cette procédure est ensuite appliquée à chacun des descendants (prpcédure récursive) jus'qu' à ce qu' il ne reste plus que des clusters ne contenant qu' une seule observation (singletons).ces méthodes sont partiquement inutilisées,la méthode la plus comme pour ce type de classification est de Williams et Lambert.

2.6 Classification des variables

La classification des variables est un peu délicaté àcause de la nature pouvant être trée différentes de ses variable d'ou la difficulte de définir une dissimilarité pour les variabe(quantitatifs)de même nature on utilise de façon naturelle$1-r$comme indice de dissimilarité.
$1-r$`Pearson :`calculé à partir du coefficient de correlation entre deux variable$r$,sans doute à l'aide de la formule

$$\delta(x,y) = 1 - r(x,y)$$

2.7 Classification non hiérarchique

La classification non hiérarchique est conçue pour grouper des individus plutôt que des variable.on un certain nombre de classes K telque $2 \leq K \leq n-1$.
On dispose d'un ensemble E de n individus,décrit par p variable statistiques.
Ces individus sont considérons comme des points de $\mathbb{R}^p$munis d'une distance euclidienne$d$,les methodes de classification non hiérarchique consistent à chercher directement une partition de E en K classes$E_1, E_2, ..., E_K$
La classification non hiérarchique permet de traitet assez rapidement un nombre d'objets tellement grand que la classification hiérarchique devient difficile sinon impossible à réaliser.

2.7.1 Critére de la classification

La classification non hiérarchique est basée sur l'optimisation locale de l'inertie
On cherche la partition qui rend l'inertié intraclasseI_W minimale afin d'avoir des classes bien homogénes ce qui en traîne(en vertu du théoréme de huyghens),une inertié interclasse$I_B$maximale et aboutissant ainsi à des classes bien distances :

$$I_{totale} = I_W + I_B$$

Calcul des inertié

En designant par :
·P_i :Le poids de l'individue$_i, i = 1, 2, ..., n$
·q_j :Le poids de la classeE_j :

$$q_j = \sum_{e_i \in E_I} P_i, j = 1, 2, ..., K$$

·g_j :Le centre de gravité de la classeE_j. :

$$g_j = \frac{1}{q_j} \sum_{e_i \in E_i} P_i e_i$$

·g :Le centre de gravité de l'ensemble$E$:

$$g = \sum_{i=1}^{n} P_i e = \sum_{j=1}^{k} q_j g_j.$$

·I_j :L'inertié interne àE_j. :

$$I_j = \frac{1}{q_j} \sum_{e_i \in E_j} P_i d^2 (e_i, g_j)$$

·I_B;L'inertié interclasse :

$$I_B = \sum_{j=1}^{k} q_j d^2 (g_j, g)$$

$$I_B = trace\,(B) \text{ où } B = \frac{1}{n} \sum_{j=1}^{k} n g_j g_j^t$$

mesure la séparation des classes.

·I_W :L'inertié intraclasse

$$I_W = \sum_{j=1}^{k} q_j I_j.$$

$$I_W = trace\,(W) \text{ où} W = \frac{1}{n} \sum_{j=1}^{k} n_j v_j$$

mesure l'homogéneité des classes.

·I_{totale} :L'inertié totale :

$$I = \sum_{i=1}^{n} p_i d^2 (e_i, g)$$

2.8 Différents algorithmes

2.8.1 Méthode des centres mobiles (Forgy)

Due à forgy en1973,cette méthode consiste à partir k points$(c_1, c_2, ..., c_k)$pris au hasard parmi les n élements de E,ces points$c_j$sont applés centres :

*Etape*1 :On désigne parE_{cj}ensemble des points de E les plus proches du centrec_j,$j = 1, 2, ..., k$ que tout autre centre,on dorme ainsi une partition E en k classesE_{c1}, E_{c2}, ..., E_{ck}.

*Etape*2 :On remplace les centrec_1, c_2, ..., c_k par les centres de gravitésg_1, g_2, ..., $g_k$de ces classes,pour fprmer une nouvelle partition de E en k classesE_{g1}, E_{g2}, ..., E_{gk},oùE_{gj}est ensemble des points de E les plus proches du centre de gravité$g_j$que tout autre centre de gravité.

*Etape*3 :On recommence le même processus jusqu' à la stabilisation où il n'y aur à plus d'affectation possible ,l'algorithme convergeassez rapidement vers la partition optimale c'est -à-dire celle qui minimiseI_W.
L'experience montre que le nombre d'iteration est faible et si aucours de cette itération une classe se vide alors on tire au hasard un nouveau centre.

2.8.2 Méthode de $k-$means (Mac Queen)

Ce type d'algorithme est basé sur la méthode des centroides (centre de gravité) :

1−On prend comme centre d'agregations les k premiére élements qui se presentent et on constitue une classe avec chacun d'eux.

2−Parmi les $n-k$ autre élement ,on prend ,le premiér qui se presente et on le réunit au centre le plus proche.

On remplace ce dernier par le centre de gravité des deux points réunis .On prend,parmi $n-k-1$autres éléments ,le premier qui se présente et on l'affecte au centre le plus proche.On remplace ce dernier par le centre de gravité de la classe foormée...et ainsi de suite jusqu' à ce que le dernier élément soit affectée.

On obtient ainsi une partition P.

3−On agglomére les points autour des centres de gravité des classes de$P$,ce qui fournit directement la partition finale de E en k classes.
La différence entre la méthode des $k-$means et celle des centres mobiles réside dans le fait que pour la méthode des $k-$means les centres de gravités sont récalculés aprés l'affectation de chaque object.alors que pour la méthode.

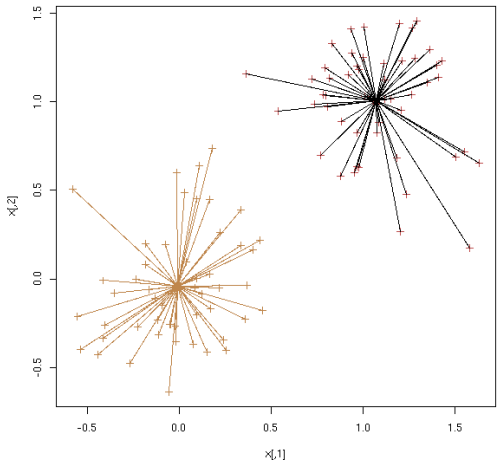


FIG. 2.8 – Méthode K-means

Chapitre 3

Application

L'analyse des données est une étude statistique qui contient plusieurs méthodes qui recherche dans les liaisons entre les variable statistique, la répresentetion graphique des ces données, dans ce travaille en peut donnée quelque défnition et qulques étapes pour fait cetté etude.

A la n en l.applique dans un exemple qui réalise classificationdes températeursmensuelles de 15 villes de algerie on utilise la logicielle R

3.1 Présentation des données

On designe par x la variable qualitative qui il répresentant la témperateur de pays algier par le méthode de classification

3.2 Statistique élémentaires

Cette section a pour but d'illustrer quelques concepts fondamentaux de la statistique inf erentielle,et de presenter les principales fonctions de R pour le traitement statistique des données recueillieslors d'un protocole experimental.

Moyenne La moyenne arithmétique d'une variable statistique X ,que l'on note par $\overline{X}$,est la somme des valeurs prises par cette variable,divisée par le nombre
La fonction summary() permet d'obtenir un r esum e statistique el ementaire du jeu de donnees.n d'observation $x_1, ..., x_n$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Médiane La médiane que l'on note M_e correspond à la valeur de la variable statistique qui partage la population en deux parties égales ,les valeurs de la série statistique étant

Ville	Jan	Fev	Mar	Apr	Mai	Jun	Jul	Aou	**Sep**	**Oct**	**Nov**	Dec
Alger	12.2	12.8	14.2	16.4	18.9	21.9	24.7	25.5	23.9	20.3	16.1	13
Annba	11	11.5	12.5	14.6	17.8	21.2	24.6	24.9	22.8	19.6	15.4	12.2
Bachar	9.7	12.6	16	19.7	24.4	29.6	33.5	33	28	21.2	14.9	10
Béjaia	11.3	11.4	12.5	14.6	17.4	20.8	23.9	25.4	22.8	18.6	15.4	11.8
Biskra	11.6	13.3	16.1	20.2	24.9	30.1	33.4	32.5	27.6	22.1	16.3	12.2
Constantine	6.9	7.2	9	11.9	16.4	21	24.4	25.1	21.1	15.8	11.2	7.4
Djelfa	4.2	5.9	9	12.6	16.7	21.6	25.6	25.1	20.6	14.9	9.5	5.8
In-Salah	14.5	16.9	20.7	25.1	30.6	35.6	36.9	36.5	32.9	26.7	20.1	14.1
Medea	6.7	6.7	9.8	12.3	15.4	20.4	24.6	24.4	20.8	15.4	11	6.4
Oran	10.2	11	13.3	15.4	18.3	21.8	24.5	25.1	22.9	18.4	14.2	11.1
Ouargla	11.1	13.6	16.9	21	25.8	31.2	33.7	34	28.6	22.2	15.8	11.1
Saida	7.8	9.1	11.1	14	17.8	22.2	27.2	27.3	23.1	17.6	11.8	8.4
Setif	4.7	6.5	8.5	11.7	15.7	20.5	24.6	24.2	20.5	14.7	9.2	5.6
Tamanrasset	12.9	15.4	19.3	22.4	26.5	29.2	28.3	28.4	26.8	22.6	17.7	13.8
Tlemcen	9	9.6	11.6	14.2	16.8	21.3	24.8	26	22.3	17.9	13.1	10

TAB. 3.1 – Table Temperateur

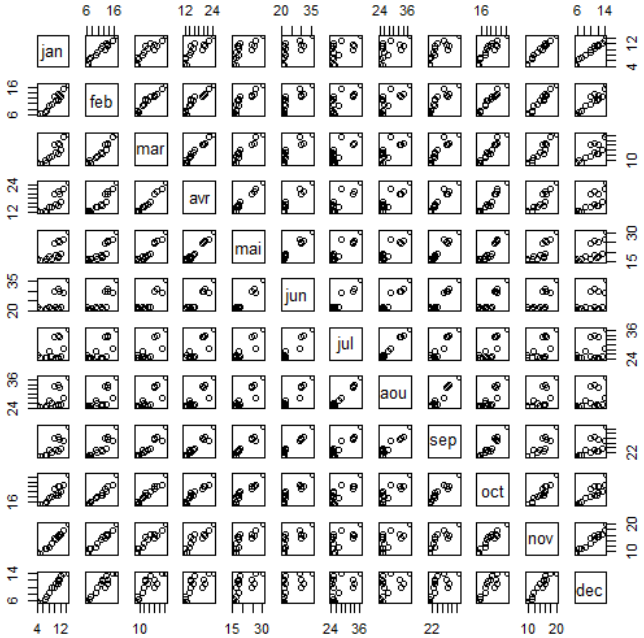


FIG. 3.1 – répesentation de table (x)

	Min	**1st Qu**	Median	Mean	3rd Qu	Max
Jan	4.200	7.350	10.200	9.587	11.450	14.5
Fev	5.90	8.15	11.40	10.90	13.05	16.90
Mar	8.50	10.45	12.50	13.37	16.05	20.70
Avr	11.70	13.30	14.60	16.41	19.95	25.10
Mai	15.40	16.75	17.80	20.23	24.65	30.60
Jun	20.40	21.10	21.80	24.56	29.40	35.60
Jul	23.90	24.60	24.80	27.65	30.85	36.90
Aou	24.20	25.10	25.50	27.85	30.45	36.90
Sep	20.50	21.70	22.90	24.31	27.20	32.90
Oct	14.70	16.70	18.60	19.20	21.65	26.70
Nov	9.20	11.50	14.90	14.11	15.95	20.10
Dec	5.60	7.90	11.10	10.19	12.20	14.10

TAB. 3.2 – Table Summmary (x)

Jan	Fev	Mar	Avr	Mai	Jun	Jul	Aou	Sep	Oct
2.995205	3.349200	3.787888	4.224498	4.819079	5.036552	4.420386	4.140715	3.644539	3.386739

TAB. 3.3 – Table Sd(x)

au préalable classées dans l'ordre croissant ou décroissant.

Quartiles Les quartiles divient l'effectif de la série ,préalablement ordonnée par ordre croissant ,en quatre parties égales.

·1^{er} quartileQ_1est tel que 25% des observations lui sont inférieures et 75% lui sont supérieures.

·2^{e}quartile Q_2est la mediane .Cette deriére apparaît donc aussi comme une caractéristique de position.

·3^{e}quartile Q_3est tel que 75% des observation lui sont inférieures et 25% lui sont supérieures.

De la même maniére ,on défint les déciles et perciles en partageant les observation ordonnées en 10 et 100 parties égales respectivement.

Le figure de scale (x) :

Le figure de moyenne :

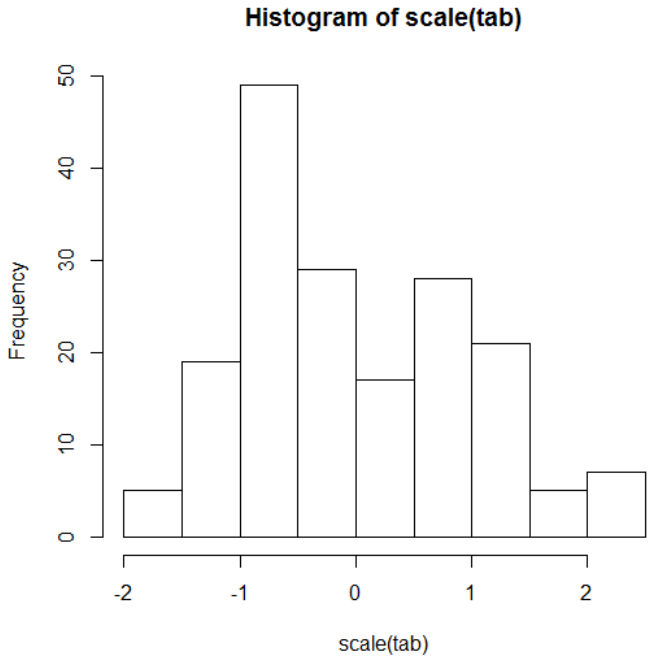


FIG. 3.2 – le scale de table température

3.3 Acp

Tableau de données

Ville	Jan	Fev	Mar	Apr	Mai	Jun	Jul	Aou	**Sep**	**Oct**	**Nov**	Dec
Alger	12.2	12.8	14.2	16.4	18.9	21.9	24.7	25.5	23.9	20.3	16.1	13
Annba	11	11.5	12.5	14.6	17.8	21.2	24.6	24.9	22.8	19.6	15.4	12.2
Bachar	9.7	12.6	16	19.7	24.4	29.6	33.5	33	28	21.2	14.9	10
Béjaia	11.3	11.4	12.5	14.6	17.4	20.8	23.9	25.4	22.8	18.6	15.4	11.8
Biskra	11.6	13.3	16.1	20.2	24.9	30.1	33.4	32.5	27.6	22.1	16.3	12.2
Constantine	6.9	7.2	9	11.9	16.4	21	24.4	25.1	21.1	15.8	11.2	7.4
Djelfa	4.2	5.9	9	12.6	16.7	21.6	25.6	25.1	20.6	14.9	9.5	5.8
In-Salah	14.5	16.9	20.7	25.1	30.6	35.6	36.9	36.5	32.9	26.7	20.1	14.1
Medea	6.7	6.7	9.8	12.3	15.4	20.4	24.6	24.4	20.8	15.4	11	6.4
Oran	10.2	11	13.3	15.4	18.3	21.8	24.5	25.1	22.9	18.4	14.2	11.1
Ouargla	11.1	13.6	16.9	21	25.8	31.2	33.7	34	28.6	22.2	15.8	11.1
Saida	7.8	9.1	11.1	14	17.8	22.2	27.2	27.3	23.1	17.6	11.8	8.4
Setif	4.7	6.5	8.5	11.7	15.7	20.5	24.6	24.2	20.5	14.7	9.2	5.6
Tamanrasset	12.9	15.4	19.3	22.4	26.5	29.2	28.3	28.4	26.8	22.6	17.7	13.8
Tlemcen	9	9.6	11.6	14.2	16.8	21.3	24.8	26	22.3	17.9	13.1	10

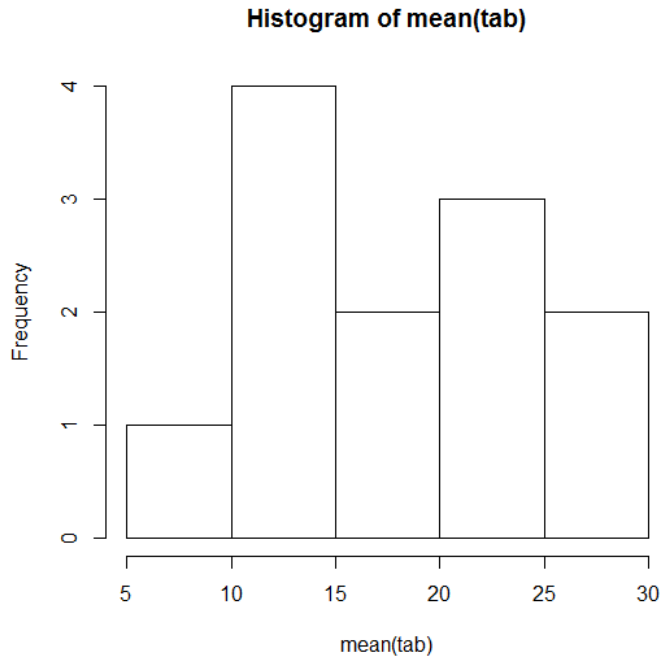


FIG. 3.3 – La moyenne de températeur

3.3.1 Nuage de variable

Matrice covariance

La matrice de covariance est une matrice carrée symétrique d'ordre p notée par V :

$$V = \frac{1}{n} y^t y = \frac{1}{n} x^t x - g g^t$$

La historgramme de matrice covariance

Matrice de corrélation

La matrice de corrélation est un matrice carrée symétrique d'ordre p notée par R :

$$r_{jj'} = r_{j'j} = \frac{s_{jj'}}{s_j s_{j'}}$$

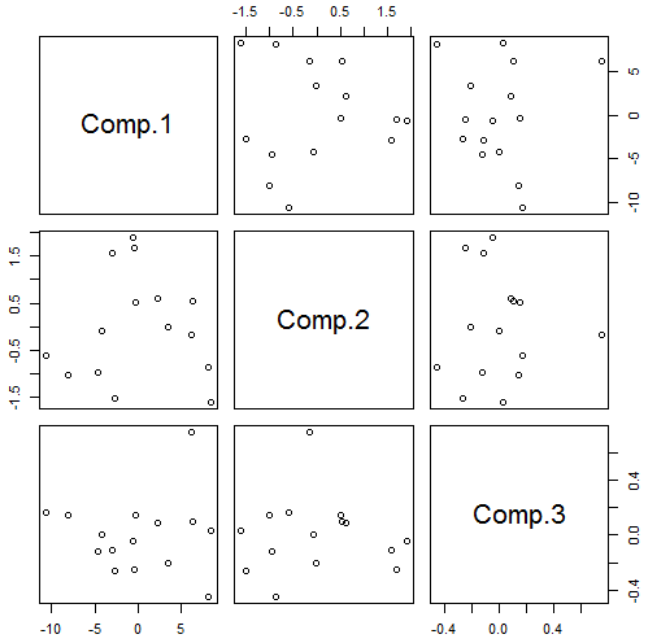


FIG. 3.4 – Tableau de données de ACP

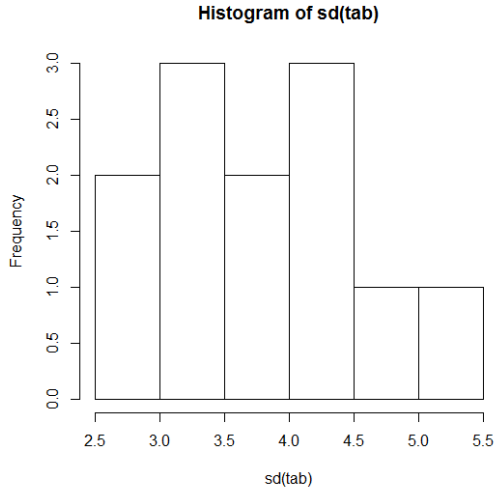


FIG. 3.5 – Historgramme de matrice covariance

3.4 Classification

3.4.1 Classification hiérarchique

Le but de cette méthode est de construire une partition d'un ensemble d'objets en classes de moins en moins fines par regroupement successif des parties de cet ensemble.Cette classification est représentée par un arbre de classification ou dendogramme qui peut être présenté de manières ascendantes ou descendantes

Methode du saut minimum

Dans cette méthode ,La distance entre deux classes est définie comme étant la plus petite distance existant entre leurs éléments pris deux à deux .
La figure de méthode saut minimum

Méthode du diamétre

On prend ici comme distance entre les parties la plus grand distance existant entre les objets de ces classes .c'est -à-dire les voisins les plus éloignés
Le figure de méthode du diamétre

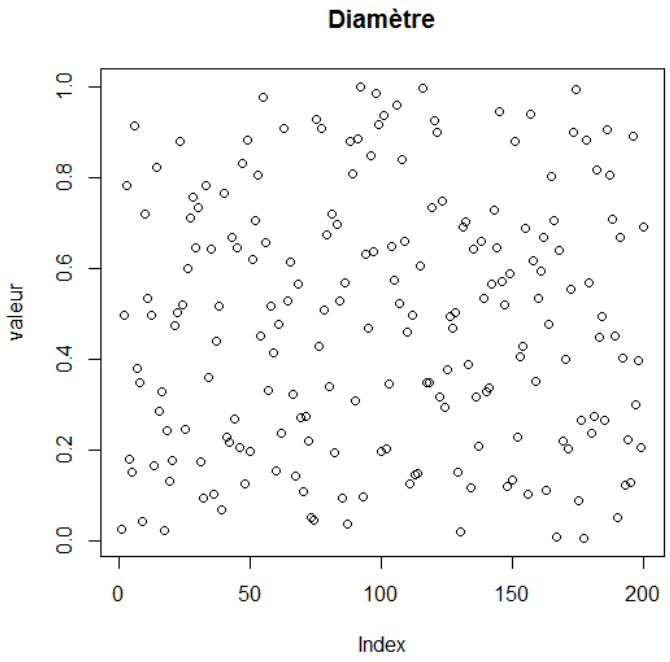


FIG. 3.6 – Méthode diamétre

Méthode du ward

La méthode de ward,égaleement appélée méthode de variance minimale,regroupe à chaque étape les deux classes qui minimisent l'inertié intraclasse

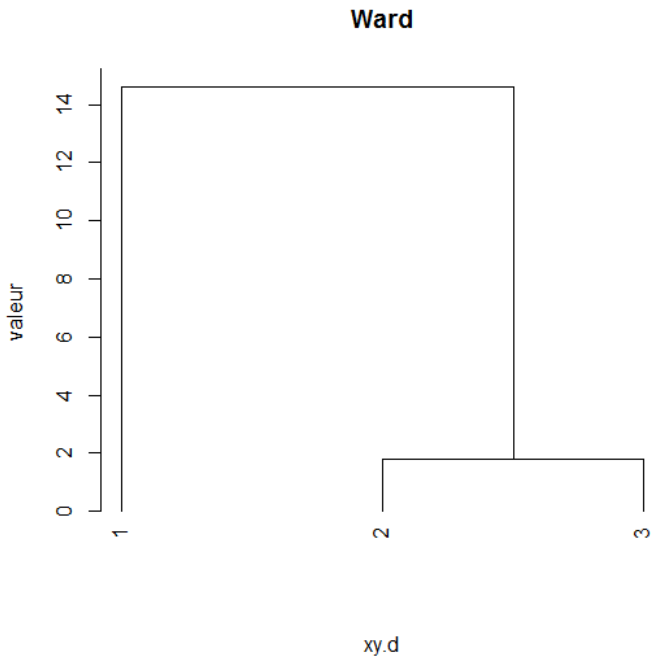
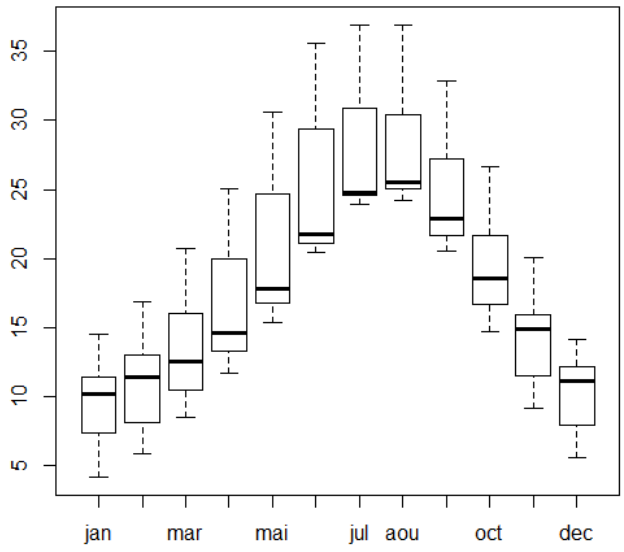


FIG. 3.7 – Méthode du ward de table témperateur

3.4.2 Classification non hiérarchique

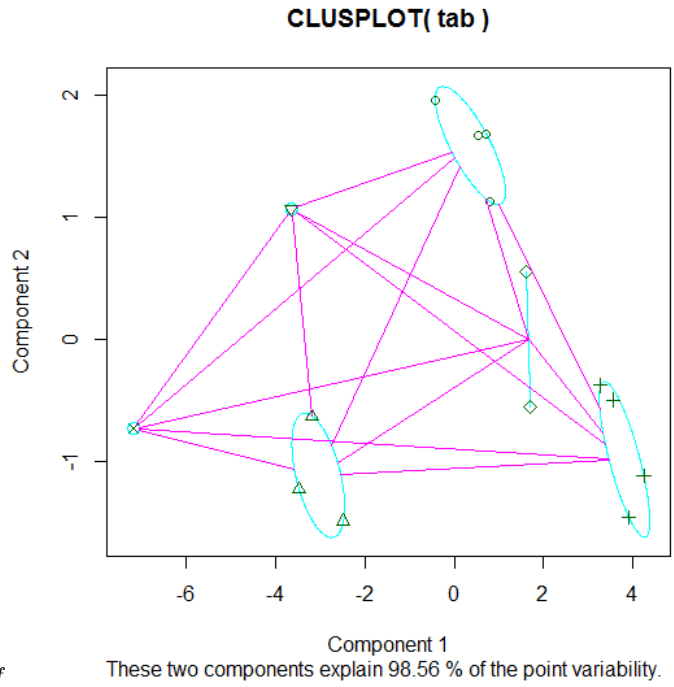
Methode de centre mobiles

(compete)



$_1$7.*pdf*
Cluster (complete) de tmprateur

clust



$_{1}8.pdf$

Mthode complete (variance)

Méthode de K-means

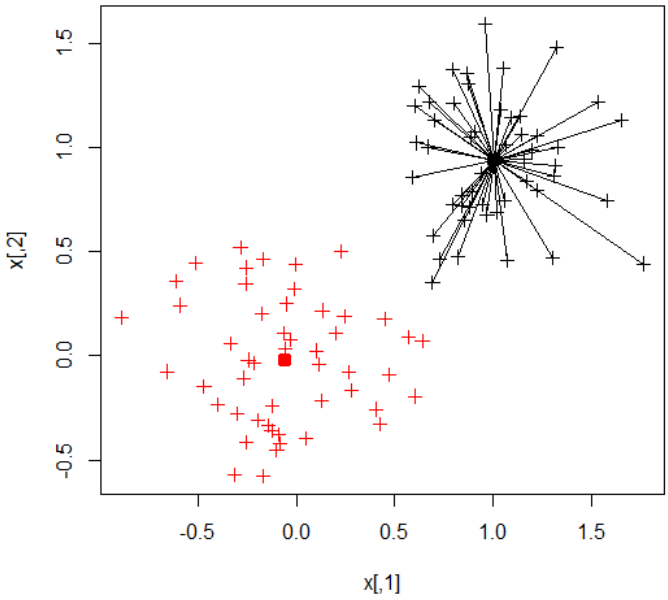


FIG. 3.8 – Méthode de K-means de la témpérateur

Conclusion

Durant cette étude sur des données réelles,on a pu vérifier certains résultats théoriques relatifs aux méthodes de classification et de discrimination. Ainsi on a constaté que :

• Les différentes méthodes de classification hiérarchique ascendante sur les individus donnent des dendogrammes incompréhensibles vu le nombre élevé d'individus.

Cependant la dichotomie obtenue par la méthode de Ward parait plus proche de la réalité que celles des autres algorithmes.

• Pour les variables dont le nombre n'est pas très grand, l'arbre de classification est facilement lisible et montre des agrégations en relation avec la nature des variables.

• Les algorithmes de classification non hiérarchique (nuées dynamiques entre autres) permettent de résoudre le problème dû au nombre élevé de victimes et donnent des résultats très satisfaisants.

Bibliographie

[1] de Lagarde, J. (1983). Initiation à l'Analyse des Données. Dunod, Paris.

[2] Dettkka Samah,Benelmir Imane(Juin 2007)Classification et Discrimation etude D'un cas reel :infarctus du MYOCARDE,Unite de Biskra Diplome d'ingenieur D'etat.

[3] Koné, Z.B., Traoré, S.A. (2008). Analyse des Donnèes d'Achat et Vente de Médicaments par l'Entreprise DIGROMED, Unité de Biskra. Mémoire de Fin d'Etudes d'Ingénieur en Statistique (Encadreur :Meraghni,D.). Université Mohamed Khider, Biskra.

[4] Escfier, B., Pagès, J. (1998). Analyses Factorielles Simples et Multiples.Dunod, Paris.Saporta, G. (1990). Probabilités analyse des données et statistique

[5] Saporta, G. (1990). Probabilités analyse des données et statistique

Annexe : Abréviations et Notations

Les diférentes abréviations et notations utilisées dans ce mémoire sont :

Cor	corrélation
Cov	covariance
CTR	contribution
QLT	qualite de représentation
$tr(.)$	trace d'une matrice
Var	variance
$d_M(.;.)$	distance par rapporta une métrique M
A^t	matrice transposée
$\overline{x}$	moyenne arithmétique
$\sum$	somme
$\|.\|_M$	norme par rapport à une métrique M
$\langle .;. \rangle$	Produit scalaire par rapport à une métrique M

ملخـــص

هذه المذ كرة مخصصة لدراسة طريقة التصنيفات، وهي طريقة إحصائية استكشافية تسمح بالحصول على تمثيل بياني وتهدف إلى صغير حجم مجموعة الأفراد بشكل مجموعات متجانسة.

Résumé

Ce mémoire est consacre à l'étude de méthode de classification, c'est une méthode statistique exploratoire qui permet d'obtenir une représentation graphique et visant à réduit la taille de de l’ensemble des individus en formant des groups homogènes.