I. Analyse de données

I.1. Introduction

L'analyse de données est un ensemble plus ou moins défini de méthodes statistiques. Ces méthodes permettent de collecter, organiser, résumer, présenter et étudier des ***données*** pour permettre d'en tirer des conclusions et de prendre des décisions. Mais à quoi donc servent toutes ces données ? Les données servent à obtenir de l'information, et l'information sert à décider, à agir.

Exemples :

1 – Le médecin analyse les données d'un patient pour effectuer un diagnostic et établir une ordonnance

2 – le politique analyse les données économiques pour connaître la situation et décider d'actions

3 - le qualiticien analyse les données d'un produit pour le tester et établir un plan d'amélioration de la qualité

4 – le gestionnaire analyse les données comptables pour connaître l'état financier de son entreprise et pour proposer, par exemple, des réductions de dépenses,….

Nous ne pouvons donc pas échapper aux données. Mais pour passer des données aux informations et de l'information à la décision, il faut de la méthode.

La première étape dans une analyse de données est essentiellement la définition de la population ou des individus à étudier. Ces individus sont décrits par des caractères ou variables. Ces individus et variables sont souvent sous forme de tableau ou matrice.
Pour un problème donné l'utilisateur doit déterminer les individus, les variables, les types associés à chaque variable, leur codage, ….

Données ➔ (utilisation des méthodes d'analyse de données) ➔ Résultats

I.2 Les variables et les individus

I.2.1 - Les variables : Définition :

« Toute caractéristique d'une personne ou d'une chose qui peut être exprimée par un nombre est appelée ***variable***. La valeur de la *variable* est le nombre réel qui décrit une personne ou une chose particulière ».

I.2.1.2 - Les types de variables :

Une fois les variables choisies, il faut leur associer un "type". On distingue deux grands types de variables:

variables quantitatives	variables qualitatives

I.2.1.2.1 - Les variables quantitatives

Une variable quantitative prend des valeurs pour lesquelles des opérations arithmétiques telles que différence et moyenne aient un sens. Dans la pratique on distingue les types suivants :

Quantitatif	Exemple
Discrète	Nombre d'enfants, Nombre de diplômes…
Continue	Poids, Température, Fréquence d'un signal, Amplitude d'un bruit thermique, Valeur boursière,...

I.2.1.2.2 - Les variables qualitatives (ou *variable de catégories*)

Une variable qualitative prend des valeurs symboliques qui désignent en fait des catégories. On distingue essentiellement les types suivants:

Qualitatif	**Exemple**
Catégorielle (Ou Nominal)	lieu géographique, catégorie socioprofessionnelle, Sexe, Nationalité, Contrôle qualitative d'une pièce, Situation de famille…
Ordinal	pas d'accord, sans opinion, Tout jugement qualitative, Mention à un examen …
Textuel	titre de film, nom d'auteur, …

I.2.2- Les individus ou entités

Dans la définition de la variable, il est dit : « Toute caractéristique d'une personne ou d'une chose qui … ». La personne ou la chose mentionnée ici est un individu (ou une entité). Cet individu appartient à une population de référence (définie par les variables dites de contrôle). Si, au lieu d'étudier toute la population, on n'en examine qu'une partie, on dit qu'on étudie un ***échantillon*** et, si cet échantillon est un modèle réduit de la population entière, on dit qu'on étudie un ***échantillon représentatif***.

Exemples :

(a) A partir du recensement de la population, on établit des échantillons représentatifs de la population sur lesquels seront ultérieurement effectués des sondages.

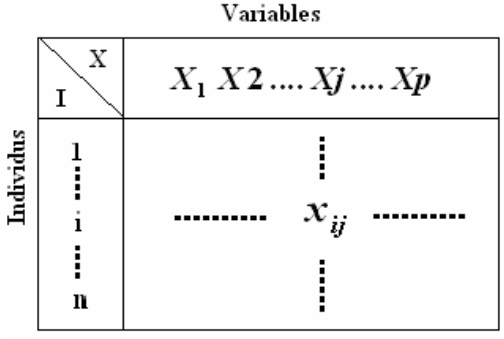
(b) A partir de l'ensemble des abonnés du téléphone, on établit des échantillons représentatifs de la population sur lesquels seront ensuite effectuées les enquêtes de satisfaction.

I.2.3- Les tableaux de données

I.2.3.1 Construction d'un tableau de données

I.2.3.1.1- Définition

N'importe quel ensemble de données non structuré n'est pas analysable par les méthodes d'analyse des données. Les objets sur lesquels on peut appliquer les méthodes dites d'analyse des données sont appelés ***tableaux de données*** **[Fig. I.1]**. Il faudra alors savoir extraire d'une situation complexe de données une situation analysable que l'on puisse exprimer sous forme de tableaux de données. Un tableau de données est un tableau à double entrée, consignant des nombres mettant en jeu deux ensembles d'objets : les lignes du tableau correspondent aux individus (ou entités) ; les colonnes du tableau correspondent aux variables.



Modèle de tableau de données
I désigne l'ensemble des individus.
X désigne l'ensemble des variables
x_{ij} désigne l'élément courant du tableau

Fig. I.1 ***tableau de données***

I.2.3.2. Exemples de tableaux de données

I.2.3.2.1 Tableau individus*variables

1. Tableau de données quantitatives: c'est le cas où toutes les variables sont quantitatives.

Exemple : Les paramètres expriment la teneur en différents minerais de chacun des sondages. x_{ij} est une mesure de la teneur du minerai pour le sondage i.

N°sondage/variable	Teneur en fer	Teneur en cuivre
sondage1	0.1	0.2
sondage2	0.3	0.3
sondage3	0.4	0.2

2. Tableau de données qualitatives ou de modalités : c'est le cas où toutes les variables sont qualitatives. Si toutes les variables sont ordinales (resp nominales) on dira que l'on a un tableau de modalités ordonnées (resp non ordonnées)

Indiv/Journal	V1	V2	V3	V4
W1	3	1	2	1
W2	2	3	1	1
W3	1	2	1	2

L'individu 2 répond qu'il lit souvent le 2ème journal

L'individu répond 1, 2 ou 3 suivant sa fréquence de lecture d'un journal.
1 → pas du tout ;
2 → quelques fois ;
3 → souvent.

3. Tableau binaire : on rencontre souvent des variables qui ne prennent que deux valeurs codées généralement 0 et 1. Elles conduisent à des tableaux binaires.

Indiv/Journal	V1	V2	V3	V4
W1	1	0	1	0
W2	0	1	1	0
W3	1	0	0	1

Chaque individu répond par oui ou par non à la question "avez-vous acheté ce journal ?"

4. Tableau de préférence : on peut par exemple disposer des préférences des personnes interrogées sur des marques de parfum

Pers/Marque	M1	M2	M3	M4	M5
W1	1	3	4	2	5
W2	3	2	5	4	1
W3	5	3	4	2	1
W4	1	5	3	4	2

Aux marques M_i sont associées des variables v_i qui peuvent être considérées comme un ensemble de variables qualitatives ordinales.
Ainsi $V_3(w_2) = 5$; signifie que le deuxième individu préfère la troisième marque de parfum à toutes les autres.

5. <u>Tableau hétérogène</u> : c'est le cas de tableau où les variables sont de types différents:

Marchandise/variable	Prix	mode transport	Fragilité
W1	7.6	avion	1
W2	10.9	bateau	2
W3	3.5	train	3

<u>I.2.3.2.2 Tableaux variables*variables</u>
<u>Tableau de contingence et tableaux de fréquence</u>

A partir de deux variables qualitatives on définit le tableau de contingence croisant les modalités de deux variables. La case à l'intersection de la ligne i et de la colonne j contient le nombre d'individus ayant choisi la modalité i de la première variable et la modalité j de la seconde variable. Si l'on divise chaque valeur de ce tableau par le cardinal de la population, on obtient le tableau de fréquences relatives que l'on appellera plus simplement tableau de fréquence.

Lecture/sexe	Garçon	Fille
Nulle	48	55
Un journal par jours	14	10
Plus qu'un journal par jours	5	3

Ce tableau de contingence permet d'étudier la fréquence de lecture des journaux selon le sexe d’une population de lycéens algériens.

II. Notions indispensables pour l'étude de l'analyse des données

Pour analyser les données collectées on a besoin de quelques chiffres pour résumer l'essentiel d'une distribution. Ces chiffres sont appelés des mesures descriptives.

Une seule variable

Caractéristiques de position centrale :

Une fois le centre est cerné, il s'agit de voir comment se développe le phénomène autour de ce centre. C'est-à-dire :

- Comment il se disperse ?
- Et sous quelle forme ?

Les caractéristiques de position centrale sont :

- Le mode
- La médiane
- Les moyennes (arithmétique, géométrique, harmonique)

Le mode : est la valeur la plus fréquente de la variable statistique, c'est-à-dire celle qui correspond au plus grand effectif.

Exemple 1 :

X_i : 8 10 **16** 20 24 32 42 : nombre de fruits (modalités[1])

n_i : 12 23 **41** 24 22 16 12 : nombre d'arbrisseaux (individus)

Le plus grand effectif est 41 donc le mode est 16

Remarque : cet exemple présente 7 modalités.

Donc le mode est la modalité dont l'effectif est plus grand.

Exemple 2 :

Nombre de points Xi : 1 2 **3** 4 **5** **6** 7 8

Nombre d'étudiants ni : 17 30 **45** 38 **45** **45** 28 2

Nous avons trois modes : les valeurs 3 5 et 6 correspondent tous les trois au plus grand effectif

La médiane : c'est la valeur de la variable statistique qui partage la population en deux populations d'effectifs égaux.

Exemple 1 :

X_i : 8 10 **16** 20 24 32 42

n_i : 12 23 41 24 22 16 12

N_i : 0 12 35 76 100 122 138 150

150 : 2 = 75 donc il est encadrée par deux valeurs de l'effectif cumulé : N_{i-1} et N_i

N_{i-1}=35 et N_i=76 donc la médiane est M=16

[1] Chaque individu de la population se situe dans une position particulière vis-à-vis du caractère étudié. Il existe pour chaque caractère, plusieurs positions sur lesquelles se répartissent tous les individus de la population. Les différentes positions que peut prendre un caractère s'appelle modalité.

La moyenne arithmétique : $\frac{1}{n}\sum_{i=1}^{n} x_i$

Exemple :

Xi : 8 10 **16** 20 24 32 42

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Remarque : On a besoin dans ce rappel de savoir comment calculer seulement la moyenne arithmétique et non les deux autres (géométrique, harmonique).

Caractéristiques de dispersion :

- **Variance :** est le paramètre qui mesure la dispersion. (c'est l'éloignement de chaque point par rapport à tous les autres points)

Var(x)= $\frac{1}{n}\sum_{i=1}^{n} n_i (x_i - \bar{x})^2$ ou Var(x)= $\left[\frac{1}{n}\sum_{i=1}^{n} n_i x_i^{\,2} \right] - \bar{x}^2$

Ecart type : noté σ_x : c'est la racine carrée de la variance $\sigma_x = \sqrt{\text{var}(x)}$

Deux variables

Caractéristiques de position et de dispersion

Les notions de moyenne ou de variance sont les mêmes qu'on a vu précédemment, il s'agit seulement de préciser pour quelle variable on les calcule.

Covariance : c'est l'outil qui précise la relation d'une variable par rapport à l'autre et dit qu'elles varient dans le même sens ou quelles varient dans le sens contraire et mesure la force de leur liaison.

Observons les nuages de points suivants [**Fig. II.1**]:

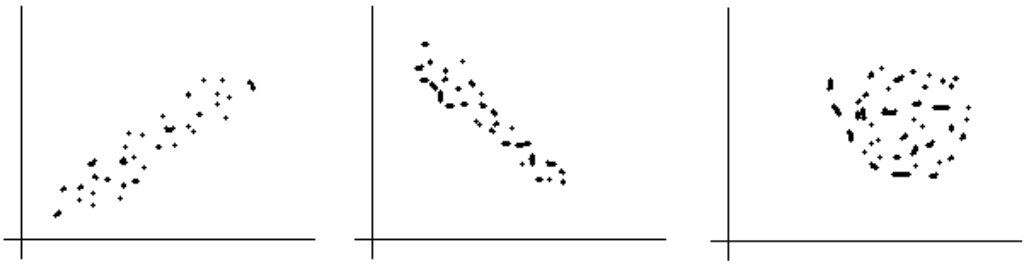


Fig. II.1 Différentes formes de nuages de points

La disposition des points nous donne déjà une idée de l'évolution des deux variables l'une par rapport à l'autre. Trois possibilités de nuages peut se poser :

La forme des deux premiers nuages est assez allongée tandis que celle du troisième est plutôt arrondie.

- Dans le premier nuage : si X croit Y croit
- Dans le second nuage : si X croit Y décroît
- Dans le troisième nuage : si X croit Y ne suit aucune forme et par suite X et Y n'ont aucune relation : dans ce troisième nuage il y a une indépendance entre les variables.

Essayons de quantifier les résultats de cette analyse, c'est-à-dire de mesurer par la quantité l'existence d'une liaison entre X et Y, d'identifier son sens (variation dans le même sens ou dans le sens contraire) et voir la force de cette liaison (forte ou faible).
Pour cela, procédons de la manière suivante :
Considérons le point du nuage (X,Y), X et Y étant les moyennes arithmétiques des variables X et Y. ce point moyen se situe au centre du nuage.

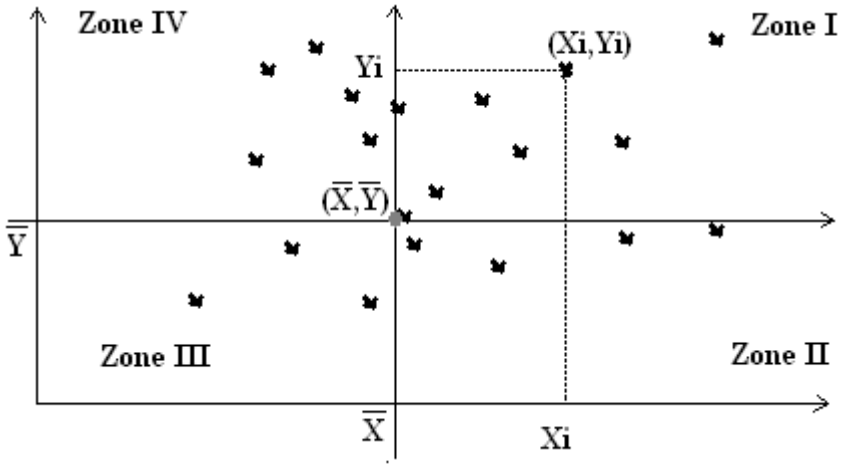


Fig. II.2 Les différentes zones des points du nuage

Quand nous déplaçons le repère sur ce point nous obtenons quatre zones.
Soit maintenant la quantité : $\alpha_i = (X_i - \overline{X}).(Y_i - \overline{Y})$

- Si le point (Xi,Yi) se trouve dans la zone I alors les quantités (Xi, $\overline{X}$) et (Yi, $\overline{Y}$) sont toutes les deux positives et donc leur produit α_i est positif

- Si le point (Xi,Yi) se trouve dans la zone III alors les quantités (Xi, $\overline{X}$) et (Yi, $\overline{Y}$) sont toutes les deux négatives et donc α_i est positif

- Si le point se trouve dans la zone II alors la quantité (Xi, $\overline{X}$) est positive tandis que la quantité (Yi, $\overline{Y}$) est négative et par conséquent, α_i est négative.

- Si le point se trouve dans la zone IV alors la quantité (Xi, $\overline{X}$) est négative tandis que la quantité (Yi, $\overline{Y}$) est positive et par conséquent, α_i est négative.

Regardons le nuage dans son ensemble et considérons la quantité :

$$\beta = (1/N)\sum_{i=1}^{N} \alpha_i$$

β tient compte de tous les points du nuage. C'est la moyenne des valeurs α_i.

- α_i est positive pour tous les points des zones I et III, et donc chaque point de ces deux zones apporte une contribution positives à β

- α_i est négative pour tous les points des zones II et IV, et donc chaque point de ces deux zones apporte une contribution négatives à β

Revenons à nos trois nuages du début.
Les points du premier nuage appartiennent presque tous aux zones I et III, et donc apportent presque tous une contribution positive à β qui sera ainsi positive et grande en valeur absolue.
Les points du deuxième nuage appartiennent presque tous aux zones II et IV, et donc apportent presque tous une contribution négative à β qui sera ainsi négative et grande en valeur absolue.
Mais les points du troisième nuage se dispersent sur les quatre zones et les contributions positives ramenées par certains points seront compensées par les contributions négatives

ramenées par les autres points. Nous devons donc nous attendre, dans ce cas, que β soit proche de zéro.
En conclusion, la quantité β nous renseigne bien sur la liaison entre les variables X et Y, son sens et sa force.
La quantité β est appelé la covariance entre les deux variables X et Y.
On la note Cov(X,Y)
Ainsi :

$$Cov(X,Y) = (1/N)\sum_{i=1}^{N}(X_i - \overline{X}).(Y_i - \overline{Y})$$
$$= [\ (1/N)\sum_{i=1}^{N} X_i.Y_i\ ] - \overline{X}.\overline{Y}$$

Ainsi la variance d'une variable statistique n'est autre qua covariance de cette variable avec elle-même.

Courbe de régression : on appelle courbe de régression de Y en x la courbe représentative des moyenne conditionnelle $\overline{y}_i$ en fonction des valeurs x_i de la variable de liaison X.
Ou
On appelle courbe de régression de X en y la courbe représentative des moyenne conditionnelle $\overline{x}_i$ en fonction des valeurs y_i de la variable de liaison Y.

Le coefficient de corrélation : est le rapport de la covariance au produit des écarts types, c'est-à-dire :

$$\rho = \frac{Cov(X,Y)}{\sigma_X.\sigma_Y}$$

Puisque $|Cov(X,Y)| \le \sigma_X.\sigma_Y$ alors $-1 \le \rho \le +1$ c'est une propriété importante de ce coefficient.
Si ρ =+1 la pente de la droite est positive, (X et Y varient dans le même sens)
Si ρ =-1 la pente de la droite est négative, (X et Y varient dans le sens contraire)
Quand ρ est proche de 1 nous comprenons que les points du nuage sont proche de la droite
Quand ρ est proche de 0, le nuage de points a une forme arrondie et cela veut dire qu'il n y a pas une liaison linéaire entre les variables X et Y.
ρ ne sert donc qu'à montrer l'existence d'une liaison linéaire entre les variables. Si ρ n'est pas proche de 1 cela veut seulement dire qu'il ne peut y avoir de liaison linéaire entre les variables X et Y mais toujours qu'elles ont une autre forme de liaison (non linéaire)

III. Régression linéaire sur deux variables

III.1 Introduction :

Les données à étudier sont décrites de la façon suivante :

- On dispose de N individus, les $(x_i; y_i)$.
- Chaque individu est décrit par n variables réelles, i.e. C'est un vecteur $X \in \mathbb{R}^n$
- Dans le cas de la ***discrimination***, chaque exemple est associé à une classe (un groupe)
- Dans le cas de la ***régression***, on cherche à expliquer y_i grâce à x_i, c'est-à-dire trouver une fonction ***f*** telle que $y_i \approx f(x_i)\ \forall i$. C'est un vecteur $Y \in \mathbb{R}^p$.

Applications pratiques :

- Prédire demain en fonction d'aujourd'hui (météo, bourse, etc.)
- Evaluer le risque de défaillance d'un emprunteur en fonction de ses caractéristiques socioprofessionnelles
- calculer le taux d'humidité du sol ou le niveau de maturité d'un champ de céréales en fonction d'une image radar de la zone concernée
- etc.

On peut reformuler les trois problèmes de l'Analyse de Données :

1. **Discrimination :** trouver un lien entre x et sa classe
2. **Régression :** trouver un lien entre x et y
3. **Classification :** construire des classes en associant des étiquettes aux individus

III.2 Régression linéaire :

On considère un couple de caractères statistiques quantitatifs et l'on cherche à expliciter la liaison linéaire entre X et Y au moyen de la méthode ci-dessus. On cherche donc un couple (a,b) tel que la quantité : $q = \sum (y_i - (a.x_i + b))^2$ soit minimale **[Fig. III.1]**.

Objectif : Trouver la droite y = ax + b des moindres carrés, c'est à dire minimisant la quantité q:

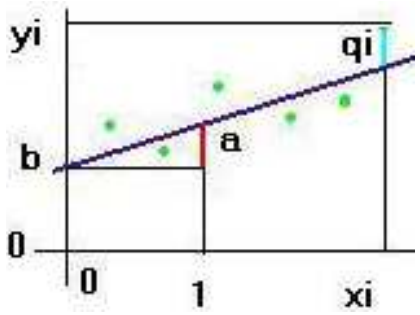


Fig. III.1 Droite des moindres carrées

Une quantité est minimale si la dérivée s'annule.

Il existe une unique droite rendant minimale la quantité ci-dessus. Cette droite **s'appelle la droite de régression de Y en X,**

III.2.1 Les différentes méthodes d'approximations linéaire d'un nuage

Notation : x pour les abscisse et y pour les ordonnées.

On s'intéresse à l'étude simultanée de deux variables quantitatives liées à un même individu i sur lequel ont été effectuées deux mesures x_i et y_i c'est-à-dire un couple d'observation. L'ensemble des couples $\{(x_i, y_i) i = 1..n\}$ s'appelle le nuage des points. On cherche très souvent à prouver une relation linéaire entre ces variables. Ce type d'étude porte le nom de : ***recherche de la régression linéaire entre x et y***. A partir d'un point i observé de coordonnées (x_i, y_i) et une droite D du plan orthonormé usuel trois types de projections sont possibles :

- P||y'y : une projection du point i parallèlement à l'axe des y (point M_i sur la figure)
- P||x'x : une projection du point i parallèlement à l'axe des x (point N_i sur la figure)
- P⊥D : une projection du point i perpendiculairement (ou orthogonalement) à la droite D (point P_i sur la figure)

Le schéma suivant précise ces différentes projections **[Fig. III.2]** :

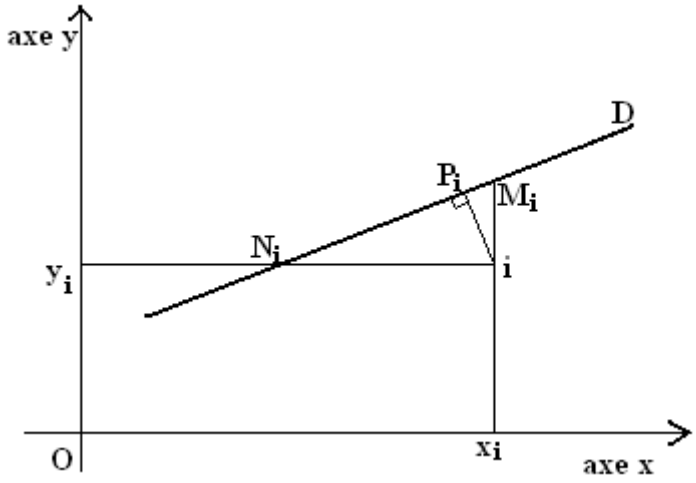


Fig. III.2 ***Les différentes projections possibles***

Pour chacune de ces projections nous chercherons une droite qui minimise une somme de carrées :

- La somme $\sum_{i=1}^{n} M_i^2$ pour la projection P||y'y
- La somme $\sum_{i=1}^{n} iN_i^2$ pour la projection P||x'x
- La somme $\sum_{i=1}^{n} iP_i^2$ pour la projection P⊥D

Ces différentes minimisations nous conduiront aux équations des droites de régressions.

Régression linéaire par la méthode des moindres carrés

III.2.1.1 Droite de régression observée de y par rapport à x

Sur le nuage traçons les distances parallèlement à l'axe des y, pour le point i on obtient la distance iM_i entre l'observation et la droite. On suppose que la droite cherchée a pour équation **y=ax+b**, on doit donc trouver les deux coefficients a (la pente) et b (l'ordonnée à l'origine). Le schéma suivant précise cette distance iM_i **[Fig. III.3]**:

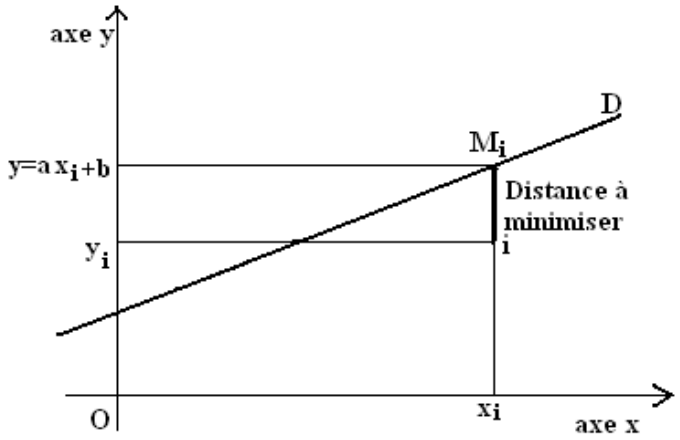


Fig. III.3 ***Pour la recherche de la droite de régression Dy/x***

La méthode des moindres carrées consiste à minimiser la somme des carrées des distances entre les ordonnées des points observés (y_i) et « théoriquement » $(ax_i + b)$, correspondant à l'abscisse (x_i) sur la droite cherchée, c'est-à-dire minimiser S(a,b) où :

$$S(a,b)=\sum_{i=1}^{n}(y_i - ax_i - b)^2$$

Les extremums de cette équation seront nécessairement des minimums (nous devrions vérifier que les solutions obtenus vérifient les conditions suffisantes à l'aide des dérivés partielles du second ordre qui doivent être >0)
Résolvant ce système :

(éq.2) $$\frac{\partial}{\partial b}\sum_{i=1}^{n}\left[y_i^2 - 2ax_i y_i - 2by_i + a^2 x_i^2 + 2abx_i + b^2\right]$$

$$\sum_{i=1}^{n}\left[2ax_i - 2y_i + 2b\right] = 0$$

$$2a\sum_{i=1}^{n}x_i - 2\sum_{i=1}^{n}y_i + 2b\sum_{i=1}^{n}1 = 0$$

d'où (éq.2') $$a\sum_{i=1}^{n}x_i + nb = \sum_{i=1}^{n}y_i \quad \text{soit} \quad b = \bar{y} - a\bar{x}$$

Reportons cette valeur b dans $S(a,b)=\sum_{i=1}^{n}(y_i - ax_i - b)^2$

$$S(a,b)=\sum_{i=1}^{n}(y_i - ax_i - b)^2 = \sum_{i=1}^{n}(y_i - ax_i - \bar{y} + a\bar{x})^2 = \sum_{i=1}^{n}\left[(y_i - \bar{y}) - a(x_i - \bar{x})\right]^2 = S(a)$$

Donc :

$$\frac{\partial(a,b)}{\partial a} = \frac{dS(a)}{da} = -2\sum_{i=1}^{n}(x_i - \bar{x})\left[(y_i - \bar{y}) - a(x_i - \bar{x})\right]$$

qui s'annule pour :

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{cov}_e(x,y)}{S_{ex}^2}$$

En notant $\mathrm{cov}_e(x,y)$ le numérateur de a, qui représente la covariance empirique entre x et y et S_{ex}^2 la variance empirique de x

Le coefficient de corrélation r=r(x,y)= $\frac{\mathrm{cov}_e(x,y)}{S_{ex}.S_{ey}}$ avec ce coefficient r la pente a s'écrit :

$a = r\frac{S_{ex}.S_{ey}}{S_{ex}^2} = r\frac{S_{ey}}{S_{ex}}$ l'équation y=ax+b s'exprime donc sous la forme

(Dy/x) : $\frac{y-\bar{y}}{S_{ey}} = r\frac{x-\bar{x}}{S_{ex}}$ ou y=ax+b avec $a = r\frac{S_{ey}}{S_{ex}}$ et $b = \bar{y} - a\bar{x}$

Cette première forme a l'avantage de présenter une parfaite symétrie, le membre de gauche représente la variable y centrée réduite, de même pour le membre de droite multiplié par le coefficient de corrélation linéaire r. cette équation est celle de la droite de régression de y par rapport à x que nous avons notée Dy/x.

III.2.1.2 <u>Droite de régression observée de x par rapport à y</u>

Sur le nuage traçons les distances parallèlement à l'axe des x, pour le point i on obtient la distance iN_i entre l'observation et la droite. On suppose que la droite cherchée a pour équation $x = \alpha y + \beta$, on doit donc trouver les deux coefficients α et β (la pente sera ici a'=$1/\alpha$ et l'ordonnée à l'origine sera ici b'=$-\beta/\alpha$ pour une équation y=a'x+b')

Le schéma suivant précise cette distance iN_i **[Fig. III.4]** :

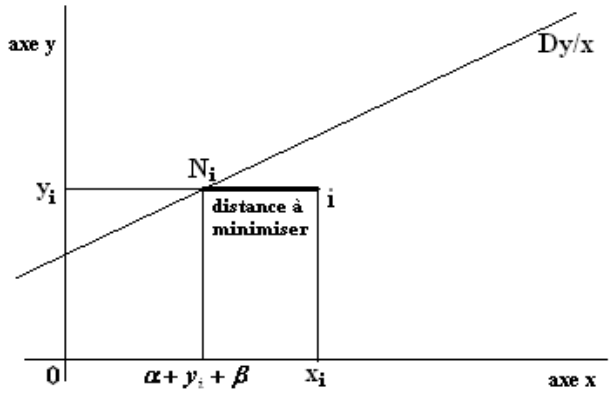


Fig. III.4 distance entre l'observation et la droite

La méthode des moindres carrées consiste à minimiser la somme des carrées des distances entre les abscisses des points observés (x_i) et celles correspondant pour l'ordonnée y_i sur la droite cherchée $(\alpha y_i + \beta)$, c'est-à-dire minimiser $\Sigma(\alpha,\beta)$ où :

$$\Sigma(\alpha,\beta) = \sum \left[x_i - (\alpha y_i + \beta)\right]^2$$

On est ainsi conduit à résoudre le système :

$\frac{\partial \sum(\alpha,\beta)}{\partial \alpha}=0$ (éq.3) $\frac{\partial \sum(\alpha,\beta)}{\partial \beta}=0$ (éq.4)

Cette résolution nous conduit aux coefficients :

$\alpha=rS_{ex}/S_{ey}$ $\beta=\bar{x} - \alpha\bar{y}$ en utilisant le coefficient r précédent.

L'équation de la droite de régression de x par rapport à y (Dx/y) aura ainsi équation sous la forme symétrie précédente :

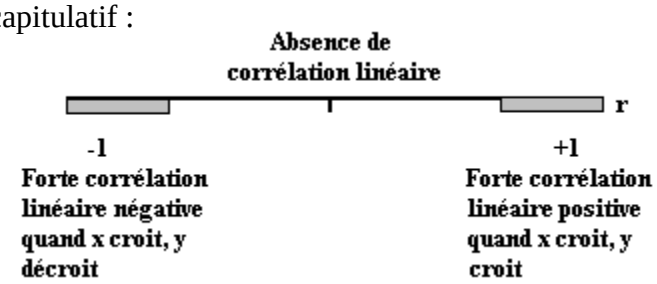
(Dx/y) : $\frac{x-\bar{x}}{S_{ex}} = r\frac{y-\bar{y}}{S_{ey}}$ où $x = \alpha y + \beta$ avec $\alpha = r\frac{S_{ex}}{S_{ey}}$ et $\beta = \bar{x} - \alpha\bar{y}$

Mise sous la forme y=a'x+b' avec a'=$1/\alpha$ et b'=$-\beta/\alpha$

<u>Interprétation de la covariance empirique entre x et y du coefficient de corrélation linéaire r(x,y) :</u>

- La covariance empirique entre x et y est un indicateur du sens de la variation simultanée entre x et y.
- Le coefficient de corrélation r est coefficient qui varie de -1 à +1 qui caractérise la linéarité du nuage que l'on étudie, plus on s'approche des valeurs extrêmes meilleure sera la linéarité du nuage.

Il est d'usage en statistique descriptive de considérer l'approximation linéaire du nuage si r^2 (appelé coefficient de détermination) est supérieur à 0.8 (un r d'environ 0.9). le schéma suivant est récapitulatif :



<u>Quel est le lien entre ces deux droites de régression ?</u>

•<u>Cas r>=0 :</u>

Ces deux droites passent par leur point moyen ($\bar{x}$, $\bar{y}$) noté G.

Les positions relatives des deux droites de régression selon la valeur de r : **[Fig. III.5]**

- Absence de corrélation linéaire r=0
- Corrélation linéaire maximum r=1 : cet angle entre les deux droites pourrait être un critère de bonne linéarité du nuage.

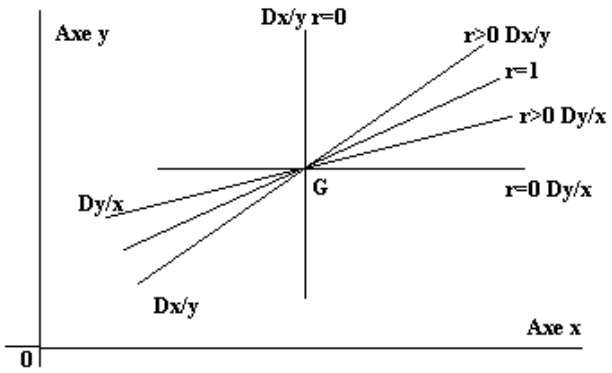


Fig. III.5 ***Les positions relatives des deux droites de régression selon la valeur de r***

Régression linéaire par la méthode des moindres rectangles

III.2.1.3– Régression orthogonale ou droite des moindres rectangles

Nous allons utiliser un troisième type de projection : projection orthogonale du point observation par rapport à une droite qui sera supposée représenter le mieux le nuage au sens des moindres rectangles.
On va essayer de développer cette méthode dans le cas simple de deux variables et sera la base des méthodes : Analyse en composantes principales (ACP) et Analyse factorielle des correspondances (AFC).

1. Rappel :
Les formules de dérivation matricielle :
Si X est un vecteur (m,1), A une matrice constante symétrique (m,m), a un vecteur constant (m,1) alors :

$$\mathbf{Y = X'AX} \qquad \frac{\partial Y}{\partial X} = 2AX$$

$$\mathbf{Y = X'X} \qquad \frac{\partial Y}{\partial X} = 2X$$

$$\mathbf{Y = X'a} = \frac{\partial Y}{\partial X} = a$$

Si A n'est pas symétrique :

$$\text{pour } Y = X'AX \quad \frac{\partial Y}{\partial X} = (A + A')X$$

2. La recherche de la droite de régression orthogonale

Notons : x_1 : variable notée précédemment x, x_2 : variable notée précédemment y
Chaque couple $(xi,yi) = (x_1(i),x_2(i))$. L'ensemble de ces couples donne le nuage de points. On cherche à approximer ce nuage par une droite **D** de vecteur $u' = [\alpha 1, \alpha 2]$, qui sera déterminée en minimisant la somme des carrées des distances mesurées perpendiculairement entre les points observées et la droite **D**.

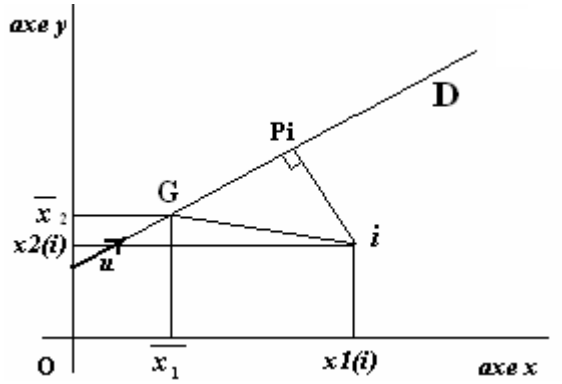


Fig. III.6 Schéma pour l'étude de la régression orthogonale

Le point ***i*** se projette orthogonalement en ***Pi*** sur la droite ***D***. **G**, de coordonnées $\overline{X} = (\overline{x_1}, \overline{x_2})$, représente le point moyen du nuage, c'est-à-dire son centre de gravité. On cherche à minimiser la somme des carrées des longueurs de ces projections, soit :

$$\sum_{i=1}^{n} iP_i^2$$

Comme le carrée de la distance de ***i*** à **G** est fixée, cette minimisation revient à maximiser le carrée de la distance de la projection **Pi** au centre de gravité **G** :

$$\text{Minimiser } \sum_{i=1}^{n} iP_i^2 \Leftrightarrow \text{maximiser } \sum_{i=1}^{n} \boldsymbol{GP_i^2}$$

Cette distance $\boldsymbol{GP_i}$ est la projection du vecteur $\overrightarrow{\mathrm{Gi}}$ sur la droite **D**, de vecteur directeur $\vec{u}$.
En notant $\boldsymbol{X'=(x_1(i),x_2(i))'}$, la longueur de cette projection est le produit scalaire :

$$\overrightarrow{iG}.\vec{u}=(\mathbf{X(i)} - \overline{X}).\vec{u}=\left[\mathbf{x_1(i)} - \overline{x}_1 , x_2(i)- \overline{x}_2 \right]\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix}=\left[\alpha_1\left(x_1(i) - \overline{x}_1\right)+\alpha_2\ (\mathbf{x_2}(i) - \overline{x}_2)\right]$$

La somme des carrées de ces longueurs vaut donc :

$$\sum_{i=1}^{n} \boldsymbol{GP_i^2} = \sum_{i=1}^{n}\ \left[\alpha_1^2\left(x_1(i) - \overline{x}_1\right)^2+\alpha_2^2\ (\mathbf{x_2}(i) - \overline{x}_2)^2+2\alpha_1\ \alpha_2\ (\mathbf{x_1}(i)- \overline{x}_1)(x_2(i) - \overline{x}_2)\right]$$

En divisant par **n** cette quantité on fait apparaître les variances et covariances empiriques de $\mathbf{x_1}$ et $\mathbf{x_2}$.

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{GP_i^2} =\left[\alpha_1^2\ \mathbf{Var(x_1)} + \alpha_2^2\ \mathrm{Var}(\mathbf{x_2}) + 2\alpha_1\ \alpha_2\ \mathrm{Cov}(\mathbf{x_1},\mathbf{x_2})\right] \qquad \textbf{(eq.1)}$$

En effet, $\mathrm{Var}(x_1)= \frac{1}{n}\sum_{i=1}^{n}\left(x_1(i)-\overline{x_1}\right)^2$ désigne la variance empirique de $\mathbf{x_1}$.
La covariance empirique entre $\mathbf{x_1}$ et $\mathbf{x_2}$ étant donnée :

$$\frac{1}{n}\sum_{i=1}^{n}\ (\mathbf{x_1}(i) - \overline{\mathbf{x}_1})(\mathbf{x_2}(i) - \overline{\mathbf{x}_2})$$

Notons **V** la matrice des variances-covariances, $\mathbf{V}=\begin{bmatrix} Var(x_1) & Cov(x_1,x_2)\\ Cov(x_1,x_2) & Var(x_2)\end{bmatrix}$

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{GP_i^2} =\left[\alpha_1\ \alpha_2\right]\begin{bmatrix} Var(x_1) & Cov(x_1,x_2)\\ Cov(x_1,x_2) & Var(x_2)\end{bmatrix}\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix}=\mathbf{u'Vu} \qquad \textbf{(éq.2)}$$

Où **u'=[α1, α2],** désigne le vecteur transposé du vecteur **u.**

Nous cherchons à rendre maximale la valeur de cette équation 2.

Remarquons qu'elle représente la « **variance des points projetés (Pi) en prenant G comme origine** » (l'observation serait la longueur algébrique GP_i), ce qui revient donc à chercher la ***variance maximale*** de la longueur GP_i de ces projections, notons alors $Var(D)$ cette quantité.

Le problème mathématique soulevé par cette maximisation est lié à la dérivation matricielle. Remarquons d'abord qu'il s'agit de trouver $[\alpha_1, \alpha_2]$, on doit donc imposer à ces deux coefficients une condition de normalisation car sinon on pourrait toujours avoir une variance aussi grande que possible en multipliant α_1 par un facteur aussi grand que l'on veut. Nous prendrons **u** de norme 1, c'est-à-dire qu'il vérifie $\alpha_1^2 + \alpha_2^2 = 1$. Il reste maintenant à trouver $u' = [\alpha_1, \alpha_2]$ qui maximise **u'Vu** avec **u'u=1**. La démonstration fait appel à la dérivation matricielle pour le calculd'extrema sous contraintes (méthode de lagrange). En notant $\mathcal{L}$ le Lagrangien $\mathcal{L}$**=u'Vu-λ(u'u-1)** où **λ** désigne le multiplicateur de Lagrange.

La dérivée par rapport au vecteur **u** de ce Lagrangien nous conduit à :

$d\mathcal{L}$**/du = 2Vu-2λu**

Cette dérivée s'annule quand **Vu-λu** = 0 c'est-à-dire **Vu = λu.** ce qui signifie que **u** est vecteur propre de **V** à la valeur propre**λ**. L'équation 2 sera donc maximale quand la valeur propre **λ** de la matrice **V** associé au vecteur propre **u** sera maximale, en effet :

$$\frac{1}{n}\sum_{i=1}^{n} GP_i^2 = \mathbf{u'Vu = \lambda u'u = \lambda}$$

L'équation de la droite de régression orthogonale de direction **u** passant par **G** sera donc :

$$\frac{x_2 - \overline{x_2}}{x_1 - \overline{x_1}} = \frac{\alpha_2}{\alpha_1} \text{ ou } x_2 = \frac{\alpha_2}{\alpha_1} x_1 + \overline{x_2} - \frac{\alpha_2}{\alpha_1}\overline{x_1}$$

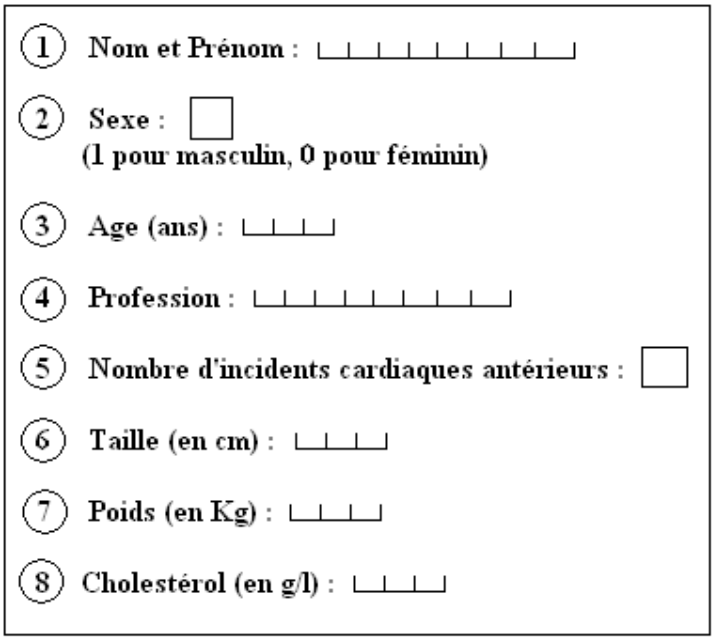
On peut montrer que la pente de cette droite de régression orthogonale est comprise entre les deux droites de régression $\mathbf{D_{y/x}}$ **et** $\mathbf{D_{x/y}}$**,** c'est-à-dire qu'elle passe par **G** et se situe entre ces deux droites.

Institut de sciences exactes
Département d'Informatique
Module : Analyse de données

Fiche de TD n° 01

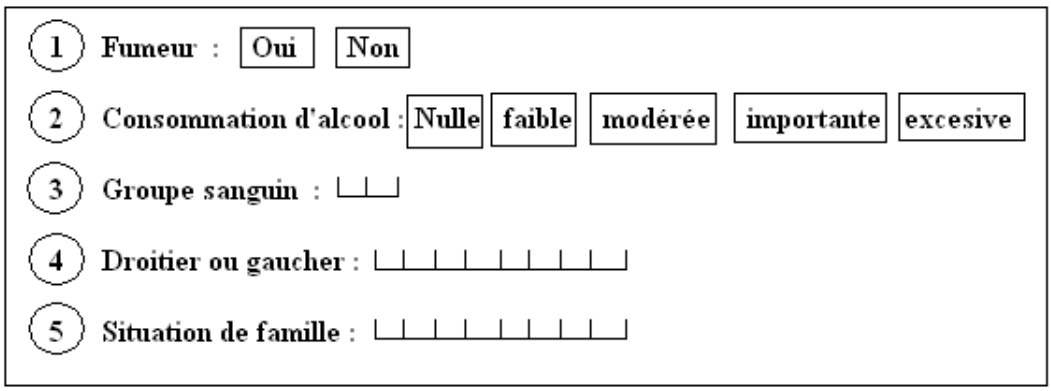
Exercice 1 :

Dites pour chacune des variables du questionnaire médical ci-dessous, si elle est qualitative ? Discrète ? Continue ?



Exercice 2 :

Pour les variables du questionnaire ci-après, dites si la variable est ordinale ? Nominale ? Dichotomique ?



Exercice 3 :

Une entreprise veut mener une étude sur la liaison entre les dépenses (hebdomadaires) mensuelles en publicité et le volume des ventes qu'elle réalise .Nous avons obtenu au cours des six derniers mois les données suivantes :

X **Dépenses publicitaires (en milliers de DA)**	70	80	30	50	35	45
Y **Volume des ventes (en milliers de DA)**	580	380	200	310	400	450

1- Tracer le nuage de points.
2- Ajuster la droite de régression.
3- Calculer le coefficient de corrélation.
4- Interpréter le coefficient de corrélation.

Exercice 4 :

On dispose de 6 boites maintenues à des températures différentes. On place une dizaine de bactéries dans chacune des boites et on compte le nombre de bactéries contenues dans chaque boite au bout de 3 minutes. On obtient les résultats suivants :

Température (en degrés(C) variable T	10	15	20	25	30	35
Nombre de bactéries Variable V	8	15	23	31	38	46

1- Calculer les moyennes T et V
2- Calculer les variances de T et de V
3- Calculer la covariance et le coefficient de corrélation linéaire
4- Déterminer l'équation de la droite de régression de V en T

Exercice 5 :

Soit le tableau suivant concernant deux variables x et y (appelées v1 et v2, et notées par la suite x1 et x2) mesurées sur 10 individus (i=1..10)

Individus	x_i	y_i	x_i^2	y_i^2	$x_i\, y_i$
1	0.5	0	0.25	0	0
2	-0.1	1.2	0.01	1.44	-0.12
3	-0.5	0.5	0.25	0.25	-0.25
4	-0.3	0.1	0.09	0.01	-0.03
5	0	2.5	0	6.25	0
6	1.6	-0.7	2.56	0.49	-1.12
7	2	2	4	4	4
8	2.4	1.2	5.76	1.44	2.88
9	0.5	3.5	0.25	12.25	1.75
10	2.7	-0.9	7.29	0.81	-2.43
Total	8.8	9.4	20.46	26.94	4.68

A. Avec la méthode des moindres carrées

1- Calculer la droite de régression observée de y par rapport à x
2- Calculer la droite de régression observée de x par rapport à y
3- Interpréter les résultats obtenus

B. Avec la méthode des moindres rectangles

1- Calculer la droite de régression observée de y par rapport à x
2- Calculer la droite de régression observée de x par rapport à y

Institut de sciences exactes
Département d'Informatique
Module : Analyse de données

Correction de la Fiche de TD n° 01

Solution exercice n° 01 :

- Variable 1 : qualitative, non discret et non continue
- Variable 2 : qualitative, non discret et non continue
- Variable 3 : non qualitative, discret et continue
- Variable 4 : qualitative, non discret et non continue
- Variable 5 : non qualitative, discret et non continue
- Variable 6 : non qualitative, non discret et continue
- Variable 7 : non qualitative, non discret et continue
- Variable 8 : non qualitative, non discret et continue

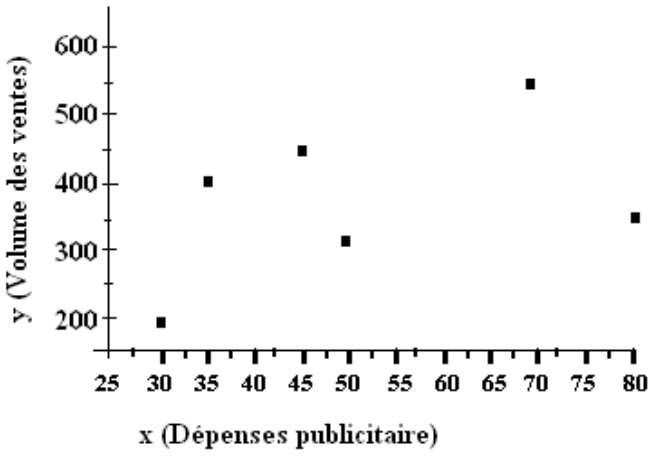
Solution exercice n° 02 :

- Variable 1 : non ordinale, non nominale et dichotomique
- Variable 2 : ordinale, non nominale et dichotomique
- Variable 3 : non ordinale, nominale et non dichotomique
- Variable 4 : non ordinale, non nominale et dichotomique
- Variable 5 : non ordinale, nominale et non dichotomique

Solution exercice n° 03 :

Question (1) :

Le graphique du nuage de points entre les dépenses en publicité et le volume des ventes que l'entreprise réalise :



A de cette représentation, nous soupçonnons l'existence d'une relation entre X et Y, mais la corrélation n'est pas très forte. (Nous la vérifions avec la question 3)

Question (2) : Ajustement de la droite de régression.

La réponse à cette question permet d'exprimer la relation à l'aide d'une équation mathématique : Y=ax+b

Xi	Yi	XiYi	X_i^2	Y_i^2
70	580	40600	4900	336400
80	380	30400	6400	14400
30	200	6000	900	4000
50	310	15500	2500	96100
35	400	14000	1225	160000
45	450	20250	2025	202500
310	2320	126750	17950	979400

$$\overline{Y} = \frac{\sum Y_i}{N} = \frac{2320}{6} = 386.66 \approx 387$$

$$\overline{X} = \frac{\sum X_i}{N} = \frac{310}{6} = 51.6 \approx 52$$

$$Var(X) = S^2(X) = \frac{\sum X_i^2}{N} - \overline{X^2} = \frac{17950}{6} - (52)^2 = 288$$

$$S(X) = \sqrt{288} = 16.97 \approx 17$$

$$Var(Y) = S^2(Y) = \frac{\sum Y_i^2}{N} - \overline{Y^2} = \frac{979400}{6} - (387)^2 = 13464$$

$$S(Y) = \sqrt{13464} = 116.03 \approx 116$$

$$Cov(X,Y) = \frac{\sum XY}{N} - \overline{X}\,\overline{Y} = \frac{126750}{6} - (52)(387) = 1001$$

Nous obtenons alors les valeurs a et b de la droite : Y=ax+b

$$a = \frac{Cov(X,Y)}{Var(X)} = \frac{1001}{288} = 3.48$$

$$b = \overline{Y} - a\overline{X} = 387 - (3.48)(52) = 387 - 180.96 \approx 387 - 181 = 206$$

d'où : **Y=3.48x+206**

Question 3 : Calcul du coefficient de corrélation

$$r = \frac{Cov(X,Y)}{S_{ex} S_{ey}} = \frac{1001}{17 * 116} \approx 0.507$$

Question 4 : Interprétation du coefficient de corrélation

Pour cette série, nous pouvons conclure qu'il y a une corrélation positive mais très forte entre le volume des ventes et les dépenses en publicité de cette entreprise. En effet, seulement : 26 % ($r^2 = 0.51^2 = 0.26$) **de la fluctuation totale de Y se trouve expliquée par le lien entre X et Y.**

Solution exercice n° 04

Les formules de calcul des moyennes et les variances de *T* et *V*, leurs covariances et le coefficient de corrélation linéaire :

Pour T : $\overline{T} = \frac{\sum T_i}{N}$, $Var(T) = \frac{\sum T_i^2}{N} - \overline{T^2}$

Pour V : $\overline{V} = \frac{\sum V_i}{N}$, $Var(V) = \frac{\sum V_i^2}{N} - \overline{V^2}$

Covariance : $Cov(T,V) = \frac{\sum TV}{N} - \overline{T}\,\overline{V}$

Coefficient de corrélation : $r = \frac{Cov(T,V)}{\sqrt{Var(T).Var(V)}}$

Les coefficients de la droite d'ajustement se calcul par les formules :

$a = \frac{Cov(T,V)}{Var(T)}$, $b = \overline{V} - a\overline{T}$

Afin d'arriver à répondre à toutes ces questions, on a besoin d'un tableau de cinq colonnes :

T_i	V_i	T_i^2	V_i^2	T_iV_i
10	**8**	**100**	**64**	**80**
15	**15**	**225**	**225**	**225**
20	**23**	**400**	**529**	**460**
25	**31**	**625**	**961**	**775**
30	**38**	**900**	**1444**	**1140**
35	**46**	**1225**	**2116**	**1610**
$\sum T_i = 135$	$\sum V_i = 161$	$\sum T_i^2 = 3475$	$\sum V_i^2 = 5339$	$\sum T_iV_i = 4290$
$\overline{T} = 22.5$	$\overline{V} = 26.83$	**579.17**	**889.83**	**715**

1- Moyenne arithmétiques : $\overline{T} = 22.5$ **et** $\overline{V} = 26.83$

2- Variances : $Var(T) = \frac{\sum T_i^2}{N} - \overline{T^2} = 579.17 - (22.5)^2 = 72.92$

$$Var(V) = \frac{\sum V_i^2}{N} - \overline{V^2} = 889.83 - (26.83)^2 = 169.9811$$

3- Covariance : $Cov(T,V) = \frac{\sum TV}{N} - \overline{T}\,\overline{V} = 715 - (22.5)(26.83) = 111.325$

Coefficient de corrélation : $r = \frac{Cov(T,V)}{\sqrt{Var(T).Var(V)}} = \frac{111.325}{\sqrt{72.92.169.9811}} \approx 0.9999 = 1$

4- Droite d'ajustement :

$$a = \frac{Cov(T,V)}{Var(T)} = \frac{111.325}{72.92} \approx 1.527$$

$$b = \overline{V} - a\overline{T} = 26.83 - 1.527.22.5 = -7.5275$$

Ainsi la droite d'ajustement a pour équation : **V= 1.527.T – 7.53**

Solution exercice n° 05

Le tableau concerne deux variables x et y mesurées sur un échantillon de taille 10 individus (i=1..10) permettant de calculer les paramètres nécessaires :

Individus	x_i	y_i	x_i^2	y_i^2	$x_i\, y_i$
1	0.5	0	0.25	0	0
2	-0.1	1.2	0.01	1.44	-0.12
3	-0.5	0.5	0.25	0.25	-0.25
4	-0.3	0.1	0.09	0.01	-0.03
5	0	2.5	0	6.25	0
6	1.6	-0.7	2.56	0.49	-1.12
7	2	2	4	4	4
8	2.4	1.2	5.76	1.44	2.88
9	0.5	3.5	0.25	12.25	1.75
10	2.7	-0.9	7.29	0.81	-2.43
Total	$\sum x_i$ =8.8	$\sum y_i$ = 9.4	$\sum x_i^2 = 20.46$	$\sum y_i^2$ = 26.94	$\sum x_i y_i$ =4.68
	$\overline{X} = \frac{\sum x_i}{N} = \frac{8.8}{10}$ =0.88	$\overline{Y} = \frac{\sum y_i}{N} = \frac{9.4}{10}$ =0.94			

A. <u>Avec la méthode des moindres carrées</u>

$\overline{X} = 0.88,\ \overline{Y} = 0.94$

$$Var(X) = \frac{\sum X_i^2}{N} - \overline{X^2} = \frac{20.46}{10} - (0.88) = 1.27$$

$$Var(Y) = \frac{\sum y_i^2}{N} - \overline{Y^2} = \frac{26.94}{10} - (0.94) = 1.81$$

$$Cov(X,Y) = \frac{\sum XY}{N} - \overline{X}\,\overline{Y} = \frac{4.68}{10} - 0.88 * 0.94 = -0.36$$

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X).Var(Y)}} = \frac{-0.36}{\sqrt{1.27 * 1.81}} = -0.24$$

1- Calculer la droite de régression observée de y par rapport à x : Dy/x

$$a = \frac{Cov(X,Y)}{Var(X)} = \frac{-0.36}{1.27} = -0.28, \quad b = \overline{Y} - a\overline{X} = 0.94 + 0.28 * 0.88 = 1.19$$

2- Calculer la droite de régression observée de x par rapport à y : Dx/y

L'équation de la droite de régression de x par rapport à y (Dx/y) aura ainsi équation sous la forme symétrie précédente (voir cours) :

(Dx/y) : $x = \alpha y + \beta$ avec $\alpha = \frac{COV(X,Y)}{Var(Y)}$ ou $\alpha = r\frac{\sqrt{Var(X)}}{\sqrt{Var(Y)}}$ et

$\beta = \overline{x} - \alpha\overline{y}$

Mise sous la forme y=a'x+b' avec a'=$\mathbf{1/\alpha}$ et b'=$-\beta/\alpha$.

α=-0.36/1.81=-0,19889503, β=0,88- α *0,94= 1,06696133

a'=1/ α=-5.02, b'=- β/ α= -5,36444444

(les résultats finaux des calculs sont données exactement, les valeurs intermédiaires prises ici sont approximatives)

Les deux droites :

Dy/x : y = -0.28 x + 1.19
Dx/y : y = -5.02 x + 5.36

3- Interprétation des résultats :

Le coefficient de corrélation linéaire r = -0.24, qui présume une très mauvaise approximation linéaire de ce nuage, sa tendance décroissante est confirmée par la présence du signe négatif.

L'angle entre ces deux droites est important, ce qui ne fait que confirmer la très mauvaise approximation linéaire du nuage (**Voir Figure ci-dessous**).

B. <u>Avec la méthode des moindres rectangles</u>

On cherche à approximer ce nuage par une droite **D** de vecteur ***u'=[α1, α2].***

Pour atteindre cet objectif, on applique la formulation de la méthode des moindres rectangles :

- Calcul de la matrice V des variances-Covariance :

$$V=\begin{bmatrix} Var(x_1) & Cov(x_1,x_2) \\ Cov(x_1,x_2) & Var(x_2) \end{bmatrix}=\begin{bmatrix} 1{,}27 & -0{,}36 \\ -0{,}36 & 1{,}81 \end{bmatrix}$$

La dérivée du Lagrangien s'annule quand **Vu-λu** = 0 ➔ **(V-λI).u=0** où I est la matrice identité 2X2 et u le vecteur propre associé à la valeur propre λ (seulement la plus grande valeur propre λ_1).

Cherchons les valeurs propres de la matrice V en résolvant l'équation du second degré :

$$P(\lambda)=Dét(V-\lambda I)=\begin{vmatrix} \mathbf{1{,}27-\lambda} & \mathbf{-0{,}36} \\ \mathbf{-0{,}36} & \mathbf{1{,}81-\lambda} \end{vmatrix}=\lambda^2-3.08\lambda+2.17$$

⇨ **λ_1=1.99 et λ_2=1.09 (max λ=λ_1)**

Pour trouver les vecteurs propres associés à ces valeurs propres, nous devons résoudre :

(V-λI).u=0

$$(V-\lambda_1 I).u_1=\begin{vmatrix} \mathbf{1{,}27-1.99} & \mathbf{-0{,}36} \\ \mathbf{-0{,}36} & \mathbf{1{,}81-1.99} \end{vmatrix}\begin{vmatrix} x_1 \\ x_2 \end{vmatrix}=\begin{vmatrix} \mathbf{-0.72} & \mathbf{-0{,}36} \\ \mathbf{-0{,}36} & \mathbf{-0.18} \end{vmatrix}\begin{vmatrix} x_1 \\ x_2 \end{vmatrix}=\begin{vmatrix} \mathbf{0} \\ \mathbf{0} \end{vmatrix}$$

Ce qui conduit à résoudre le système :

$-0.72x_1-0.36x_2=0$ eq.1
$-0.36x_1-0.18x_2=0$ eq.2

Comme l'équation (2) est la double de l'équation (1) on aura $2x_1+x_2=0$ c'est-à-dire :

$\boldsymbol{u_1' = (x_1, -2x_1)}$

Il reste à nommer ce vecteur à 1 : $\|\boldsymbol{u_1}\|^2 = \mathbf{1} = (x_1^2 + 4x_1^2) = 5x_1^2$, on en déduit que : $\boldsymbol{x_1 = 1/\sqrt{5}}$ ainsi que :

$$\boxed{\boldsymbol{u_1' = (0.4472, -0.8944)}}$$

Le vecteur propre u_2 associé à la valeur propre λ_2.

$$(V-\lambda_2 I).u1 = \begin{vmatrix} 1{,}27 - 1.09 & -0{,}36 \\ -0{,}36 & 1{,}81 - 1.09 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} 0.18 & -0{,}36 \\ -0{,}36 & 0.72 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} 0 \\ 0 \end{vmatrix}$$

ce qui conduit à résoudre le système :

$0.18x_1 - 0.36x_2 = 0$ éq.1

$-0.36x_1 + 0.72x_2 = 0$ éq.2

Comme l'équation 2 est la double de l'équation 1 on aura $x_1-2x_2=0$, c'est-à-dire

$\boldsymbol{u_2' = (x_1, -x_1/2)}$

Il reste à remplacer à nommer ce vecteur à 1. $\|\boldsymbol{u_2}\|^2 = \mathbf{1} = (x_1^2 + x_1^2/4) = 5x_1^2/4$, on en déduit que $\boldsymbol{x_1 = 2/\sqrt{5}}$ ainsi que :

$$\boxed{\boldsymbol{u_2 = (0.8944, -0.4472)'}}$$

On obtient ainsi la droite de régression orthogonale, portée par le vecteur :

$\boldsymbol{u_1' = (0.4472, -0.8944)}$ correspondant à la plus grande valeur propre $\lambda_1=1.99$ en résolvant :

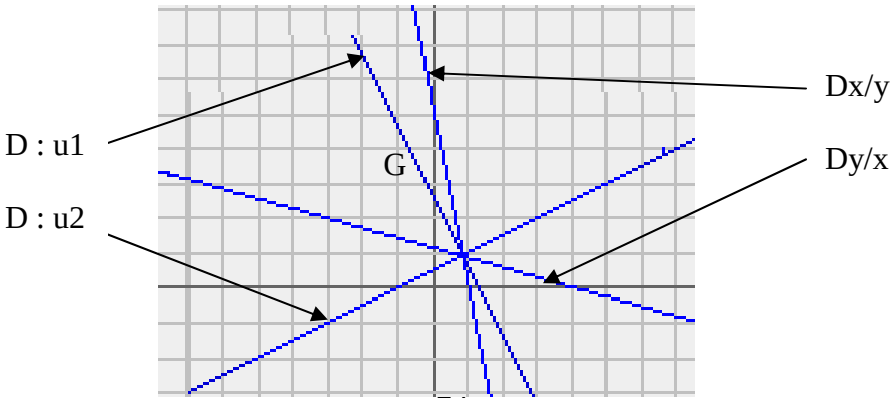
$$\boldsymbol{x_2 = \frac{-0.8944}{0.4472} x_1 + 0.94 - \frac{-0.8944}{0.4472} 0.88}$$

⇨ Soit $\mathbf{x_2=-2x_1+2.7}$ **D : u1**

L'équation de la deuxième droite de régression, associée à la deuxième valeur propre, orthogonale à la précédente :

$$\boldsymbol{x_2 = \frac{-0.4472}{0.8944} x_1 + 0.94 - \frac{-0.4472}{0.8944} 0.88}$$

⇨ Soit $\mathbf{x_2=0{,}5x_1+0.5}$ **D : u2**



On peut montrer que la pente de cette droite de régression orthogonale est comprise entre les deux droites de régression $\mathbf{D_{y/x}}$ **et** $\mathbf{D_{x/y}}$**,** c'est-à-dire qu'elle passe par **G** et se situe entre ces deux droites.

IV. L'ANALYSE GENERALE

IV.1 Présentation de l'analyse générale

La méthode générale est une méthode commune à l'ensemble des méthodes factorielles (ACP, AFC…). Quelques modifications suffisent pour passer de l'analyse générale aux méthodes factorielles.

IV.1.1 Matrice de données :

La matrice des données est un tableau, contenant des valeurs numériques, notée R formé de n lignes et p colonnes (p variables statistiques composées de n individus).
Notons $R=(r_{ij})i=1..n,\ j=1..p$

$$\boldsymbol{R}=\begin{bmatrix} \boldsymbol{r_{11}} & \boldsymbol{\ldots} & \boldsymbol{r_{1p}} \\ \boldsymbol{\ldots.} & \boldsymbol{r_{ij}} & \boldsymbol{\ldots} \\ \boldsymbol{r_{n1}} & \boldsymbol{\ldots} & \boldsymbol{r_{np}} \end{bmatrix}$$

Exemple : La matrice des données suivante est composée de 10 individus (I_1,..,I_{10}) et 2 variables (v_1 et v_2) :

Individus	v_1	v_2
I_1	0.5	0
I_2	-0.1	1.2
I_3	-0.5	0.5
I_4	-0.3	0.1
I_5	0	2.5
I_6	1.6	-0.7
I_7	2	2
I_8	2.4	1.2
I_9	0.5	3.5
I_{10}	2.7	-0.9

Les analyses dérivant de l'analyse générale que nous verrons plus tard transforment les données initiales de la matrice R, la nouvelle matrice ainsi créée sera appelée X. Comme dans la méthode générale la matrice R n'est pas transformée nous aurons R=X.

IV.1.2 Généralités :

On cherche à résoudre le problème d'approximation numérique suivant. Soit une matrice à n lignes et p colonnes $X=(x_{ij})i=1..n,\ j=1..p$, est il possible de reconstituer les ***n.p*** valeurs à $\boldsymbol{x_{ij}}$ partir d'un plus petit nombre de valeurs numériques ? Autrement dit, cela revient à approximer le nuage des n points dans un sous espace $\mathbb{R}^q$ avec q<p variables, ou plus simplement dans notre exemple à approximer le nuage des 10 points par seulement une variable (rappelons à ce stade que c'est exactement ce que nous avons fait dans le l'approximation par la droite de régression orthogonale correspondant à la plus grande valeur propre).

Ainsi, on cherche une approximation de rang q (q<p) pour X de la forme :

$X=v_1u_1'+v_2u_2'+\ldots.+v_qu_q'+E$, où v_i est (n,1), ui est (p,1) et E est (n,p).

E est une matrice résiduelle (n,p) dont les termes suffisamment petits pour que X soit bien approximée par seulement les q(n+p) valeurs des q premiers termes de la décomposition de X. afin de se fixer les idées, pour n=1000 et p=100 si q=10 on pourrait remplacer les 1000.100=100 000 valeurs de X par seulement 10(1000+100)=11 000 valeurs.

Deux ajustements sont concevables selon la lecture de X en lignes ou en colonnes :

- Les n lignes peuvent être considérées comme n points d'un espace euclidien $\mathbb{R}^p$ ($\mathbb{R}$ désigne l'ensemble des réels). C'est-à-dire la ligne i de X : ($x_{i1},\ldots,x_{ip}$) $\in \mathbb{R}^p$ est une observation dans $\mathbb{R}^p$. Pour notre exemple les 10 lignes de X sont des points de $\mathbb{R}^2$, le plan euclidien usuel , déjà vu et libellés par $i_1,i_2,\ldots,i_{10}$.
- Les p colonnes de X peuvent être considérées comme p points d'un espace euclidien $\mathbb{R}^n$. C'est-à-dire que la colonne j de X (x1j,…,xnj) $\in \mathbb{R}^n$ est une observation dans $\mathbb{R}^n$. Pour notre exemple les 2 colonnes de X sont des points de $\mathbb{R}^{10}$, ce qui est bien sûr beaucoup plus difficile à imaginer.

A chacun de ces lectures correspondra deux ajustements : l'ajustement du nuage des n observations (les individus) par un sous-espace de $\mathbb{R}^p$, l'ajustement du nuage des p observations (les variables) par un sous-espace de $\mathbb{R}^n$.
Tous les espaces sur lesquels nous travaillons sont munis de la distance euclidienne usuelle qui est, rappelons-le par exemple dans $\mathbb{R}^p$.

La distance de M=($x_1,x_2,\ldots,x_p$)' à N=($y_1,y_2,\ldots,y_p$)' est MN=$\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2}$ ce qui donne dans $\mathbb{R}^2$
MN=$\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$.

IV.2 L'ajustement du nuage des n observations (les individus) par sous-espace de $\mathbb{R}^p$

IV.2.1 Cas particulier de 2 variables (p=2)

On reprend ce que nous avons vu dans l'ajustement par la méthode des moindres rectangles jusqu'au niveau où on prit le point moyen G comme origine **[Fig. IV.1]**. Dans l'analyse générale, le point O sera considéré comme origine O=(0,0), l'origine usuelle de $\mathbb{R}^2$.
Le schéma suivant précise ce nouveau critère et les notations :

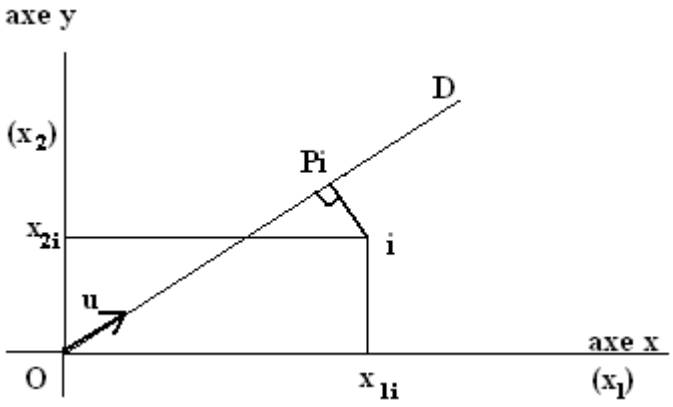


Fig. IV.1 Droite des moindres rectangles

Le point i se projette orthogonalement en Pi sur la droite D.

On cherche à minimiser la somme des carrées des longueurs de ces projections, soit $\sum_{i=1}^{n} iP_i^2$.

Cette minimisation revient à maximiser le carré de la distance de la projection Pi à l'origine O :

$$\text{Minimiser } \sum_{i=1}^{n} iP_i^2 \Leftrightarrow \text{maximiser } \sum_{i=1}^{n} OP_i^2$$

Cette distance OPi est la projection du vecteur $\overrightarrow{Oi}$ sur la droite D, de vecteur directeur $\vec{u}$.
La longueur de cette projection est le produit scalaire $X_i\vec{u}$, où Xi représente la ligne i de X :

$$X_i\vec{u}=[x_{i1}, x_{i2}]\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix}=[\alpha_1 . x_{i1}+\alpha_2 . x_{i2}]$$

La somme des carrées de ces longueurs vaut donc :

$$\sum_{i=1}^{n} OP_i^2 = \sum_{i=1}^{n} [X_i.u].[X_i.u] = [X_1.u,...,X_n.u].\begin{bmatrix}X_1.u\\ ...\\ X_n.u\end{bmatrix}$$

$$\begin{bmatrix}X_1.u\\ ...\\ X_n.u\end{bmatrix}=\begin{bmatrix}x_{11} & x_{12}\\ ... & ...\\ x_{n1} & x_{n2}\end{bmatrix}.\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix}=X.u$$, matrice (n,1) et que $\begin{bmatrix}X_1.u\\ ...\\ X_n.u\end{bmatrix}$ admet pour transposé

$[X_1.u \quad ... \quad X_n.u]$, c'est-à-dire (Xu)', matrice (1,n) on aura la formule fondamentale :

$$\sum_{i=1}^{n} OP_i^2 = (Xu)'(Xu)$$

Cette écriture semble un peu compliquée pour notre problème où nous avons seulement 2 variables, et nous aurions, bien sûr, pu l'écrire plus simplement. Son intérêt réside dans le fait qu'elle se généralise au cas de plus de deux variables en remplaçant les 2 variables précédentes par p variables. Il suffit maintenant que nous trouvions l'axe, que nous appellerons F1 (F comme factoriel), porté par u (cet axe correspond à la droite D précédente).

Cela revient à trouver le vecteur $u=[\alpha_1, \alpha_2]$ qui rend maximale la valeur $(Xu)'(Xu) = u'X'Xu$.

Le problème mathématique soulevé par cette maximisation est lié à la dérivation matricielle. Remarquons tout d'abord qu'il s'agit de trouver $[\alpha_1, \alpha_2]$', on doit donc imposer à ces deux coefficients une condition de normalisation car sinon on pourrait toujours avoir une valeur aussi grande que possible en multipliant α_1 par un facteur aussi grand que l'on veut. Nous prendrons u de norme 1, c'est-à-dire qu'il vérifie $\alpha_1^2+\alpha_2^2=1$. Il reste maintenant à trouver u'=$[\alpha_1, \alpha_2]$ qui maximise $(Xu)'(Xu)$, avec u'u=1. La démonstration fait appel à la dérivation matricielle pour le calcul d'extrema sous contraintes (méthode de Lagrange).
En notant $\mathcal{L}$ le Lagrangien $\mathcal{L}$=u'X'Xu-λ(u'u-1) où λ désigne le multiplicateur de Lagrange.

La dérivée par rapport au vecteur u de ce Lagrangien nous conduit à :

$d\mathcal{L}/du=2X'Xu-2\lambda u$

Cette dérivée s'annule quand $X'Xu-\lambda u=0$ c'est-à-dire $X'Xu=\lambda u$. ce qui signifie que u est vecteur propre de X'X à la valeur propre λ (Remarquons que X'X est symétrique). La somme des carrées des longueurs de ces projections, étant égale à λ, qui représente l'inertie maximum du nuage, sera donc maximale quand la valeur propre λ de la matrice X'X associée au vecteur propre u sera maximale. Nous noterons λ_1 la valeur propre maximale de X'X et u_1 le vecteur propre normé associé à cette valeur propre.

Résumons les résultats de cette approche théorique :

F_1 est le sous espace de $\mathbb{R}^2$ associé au vecteur propre u_1 correspondant à la plus grande valeur propre λ_1 de X'X.
Nous allons maintenant réitérer ce procédé pour chercher un deuxième axe, porté par le vecteur unitaire u_2 (u_2 est de norme égale à 1), orthogonal à u_1. Nous appellerons F_2 ce deuxième axe.

La démonstration fait appel à nouveau à la dérivation matricielle pour le calcul d'extrema sous contraintes. Ici nous cherchons le vecteur u_2 sous les contraintes : $u_2'u_1=0$ (orthogonalité de u_1 et u_2) et $u_2'u_2=1$ (normalisation de u_2).

En notant $\mathcal{L}$ le lagrangien $\mathcal{L}=u_2'X'Xu_2-\lambda(u_2'u_2-1)-\mu(u_2'u_1)$, où λ et μ désignent les multiplicateurs de Lagrange. La dérivée par rapport au vecteur u_2 de ce Lagrangien nous conduit à :

$d\mathcal{L}/du_2=2X'Xu_2-2\lambda u_2-\mu u_1$

Cette dérivée s'annule quand $2X'Xu_2-2\lambda u_2-\mu u_1=0$
En multipliant à gauche par u_1' : $2u_1'X'Xu_2-2\lambda u_1'u_2-\mu u_1'u_1=0$
Ou $2[u_1'X'X-\lambda u_1']u_2-\mu u_1'u_1=0$, ainsi :

- Pour $\lambda=\lambda_1$, $X'Xu_1=\lambda_1 u_1$ ou par transposition $u_1'X'X=\lambda_1 u_1'$ (X'X est symétrique), le terme entre crochets est donc nul et on obtient $\mu u_1'u_1=0$, soit $\mu=0$;
- Pour $\lambda\neq\lambda_1$, $X'Xu_2-\lambda u_2=0$, c'est-à-dire $X'Xu_2=\lambda u_2$. Ce qui signifie que u_2 est vecteur propre de X'X associé à la deuxième valeur propre $\lambda=\lambda_2$.

Résumons nos résultats :

F_2 est le sous espace de $\mathbb{R}^2$ associé au vecteur propre u_2 correspondant à la deuxième valeur propre λ_2 de X'X.
On a ainsi constitué une nouvelle base orthonormée de $\mathbb{R}^2$, la base $\{u_1,u_2\}$ formée des vecteurs propres de X'X associé respectivement aux deux valeurs propres λ_1 et λ_2 ($\lambda_1>\lambda_2$).

IV.2.2 Cas général de p variables :

Il suffit de généraliser le cas de 2 variables au cas où p est quelconque. X est cette fois une matrice (n,p) et la matrice X'X une matrice (p,p). On cherche le vecteur normé, qui maximise (Xu)'(Xu) et la méthode précédente revient à trouver une valeur propre de X'X qui correspond à la plus grande valeur propre. Nous noterons λ_1 à la valeur propre de X'X et u_1 le vecteur propre normé associé à cette valeur propre.

- F_1 est le sous espace de $\mathbb{R}^p$ associé au vecteur propre u_1 correspondant à la plus grande valeur propre λ_1 de X'X.

Nous allons maintenant réitérer ce procéder pour chercher un deuxième axe, porté par le vecteur unitaire u_2 (u_2 est de norme égale à 1), orthogonal à u_1. Nous appellerons F_2 ce deuxième axe.

- F_2 est le sous espace de $\mathbb{R}^p$ associé au vecteur propre u_2 correspondant à la deuxième valeur propre λ_2 de X'X ($\lambda_2 \geq \lambda_2$).

En poursuivant cette procédure p fois, on constitue une nouvelle base orthonormé de $\mathbb{R}^n$, la base $\{u_1, u_2, \ldots, u_p\}$ formée des p vecteurs $\lambda_1, \lambda_2, \ldots, \lambda_p$ ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$). En se restreignant aux q premiers vecteurs propres associé aux q premières valeurs propres ($q<p$) nous avons résolu le problème de l'approximation du nuage des n points individus dans $\mathbb{R}^n$ par le sous espace $\mathbb{R}^q$ de $\mathbb{R}^p$.

Institut de sciences exactes
Département d'Informatique
Module : Analyse de données

Fiche de TD n°02 : Analyse générale

Exercice 01 :

Soit le tableau suivant concernant deux variables x_1 et x_2 mesurées sur 10 individus.

X=

Individus i	x_1	x_2
1	0.5	0
2	-0.1	1.2
3	-0.5	0.5
4	-0.3	0.1
5	0	2.5
6	1.6	-0.7
7	2	2
8	2.4	1.2
9	0.5	3.5
10	2.7	-0.9
$\sum$	**8.8**	**9.4**

En utilisant les étapes de l'analyse générale, calculez les valeurs et les vecteurs propres de ce tableau de données.

Exercice 02 :

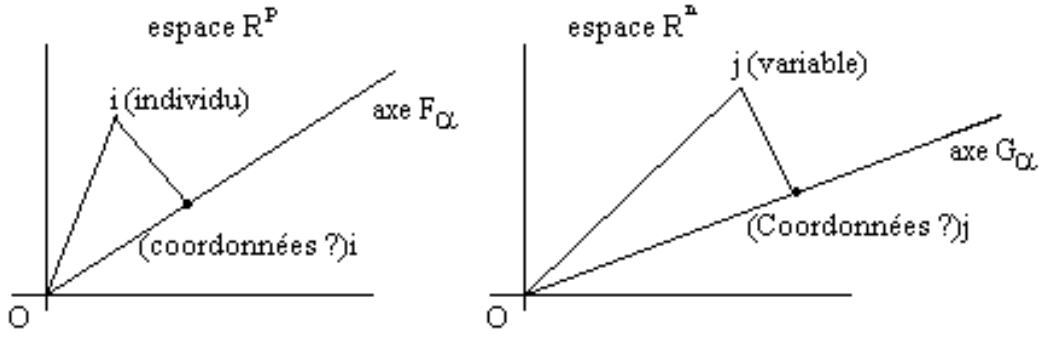
Suivant les mêmes étapes de l'ajustement du nuage des n observations (les individus) par un sous-espace de $\mathbb{R}^p$ vues en analyse générale, effectuez l'ajustement du nuage des p observations (les variables) par sous-espace de $\mathbb{R}^n$

Exercice 03 :

Donner, dans $\mathbb{R}^p$ et dans $\mathbb{R}^n$, le système de relation entre les deux sous espaces $\mathbb{R}^q$ de $\mathbb{R}^p$ et de $\mathbb{R}^n$, et ceci, suivant les règles de l'analyse générale.

Exercice 04 :

A l'aide des résultats de l'exercice précédent, déduisez les coordonnées des points dans les deux espaces.



Exercice 05 :

Définir les coordonnées du point n° 07 ([2,2]), de l'exercice 01 (ci-dessus), dans les deux sous espaces $\mathbb{R}^q$ de $\mathbb{R}^p$ et de $\mathbb{R}^n$ sur les axes F1 et F2 puis les coordonnées de la variable n° 01 sur les axes G1 et G2.

Exercice 06 :

A partir des résultats de l'exercice n° 03 ci-dessus, extraire les formules de reconstitution approchée du tableau X de départ, sachant qu'on se limite seulement à q premiers axes factoriels ($q<p$).

Institut de sciences exactes
Département d'Informatique
Module : Analyse de données

Fiche de TD n°02 : Analyse générale

Exercice 01 :

Soit le tableau suivant concernant deux variables x_1 et x_2 mesurées sur 10 individus.

X=

Individus i	x_1	x_2
1	0.5	0
2	-0.1	1.2
3	-0.5	0.5
4	-0.3	0.1
5	0	2.5
6	1.6	-0.7
7	2	2
8	2.4	1.2
9	0.5	3.5
10	2.7	-0.9
$\sum$	**8.8**	**9.4**

En utilisant les étapes de l'analyse générale, calculez les valeurs et les vecteurs propres de ce tableau de données.

Exercice 02 :

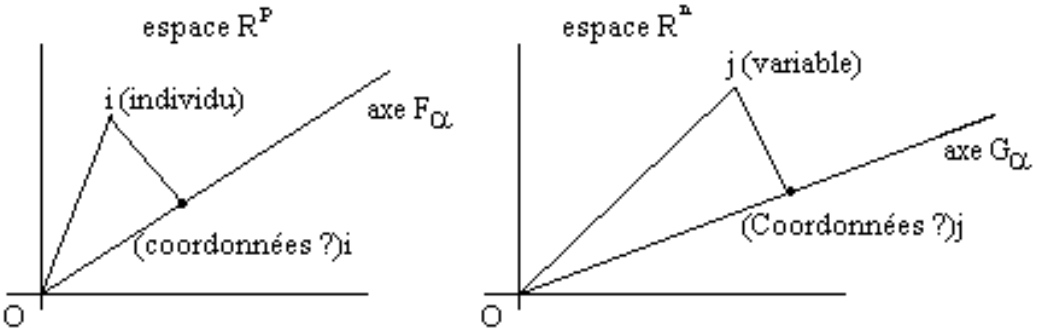
Suivant les mêmes étapes de l'ajustement du nuage des n observations (les individus) par un sous-espace de R^p vues en analyse générale, effectuez l'ajustement du nuage des p observations (les variables) par sous-espace de R^n

Exercice 03 :

Donner, dans R^p et dans R^n, le système de relation entre les deux sous espaces R^q de R^p et de R^n, et ceci, suivant les règles de l'analyse générale.

Exercice 04 :

A l'aide des résultats de l'exercice précédent, déduisez les coordonnées des points dans les deux espaces.



Exercice 05 :

Définir les coordonnées du point n° 07 ([2,2]), de l'exercice 01 (ci-dessus), dans les deux sous espaces R^q de R^p et de R^n sur les axes F1 et F2 puis les coordonnées de la variable n° 01 sur les axes G1 et G2.

Exercice 06 :

A partir des résultats de l'exercice n° 03 ci-dessus, extraire les formules de reconstitution approchée du tableau X de départ, sachant qu'on se limite seulement à q premiers axes factoriels (q<p).

Institut de sciences exactes
Département d'Informatique
Module : Analyse de données

Correction de la Fiche de TD n° 02 : Analyse générale

Exercice n 01 :

Le calcul des valeurs propres et des vecteurs propres :

$$\begin{bmatrix} 0.5 & -0.1 & -0.5 & -0.3 & 0 & 1.6 & 2 & 2.4 & 0.5 & 2.7 \\ 0 & 1.2 & 0.5 & 0.1 & 2.5 & -0.7 & 2 & 1.2 & 3.5 & -0.9 \end{bmatrix} * \begin{bmatrix} 0.5 & 0 \\ -0.1 & 1.2 \\ -0.5 & 0.5 \\ -0.3 & 0.1 \\ 0 & 2.5 \\ 1.6 & -0.7 \\ 2 & 2 \\ 2.4 & 1.2 \\ 0.5 & 3.5 \\ 2.7 & -0.9 \end{bmatrix} = \begin{bmatrix} 20.46 & 4.68 \\ 4.68 & 26.94 \end{bmatrix}$$

$$Donc\ X'X = \begin{bmatrix} 20.46 & 4.68 \\ 1.68 & 26.94 \end{bmatrix}$$

Cherchons les valeurs propres de cette matrice en résolvant l'équation du second degré :

$$P(\lambda)=Dét(X'X-\lambda I)=\begin{vmatrix} 20.46 & 4.68 \\ 4.68 & 26.94 \end{vmatrix} = \lambda^2-\lambda trace(X'X)+Dét(X'X)$$

$$=\lambda^2-\lambda(20.46+26.94)+[20.46*26.94-(4.68)^2]=\lambda^2-47.4\lambda+529.29$$

Cette équation admet les deux racines réelles $\lambda_1=29.39$ et $\lambda_2=18.01$

Pour trouver les deux vecteurs propres à ces deux valeurs propres, nous devons résoudre :
$(X'X-\lambda I)u=0$, où I est la matrice identité 2x2 et u le vecteur propre associé à la valeur propre λ.
Le vecteur propre u_1 associé à la valeur propre λ_1 :

$$(X'X-\lambda_1 I).u_1=\begin{bmatrix} 20.46-29.39 & 4.68 \\ 4.68 & 26.94-29.39 \end{bmatrix}.\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} -8.93 & 4.68 \\ 4.68 & -2.45 \end{bmatrix}.\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Ce qui nous conduit à résoudre le système :

$$-8.93x_1+4.68x_2=0 \quad éq.(1)$$
$$4.68x_1-2.45x_2=0 \quad éq.(2)$$

L'équation (2) est proportionnelle à l'équation (1) (et c'est toujours le cas) on aura :
$x_1-0.52x_2=0$, c'est-à-dire $u_1'=(0.52x_2,x_2)$.
Il reste à normé ce vecteur à 1 : $\|u_1\|^2=1=((0.52x_2)^2+x_2^2)=1.27x^2=(1.13x^2)^2$, on en déduit que $x_2=1/1.13=0.89$. Ainsi $u_1'=(0.46\ ;0.89)$.
Remarquons que ce vecteur est défini à une orientation des axes près, ce sera toujours le cas des vecteurs propres.

Le vecteur propre u_2 associé à la valeur propre λ_2 :

$$(X'X-\lambda_2 I).u_2=\begin{bmatrix}\mathbf{20.46-18.01} & \mathbf{4.68}\\ \mathbf{4.68} & \mathbf{26.94-18.01}\end{bmatrix}.\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}\mathbf{2.45} & \mathbf{4.68}\\ \mathbf{4.68} & \mathbf{8.93}\end{bmatrix}.\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

Ce qui nous conduit à résoudre le système :

$2.45x1+4.68x2=0$ (éq.1)
$4.68x1+8.93x2=0$ (éq.2)

Comme l'équation (2) est proportionnelle à l'équation (1) on aura $x_1+1.91x_2=0$, c'est-à-dire $u1=(-1.91x_2,x_2)'$.
Il reste à normer ce vecteur à 1 : $\|\boldsymbol{u_1}\|^2=1=((-1.91x_2)^2+x_2^2))=4.65x_2^2=(2.16x_2)^2$, on en déduit que $x_2=1/2.16=0.46$
Ainsi $u_2=(-0.89\ ;0.46)'$

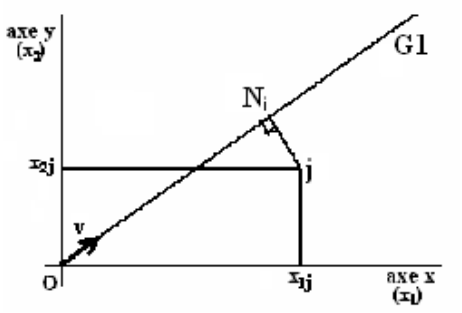
Remarque :

Comme les vecteurs propres d'une matrice réelle symétrique sont orthogonaux deux à deux il n'était pas nécessaire de calculer u_2, on aurait pu le déduire à partir de u_1, on peut ailleurs le vérifier en effectuant leurs produit scalaire : $u_1.u_2=(0.46)(-0.89)+(0.89)(0.46)=0$

Exercice n° 02 :

Après avoir vu dans le cours les résultats de l'ajustement du nuage des n observations (les individus) par un sous espace de R^p, essayez de suivre les mêmes étapes, donner les résultats de l'ajustement des p observations (les variables) par un sous espace de R^n.

Solution :



On cherche une droite G1 de R^n, qui passe par l'origine O, de vecteur unitaire v (v est de (n,1)). Notons X^j, le vecteur de R^n, qui représente la colonne j de la matrice des données X. La projection de la variable j (point Nj) sur G1 est : $v'X^j=\sum_{i=1}^{n} z_i x_{ij}$ ou $X^{j'}v=\sum_{i=1}^{n} x_{ij} z_i$, maximiser la somme des carrées de ces projections revient donc à

$$\text{maximiser}\sum_{j=1}^{p} ON_j^2=\sum_{j=1}^{p}\left[v'X^j\right]\left[v'X^j\right]=\left[v'X^1, v'X^2,..., v'X^p\right]\begin{bmatrix}v'X^1\\ ...\\ v'X^p\end{bmatrix}=v'XX'v \text{ où } v=\begin{bmatrix}z_1\\ ...\\ z_n\end{bmatrix}$$

Ou
$$\sum_{j=1}^{p} ON_j^2=\sum_{j=1}^{p}\left[X^{j'}v\right]\left[X^{j'}v\right]=\left[X^{1'}v, X^{2'}v,..., X^{p'}v\right]\begin{bmatrix}X^{1'}v\\ ...\\ X^{p'}v\end{bmatrix}$$

$$\begin{bmatrix} X^1{}'v \\ \dots \\ X^p{}'v \end{bmatrix} = X'v \quad \text{et} \quad \left[X^1{}'v, X^2{}'v, \dots, X^p{}'v\right] = \begin{bmatrix} v'X^1 \\ \dots \\ v'X^p \end{bmatrix} = v'X$$

$$\sum_{j=1}^{p} ON_j^2 = \sum_{j=1}^{p} \left[X^{j'}v\right]\left[X^{j'v}\right] = \left[X^1{}'v, X^2{}'v, \dots, X^p{}'v\right] \begin{bmatrix} X^1{}'v \\ \dots \\ X^p{}'v \end{bmatrix} = v'XX'v$$

Ainsi le vecteur v de R^n doit rendre maximal v'XX'v sous la contrainte v'v=1. La démarche de l'ajustement précédente s'applique à nouveau (en remplaçant u par v, Fj par Gi, X'X (matrice symétrique (p,p) par XX' (matrice symétrique (n,n)).
En poursuivant cette procédure n fois, on constitue une nouvelle base orthonormée de R^n, la base {v1,v2,…,vn} formée des n vecteurs propres de XX' associés aux n valeurs propres μ1, μ2,…, μn (μ1≥μ2≥…≥μn). en se restreignant aux q premiers vecteurs propres associés aux q premières valeurs propres (q<p) nous avons résolu le problème de l'approximation du nuage des p points variables dans R^n par le sous espace R^q de R^n.

Exercice 3 :

Cherchons les relations qui lient les deux espaces R^q qui approximent R^p et R^n. notons u_α et λ_α, α=1,..,q les q premiers vecteurs et valeurs propres de X'X et de même v_α et μ_α, α=1,..,q les q premiers vecteurs propres et valeurs propres de X'X (nous supposons dans ce qui suit que les valeurs propres λα et uα sont non nulles).
Nous avons les relations :
Dans R^p $X'Xu_\alpha = \lambda_\alpha u_\alpha$ (éq.1)
Dans R^n $XX'v_\alpha = \mu_\alpha v_\alpha$ (éq.2)
Si on note r le rang de X'X (resp. de X'X), on sait que r est aussi le rang de X et r≤min(n,p).
Comme

$XX'v_\alpha = \mu_\alpha v_\alpha$
$X'[XX'v_\alpha] = X'[\mu_\alpha v_\alpha]$
$[X'X]\,[X'v_\alpha] = \mu_\alpha\,[X'v_\alpha]$
ce qui signifie qu'à chaque vecteur propre v_α (α≤r) de XX' correspond un vecteur propre de $X'v_\alpha$ X'X relatif à la même valeur propre μ_α. Toute valeur propre de X'X est donc valeur propre de XX'.
Comme tout vecteur propre est défini à une homothétie près, les vecteurs propres u_α et $X'v_\alpha$ sont proportionnels, soit k_α ce coefficient de proportionnalité :
$u_\alpha = k_\alpha X'v_\alpha$ (éq.3)
or $X'Xu_\alpha = \lambda_\alpha u_\alpha$ (éq.1)
donc $X[X'Xu_\alpha] = X[\lambda_\alpha u_\alpha]$
$[XX'][Xu_\alpha] = \lambda_\alpha[Xu_\alpha]$
ce qui signifie qu'à chaque vecteur propre u_α (α≤r) de X'X correspond un vecteur propre Xu_α de X'X relatif à la même valeur propre λ_α.Toute valeur propre de X'X est donc valeur propre de X'X.
Ces vecteurs propres v_α et Xu_α sont donc proportionnalité :
$v_\alpha = k'_\alpha Xu_\alpha$ (éq.4)
Ainsi pour α≤r, $\mu_\alpha=\lambda_\alpha$,cherchons le lien entre ces deux constantes k_α et k'_α. Les vecteurs propres étant normés on a :=1
$u'_\alpha.u_\alpha = v'_\alpha v_\alpha = 1$, en remplaçant par les équations 3 et 4 on obtient
$[k_\alpha X'v_\alpha]'[k_\alpha X'v_\alpha] = [k'_\alpha Xu_\alpha]'[k'_\alpha Xu_\alpha]$
$k^2_\alpha\, v_\alpha'\, XX'v_\alpha = k'^2_\alpha u'_\alpha X'Xu_\alpha$
$k^2_\alpha\, \mu_\alpha = k'^2_\alpha\, \lambda_\alpha = 1$ et comme $\mu_\alpha = \lambda_\alpha$
$k^2_\alpha = k'^2_\alpha = 1/\lambda_\alpha$ et donc $k_\alpha = k'_\alpha = 1/\sqrt{\lambda_\alpha}$

Obtient ainsi le système de relations fondamentales suivantes qui lient les deux espaces :

Dans R^p : $u_\alpha = X'v_\alpha / \sqrt{\lambda_\alpha}$	**et dans R^n : $v_\alpha = Xu_\alpha / \sqrt{\lambda_\alpha}$**

L'axe F_α qui porte le vecteur unitaire u_α est appelé α–ième axe factoriel de R^p, l'axe G_α qui porte le vecteur unitaire v_α est appelé α-ième axe factoriel de R^n.

Exercice 04 :

A l'aide des résultats précédents nous pouvons déduire les coordonnées des points dans les deux espaces :

- Sur l'axe F_α dans R^p, les coordonnées des n points individus sont par construction les composantes (les lignes) de Xu_α matrice (n,1)=(n,p)x(p,1) comme $Xu_\alpha = \sqrt{\lambda_\alpha}\, v_\alpha$ la ligne i de Xu_α sera $(Xu_\alpha)_i = \sqrt{\lambda_\alpha}\, (v_\alpha)_i$
- Sur l'axe G_α dans R^n, les coordonnées des p points variables sont par construction les composantes (les lignes) de $X'v_\alpha$ matrice (p,1)=(p,n)x(n,1),

Comme $X' v_\alpha = \sqrt{\lambda_\alpha}\, u_\alpha$ la ligne j de $X' v_\alpha$ $(X' v_\alpha)_j = \sqrt{\lambda_\alpha}\, (u_\alpha)_j$

(ce sont ces formules que nous utiliserons pour le calcul des coordonnées des p points variables).

Il y a ainsi proportionnalité entre les coordonnées d'un point sur un axe α dans un espace et les composantes unitaires de l'axe α dans l'autre espace : par exemple la composante i du point individu sur l'axe F_α dans R^p est proportionnelle à la i-ième composante du vecteur unitaire v_α de l'axe G_α dans R^n.

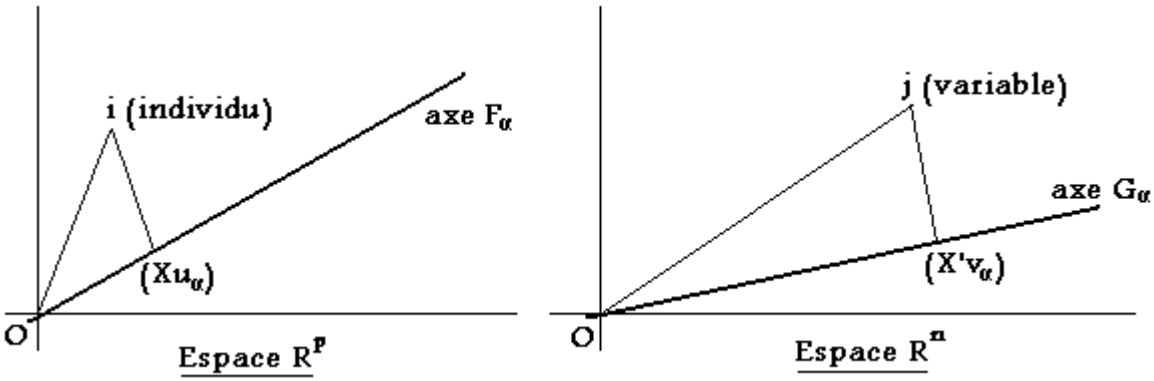


Schéma pour le calcul des coordonnées sur les axes factoriels

Exercice 05 :

- coordonnées de l'individu 7 sur les axes F_1 et F_2 :

  - Pour l'axe F_1 ($7^{ème}$ ligne de X)x(les 2 lignes de u_1)=[2 , 2] $\begin{bmatrix} 0,46 \\ 0,89 \end{bmatrix}$ =2,7

  - Pour l'axe F_2 ($7^{ème}$ ligne de X)x(les 2 lignes de u_2)=[2 , 2] $\begin{bmatrix} -0,89 \\ 0,46 \end{bmatrix}$ =-0,86

Ainsi l'individu 7 aura (2,7;-0,86) comme nouvelles coordonnées sur le repère $\{F_1,F_2\}$ passant par l'origine.

- Les coordonnées des variables sur les axes G_1 et G_2 :

$(X'v_\alpha) = \sqrt{\lambda_\alpha}\, (u_\alpha)$ et α varie entre 1 et P=2.

Les coordonnées de la $1^{ère}$ variable : $[\sqrt{\lambda_1}(u_1)_1 , \sqrt{\lambda_2}\, (u_2)_1]$

Les coordonnées de la $2^{ème}$ variable : $[\sqrt{\lambda_1}(u_1)_2 , \sqrt{\lambda_2}\, (u_2)_2]$

$u_1 = \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix}$ $u_2 = \begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix}$

Coordonnées de la variable 1 sur les axes G_1 et G_2 :

- La première coordonnée est calculée par $(X'v_1)_1=\sqrt{\lambda_1}\ (u_1)_1=\sqrt{29,39}$ x0,46=2,49
- La deuxième coordonnée est calculée par $(X'v_2)_1=\sqrt{\lambda_2}\ (u_2)_1=\sqrt{18,01}$ x(-0,89)=-3,77

Exercice 06 :

Reconstitution et reconstitution approchée du tableau X de départ

De la relation $Xu_\alpha=\sqrt{\lambda_\alpha}\ v_\alpha$ établie précédemment on déduit $(X\ u_\alpha)u_\alpha' =\sqrt{\lambda_\alpha}\ v_\alpha\ u_\alpha'$.

En sommant sur l'ensemble des vecteurs propres dans R^p, α=1..p, il vient :

$$X\sum_{\varepsilon=1}^{p} u_\alpha u_\alpha' = \sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha} v_\alpha u_\alpha'$$

Or $\sum_{\alpha=1}^{p} u_\alpha u_\alpha' = u_1u_1'+\ldots+ u_pu_p'$, matrice (p,1)x(1,p)=(p,p),

En multipliant à gauche par u_1', on obtient :

$u_1'(u_1u_1'+\ldots+ u_pu_p')= u_1'u_1u_1'+ u_1'u_2u_2'+\ldots+ u_1'u_pu_p')= u_1'1+0+\ldots+0= u_1'$

Car $u_1'u_1=1$ (vecteur $u_1$normé) et $u_1'u_i=0$ (vecteur u_1 et u_i orthogonaux pour $i\neq1$),

ainsi $u_1u_1'+..+ u_pu_p'=I_p$, matrice unité d'ordre p.

Donc : $X=\sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha} v_\alpha u_\alpha'$, matrice (n,1)x(1,p), qui est la formule de reconstitution du tableau X.

Si on se limite aux q axes (q≤p) et si les valeurs propres $\lambda_{q+1},\ldots, \lambda_p$ sont négligeables par rapport aux q premières, on obtient la reconstitution approchée X^* du tableau de départ X, composé de q termes :

$$X^*=\sum_{\alpha=1}^{q}\sqrt{\lambda_\alpha} v_\alpha u_\alpha'$$

Les np valeurs initiales de X sont remplacées par les q(n+p) valeurs de X^*. La quantité globale de la reconstitution du tableau de données par q axes est :

$\tau_q = \dfrac{\sum_{\alpha=1}^{q}\lambda_\alpha}{\sum_{\alpha=1}^{p}\lambda_\alpha}$ Appelée taux d'inertie ou part de variance qui mesure la part de la "***dispersion***" du nuage expliquée par le sous espace à q dimensions.

Exemple : si on résumait le nuage de l'exercice n° 01 à deux dimensions, correspondant aux variables x_1 et x_2, à une seule dimension, correspondant à l'axe factoriel u_1, on conserverait dans cette approximation 62 % de l'information.

V. L'Analyse en Composantes Principales - ACP

Nous allons nous appuyer sur les résultats de l'analyse générale. Nous traiterons parallèlement au cas général d'une matrice (n,p) (n lignes qui correspondent aux individus, p colonnes qui correspondent aux variables).

Nous allons donc simplifier un tableau de données statistique où les variables seront des variables quantitatives. Le but n'étant pas seulement de réduire ce tableau, comme dans le cas de l'analyse générale, mais aussi d'interpréter statistiquement et graphiquement ce tableau réduit.

Présentons d'abord quelques généralités sur l'analyse en composantes principales.

Généralités

La matrice des données est un tableau contenant les valeurs numériques qui serviront aux calculs. Cette matrice R, contient n lignes et p colonnes représentent p variables statistiques composées de n individus ou observations.

On note $R = (r_{ij}) = 1….n, j = 1…..p$ cette matrice nxp où i est l'indice des lignes et j l'indice des colonnes.

$$R = \begin{bmatrix} r_{11} & \dots & r_{1j} & \dots & r_{1p} \\ \dots & & & & \\ r_{i1} & \dots & r_{ij} & & r_{ip} \\ \dots & & & & \\ r_{n1} & & r_{nj} & & r_{np} \end{bmatrix}$$

r_{ij} est l'observation de la variable j sur l'individu i.

Comme dans l'analyse générale nous allons effectuer l'analyse dans les deux espaces : d'abord l'étude du nuage des n individus dans l'espace $\mathbb{R}^p$ des p variable, puis l'étude du nuage des p variable dans l'espace $\mathbb{R}^n$ des individus.

V.1 Analyse du nuage des n individus dans l'espace R^p des variables

V.1.1 Généralités

On souhaite obtenir une représentation des projections n points individus dans un sous-espace de l'espace R^p des variables, c'est-à-dire un espace de dimension inférieure à p. Le schéma suivant illustre notre approche **[Fig. V.1**] :

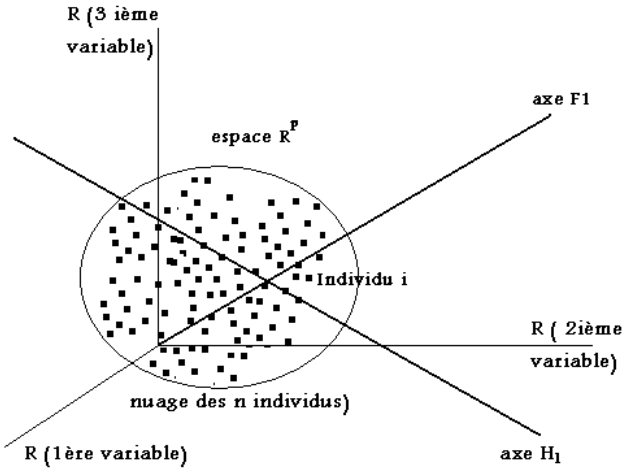


Fig. V.1 Schéma pour la recherche d'un sous–espace de $\mathbb{R}^p$

La démarche étudiée dans l'analyse générale nous a conduit à chercher un axe F_1, passant par l'origine O. Il semble plus intéressant de chercher un axe H_1 qui passe par le centre de gravité du nuage, c'est-à-dire le point moyen :

$$\bar{r} = (\bar{r}_1, \bar{r}_2, ..., \bar{r}_p)' \text{ où } \bar{r}_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij} \quad j = 1..p$$

Puisque seule la forme du nuage nous intéresse et non sa position relative par rapport à l'origine.

Le schéma suivant montre en deux dimensions la position du centre de gravité **G [Fig. V.2]** :

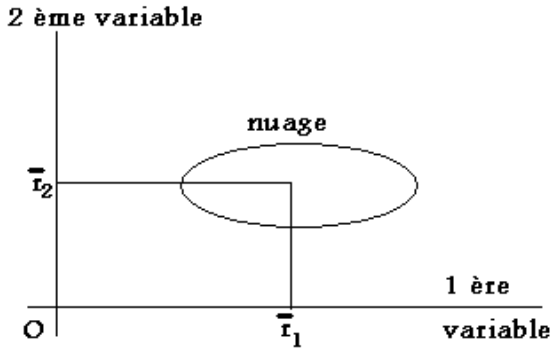


Fig. V.2 Schéma pour la recherche de l'axe H_1

Ainsi nous allons prendre comme nouvelle origine le centre de gravité G, ou point dont les coordonnées sont les moyennes des variables.

Avec ce changement d'origine, l'analyse générale précédente et ses résultats sont-ils respectés ? Précisément, va-t-on à nouveau maximiser la somme des carrés des longueurs des projections ?

Nous allons montrer que si nous projetons deux points individus i et j sur un axe H quelconque d'origine Ω, les mesures algébriques de ces projections sur l'axe étant respectivement h_i et h_j alors on pourra proposer comme critère de rendre maximale la somme (Ce n'est plus la somme des carrées des distances à l'origine en projection qu'il faut rendre maximum, mais la somme des carrées des distances entre tous les couples d'individus) :

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(h_i - h_j)^2$$

qui correspond à rendre minimale les longueurs des carrés des projections iH_i^2 ou jH_j^2 comme le montre le schéma suivant **[Fig. V.3]** :

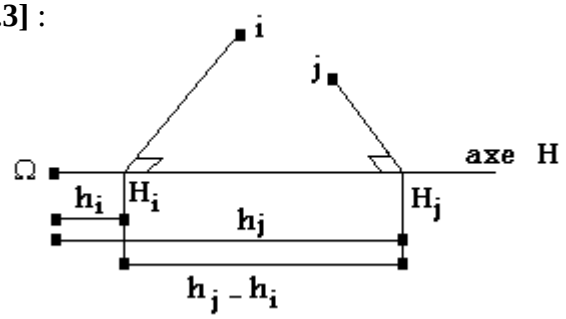


Fig. V.3 Schéma pour le critère de maximisation

Nous allons démontrer que:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(h_i - h_j)^2 = 2n\sum_{i=1}^{n}(h_i - \bar{h})^2 \qquad \text{où} \quad \bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_i$$

C'est-à-dire que si on prend comme origine le point moyen des projections, la maximisation reviendra à la maximisation d'une somme de carrés, précisément la somme

$$2n\sum_{i=1}^{n}(h_i - \bar{h})^2 \quad \text{ou, ce qui revient au même} \quad \sum_{i=1}^{n}(h_i - \bar{h})^2$$

On se retrouve bien ainsi dans le cadre de l'analyse générale, il nous suffira de prendre comme terme générique de la matrice X la valeur:

$x_{ij} = r_{ij} - \overline{r_j}$ (plus précisément un multiple de ce terme comme nous le verrons ci-après)

Afin d'évacuer toute ambiguïté nous donnons ci-dessous un schéma qui reprend la figure de l'étude générale **[Fig. V.4]** où nous avons surmonté les anciennes notations des notations correspondantes à notre analyse actuelle :

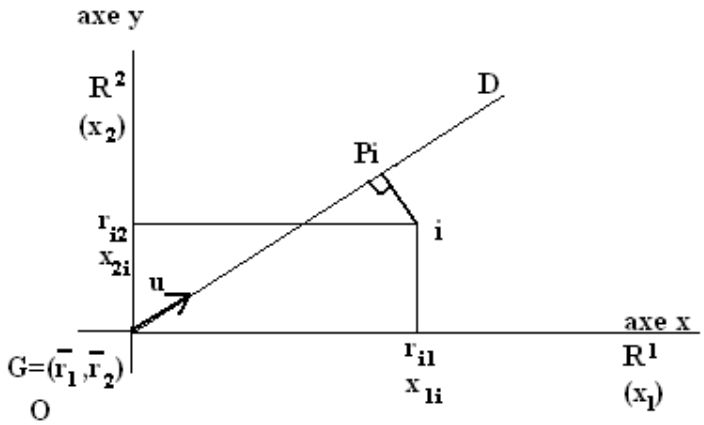


Fig. V.4 Schéma de la droite des moindres rectangles

Démontrons la formule $\sum_{j=1}^{n}\sum_{i=1}^{n}(h_i - h_j)^2 = 2n\sum_{i=1}^{n}(h_i - \bar{h})^2$ où $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_i$

Annoncée plus haut :

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(hi - h_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(h_i^2 - 2h_i h_j + h_j^2) = \sum_{i=1}^{n}(nh_i^2 - 2h_i\sum_{j=1}^{n}h_j + \sum_{j=1}^{n}h_j^2)$$

En remplaçant : $\sum_{j=1}^{n} h_j$ par $n\bar{h}$ on aura : $\sum_{i=1}^{n}(nh_i^2 - 2nh_i\bar{h} + \sum_{j=1}^{n}h_j^2) =$

$$n\sum_{i=1}^{n}(h_i^2 - 2h_i\bar{h} + \frac{1}{n}\sum_{j=1}^{n}h_j^2) \text{ [eq.1]}$$

Ce dernier terme est proche de la forme à atteindre qui est : $\sum_{i=1}^{n}(h_i - \bar{h})^2$

Il suffit seulement de trouver un équivalent de $\sum_{j=1}^{n} h_j^2$ pour effectuer une transformation de l'expression.

En partant de $\sum_{j=1}^{n}(h_j-\bar{h})^2=\sum_{j=1}^{n}(h_j^2-2h_j\bar{h}+\bar{h}^2)=\sum_{j=1}^{n}h_j^2-2n\bar{h}^2+n\bar{h}^2=\sum_{j=1}^{n}h_j^2-n\bar{h}^2$

Donc $\sum_{j=1}^{n}h_j^2=n(\bar{h}^2)+\sum_{j=1}^{n}(h_j-\bar{h})^2$

Remplaçant ce résultat dans [eq.1] on a :

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(h_i-h_j)^2=n\sum_{i=1}^{n}\left[h_i^2-2h_i\bar{h}+(\bar{h})^2\right]+\sum_{i=1}^{n}\sum_{j=1}^{n}(h_j-\bar{h})^2$$

$$=n\sum_{i=1}^{n}(h_i-\bar{h})^2+\sum_{i=1}^{n}\sum_{j=1}^{n}(h_j-\bar{h})^2$$

$$=n\sum_{i=1}^{n}(h_i-\bar{h})^2+n\sum_{j=1}^{n}(h_j-\bar{h})^2=2n\sum_{i=1}^{n}(h_i-\bar{h})^2$$

Résumons nos résultats : le tableau R=(r_{ij}) est remplacé par le tableau X=(x_{ij}) ou $x_{ij}=r_{ij}-\bar{r}_j$.

Afin de retrouver des formules statistiques usuelles nous effectuerons en fait la transformation :

$$x_{ij}=\frac{r_{ij}-\bar{r}}{\sqrt{n}},\ i=1\ldots\ldots n, j=1\ldots\ldots..p.$$

En effet d'après l'analyse générale, la matrice dont on cherche les valeurs propres est la matrice X'X de taille(p,n)x(n,p)=(p,p).Calculons le terme(k,j) de cette matrice:

$$(X'X)_{kj}=\left[x_{1k}\ldots x_{ik}\ldots x_{nk}\right]\begin{bmatrix}x_{1j}\\ \ldots\\ x_{ij}\\ \ldots\\ x_{nj}\end{bmatrix}=\sum_{i=1}^{n}x_{ik}x_{ij}=\sum_{i=1}^{n}\left(\frac{r_{ik}-\bar{r}_k}{\sqrt{n}}\right)\left(\frac{r_{ij}-\bar{r}_j}{\sqrt{n}}\right)=\frac{1}{n}\sum_{i=1}^{n}\left(r_{ik}-\bar{r}_k\right)\left(r_{ij}-\bar{r}_j\right)=\mathrm{cov}_e(R^k,R^j)$$

obtenu par la multiplication terme à terme de la k-ième ligne de X' (k-ième colonne de X) par la j-ième colonne de X.

Le terme multiplicatif $1/\sqrt{n}$ permet de donner au terme (k,j) de X'X la valeur de la covariance empirique entre les variables j (notée R^j) et k (notée R^k), qui correspondent respectivement aux colonnes j et k de la matrice R. La matrice X'X est donc la matrice (p,p) (à p lignes et p colonnes) des variances-covariance expérimentales (on dit aussi «matrice des covariance expérimentales»):

$$X'X=\begin{bmatrix}\mathrm{var}(R^1)=s_{e1}^2 & & & \\ \mathrm{cov}_e(R^2,R^1) & \mathrm{var}(R^2)=s_{e2}^2 & & \\ \ldots & \ldots & & \\ \mathrm{cov}_e(R^p,R^1) & \ldots & \mathrm{cov}_e(R^p,R^{p-1}) & \mathrm{var}(R^p)=s_{ep}^2\end{bmatrix}$$

Nous avons écrit que les terme situés sous la diagonale principale car cette matrice est symétrique, en effet $\text{cov}_e(R^j, R^k) = \text{cov}_e(R^k, R^j)$ et $\text{cov}_e(R^j, R^j) = \text{var}(R^j) = s_{ej}^2$, ces variances empirique formant les termes de la diagonale.

Il peut arriver que les dispersions des données entre colonnes, c'est-à-dire les écarts types entre les variables soient très différents les uns des autres, ce qui arrive notamment si on utilise des données numériques d'unités différentes (dinars, livre, heure, kilogrammes,…). Cette possibilité fera apparaître des distorsions dans les représentations. Pour y remédier on effectue une nouvelle transformation en divisant chacune des coordonnées d'une variable j (j-ième colonne de R) par son écart type s_{ej} : précisément on remplace x_{ij} par :

$$x_{ij} = \frac{r_{ij} - \bar{r}}{s_{ej}\sqrt{n}}, \text{ i=1……n,j=1……..p.}$$

Avec cette nouvelle transformation calculons le terme (k,j) de X'X :

$$(X'X)_{kj} = \frac{1}{n}\sum_{i=1}^{n} \frac{(r_{ik} - \bar{r_k})}{S_{ek}} . \frac{(r_{ij} - \bar{r_j})}{S_{ej}} = \frac{\text{cov}_e(R^k, R^j)}{S_{ek} S_{ej}} = corr(R^j, R^k)$$

Le terme $(X'X)_{kj}$ est donc le coefficient de corrélation linéaire entre les variables k et j, noté $corr(R^j,R^k)$. Comme $corr(R^j,R^k)=1$ (une variable est en corrélation linéaire parfaite avec elle-même) la matrice X'X est la matrice (p,p) des corrélations expérimentales entre les variables, notées C :

$$C = X'X = \begin{bmatrix} 1 & & \\ corr(R^2,R^1) & 1 & \\ \ldots & \ldots & \\ corr(R^p,R^1) & \ldots & corr(R^p,R^{p-1}) \quad 1 \end{bmatrix}$$

Nous avons écrit que les termes situés sous la diagonale principale car cette matrice est symétrique, en effet $corr(R^j,R^k)=corr(R^k,R^j)$ et $corr(R^j,R^j)=1$ formant les termes de la diagonale.

Remarque :

Quand on travaille sur la matrice des corrélations C=X'X, la diagonale de C est formée de 1. La somme des éléments de cette diagonale, la trace de C, est donc tr(C)=1+…+1=p (il y a en effet p valeurs propres si les variables initiales n'ont pas de dépendance linéaire entre elles). Des résultats d'algèbre linéaire nous confirment que la matrice C est diagonalisable dans la base des vecteurs propres $\{u_1,u_2,\ldots,u_p\}$; c'est-à-dire qu'il existe une matrice de passage P telle que :

$$C = X'X = P\Lambda P^{-1} = P\begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \lambda_p \end{bmatrix} P^{-1}$$

Cette matrice Λ est la matrice diagonale, dont la diagonale est formée des valeurs propres de X'X, les autres termes étant nuls. On montre par ailleurs que $tr(C)=Tr((P\ \Lambda\ P^{-1})=Tr(\Lambda)$. Ainsi

la somme des valeurs propres est $\lambda_1+\ldots+\lambda_p=p$. La transformation qui aboutit à utiliser la matrice des corrélations donne à chacune des variables la même variance (le même poids), précisément 1/n. Dans l'espace des <<variables>> vecteurs propres les poids sont transformés : λ_α est la variance expliquée par l'axe α et λ_α/p est la part de la variance totale expliquée par cet axe F_α qui correspond au α-ième axe principal.

Conclusion : l'analyse du nuage des n points individus dans $\mathbb{R}^p$ nous a conduit à effectuer :

- **Cas 1 :** une translation de l'origine au centre de gravité G du nuage multipliée par le coefficient $1/\sqrt{n}$ (***analyse en composantes principales non normée***);

- **Cas 2 :** un changement d'échelles, en divisant la transformation linéaire précédente par l'écart type de chacune des variables (***analyse en composantes principales normée***).

La théorie générale nous conduit à chercher les vecteurs propres u_α de X'X où :

- Cas 1 : $\boldsymbol{X'X}$ est la matrice des ***variances-covariances***;
- Cas 2 : $\boldsymbol{X'X}$ est la matrice des ***corrélations***.

Les coordonnées des points individus sur l'axe α (axe factoriel α ou composante principale α) porté par le vecteur propre u_α seront donc données par $X\,u_\alpha$.

V.2 Analyse du nuage des p variables dans l'espace $\mathbb{R}^n$ des individus

V.2.1 Généralités : contrairement à l'analyse générale, la transformation effectuée sur le tableau $R=(r_{ij})$ qui a conduit à $X=(x_{ij})$ où :

$$x_{ij}=\frac{r_{ij}-\bar{r}}{\sqrt{n}} \text{ (ACP non normée)}, \quad x_{ij}=\frac{r_{ij}-\bar{r}}{s_{ej}\sqrt{n}} \text{ (ACP normée), i=1……n,j=1……..p.}$$

Dans le cas de l'ACP montre que les indices i et j ne jouent pas un rôle symétrique dans $\mathbb{R}^p$ et dans $\mathbb{R}^n$.

Etudions tout d'abord la transformation $x_{ij}=r_{ij}-\bar{r}$, qui représentait une translation à la nouvelle origine "***centre de gravité du nuage***" dans $\mathbb{R}^p$, dans l'espace $\mathbb{R}^n$ des individus :

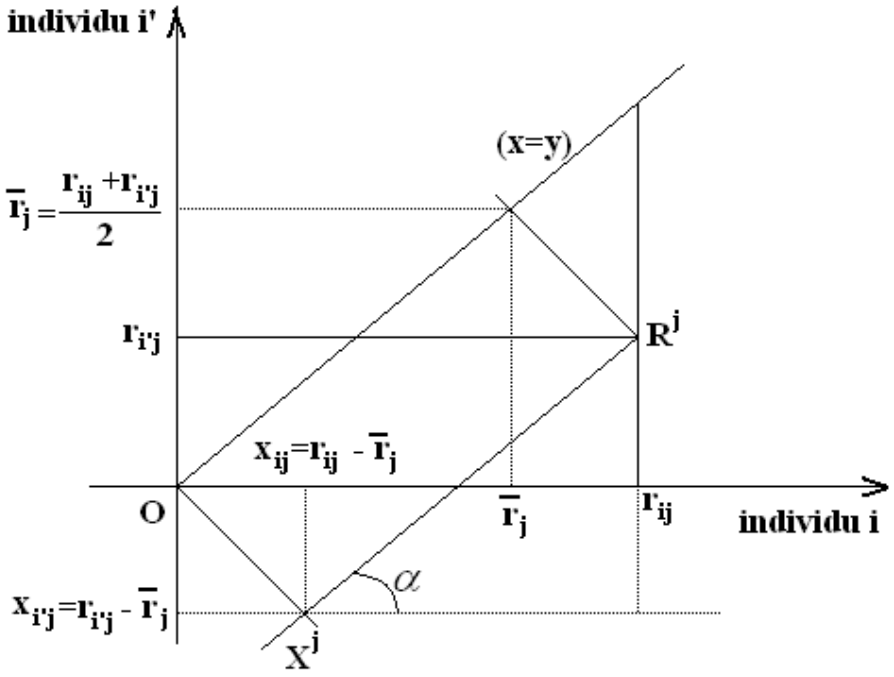


Fig. V.5 Schéma pour l'étude des variables dans l'espace des individus

A partir de ce schéma calculons la tangente de l'angle α :

$$tg(\alpha) = \frac{r_{i'j} - x_{i'j}}{r_{ij} - x_{ij}} = \frac{\bar{r}_j}{\bar{r}_j} = 1$$

Ainsi R^j est transformé en X^j par une translation parallèle à la première bissectrice du plan (i,i') de $\mathbb{R}^n$. En étendant cette observation sur deux dimensions à l'espace $\mathbb{R}^n$ à n dimensions, on en déduit que la transformation qui remplace r_{ij} par $x_{ij} = r_{ij} - \bar{r}$ est une projection parallèle à la première bissectrice de cet espace sur l'hyperplan orthogonal à cette première bissectrice, encore appelée droite des constantes (cette observation est identique si on multiplie, comme nous l'avons fait précédemment, x_{ij} par $1/\sqrt{n}$). Regardons maintenant la transformation qui modifie l'échelle des axes, c'est-à-dire la transformation qui à r_{ij} fait correspondre :

$$x_{ij} = \frac{r_{ij} - \bar{r}_j}{s_{ej}\sqrt{n}}, \text{ i=1……n,j=1……..p}$$

Pour cela, calculons le carrée de la distance, que nous notons $d^2(j,O)$, d'un point variable j à l'origine :

$$d^2(j,O) = \sum_{i=1}^{n} (x_{ij} - 0)^2 = \sum_{i=1}^{n} \frac{(r_{ij} - \bar{r}_j)^2}{nS_{ej}^2} = \frac{S_{ej}^2}{S_{ej}^2} = 1$$

On obtient ainsi dans $\mathbb{R}^n$, une déformation du nuage des p points variables qui sont tous ramenés à une distance unité de l'origine : situés sur un cercle de rayon 1 si n=2, une sphère de rayon 1 si n=3 et de façon générale une ***hypersphères*** de rayon 1 de $\mathbb{R}^n$.

Calculons le carré de la distance entre deux points variables j et j':

$$D^2(j,j')=\sum_{i=1}^{n}(x_{ij}-x_{ij'})^2=\frac{1}{n}\sum_{i=1}^{n}(\frac{(r_{ij}-\overline{r_j})}{S_{ej}}-\frac{(r_{ij}-\overline{r_j})}{S_{ej}})^2$$

$$D^2(j,j')=\frac{S_{ej}^2}{S_{ej}^2}+\frac{S_{ej'}^2}{S_{ej'}^2}-2.\frac{\frac{1}{n}\sum_{i=1}^{n}(r_{ij}-\overline{r_j})(r_{ij'}-\overline{r_{j'}})}{S_{ej}S_{ej'}}=\quad \mathbf{2-2corr(j,j')=2(1-corr(j,j'))}$$

Où ***corr(j,j')*** est le coefficient de corrélation linéaire entre les variable j et j'.
La distance entre deux points variables s'interprétera donc en termes de corrélation linéaire.

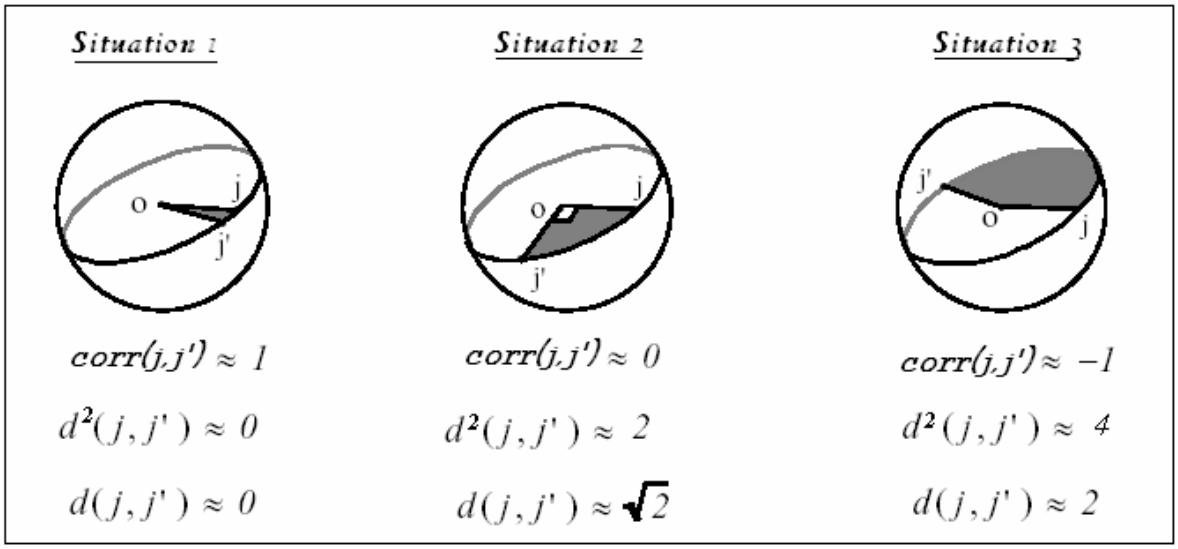


Fig. V.6 Corrélations et distances entre points-variables

- **Situation 1** : Les variables j et j' sont fortement corrélées positivement donc les points sont très proches sur la sphère
- **Situation 2 :** Les variables j et j' ne sont pas corrélées donc les points sont les sommets d'un triangle rectangle en O. Si ces points sont sur des axes de $\mathbb{R}^n$, cela signifie que ces axes ne sont pas corrélés (rappelons que les axes factoriels sont orthogonaux par construction)
- **Situation 3 :** Les variables j et j' sont fortement corrélées négativement donc les points sont très presque diamétralement opposés sur la sphère

Remarques :
Calcul des coordonnées des points variable (Dans le cas d'une analyse en composantes principales):

- D'après la théorie de l'analyse générale : sont les p composantes du vecteur $X'v_\alpha$.
- Sont les coefficients de corrélation de la matrice de corrélation des variables et leurs composantes principales (les variables sont représentées dans un cercle de rayon 1).

Corr(x_j,y_k) $=\frac{\sqrt{\lambda_k}}{S_{ej}}u_k$ Donc il faudra tout simplement de diviser la coordonnées de la variable étudiée par l'écart type de la variable réelle pour la représenter sur le cercle.

V.2.2 Reconstitution et reconstitution approchée du tableau X de départ :

Donc : $X = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha} v_\alpha u_\alpha'$, matrice (n,1)x(1,p), qui est la formule de reconstitution du tableau X.

Si on se limite aux q axes (q≤p) et si les valeurs propres $\lambda_{q+1},\ldots, \lambda_p$ sont négligeables par rapport aux q premières, on obtient la reconstitution approchée X^* du tableau de départ X, composé de q termes :

$$X^* = = \sum_{\alpha=1}^{q} \sqrt{\lambda_\alpha} v_\alpha u_\alpha'$$

Les np valeurs initiales de X sont remplacées par les q(n+p) valeurs de X^*. La quantité globale de la reconstitution du tableau de données par q axes est : $\tau_q = \frac{\sum_{\alpha=1}^{q} \lambda_\alpha}{\sum_{\alpha=1}^{p} \lambda_\alpha}$ appelée ***taux d'inertie*** ou ***part de variance*** qui mesure ***la part de la " dispersion"*** du nuage expliquée par le sous espace à q dimensions.

L'exemple suivie :

$(\lambda_1, \lambda_2)=(29.39;18.01)$

$\lambda_1+\lambda_2=47.4$

$\lambda_1/(\lambda_1+\lambda_2) =29.39/47.4=0.62=62\%$

$\lambda_2/(\lambda_1+\lambda_2) =18.01/47.4=0.38=38\%$

Le taux d'inertie de l'axe1 représente 62% de l'inertie totale, donc l'axe2 seulement 38% de cette inertie (τ_1=62%). C'est-à-dire que si l'on résumait notre nuage à deux dimensions, correspondant aux variables v_1 et v_2, à une seule dimension, correspondant à l'axe factoriel u_1, on conserverait dans cette approximation 62% de l'information.

V.3 Qualité de la représentation des individus et variables,

V.3.1 Contribution relative des individus :

Soit le schéma suivant représentant trois variables dans un espace de trois axes factoriels :

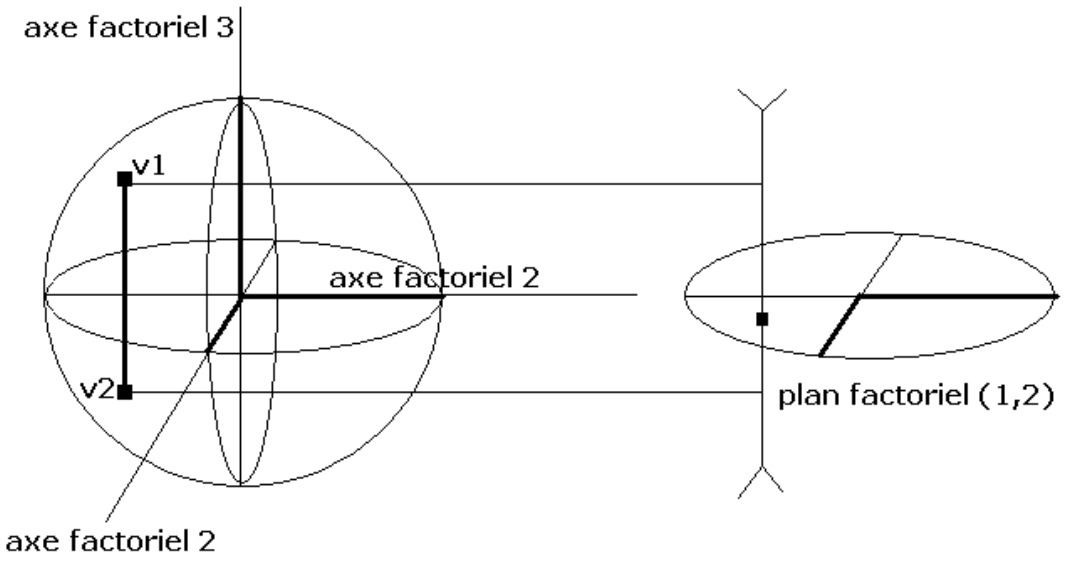


Fig. V.7 Schéma pour l'étude des projections sur un plan factoriel

Sur ce schéma nous voyons que les deux points v1 et v2 portés par le sphère, qui sont éloignés dans l'espace sont confondus dans la projection sur le plan factoriel formé par les axes 1 et 2. On ferait une erreur d'interprétation en affirmant que ces variables sont proches, elles sont en fait opposées et explicatives du plan (1,3). Le problème se pose de même pour les individus.

V.3.1.1 Contribution relative d'un individu à la formation d'un axe :

La variance des coordonnées de tous les individus sur l'axe α est proportionnelle à la valeur propre de cet axe et vaut : λ_α /n où n est le nombre d'individus. La part d'un individu à la variance de cet axe est donnée : $cr_\alpha(i)=(Xu_\alpha)^2/ \lambda_\alpha$.Donc $cr_\alpha(i)$ est la contribution relative de l'individu i à l'axe α.

Utilité de ce paramètre : sert à repérer facilement les individus qui ont le plus participé à la formation de l'axe.

Exemple : calcul de la contribution relative des individus 5 et 7 aux axes F_1 et F_2 (choix de la matrice des corrélations)

	Individu 5	Individu 7
Coordonnées sur F_1	-0.43	0.05
Coordonnées sur F_2	0.08	0.4
Contribution relative à l'axe F_1	$cr_1(5)=(-0.43)^2/1.237=0.15$	$Cr1(7)=(0.05)^2/1.237=0.009$
Contribution relative à l'axe F_2	$Cr2(5)=(0.08)^2/0.76=0.008$	$Cr2(7)=(0.4)^2/0.76=0.21$

V.3.1.2 Qualité de la représentation d'un individu i sur l'axe factoriel α :

Le calcul du COSINUS de l'angle β, formé entre le segment Oi et l'axe ou le plan associé (O le centre de gravité du nuage des individus), permet de savoir si un point individu i est proche d'un axe principal ou d'un plan principal.

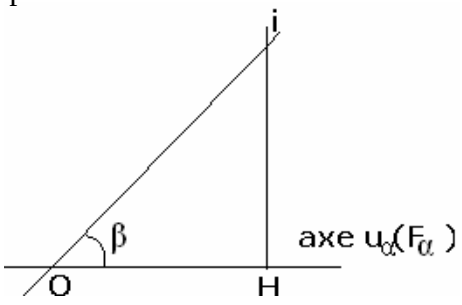


Fig. V.8 Schéma de représentation d'un individu i sur l'axe factoriel α

D'après le schéma cos β = OH/Oi,

- OH est donné par la coordonnée xi_α de l'individu i sur l'axe F_α (I-ième coordonnées de $Xu_\alpha)_i$
- Oi est la distance de l'individu i à l'origine $Oi^2=\sum_{j=1}^{p}(Xu_j)_i^2$

Les quantités à calculées dans le logiciel sont généralement les carrées des cosinus :

$$\cos^2\beta=\frac{(Xu_j)_i^2}{\sum_{j=1}^{p}(Xu_j)_i^2}$$

Si le cos β est plus grand (proche de + ou – 1), le point est proche de l'axe
Si le cos β est petit (proche de 0), le point est éloigné de l'axe.

V.3.1.3 Qualité de la représentation d'un individu sur le plan factoriel (F1,F2) :

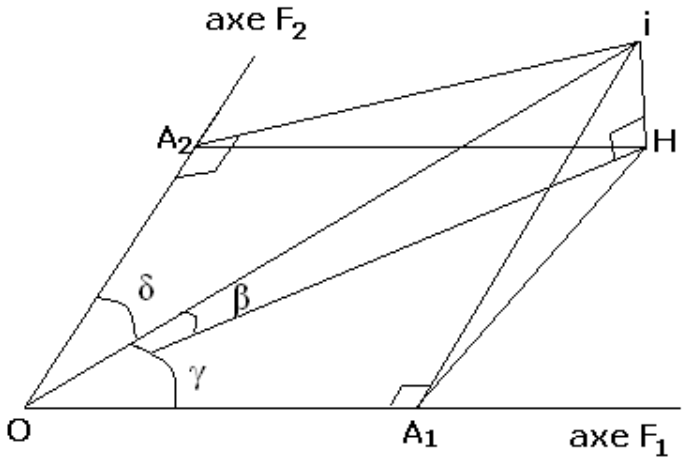


Fig. V.9 Schéma pour le calcul de la qualité de la représentation d'un point individu sur un plan factoriel

Comme cos β= OH/Oi,
$\cos^2\beta = OH^2/Oi^2 = (OA_1+OA_2)^2/Oi^2$
$= (OA_1^2/Oi^2)+(OA_2^2/Oi^2)+(2OA_2OA_1)/Oi^2$
$= (OA_1^2/Oi^2)+(OA_2^2/Oi^2)$
$= \cos^2\delta + \cos^2\gamma$

Cette quantité, le cosinus carré est appelée qualité de la représentation d'un individu sur un axe engendré par une composante principale, sur un plan engendré par deux composantes principales.
Une fois assuré de la bonne qualité de la représentation d'un individu sur un axe plan, nous pourrons regrouper les individus semblables c'est-à-dire ceux qui sont proches : ces individus auront des caractéristiques voisins, caractéristiques définies par les variables représentant les axes principaux.

Exemple : calcul de la qualité de la représentation des individus 5 et 7 aux axes F_1 et F_2 (choix de la matrice des corrélations)

	Individu 5	Individu 7
Coordonnées sur F_1	-0.43	0.05
Coordonnées sur F_2	0.08	0.4
Contribution relative à l'axe F_1	$Cos^2(\beta)=(-0.43)/((-0.43)^2+0.08^2))$ $=0.97$	$Cos^2(\beta)=(0.05)/((0.4)^2+0.05^2))$ $=0.015$
Contribution relative à l'axe F_2	$Cos^2(\beta)=(-0.8)/((-0.43)^2+0.08^2))$ $=0.033$	$Cos^2(\beta)=(0.4)/((0.4)^2+0.05^2))$ $=0.984$

V.3.1.4 Interprétation des proximités entre individus :

On remarque que l'individu 5 est très bien représenté sur l'axe F1 (sa qlt (son cosinus) est 0.96) et que l'individu 7 est très bien représenté sur l'axe F2 (sa qlt est 0.99), alors que ces individus sont mal représentés sur l'autre axe. Pour l'ensemble des 10 individus, on aura la classification

suivante, en prenant comme critère qu'un point est bien représenté si sa qlt est supérieure à 0.64 (ou une qlt en valeur absolue de 0.8) :

- sur l'axe F_1 : les individus 2,5,6,9 et 10
- sur l'axe F_2 : les individus 1,3,4,7 et 8

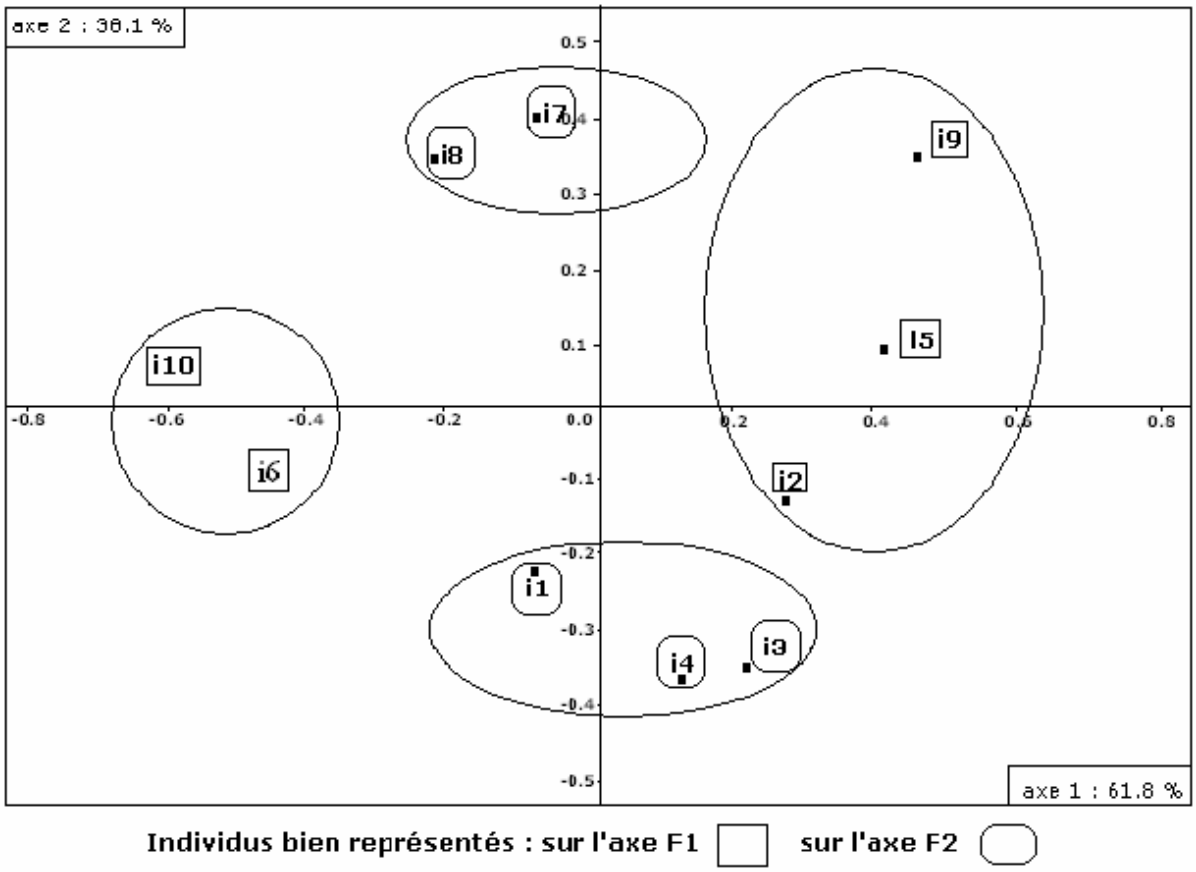


Fig. V.10 Schéma de proximité des individus

Référons nous maintenant aux données initiales des variables v_1 et v_2 et interprétons ces deux axes :

- pour l'axe F_1 : deux groupes s'opposent, le groupe i6 et i10 où les valeurs de v_1 et celles de v_2 négative et moyenne, et le groupe i2, i5 et i9 où la valeur de v_1 est petite et celle de v_2 grande et moyenne
- pour l'axe F_2 : deux groupes s'opposent, le groupe i3 et i4 où les valeurs de v_1 et celle de v_2 sont petites, et le groupe i8 et i7 où les valeurs de v1 et celles de v2 sont grandes. Le point i1 est un point proche du point moyen du nuage.

On obtient ainsi une géographie de nos individus dans les axes principaux exprimés en termes des variables initiales. Il faut bien sûr connaître le sens à donner aux axes principaux pour terminer cette interprétation.

V.3.1.5 Qualité de la représentation d'une variable :

Pour savoir si un point variable j est proche d'un axe principal ou d'un plan principal nous calculons le cosinus de l'angle β formé entre le segment Oj et l'axe ou le plan associé (O représente l'origine du repère). On sait que la coordonnée d'une variable sur un axe principal, représentée sur le cercle des corrélations, était le coefficient de corrélation linéaire entre la variable et cet axe. Le calcul est en fait un produit scalaire entre les vecteurs associés et est donc le cosinus de l'angle formé entre la variable et l'axe choisi (ou le plan). Le critère choisi, sera,

comme précédemment, de prendre le carré de ce cosinus ou le carré du coefficient de corrélation linéaire. Les calculs peuvent être exécutés de deux façons :

- soit en calculant directement le cosinus au carré de l'angle entre la variable et l'axe choisi
- ou bien en prenant le coefficient de corrélation linéaire entre la variable et cet axe.

Exemple :

Calcul de la qualité de la représentation des variables 1 et 2 sur les deux axe factoriels G1 et G2 (coordonnées utilisées sont celles de la matrice des variances-covariances)

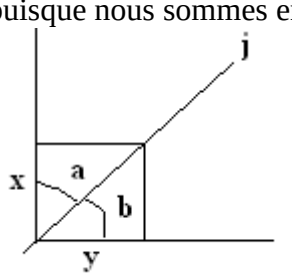
1 – avec les coordonnées des points variables :

	Variable 1	variable 2
Coordonnées sur F_1	-0.559	0.9377
Coordonnées sur F_2	-0.8289	0.3473
Qualité de la représentation sur l'axe G1	$Cos^2(\beta)=(-0.559)/((-0.559)^2 + 0.8289^2)) = 0.31$	$Cos^2(\beta)=(0.9377)/((0.9377)^2+0.3473^2)) =0.88$
Qualité de la représentation sur l'axe G2	$Cos^2(\beta)=(0.8289)/((-0.559)^2+0.8289^2)) =0.69$	$Cos^2(\beta)=(0.3473)/((0.9377)^2+0.^{34732})) =0.12$

Remarque :

Pour la variable 1, sommant ses cosinus carrés pour chacun des axes, on obtient : 0.31+0.69=1, la valeur exacte est évidemment 1 puisque nous sommes en dimension 2.

$(Cos\ a)^2+(cos\ b)^2=1$



2 – Avec les coefficients de corrélation linéaire :

En élevant au carré chacun de ces termes on obtient :

Variables	Axe G1	Axe G2
v1	-0.559	-0.829
v2	0.9377	0.347

	G1	G2
v1	$(-0.559)^2=0.31$	$(-0.829)^2=0.69$
v2	$(0.9377)^2=0.88$	$(0.347)^2=0.12$

Les calculs sont évidemment les mêmes en utilisant la matrice de corrélation.

Conclusion :

Il est d'usage de prendre les critères suivants de bonne représentation à partir de la valeur des cosinus carrés ou qualités de la représentation. En notant qlt cette valeur on dira :

- Très bonne représentation si qlt >0.8
- Bonne représentation si 0.65<qlt<=0.8
- Représentation moyenne si 0.4<qlt<=0.65
- Médiocre représentation si qlt<=0.4

Un résumé pour l'interprétation d'une analyse en composantes principales

Voici les grandes lignes à respecter pour une interprétation d'une analyse ACP.

Etape 1 : Regarder le plan 1 et 2, formé par les axes principaux 1 et 2. Sur ce plan il faut voir comment se répartissent les individus. Trois situations typiques peuvent se présenter comme le montre le schéma suivant :

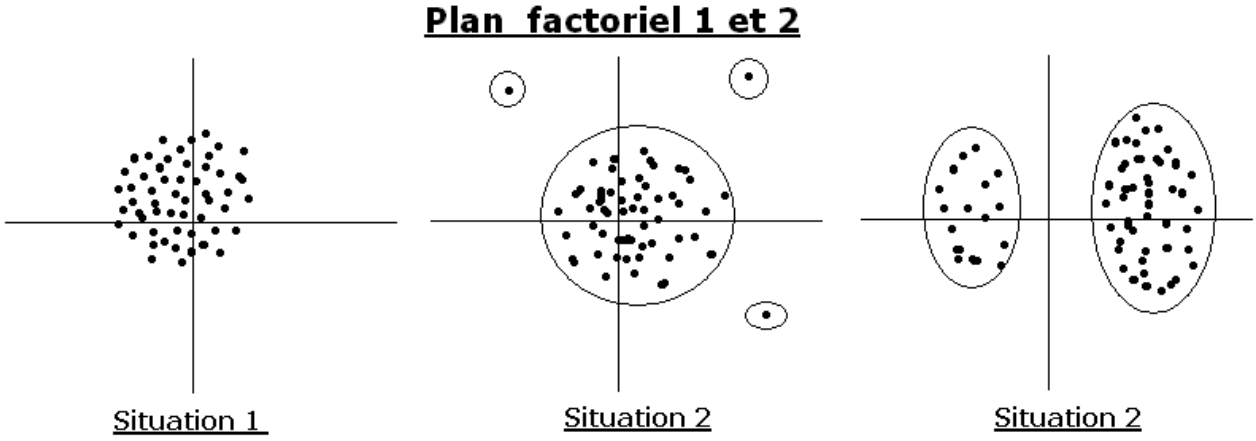


Fig. V.11 Situations typiques de répartition des individus

- **Situation 1** : les individus se répartissent uniformément sur le plan, c'est la situation idéale pour examiner de plus près les résultats de l'ACP

- **Situation 2** : quelques individus sont isolés du nuage des autres individus, ces individus faussent l'ACP, il est recommandé de les supprimer de l'analyse et de la refaire sans ces individus, lesquels pourront être placés en supplémentaire (partie d'ACP non traité)

- **Situation 3** : quelques groupes d'individus forment des nuages bien distincts, cela signifie que des sous populations ont été mises en évidence. Si cette image représente le but recherché, c'est-à-dire mettre en évidence des sous groupes d'individus, on peut continuer l'analyse. Sinon, cette existence de sous groupes fausse les corrélations entre les variables et il faudrait recommencer l'analyse sur chacun des sous groupes.

Etape 2 :
Regarder les valeurs propres de chaque axe et les pourcentages de variation expliquée par chaque composante. Le bon sens nous conduit à retenir les axes jusqu'à un seuil suffisant d'information en tenant compte de la valeur dégressive des valeurs propres. N'oublions pas que l'ACP effectuée sur la matrice des corrélations attribue la valeur 1 à une variable initiale et donc on ne retiendra que les axes dont la part de variation est supérieure à celle d'une variable initiale, c'est-à-dire dont la valeur propre est supérieure à 1. Enfin, il peut apparaître qu'un axe éloigné de faible inertie soit intéressant car il est lié une variable particulièrement importante pour l'étude.

Etape 3 : Etudier les variables.

L'interprétation des variables se fait à partir du cercle des corrélations et de leur qualité de représentation (leur cos^2 ou $corr^2$ noté aussi **qlt**) qui donne la part de variation de la variable expliquée par l'axe.

Rappelons les critères usuels :

- Très bonne représentation si qlt >0.8
- Bonne représentation si 0.65<qlt<=0.8
- Représentation moyenne si 0.4<qlt<=0.65
- Médiocre représentation si qlt<=0.4

Une variable est d'autant mieux représentée qu'elle se situe près du cercle des corrélations. Les variable les plus liées à un axe (plan) donne typologie de cet axe (de ce plan) qui le résument. Il faut évidemment tenir compte du signe des coordonnées des variables pour mettre en évidence les corrélations positives ou négative (anti-corrélations) et l'angle entre les variables (un angle droit signant l'absence de corrélation linéaire).

Etape 4 : Etudier les individus.

La variance des coordonnées des individus sur un axe est la valeur propre de cet axe. Les individus qui participent le plus à la formation de cet axe sont ceux les plus éloignés de l'origine (0,0), c'est-à-dire ceux dont les coordonnées sont les plus fortes en valeur absolue, soit encore ceux dont la contribution relative **cr** est la plus élevée. Deux cas se présentent :

1 – soit l'axe est représenté par seulement quelques individus, ces individus "suspect" doivent retirés et l'ACP doit être recommencée sans ces individus.

2 – soit l'axe est conduit par un ensemble homogène d'individus, on dit alors que l'axe est stable et sa signification n'est pas modifiée par la présence de quelques individus en plus ou en moins.

Pour interpréter correctement la proximité ou l'éloignement entre individus il faut tenir compte de leur qualité de représentation (qlt) sur l'axe (sur le plan) étudié et cela par l'intermédiaire de leur cos^2. Nous prendrons comme critère de représentation :

- Très bonne représentation si qlt >0.8
- Bonne représentation si 0.65<qlt<=0.8
- Représentation moyenne si 0.4<qlt<=0.65
- Médiocre représentation si qlt<=0.4

On peut ainsi créer des regroupements d'individus proches bien ou très bien représentés, ces regroupements auront des caractéristiques voisines.

Etape 5 : Etudier les représentations graphiques .

L'étude de ces représentations se fait parallèlement aux étapes précédentes. Les graphes proposés sont de deux types : le plan des individus et le plan de variables. Il ne faut pas oublier alors que les variables doivent être "pensées" comme des vecteurs joignant le "point variable" à l'origine le centre du repère, l'interprétation des variables se faisant à partir de leur direction. Il faut tenir compte de cette remarque dans l'étude de la représentation simultanée des variables et des individus.

Institut de sciences exactes
Département d'Informatique
Module : Analyse de données

Fiche de TD n°04 (ACP)

Exercice n° 01 : Soit la matrice des données suivante **:**

$$R=\begin{bmatrix} 0.5 & 0 \\ -0.1 & 1.2 \\ -0.5 & 0.5 \\ -0.3 & 0.1 \\ 0 & 2.5 \\ 1.6 & -0.7 \\ 2 & 2 \\ 2.4 & 1.2 \\ 0.5 & 3.5 \\ 2.7 & -0.9 \end{bmatrix}$$

Calculer pour cette matrice :

1- Pour l'analyse R^p

- La matrice des variances - covariances
- Les valeurs propres et les vecteurs propres de la matrice des variance - covariances
- Les coordonnées des points individus dans l'espace des vecteurs propres
- La matrice des corrélations
- les valeurs propres et les vecteurs propres de la matrice des corrélations
- Les coordonnées des points dans l'espace des vecteurs propres

2- Pour l'analyse R^n

- Calculer pour les deux cas de la matrice X'X des (variances-covariances et corrélations) les coordonnées des variables sur les axes factoriels.

Exercice n° 02 :

Soit le tableau de données : (individus/variables : n=6/p=3)

$$R=\begin{bmatrix} 8 & 1 & 0 \\ 4 & 6 & 5 \\ 6 & 8 & 7 \\ 10 & 4 & 7 \\ 8 & 2 & 5 \\ 0 & 3 & 6 \end{bmatrix}$$

Appliquer l'A.C.P.(non normée) sur ce tableau de données.

Université de Béchar
Département d'Informatique
Module : Analyse de données

Solution type de la fiche de TD n° 04

Exercice n° 01

A. Cas 1: X'X est la matrice des variances-covariances :

On calcule les valeurs propres et les vecteurs correspondant aux données de l'exemple 1 de la matrice R.

- La matrice X : comme X'X est la matrice des **variances-covariances,** nous devons tout d'abord modifier la matrice R en X en effectuant la transformation :

$$x_{ij} = \frac{r_{ij} - \overline{r_j}}{\sqrt{n}}, \; i=1\ldots.10, \; j=1,2.$$

Cette matrice X est donc :

A partir de R(10,2) on obtient X(10,2)

i	v_1	v_2	v_1	v_2
i_{11}	0,5	0	-0,12	-0,3
i_2	-0,1	1,2	-0,31	0,08
i_3	-0,5	0,5	-0,44	-0,14
i_4	-0,3	0,1	-0,37	-0,27
i_5	0	2,5	-0.28	0,49
i_6	1,6	-0,7	0,23	-0,52
i_7	2	2	0,35	0,34
i_8	2,4	1,2	0,48	0,08
i_9	0,5	3,5	-0,12	0,81
i_{10}	2,7	-0,9	0,58	-0,58

Total :	8,8	9,4
Moyenne :	0,88	0,94
Ecart-type :	1,1277	1,346
Taille :	10	

Où $i_1 \ldots i_{10}$ représentant les 10 et v1, v2 les deux variables (colonnes R^1 et R^2). Par exemple le terme r_{32} de R devient x_{32} de X où :

$$x_{32} = \frac{0,5 - 0,94}{\sqrt{10}} = -0,14$$

- **La matrice X'X :** on en déduit la matrice des variances-covariances X'X :

X'X=

X' (2,10)

-0,12	**-0,31**	**-0,44**	**-0,37**	**-0,28**	**0,23**	**0,35**	**0,48**	**-0,12**	**0,58**
-0,3	**0,08**	**-0,14**	**-0,27**	**0,49**	**-0,52**	**0,34**	**0,08**	**0,81**	**-0,58**

x

X (10,20)

-0,12	-0,3
-0,31	0,08
-0,44	-0,14
-0,37	-0,27
-0.28	0,49
0,23	-0,52
0,35	0,34
0,48	0,08
-0,12	0,81
0,58	-0,58

D'où X'X =

1,27	**-0,36**
-0,36	**1,81**

On aurait pu obtenir directement cette matrice X'X, de taille (2,2), en utilisant les résultats vu dans la solution de l'exercice n° 05 de la 1ère fiche de TD (moindres carrées) où nous avons calculé les variances des variables (var(R^1)=1,1277^2=1,27, var(R^2)=1,346^2=1,81) et la covariance cov(R^1, R^2)=-0,36.

- Les valeurs propres et les vecteurs propres :

Le calcul des valeurs propres et des vecteurs propres de cette matrice sont aussi vu dans la solution de l'exercice n° 05 de la 1ère fiche de TD (moindres rectangles), rappelons les résultats: à la plus grande valeur propre λ_1=1,99 est associé le vecteur propre u_1'=(0,4472, -0,8944), de norme 1 et à la seconde valeur propre λ_2=1,09 est associé le vecteur propre u_2'=(0,8944 , 0,4472), de norme 1. Ces vecteurs sont définis à une orientation près.

- Les coordonnées des points individus dans l'espace des vecteurs propres :

Pour obtenir les coordonnés des points individus dans l'espace des vecteurs propres de base $\{u_1, u_2\}$qui est aussi appelé le plan factoriel $\{F_1, F_2\}$ nous calculerons X_{u_1} et X_{u_2} qui donnent les coordonnées sur chacun de ces axes.

$X_{(10,2)}$			**$U_1(2,1)$**		**$Xu_{1\ (10,1)}$**		**$X_{(10,2)}$**		**$U_2(2,1)$**		**$Xu_{2(10,1)}$**
-0,12	**-0,3**				**0,21**		**-0,12**	**-0,3**			**-0,24**
-0,31	**0,08**				**-0,21**		**-0,31**	**0,08**			**-0,24**
-0,44	**-0,14**				**-0,07**		**-0,44**	**-0,14**			**-0,45**
-0,37	**-0,27**				**0,07**		**-0,37**	**-0,27**			**-0,45**
-0.28	**0,49**				**-0,57**		**-0.28**	**0,49**			**-0,03**
0,23	**-0,52**	x	**0,4472**	=	**0,57**	x	**0,23**	**-0,52**	x	**0,8944**	**-0,03**
0,35	**0,34**		**-0,8944**		**-0,14**		**0,35**	**0,34**		**0,4472**	**0,47**
0,48	**0,08**				**0,14**		**0,48**	**0,08**			**0,47**
-0,12	**0,81**				**-0,78**		**-0,12**	**0,81**			**0,25**
0,58	**-0,58**				**0,78**		**0,58**	**-0,58**			**0,25**

B. Cas 2: X'X est la matrice des corrélations

Nous allons reprendre les calculs précédents mais nous devons auparavant transformer <u>la matrice R :</u>

la matrice X: comme X'X est la matrice des corrélation, nous devons tout d'abord modifier la matrice R en X en effectuant la transformation :

$$\mathbf{x}_{ij} = \frac{r_{ij} - \overline{r_j}}{s_{ej}\sqrt{n}}\text{,}\ \text{i=1…..10, j=1,2.}$$

Cette matrice X est donc :

A partir de R(10,2) on obtient X(10,2)

i	v_1	v_2	v_1	v_2
i_{1}	0,5	0	-0,11	-0,22
i_2	-0,1	1,2	-0,27	0,06
i_3	-0,5	0,5	-0,39	-0,1
i_4	-0,3	0,1	-0,33	-0,2
i_5	0	2,5	-0.25	0,37
i_6	1,6	-0,7	0,2	-0,39
i_7	2	2	0,31	0,25
i_8	2,4	1,2	0,43	0,06
i_9	0,5	3,5	-0,11	0,6
i_{10}	2,7	-0,9	0,51	-0,43

Total :	8,8	9,4
Moyenne :	0,88	0,94
Ecart-type :	1,1277	1,346
Taille :	10	

Où $i_1 \ldots i_{10}$ représentant les 10 individus et v1, v2 les deux variables (colonnes R^1 et R^2). Par exemple le terme r_{32} de R devient x_{32} de X où :

$$x_{32} = \frac{0,5 - 0,94}{1,346\sqrt{10}} = -0,1$$

- **La matrice X'X :** on en déduit la matrice des corrélations C = X'X :

$C_{(2,2)}$=X'X=

$X'_{(2,10)}$

-0,11	**-0,27**	**-0,39**	**-0,33**	**-0,25**	**0,2**	**0,31**	**0,43**	**-0,11**	**0,51**
-0,22	**0,06**	**-0,1**	**-0,2**	**0,37**	**-0,39**	**0,25**	**0,06**	**0,6**	**-0,43**

x

$X_{(10,2)}$

-0,11	**-0,22**
-0,27	**0,06**
-0,39	**-0,1**
-0,33	**-0,2**
-0,25	**0,37**
0,2	**-0,39**
0,31	**0,25**
0,43	**0,06**
-0,11	**0,6**
0,51	**-0,43**

D'où $C=X'X=\begin{bmatrix} 1 & -0,237 \\ -0,237 & 1 \end{bmatrix}$

On aurait pu obtenir directement cette matrice X'X en utilisant les résultats du paragraphe 1.1.4 ou nous avons calculé le coefficient de corrélation linéaire $r=corr(R^1,R^2)=-0,237$.

- **les valeurs propres et les vecteurs propres:** cherchons les valeurs propres de cette matrice en résolvant l'équation du second degré:

$$P(\lambda)=Dét(C-\lambda I)=\begin{vmatrix} 1-\lambda & -0,237 \\ -0,237 & 1-\lambda \end{vmatrix} = \lambda^2 - 2\lambda + 0.944$$

Cette équation admet les deux racines réelles λ_1 =1,237 et λ_2 =0,763, en convenant d'associer λ_1 à la plus grande des deux valeurs propres. Remarquez que la somme des deux valeurs propres (la trace de C) est bien p=2, le nombre des variables.
Pour trouver les vecteurs propres associés à ces valeurs propres, nous devons résoudre $(C-\lambda I)u=0$, où I est la matrice identité 2×2 et u le vecteur propre associé à la valeur propre λ.

Calculons les deux vecteurs propres u_1 et u_2 associés aux valeurs propres λ_1 et λ_2.

1- Le vecteur propre u_1 associés à la valeur propre λ_1 :

$$(C-\lambda_1 I).\, u_1 = \begin{bmatrix} 1-1,237 & -0,237 \\ -0,237 & 1-1,237 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -0,237 & -0,237 \\ -0,237 & -0,237 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Ce qui nous conduit à résoudre le système :

$$\begin{cases} 0{,}237x_1 - 0{,}237x_2 = 0 & (\text{éq.1}) \\ -0{,}237x_1 + 0{,}237x_2 = 0 & (\text{éq.2}) \end{cases}$$

Comme l'équation (2) est l'opposée de l'équation (1) on aura x1-x2=0, c'est-à-dire u'2=(x_1,x_1). Il reste à normer ce vecteur à 1 : $\|u_1\|^2 = 1 = (x_1^2 + x_1^2) = 2x_1^2$, on en déduit que $x_1 = 1/\sqrt{2}$. Ainsi : $u_1^{'}$=(0,71;-0,71),

2- Le vecteur propre u_2 associés à la valeur propre λ_2:

$$(C-\lambda_2 I).\, u_2 = \begin{bmatrix} 1-0{,}763 & -0{,}237 \\ -0{,}237 & 1-0{,}763 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0{,}237 & -0{,}237 \\ -0{,}237 & 0{,}237 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Ce qui nous conduit à résoudre le système :

$$\begin{cases} 0{,}237x_1 - 0{,}237x_2 = 0 & (\text{éq.1}) \\ -0{,}237x_1 + 0{,}237x_2 = 0 & (\text{éq.2}) \end{cases}$$

Comme l'équation (2) est l'opposée de l'équation (1) on aura x1-x2=0, c'est-à-dire u'2=(x_1,x_1). Il reste à normer ce vecteur à 1 : $\|u_1\|^2 = 1 = (x_1^2 + x_1^2) = 2x_1^2$, on en déduit que $x_1 = 1/\sqrt{2}$. Ainsi : $u_2^{'}$=(0,71;0,71),
Ainsi $u_2^{'}$=(0,71;0,71).

<u>Résumons ces résultats :</u> à la plus grande valeur propre $\lambda_1 = 1{,}237$ est associé le vecteur propre $u_1^{'}$=(0,71;0,71), de norme 1 et à la seconde valeur propre $\lambda_2 = 0{,}763$ est associé le vecteur propre $u_2^{'}$=(0,71;0,71), de norme 1. Ces vecteurs sont définis à une orientation près.

- Les coordonnées des points individus dans l'espace des vecteurs propres de base {u_1,u_2}qui est aussi appelé le plan factoriel {F_1,F_2}nous calculerons Xu_1 et Xu_2 qui donnent les coordonnées sur chacun de ces axes :

$X_{(10,2)}$

-0,11	**-0,22**
-0,27	**0,06**
-0,39	**-0,1**
-0,33	**-0,2**
-0,25	**0,37**
0,2	**-0,39**
0,31	**0,25**
0,43	**0,06**
-0,11	**0,6**
0,51	**-0,43**

x $U_{1(2,1)}$

0, 71
-0,71

= $Xu_{1(10,1)}$

0,08
-0,24
-0,2
-0,09
-0,43
0,42
0,05
0,26
-0,5
0,67

x $X_{(10,2)}$

-0,11	**-0,22**
-0,27	**0,06**
-0,39	**-0,1**
-0,33	**-0,2**
-0.25	**0,37**
0,2	**-0,39**
0,31	**0,25**
0,43	**0,06**
-0,11	**0,6**
0,51	**-0,43**

x $U_{2(2,1)}$

0,71
0,71

= $Xu_{2(10,1)}$

-0,23
-0,15
-0,35
-0,37
0,08
-0,13
0,4
0,34
0,35
0,06

Exercice n° 02

$$R = \begin{bmatrix} 8 & 1 & 0 \\ 4 & 6 & 5 \\ 6 & 8 & 7 \\ 10 & 4 & 7 \\ 8 & 2 & 5 \\ 0 & 3 & 6 \end{bmatrix}$$

ACP non normé nécessite la réalisation des étapes suivantes :

1. Calcul de la matrice X'X (matrice de variances-covariances)
2. Calcul des valeurs et vecteurs propres :
3. Les axes principaux d'inertie :
4. Composantes principales
5. Représentation graphique
6. Corrélation des variables
7. Contribution des individus aux inerties des axes factoriels

1 -Calculons la matrice X à partir de la transformation :

$x_{ij} = \frac{r_{ij} - \overline{r_j}}{\sqrt{n}}$, i=1….6, j=1,3.

Cette matrice X est donc :

A partir de R(10,2) on obtient $1/\sqrt{6}$ .X(10,2)

i	v_1	v_2		v_1	v_2
i_1	**8**	**1**	**0**	**2**	**-3**	**-5**
i_2	**4**	**6**	**5**	**-2**	**2**	**0**
i_3	**6**	**8**	**7**	**0**	**4**	**2**
i_4	**10**	**4**	**7**	**4**	**0**	**2**
i_5	**8**	**2**	**5**	**2**	**-2**	**0**
i_6	**0**	**3**	**6**	**-6**	**-1**	**1**

Total :	36	24	30		0	0	0
Moyenne :	6	4	5		0	0	0

- **La matrice X'X :** on en déduit la matrice des variances-covariances X'X :

X' (3,6)

2	**-2**	**0**	**4**	**2**	**-6**
-3	**2**	**4**	**0**	**-2**	**-1**
-5	**0**	**2**	**2**	**0**	**1**

X (3,6)

2	**-3**	**-5**
-2	**2**	**0**
0	**4**	**2**
4	**0**	**2**
2	**-2**	**0**
-6	**-1**	**1**

X'X=1/6 x

D'où C = X'X =1/6

64	-8	-8
-8	34	22
-8	22	34

2 -Calcul des valeurs et vecteurs propres :

2.1 - Valeurs propres :

Det(C-λI)=0

$$\begin{bmatrix} 32-\lambda & -4 & -4 \\ -4 & 34-\lambda & 22 \\ -8 & 22 & 34-\lambda \end{bmatrix} = (64-\lambda)\begin{bmatrix} 34-\lambda & 22 \\ 22 & 34-\lambda \end{bmatrix} + 8\begin{bmatrix} -8 & 22 \\ -8 & 34-\lambda \end{bmatrix} - 8\begin{bmatrix} -8 & 34-\lambda \\ -8 & 22 \end{bmatrix} = 0$$

$(64-\lambda)[(34-\lambda)^2-22^2]+8[-8(34-\lambda)+8*22]-8[-8*22+8(34-\lambda)]=0$

On obtient l'équation caractéristique : $\lambda^3-22\lambda^2+136\lambda-192=0$ ➔ ($\lambda_1=12$, $\lambda_2=8$, $\lambda_3=2$)

2.2 - Vecteurs propres :

$$\begin{cases} X'Xu_1 = \lambda_1 u_1 \ \text{(éq.1)} \\ X'Xu_2 = \lambda_2 u_2 \ \text{(éq.2)} \\ X'Xu_3 = \lambda_3 u_3 \ \text{(éq.3)} \end{cases}$$

Equation 1 :

$X'Xu_1 = \lambda_1 u_1$ (éq.1)

Et $\|u_1\|=u_1'.u_1=1$ ($x^2+y^2+z^2=1$)

$$1/6 \begin{bmatrix} 64 & -8 & -8 \\ -8 & 34 & 22 \\ -8 & 22 & 34 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 12 \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

$$\begin{cases} 1/6(64x-8y-8z)=12x \\ 1/6(-8x+34y+22z)=12y \\ 1/6(-8x+22y+34z)=12z \end{cases} \rightarrow \begin{cases} 64x-8y-8z=72x \\ -8x+34y+22z=72y \\ -8x+22y+34z=72z \end{cases} \rightarrow \begin{cases} -8x-8y-8z=0 \\ -8x+34y+22z=72y \\ -8x+22y-38z=0 \end{cases} \rightarrow \begin{cases} x+y+z=0 \\ -4x-19y+11z=0 \\ -4x+11y+19z=0 \end{cases} \rightarrow$$

$$\begin{cases} x+y+z=0 \\ 8y-8z=0 \end{cases} \rightarrow \begin{cases} x+y+z=0 \\ y=z \\ x^2+y^2+z^2=1 \end{cases} \rightarrow \begin{cases} x+2y=0 \\ y=z \\ x^2+2y^2=1 \end{cases} \rightarrow \begin{cases} x=-2y \\ (-2y)^2+2y^2=1 \end{cases} \rightarrow \begin{cases} x=-2y \\ 6y^2=1 \end{cases} \rightarrow \begin{cases} x=-2y \\ y=z=1/\sqrt{6} \end{cases}$$

$$\rightarrow \begin{cases} x=-2/\sqrt{6} \\ y=z=1/\sqrt{6} \end{cases} \rightarrow \text{ et par suite } u_1=1/\sqrt{6}\begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}$$

Et ainsi de la même façon pour l'équation (2) , on aura : $u_2=1/\sqrt{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

3. Les axes principaux d'inertie :

- Si on prend un axe principal, on obtient $Q1=\lambda_{max}/\sum_{i=1}^{3}\lambda_i = 12/22=0.54=54\%<60\%$ d'inertie totale
- Si on prend 2 axes principaux, on obtient $Q2=(\lambda_1+\lambda_2)/\sum_{i=1}^{3}\lambda_i = (12+8)/22=20/22=0.9=90\%$

On est pas obliger de calculer u_3 parce que Q2>80%, Donc on s'arrête.

u1,u2 : les axes principaux car Q>80%

λ_1 l'inertie par l'axe u_1

On choisis q=2 (q<p) : 2 colonnes dans y et non une colonnes.

4. Les composantes principales :

$Xu_1=(C_1^1, C_1^2, .., C_1^6)$ et $Xu_2=(C_2^1, C_2^2, .., C_2^6)$

$$Xu_1 = \begin{bmatrix} \mathbf{2} & \mathbf{-3} & \mathbf{-5} \\ \mathbf{-2} & \mathbf{2} & \mathbf{0} \\ \mathbf{0} & \mathbf{4} & \mathbf{2} \\ \mathbf{4} & \mathbf{0} & \mathbf{2} \\ \mathbf{2} & \mathbf{-2} & \mathbf{0} \\ \mathbf{-6} & \mathbf{-1} & \mathbf{1} \end{bmatrix} \frac{1}{\sqrt{6}}\begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}$$

Xu_2=

2	**-3**	**-5**
-2	**2**	**0**
0	**4**	**2**
4	**0**	**2**
2	**-2**	**0**
-6	**-1**	**1**

$$\frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}$$

$$C_1^1 = X_1.u_1 = \begin{pmatrix}2 & -3 & -5\end{pmatrix}\frac{1}{\sqrt{6}}\begin{pmatrix}2\\-1\\-1\end{pmatrix} = \frac{4+3+5}{\sqrt{6}} = 12/\sqrt{6}$$

De la même façon, on obtient :

$$Xu_1 = \frac{1}{\sqrt{6}}\begin{pmatrix}12\\-6\\-6\\6\\6\\-12\end{pmatrix} = \sqrt{6}\begin{pmatrix}2\\-1\\-1\\1\\1\\-2\end{pmatrix}$$

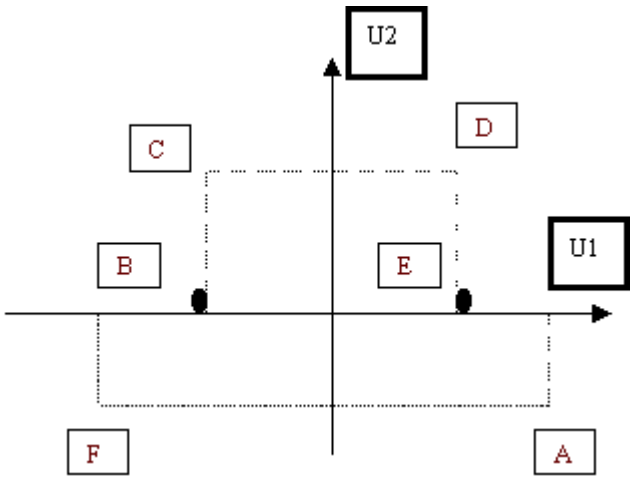
$$C_2^1 = X_1.u_2 = \begin{pmatrix}2 & -3 & -5\end{pmatrix}\frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix} = \frac{2-3-5}{\sqrt{3}} = \frac{-6}{\sqrt{3}} = \frac{-2*\sqrt{3}*\sqrt{3}}{\sqrt{3}} = -2*\sqrt{3} = -\sqrt{2}*\sqrt{2}*\sqrt{3} = -\sqrt{2}*\sqrt{6}$$

De la même façon, on obtient :

$$Xu_2 = \frac{1}{\sqrt{3}}\begin{pmatrix}-6\\0\\6\\6\\0\\-6\end{pmatrix} = \sqrt{6}\begin{pmatrix}\sqrt{2}\\0\\\sqrt{2}\\\sqrt{2}\\0\\-\sqrt{2}\end{pmatrix}$$

5. Représentation graphique :

Si l'on désire représenter les individus dans le plan formé par les deux premiers axes factoriels on aura:



Les parts d'inertie expliquée par les deux premiers axes factoriels sont:

$$I_E(u_1)=\frac{12}{22}\ ;I_E(u_2)=\frac{8}{22}\ ;I_E(u_1\ u_2)=\frac{20}{22}=91\%$$

Le dernier terme correspondant à la part d'inertie expliquée par le plan formé de ces deux vecteurs.
Si l'on cherche la part de l'inertie du point A restituée par l'axe 1 (ou encore le cosinus carré) on a:

$$COS2_\alpha\ (i)=\frac{\mu_{u\alpha}\ (i)^2}{\|X_i\|^2}=\frac{\left(2\sqrt{6}\right)^2}{4+9+25}$$

6. Corrélation des variables
7. Contribution des individus aux inerties des axes factoriels

VI. Analyse Factorielle des Correspondances (A.F.C.)

Présentation :

Le tableau de données statistiques traité, contrairement à l'ACP, les lignes individus et les colonnes variables seront des variables qualitatives et il ne s'agit pas par suite d'individus en lignes ni de variables en colonnes mais d'effectifs d'individus vérifiant simultanément la modalité de la variable écrite en ligne et celle de la variable écrite en colonne. Le but ne sera pas seulement de réduire ce tableau, mais aussi d'interpréter statistiquement et graphiquement ce tableau réduit.

VI.1 Généralités :

VI.1.1 Le tableau K des effectifs :

La matrice des données est un tableau qui croise deux variables qualitatives qui serviront aux calculs. Cette matrice K, contient n lignes, les n modalités de la variable X correspondant aux lignes et p colonnes, les p modalités de la variable Y correspondant aux colonnes. A l'intersection de ligne i et de la colonne j se trouve l'effectif k_{ij} correspondant au nombre des individus de l'échantillon étudié qui vérifient simultanément le caractère x_i de X et le caractère y_j de Y. Ce tableau K est appelé tableau de contingence ou tableau croisé.
On note K=(k_{ij})i=1..n,j=1..p cette matrice nxp où i est l'indice des lignes et j l'indice des colonnes.

$$K = \begin{bmatrix} k_{11} & \dots & k_{1j} & \dots & k_{1p} \\ \dots & & \dots & & \dots \\ k_{i1} & \dots & k_{ij} & \dots & k_{ip} \\ \dots & & \dots & & \dots \\ k_{n1} & \dots & k_{nj} & \dots & k_{np} \end{bmatrix}$$

k_{ij} est l'effectif correspondant à ($X=x_i$) et ($Y=y_j$)

Le nombre total d'individus étudiés donne la talle N de l'échantillon : $N = \sum_{i=1}^{n} \sum_{j=}^{p} k_{ij}$

<u>Tableau exemple :</u> dimension (4,2)

Un disquaire qui a fait un test de vente sur un échantillon de 655 disques destinés à diverses catégories (ou classes) de clientèle.

Le disquaire a réparti ses disques en ***2 catégories*** : Chanson orientale (ChO) et Chanson islamiques (CIS) et la population des utilisateurs en ***4 catégories*** (jeunes sans distinction de sexe, adultes féminins, adultes masculins et vieux sans distinction de sexe). Il a obtenu le tableau suivant :

K=

k_{ij}	ChO	CIS
Jeunes	69	41
Ad Fém	172	84
Ad masc	133	118
Vieux	27	11

VI.1.2 Le tableau F des fréquences :

A partir du tableau K, on crée le tableau des fréquences F=(f_{ij}) en divisant l'effectif k_{ij} par la taille N de l'échantillon : $f_{ij}=k_{ij}/N$.

En effectuant la somme des lignes on obtient les fréquences marginales $f_{.j}$ de la variable Y (en marge de la variable Y) et en effectuant la somme des colonnes on obtient les fréquences marginales $f_{i.}$ de la variable X (en marge de la variableX) : N=655

Exemple : fréquence relative de X=Jeunes et Y=**ChO** est f_{11}=69/655=0,105 et la fréquence marginale de X=jaunes est $f_{1.}$=(0.105+0,063)=0,168

F= **Variable X** / **Variable Y**

f_{ij}	ChO	CIS	fréquence Marginale de X=$f_{i.}$
Jeunes	0,105	0,063	$f_{1.}$=0,168
Ad Fém	0,263	0,128	$f_{2.}$=0,391
Ad masc	0,203	0,180	$f_{3.}$=0,383
Vieux	0,041	0,017	$f_{4.}$=0,058
Fréquence marginale de y=$f_{.j}$	$f_{.1}$=0.612	$f_{.2}$=0,388	1

Fig. VI.1 : Le tableau F des fréquences et des fréquences marginales de X et Y

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

VI.1.3 Les tableaux profils-lignes et profils-colonnes :

En AFC on n'analyse pas le tableau K des effectifs et non plus le tableau des fréquences relatives F. On crée des tableaux de profils qui diminuent l'influence d'une ligne ou d'une colonne de fort effectif en divisant la fréquence de la ligne ou de la colonne par sa fréquence marginale.

VI.1.3.1 Le tableau des profils-lignes : (PFL) :

Pour obtenir le tableau des profils-lignes nous divisons la fréquence f_{ij} de la ligne i du tableau F par la fréquence marginale de cette ligne $f_{i.}$ (ce rapport est aussi égal à $k_{ij}/k_{i.}$) :
$f_{11}/f_{1.}$= 69/110=0.627

PFL= **Variable X** / **Variable Y**

	ChO	CIS	
Jeunes	0.627=69/110	0.373=41/110	1
Ad Fém	0.672=172/256	0.328=84/256	1
Ad masc	0.53=133/251	0.47=118/251	1
Vieux	0.711=27/38	0.289=11/38	1

Fig. VI.2 Le tableau PFL des profils lignes

La proximité de deux lignes apparaîtrait seulement dans le tableau des profils-lignes. Les modalités Ad Fém et Ad masc de la variable X apparaissaient avec des valeurs très proches dans ce tableau : elles ont le même profil c'est-à-dire qu'elles sont proches et elles sont tout les deux éloignées des autres modalités.

VI.1.3.2 Le tableau des profils-colonnes : (PFC)

Pour obtenir le tableau des profils-colonnes nous divisons la fréquence f_{ij} de la colonne j du tableau F par la fréquence marginale de cette colonne $f_{.j}$ (ce rapport est aussi égal à $k_{ij}/k_{.j}$) :

PFC=

		Variable Y	
Variable X	$f_{ij}/f_{.j}$	**ChO**	**CIS**
	Jeunes	0.172	0.161
	Ad Fém	0.429	0.331
	Ad masc	0.332	0.465
	Vieux	0.067	0.043
		1	1

Fig. VI.3 Le tableau PFC des profils colonnes

Résumé :

Dans l'AFC, nous cherchons à mettre en évidence les proximités existant entre les lignes et les colonnes en tenant compte des poids respectifs de ces lignes et colonnes par utilisation des tableaux de profils lignes et profils colonnes

Comme dans l'ACP, on effectue une analyse dans les deux espaces :

1- L'étude du nuage des n lignes (les individus de l'ACP) dans $\mathbb{R}^p$ des p colonnes (les variables de l'ACP)
2- L'étude du nuage des p colonnes dans l'espace $\mathbb{R}^n$ des lignes.

VI.2 Analyse du nuage des n lignes dans l'espace $\mathbb{R}^p$ des colonnes et des p colonnes dans l'espace $\mathbb{R}^n$ des lignes

VI.2.1 Le nuage des n lignes dans l'espace $\mathbb{R}^p$ des colonnes :

On souhaite obtenir une représentation des proximités entre les n points-lignes i dans un sous-espace de l'espace $\mathbb{R}^p$ des colonnes, c'est-à-dire un espace de dimension inférieure à p.
L'ensemble des profils lignes forme un nuage de n points dans l'espace des p colonnes et représente ici le nuage des 4 modalités de la population des utilisateurs.
Chaque point i ayant pour coordonnées dans $\mathbb{R}^p$ les quantités :

$$\left\{ \frac{f_{ij}}{f_{i.}} \; ; \; j=1..p \right\}$$

qui correspondent aux n lignes i du tableau PFL des profils-lignes affectées des poids ou masses $f_{i.}$ des lignes correspondantes qui correspond à sa fréquence relative.

Puisque $\sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}} = 1$, les n points du nuage sont situés dans un sous-espace à p-1 dimensions.

Le centre de gravité de ce nuage est la moyenne des profils lignes affectés de leurs masses et correspond au profil moyen, c'est-à-dire au profil de la catégorie de disque sur l'ensemble de la population. Sa $j^{ème}$ composante vaut :

$$\sum_{i=1}^{n} f_{i.} \frac{f_{ij}}{f_{i.}} = f_{.j} \quad : \text{ C'est la fréquence marginale des colonnes.}$$

VI.2.2 Le nuage des p colonnes dans l'espace $\mathbb{R}^n$ des lignes :

On souhaite obtenir une représentation des proximités entre les p points-colonnes j dans un sous-espace de l'espace $\mathbb{R}^n$ des lignes, c'est-à-dire un espace de dimension inférieure à n.
L'ensemble des p profils colonnes constitue un nuage de p points dans l'espace des n lignes et représente ici le nuage des 2 modalités de la catégorie de chanson.
Chaque j ayant pour coordonnées dans $\mathbb{R}^n$ les quantités :

$$\{\frac{f_{ij}}{f_{.j}} \text{ ; i=1..n}\}$$

qui correspondent aux p colonnes j du tableau PFC des profils-colonnes affectées des poids ou masses $f_{.j}$ des colonnes correspondantes.

Les p points du nuage sont situés dans un sous-espace à n-1 dimensions puisque : $\sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}} = 1$

Le centre de gravité de ce nuage des profils colonnes est le profil moyen de la population des utilisateurs. Sa i^{ème} composante vaut :

$$\sum_{j=1}^{p} f_{.j} \frac{f_{ij}}{f_{.j}} = f_{i.}$$: C'est la fréquence marginale des lignes.

On cherche à représenter géométriquement les similitudes entre les différentes modalités d'une même variable, ce qui nous conduit à représenter les proximités entre les profils et le profil moyen défini sur l'ensemble de la population. Ceci nous amène, comme en ACP dans le cas des points-individus, à considérer le nuage de points centré sur son centre de gravité.

VI.2.3 Le choix des distances :

C'est la distance de **KHI-2** qui est utilisée en AFC pour déterminer les proximités entre les points des nuages étudiés. La distance euclidienne usuelle entre deux profils-lignes :

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

Cependant cette distance favorise les colonnes qui ont une masse $f_{.j}$ importante c'est-à-dire les catégories de chanson qui sont bien représentées dans la population étudiée.

Exemple : Ajoutant à notre tableau des effectifs une colonne définissant la chanson arabe "CAR".

K=

k_{ij}	ChO	CIS	CAR
Jeunes	69	41	10
Ad Fém	172	84	21
Ad masc	133	118	32
Vieux	27	11	2

L'effectif de la colonne j_0 : " **ChO** " est assez considérable, en tout cas beaucoup plus important que celui de la colonne "CAR". Dans un tel cas, la différence $\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$ joue un rôle excessif dans le calcul de : $d^2(i,i')$

La première modalité écrase la dernière modalité.

	ChO	CIS	CAR
D^2(Ad Fém, Jeune)	0.011	0.006	0.001

Ce qui est vrai : 0.011 est plus grande que 0.001

Pour remédier à cela, on pondère chaque écart par l'inverse de la masse de la colonne (ou l'inverse de la moyenne des profils-lignes) et l'on calcule une nouvelle distance appelée la distance de KHI-2.

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

	ChO	**CIS**	**CAR**
$D^2_{\mathbf{x}}$(Ad Fém, Jeune)	**0.020**	0.016	**0.010**

- La distance entre le profil-ligne i et le profil-ligne i' est définie par :

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

Un intérêt de la métrique du χ^2 est qu'elle vérifie le principe d'équivalence distributionnelle. Grâce à cette distance, si deux modalités ont le même profil, l'AFC fournira des résultats identiques lorsqu'on les distinguera toutes deux et lorsqu'on les réunira pour former une modalité.

En procédant à une ***transformation du tableau de profils lignes PFL***, on peut définir un tableau Y tel que la distance euclidienne usuelle calculée entre les points du nuage associé aux lignes de ce tableau soit égale à la distance du KHI-2 entre les points cordonnées du nuage associé au tableau PFL.

$$\left(\text{de } \sum_{j=1}^{p}\frac{1}{f_{.j}}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2 \text{ à } \sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i.}}\cdot\frac{1}{(f_{.j})^{1/2}} - \frac{f_{i'j}}{f_{i'.}}\cdot\frac{1}{(f_{.j})^{1/2}}\right)^2\right) \rightarrow \left(\text{de } \frac{f_{ij}}{f_{i.}} \text{ à } \frac{f_{ij}}{f_{i.}}\cdot\frac{1}{(f_{.j})^{1/2}}\right)$$

Le terme général de ce tableau Y est : $y_{ij} = \dfrac{f_{ij}}{f_{i.}}\cdot\dfrac{1}{(f_{.j})^{1/2}}$ [**I**]

Donc on a : $\sum_{j=1}^{p}(y_{ij} - y_{i'j})^2 = d^2{}_{\chi^2}(i,i') = \sum_{j=1}^{p}\frac{1}{f_{.j}}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2$

En raison de cette dernière relation, il est équivalent d'analyser le nuage associé aux lignes du tableau PFL en utilisant du KHI-2 et d'analyser le nuage associé aux lignes du tableau Y en recourant à la distance euclidienne usuelle. En adoptant cette procédure on s'engage dans une analyse basée sur la distance du KHI-2, tout en se donnant la possibilité de continuer la présentation de l'AFC en se référant à la distance euclidienne usuelle.

En utilisant la formule [I] on définit donc un tableau Y de profils lignes que l'on nommera profils lignes ***PFL "transformés"***.

En utilisant la formule suivante $y_{ij} = \dfrac{f_{ij}}{f_{i.}}\cdot\dfrac{1}{(f_{.j})^{1/2}}$, on obtient le tableau Y suivant :

		Variable Y	
	$f_{ij}/f_{.j}$	ChO	CIS
Y= Variable X	Jeunes	**0,802**	**0,599**
	Ad Fém	**0,859**	**0,527**
	Ad masc	**0,677**	**0,755**
	Vieux	**0,908**	**0,474**

Exemple de calcul : $y_{11} = 0.105/(0.168*\sqrt{0.612}) = 0.802$

Remarque :

Cette distance du χ^2 a la propriété de vérifier le principe de ***l'équivalence distributionnelle***. Selon ce principe si deux points-lignes i et l sont confondus dans R^p et si on les considère comme un seul point I de $\mathbb{R}^p$ affecté de la somme des poids de i et de l alors les distances entre tous les couple de points de $\mathbb{R}^p$ et de $\mathbb{R}^n$ restent inchangés. Tout cela signifie que si deux profil-lignes ou deux profil-colonnes sont presque confondus ou identiques (ou respectivement deux colonnes ou deux lignes sont quasiment proportionnelles) les remplacer par leur somme ne modifie presque pas les résultats de l'analyse factorielle des correspondances. Ce qui mène à diminuer les dimensions des espaces étudiés.

Définition du centre de gravité et centrage :

Le ***centre de gravité*** du nuage, munis des poids $f_{i.}$ se calcule comme une ***moyenne pondérée*** des profils-lignes :

en effet : $\sum_{i=1}^{n} f_{i.}\left(f_{ij} / f_{i.}\right) = f_{.j}$

Dans l'exemple, G est le profil des choix Chanson orientale et Chanson Islamique de l'ensemble de la population, tous les individus Jeunes, Ad Fém, Ad masc et vieux étant cumulés. Ce barycentre G servira de référence dans l'étude des lignes du tableau.
Par suite, le centre de gravité du nuage des points associés aux lignes du tableau ***PFL "transformés" Y*** est :

$G_I=((f_{.1})^{1/2}\ldots (f_{.j})^{1/2}\ldots(f_{.p})^{1/2})$ avec $G_{Ij} = \sum_{i=1}^{n} f_{i.} \frac{f_{ij}}{f_{i.}\left(f_{.j}\right)^{1/2}} = \left(f_{.j}\right)^{1/2}$

Cela correspond ici au vecteur suivant :

0,78244089	**0,62272487**	**$=G_{I.}$**

En AFC on cherche, comme en ACP, à obtenir une image du nuage accessible à nos sens, tout en veillant à déformer le moins possible l'ensemble des distances entre les points. Pour que l'on puisse obtenir ce résultat il est "***théoriquement***" nécessaire que l'ajustement que l'on va pratiquer soit réalisé à partir d'un tableau de données préalablement centrées.
On opérant ce centrage on obtient le tableau L des ***PFL "transformés"*** puis ***centrés*** dont le terme général est : $l_{ij} = y_{ij} - \left(f_{.j}\right)^{1/2} = \frac{f_{ij}}{f_{i.}\left(f_{.j}\right)^{1/2}} - \left(f_{.j}\right)^{1/2}$

Ce tableau L est ici le suivant :

L= **Variable X**

	Variable Y	
	ChO	**CIS**
Jeunes	**0,019**	**-0,024**
Ad Fém	**0,076**	**-0,096**
Ad masc	**-0,105**	**0,132**
Vieux	**0,126**	**-0,149**

Exemple de calcul $\mathbf{l_{11}= y_{11}-(f_{.1})^{1/2}=0,802-\sqrt{0.612}=0.019}$

Définition des axes factoriels :

L'ajustement du nuage des points associés aux lignes du tableau L consiste à définir les axes factoriels qui, avec une déformation minimale, permettant de réduire le nuage à sa projection dans un espace de plus faible dimension.

La logique de cet ajustement est formellement identique à celle qui est appliquée au nuage des individus en ACP non normée. Pour déterminer les axes factoriels on doit théoriquement diagonaliser la matrice des variances-covariance associée au tableau analysé. Le $\alpha^{ème}$ axe factoriel est en effet engendré par le vecteur propre du u_α associé à la $\alpha^{ème}$ plus grande valeur propre de cette matrice des variances-covariance que l'on notera T.

Quand on analyse le nuage associé au tableau des profils lignes "transformés" centrés, avec les poids ($f_{1.}$ …$f_{i.}$ … $f_{n.}$), la matrice T a pour terme général :

$$t_{jj'} = \sum_{i=1}^{n} f_{i.}\left(\frac{f_{ij}}{f_{i.}(f_{.j})^{1/2}} - (f_{.j})^{1/2}\right)\left(\frac{f_{ij'}}{f_{i.}(f_{.j'})^{1/2}} - (f_{.j'})^{1/2}\right)$$

Elle est définie par le produit matriciel X'X=T avec : $x_{ij} = \dfrac{f_{ij} - (f_{i.} f_{.j})}{(f_{i.} f_{.j})^{1/2}}$

Dans la pratique, on obtient généralement les vecteurs propres recherchés en diagonalisant non pas la matrice T mais une matrice T* qui a pour terme général :

$t^*_{jj'} = \sum_{i=1}^{n} \dfrac{f_{ij} f_{ij'}}{f_{i.}(f_{.j} f_{.j'})^{1/2}}$ qui est définie par le produit matriciel X*'X*=T* avec : $x^*_{ij} = \dfrac{f_{ij}}{(f_{i.} f_{.j})^{1/2}}$

Il est possible de procéder ainsi parce que les matrices T=X'X et T*=X*'X* ont les mêmes vecteurs propres associés aux mêmes valeurs propres, l'un de ces vecteurs propres, u_0, étant associé à une valeur propre λ_0 qui est nulle pour T et vaut 1 pour T*.

La différence concernant la valeur propre associée à u_0 est sans conséquence car ce vecteur propre correspond à un facteur "trivial" que l'on ne saurait utiliser.

Ce facteur trivial devant être éliminé de l'analyse, on peut indifféremment diagonaliser T et T* lorsque l'on cherche à déterminer les axes factoriels.

La matrice T* correspond à la matrice que l'on aurait dû diagonaliser si l'on avait analysé le nuage sans procéder préalablement à un centrage des données.

Les principes établis dans le cadre de l'ACP indique que le centrage est théoriquement nécessaire pour que la "réduction" du nuage minimise la déformation des distances entre les points.

Celui-ci n'est cependant pas indispensable en AFC : ce type d'analyse, que l'on centre ou non les données, on parvient aux mêmes axes factoriels et aux mêmes coordonnées factorielles.

Calcul des valeurs propres de la matrice T=X'X

X=

Variable X	**Variable Y**	
	ChO	**CIS**
Jeunes	**0,008**	**-0,010**
Ad Fém	**0,048**	**-0,060**
Ad masc	**-0,065**	**0,082**
Vieux	**0,030**	**-0,038**

Par exemple : $x_{12} = \dfrac{f_{12} - f_{1.} f_{.2}}{\sqrt{f_{1.} f_{.2}}} = \dfrac{0.063 - 0.168 * 0.388}{\sqrt{0.168 * 0.388}} = -0.010$

T=X'X=

0,008	0,048	-0,065	0,03
-0,01	-0,06	0,082	-0,038

•

0,008	-0,010
0,048	-0,060
-0,065	0,082
0,030	-0,038

=

0,007	-0,009
-0,009	0,012

T=X'X=

X'X(2x2)	**ChO**	**CIS**
ChO	**0,007**	**-0,009**
CIS	**-0,009**	**0,012**

Calculons les valeurs propres de T. les valeurs propres de cette matrice vérifient l'équation du second degré :

$$P(\lambda) = Det(T - \lambda I) = \begin{vmatrix} 0.007 - \lambda & -0.009 \\ -0.009 & 0.012 - \lambda \end{vmatrix} = \lambda^2 - 0.019\lambda$$

$\lambda_1 = 0.019$ et $\lambda_2 = 0$. La valeur propre triviale de T est bien 0.

Calcul des valeurs propres de la matrice T*=X*'X* ?

Calcul des valeurs propres de la matrice T*=X*'X* :

Reprenons les calculs pour la matrice T* :

$$x_{ij}^{*} = \frac{f_{ij}}{(f_{i.} f_{.j})^{1/2}},$$

Exemple :

$$x_{12} = \frac{f_{11}}{\sqrt{f_{1.} f_{.1}}} = \frac{0.063}{\sqrt{0.168 * 0.612}} = 0.329$$

X*=

Variable	**Variable Y**	
	ChO	**CIS**
Jeunes	0,329	0,245
Ad Fém	0,537	0,329

X

Ad masc	0,419	0,467
Vieux	0,219	0,112

T*=X*'X*=

0,329	0,537	0,419	0,219
0,245	0,329	0,467	0,112

•

0,329	0,245
0,537	0,329
0,419	0,467
0,219	0,112

T*=X*'X*=

X*'X*(2x2)	**ChO**	**CIS**
ChO	**0,620**	**0,478**
CIS	**0,478**	**0,399**

Calcul des valeurs propres de la matrice T*=X*'X* :

$$P(\lambda) = Det(T^* - \lambda I) = \begin{vmatrix} 0.620 - \lambda & 0.478 \\ 0.478 & 0.399 - \lambda \end{vmatrix} = \lambda^2 - 1.019\lambda + 0.019$$

$\lambda_1 = 1$ et $\lambda_2 = 0.0193$. La valeur propre triviale de T* est bien 1.

Calcul des vecteurs propres :

Il n'y a dans cet exemple qu'un seul (1) vecteur propre u_1 à chercher. Le vecteur propre u_1 associé à la valeur propre λ_2 vérifie :

$$(T^* - \lambda_2 I).u_1 = \begin{bmatrix} 0.620 - 0.0193 & 0.478 \\ 0.478 & 0.399 - 0.0193 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.601 & 0.478 \\ 0.4778 & 0.380 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Ce qui nous conduit à résoudre le système :

$0.601x_1+0.478x_2=0$ (eq. 1)

$0.4778x_1+0.380x_2=0$ (eq. 2)

Comme l'équation (2) est proportionnelle à l'équation (1) on aura $1.26x_1+x_2=0$, c'est-à-dire $u_1=(x_1,-1.26\ x_1)'$. Il reste à normé ce vecteur à 1 :

$\| u_1\|^2=1=(x_1^2+(-1.26\ x_1)^2)=2.58\ x_1^2$, on en déduit que $x_1=0.63$

Ainsi $u_1=(0.623,-0.782)'$

Calcul du taux d'inertie :

Le calcul du taux d'inertie permet d'évaluer la qualité globale de l'ajustement.

Le taux d'inertie associé à l'axe factoriel α indique la part d'inertie totale du nuage que "restitue" cet axe. Il est défini par : $\tau = \dfrac{\lambda_\alpha}{\sum_\alpha \lambda_\alpha}$

N.B. : La somme figurant au dénominateur ne prend pas en compte la valeur propre λ_0 de T* qui correspond au facteur trivial éliminé de l'analyse.

On calcul les taux d'inertie on obtient :

N°	Valeur propre	% d'inertie	% cumulé
1	0,0193	100%	0,0193

Du moment qu'il y a deux valeurs propres et que la première valeur propre n'est pas prise en compte, on aura par suite, seulement cette unique valeur propre. On voit ici que l'inertie du nuage des points lignes (ou colonne) sur le premier plan factoriel représente 100% de l'inertie totale du nuage.

Calcul des coordonnées factorielles des points profils-lignes :

A l'aide du vecteur propre on peut calculé les coordonnées des projections des points profils-lignes sur l'axe factoriel. On utilisant le tableau des profils-lignes "transformé" centré et en opérant le produit matriciel LU on obtient les coordonnées des projections des points lignes sur l'axe factoriel. On obtient les mêmes résultats en utilisant les profils "transformés" non centrés et en réalisant le produit matriciel YU.
Le produit LU fournit ici les coordonnées factorielles suivantes :

1- LU

u1=(0.623,-0.782)

0,019	**-0,024**
0,076	**-0,096**
-0,105	**0,132**
0,126	**-0,149**

$$\bullet \begin{bmatrix} 0.623 \\ -0.782 \end{bmatrix}$$

=0.019*0.623+(-0.024)*(-0.782)=0.031

2- ou YU

0,802	**0,599**
0,859	**0,527**
0,677	**0,755**
0,908	**0,474**

$$\bullet \begin{bmatrix} 0.623 \\ -0.782 \end{bmatrix}$$

=0,802*0,623+0,599*(-0,782)=0.031

On aura dans les deux cas le tableau des coordonnées:

0.031
0.122
-0.169
0.202

Calcul des contribution absolue des points profils-lignes CTA:

Afin de faciliter l'interprétation des axes factoriels il est utile de calculer les contributions "absolue" (CTA) des points lignes définies par :

$$\text{CTA}(i,\alpha)=\frac{f_{i.}Coord^2(i,\alpha)}{\lambda_\alpha} \quad \text{avec} \sum_{i=1}^{n} f_{i.}Coord^2(i,\alpha)=\lambda_\alpha$$

La contribution absolue CTA(i,α) représente la part de l'inertie de la projection du point i sur l'axe factoriel α dans l'inertie des projections de l'ensemble du nuage sur ce même axe.
Les points auxquels est attachée une CTA(i,α) élevée ont fortement "attiré" l'axe factoriel α et ont joué un grand rôle dans la fixation de sa direction. Ce sont ces points qu'il faut prendre en compte lorsque l'on procède à l'interprétation de l'axe factoriel α. Le calcul donne les CTA suivantes exprimé en % :
Exemple : CTA(Ad Fém,1)=(0,391*$\mathbf{0.122}^2$/0.0193)=0.302

Institut de sciences exactes
Département d'Informatique
Module : Analyse de données

Fiche de TD
Analyse Factorielle des Correspondances

Exercice 1 :

Tableau exemple : dimension (3,4)

Un disquaire qui a fait un test de vente sur un échantillon de 1000 disques destinés à diverses catégories (ou classes) de clientèle.

Le disquaire a réparti ses disques en ***3 catégories*** : Chansons (Ch), Islamiques (IS) et musique classiques (MC) et la population des utilisateurs en ***4 catégories*** (jeunes sans distinction de sexe, adultes féminins, adultes masculins et vieux sans distinction de sexe). Il a obtenu le tableau suivant :

K=

k_{ij}	Ch	IS	MC
Jeunes	69	41	18
Ad Fém	172	84	127
Ad masc	133	118	157
Vieux	27	11	43

Cet exercice contient en grande partie les données de l'exemple du cours, une troisième colonne, qui correspond à l'achat de disques de musique classique par les quatre types de consommateurs.

Effectué une Analyse Factorielle des Correspondances (AFC) pour la matrice des effectifs ci-dessus?

Exercice 2 :

Au cours d'une enquête 264 personnes ont répondu à deux questions Q1 et Q2 en choisissant à chaque fois une modalité (et une seule) parmi les quatre qui leur étaient proposées. Un "tri croisé" a été effectué à partir des réponses relatives à ces deux questions. Ce traitement a permis d'établir le tableau de contingence suivant K.

K=

	q2m1	q2m2	q2m3	q2m4	Total
Q1M1	20	15	9	7	51
Q1M2	**8**	11	30	25	74
Q1M3	18	16	12	10	56
Q1M4	6	50	14	13	83
Total	52	92	65	55	264

Exemple de lecture de ce tableau : il y a 8 personnes qui ont choisi la deuxième modalité en répondant à la question Q1 et ont opté pour la première modalité en répondant à la question Q2.

Réalisez une **Analyse Factorielle des Correspondances (AFC)** à partir de ce tableau de contingence et vous pouvez prendre les éléments suivants :

1- valeurs propres :

$\lambda_0 = 1.0000$

$\lambda_1 = 0.146445$

$\lambda_2 = 0.111375$

$\lambda_3 = 0.00014$

2- Matrice des vecteurs propres u_α

u_0	u_1	u_2	u_3
0.443812684	-0.146669952	0.883549918	0.029287701
0.590326051	0.75965758	-0.164436558	-0.030298142
0.496197622	-0.459857338	-0.3033366091	-0.671048705
0.456435510	-0.378762933	-0.316681823	0.740214758

Institut de sciences exactes
Département d'Informatique
Module : Analyse de données

Solution de la fiche de TD n°05 (AFC)

Exercice 1 :

Tableau exemple : dimension (3,4)

Un disquaire qui a fait un test de vente sur un échantillon de 1000 disques destinés à diverses catégories (ou classes) de clientèle.

Le disquaire a réparti ses disques en ***3 catégories*** : Chansons (Ch), Islamiques (IS) et musique classiques (MC) et la population des utilisateurs en ***4 catégories*** (jeunes sans distinction de sexe, adultes féminins, adultes masculins et vieux sans distinction de sexe). Il a obtenu le tableau suivant :

K=

k_{ij}	**Ch**	**IS**	**MC**
Jeunes	69	41	18
Ad Fém	172	84	127
Ad masc	133	118	157
Vieux	27	11	43

Cet exercice contient en grande partie les données de l'exemple du cours, une troisième colonne, qui correspond à l'achat de disques de musique classique par les quatre types de consommateurs.

Effectué une Analyse Factorielle des Correspondances (AFC) pour la matrice des effectifs ci-dessus?

Solution :

K=

k_{ij}	**Ch**	**IS**	**MC**	**Total : $k_{i.}$**
Jeunes	69	41	18	128
Ad Fém	172	84	127	383
Ad masc	133	118	157	408
Vieux	27	11	43	81
Total : $k_{.j}$	401	254	345	1000

Le tableau F des fréquences et des fréquences marginales de X et Y

F=

		Variable Y			**fréquence Marginale de X**
		Ch	**IS**	**MC**	
Variable X	Jeunes	0,069	0,041	0,018	$f_{1.}$=0,128
	Ad Fém	0,172	0,084	0,127	$f_{2.}$=0,383
	Ad masc	0,133	0,118	0,157	$f_{3.}$=0,408
	Vieux	0,027	0,011	0,043	$f_{4.}$=0,081
	Fréquence marginale de y	$f_{.1}$=0,401	$f_{.2}$=0,254	$f_{.3}$=0,345	1

Le tableau PFL des profils lignes

PFL=

		Variable Y			
		Ch	**IS**	**MC**	Total ligne
Variable X	Jeunes	0,539	0,320	0,141	1
	Ad Fém	0.449	0,219	0,332	1
	Ad masc	0,326	0,289	0,385	1
	Vieux	0,333	0,136	0,531	1

Le tableau PFC des profils colonnes

PFC=			Variable Y		
			Ch	**IS**	**MC**
	Variable X	Jeunes	0,172	0,161	0,052
		Ad Fém	0,429	0,331	0,368
		Ad masc	0,332	0,465	0,455
		Vieux	0,067	0,043	0,125
	Total colonne		1	1	1

Le tableau X :

X =			Variable Y		
			Ch	**IS**	**MC**
	Variable X	Jeunes	0.08	0.05	-0.12
		Ad Fém	0.05	-0.04	-0.01
		Ad masc	-0.08	0.04	0.04
		Vieux	-0.03	-0.07	0.09

X*=Y=			Variable Y		
			Ch	**IS**	**MC**
	Variable X	Jeunes	0.30	0.23	0.09
		Ad Fém	0.44	0.27	0.35
		Ad masc	0.33	0.37	0.42
		Vieux	0.15	0.08	0.26

T* =		**Ch**	**IS**	**MC**
	Ch	0.4159	0.319	0.356
	IS	0.3195	0.264	0.287
	MC	0.3556	0.287	0.371

Calcul des valeurs propres :

Det(T*- λ I)=0

λ_0=1 (valeur propre triviale), λ_1=0.0397, λ_2=0.0114

Exercice 2 :

le tableau de contingence suivant K.

K=

	q2m1	**q2m2**	**q2m3**	**q2m4**	**Total**
Q1M1	20	15	9	7	51
Q1M2	**8**	11	30	25	74
Q1M3	18	16	12	10	56
Q1M4	6	50	14	13	83
Total	52	92	65	55	264

Exemple de lecture de ce tableau : il y a 8 personnes qui ont choisi la deuxième modalité en répondant à la question Q1 et ont opté pour la première modalité en répondant à la question Q2.

1- valeurs propres :

$\lambda_0 = 1.0000$

$\lambda_1 = 0.146445$

$\lambda_2 = 0.111375$

$\lambda_3 = 0.00014$

2- Matrice des vecteurs propres u_α

u_0	u_1	u_2	u_3
0.443812684	-0.146669952	0.883549918	0.029287701
0.590326051	0.75965758	-0.164436558	-0.030298142
0.496197622	-0.459857338	-0.3033366091	-0.671048705
0.456435510	-0.378762933	-0.316681823	0.740214758

Solution :

K=

	q2m1	q2m2	q2m3	q2m4	Total
Q1M1	20	15	9	7	51
Q1M2	**8**	11	30	25	74
Q1M3	18	16	12	10	56
Q1M4	6	50	14	13	83
Total	52	92	65	55	264

Le tableau F des fréquences et des fréquences marginales de X et Y

F=

	q2m1	q2m2	q2m3	q2m4	fréquence Marginale de X
Q1M1	0.0758	0.0568	0.0341	0.0265	**0,1932**
Q1M2	**0.0303**	0.0417	0.1136	0.0947	**0,2803**
Q1M3	0.0682	0.0606	0.0545	0.0379	**0,2121**
Q1M4	0.0227	0.1894	0.0530	0.0492	**0,3144**
	0,1970	**0,3485**	0,2462	0,2083	1

Le tableau PFL des profils lignes

PFL=

	q2m1	q2m2	q2m3	q2m4	fréquence Marginale de X
Q1M1	0,3922	0,2941	0,1765	0,1373	1,0000
Q1M2	0,1081	0,1486	0,4054	0,3378	1,0000
Q1M3	0,3214	0,2857	0,2143	0,1786	1,0000
Q1M4	0,0723	0,6024	0,1687	0,1566	1,0000

Le tableau PFLT des profils lignes transformé

Y=

	q2m1	q2m2	q2m3	q2m4	fréquence Marginale de X
Q1M1	0,8836	0,4982	0,3556	0,3007	1,0000
Q1M2	0,2436	0,2518	0,8170	0,7402	1,0000
Q1M3	0,7242	0,4840	0,4319	0,3912	1,0000
Q1M4	0,1629	1,0205	0,3399	0,3432	1,0000

Profil Ligne transformé puis centré

L=

	q2m1	q2m2	q2m3	q2m4
Q1M1	**0,4398**	**-0,0921**	**-0,1406**	**-0,1557**
Q1M2	**-0,2002**	**-0,3385**	**0,3208**	**0,2837**
Q1M3	**0,2804**	**-0,1063**	**-0,0643**	**-0,0652**
Q1M4	**-0,2809**	**0,4301**	**-0,1563**	**-0,1133**

X=

	q2m1	q2m2	q2m3	q2m4
Q1M1	0,1933	-0,0405	-0,0618	-0,0684
Q1M2	-0,1060	-0,1792	0,1699	0,1502
Q1M3	0,1292	-0,0490	-0,0296	-0,0300
Q1M4	-0,1575	0,2412	-0,0876	-0,0635

X'=

	q2m1	q2m2	q2m3	q2m4
Q1M1	0,1933	-0,1060	0,1292	-0,1575
Q1M2	-0,0405	-0,1792	-0,0490	0,2412
Q1M3	-0,0618	0,1699	-0,0296	-0,0876
Q1M4	-0,0684	0,1502	-0,0300	-0,0635

T=X'X=

	q2m1	q2m2	q2m3	q2m4
Q1M1	0,0901	-0,0331	-0,0200	0,0901
Q1M2	-0,0331	0,0943	-0,0476	-0,0331
Q1M3	-0,0200	-0,0476	0,0412	-0,0200
Q1M4	-0,0230	-0,0380	0,0362	-0,0230

X*=

	q2m1	q2m2	q2m3	q2m4
Q1M1	0,3884	0,2190	0,1563	0,1322
Q1M2	0,1290	0,1333	0,4326	0,3919
Q1M3	0,3336	0,2229	0,1989	0,1802
Q1M4	0,0913	0,5722	0,1906	0,1924

X*'=

	q2m1	q2m2	q2m3	q2m4
Q1M1	0,3884	0,1290	0,3336	0,0913
Q1M2	0,2190	0,1333	0,2229	0,5722
Q1M3	0,1563	0,4326	0,1989	0,1906
Q1M4	0,1322	0,3919	0,1802	0,1924

T*=X*'X*=

	q2m1	q2m2	q2m3	q2m4
Q1M1	0,2871	0,2289	0,2002	0,1795
Q1M2	0,2289	0,4428	0,2453	0,2314
Q1M3	0,2002	0,2453	0,2874	0,2627
Q1M4	0,1795	0,2314	0,2627	0,2405

V ecteur propre

u_1	u_2	u_3
-0.1467	0.8834	0.0293
0,7597	-0,1644	-0,0303
-0,4599	-0,3033	-0,6710
-0,3788	-0,3167	0,7402

Taux d'inerie

N°	V aleur propre	% inertie
1	0,1464	56,7704
2	0,1114	43,1753
3	0,0001	0,0543

Calcul des coordonnées factorielles des points profils lignes sur les axes factoriels :

En utilisant le tableau des profils lignes "transformé" centrés et on opère le produit : LU on obtient les coordonnées des projections des points lignes sur les axes factoriels. Et on obtient le même résultat en utilisant les profils "transformés" non centrés et en réalisant le produit matriciel : YU.

Le produit LU est :

-0,0109	0,4957	-0,0053
-0,4828	-0,3084	-0,0009
-0,0676	0,3054	0,0063
0,4827	-0,2357	-0,0003

Calcul des contributions absolue des points profils lignes :

$$\text{CTA(i,}\alpha\text{)} = \frac{f_{i.}\,coord^2(i,\alpha)}{\lambda_\alpha}$$

Avec : $\sum_{i=1}^{n} f_{i.}\,coord^2(i,\alpha) = \lambda_\alpha$

Cette contribution représente la part de l'inertie de la projection du point i sur l'axe α dans l'inertie des projections de l'ensemble du nuage sur ce même axe :

	Cta(i,1)	Cta(i,2)	Cta(i,3)
Q1M1	0,02	42,62	38,5
Q1M2	46,51	23,94	1,54
Q1M3	0,73	17,77	61,01
Q1M4	52,74	15,68	0,15
Total	100	100	100

VII. Classification automatique

VII.1 Introduction

La nature offre un grand nombre de populations donc il y a nécessité d'une éventuelle répartition en classes.

Exemples :

1. En médecine les regroupements de malades ayant le même comportement vis à vis de certaines maladies.
2. Répartition d'une population de personnes suivant des critères tel que sexe, activité, état matrimonial ....

La même population peut aussi être soumise, suivant le besoin, à une autre classification comme par exemple le sexe, la nature du travail... .

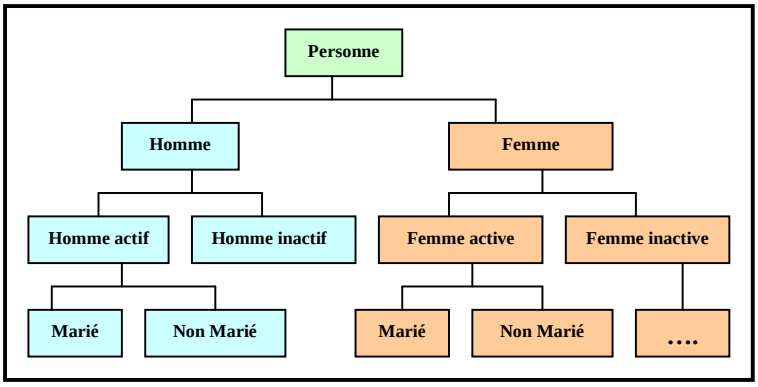


Fig. VII.1 Exemple de classification

VII.2 But de la classification

Classifier, c'est regrouper entre eux des objets similaires selon tel ou tel critère. Les diverses techniques visent toutes à répartir *n* individus, caractérisés par *p* variables *X1*, *X2*, ..., *Xp* en un certain nombre m de sous-groupes aussi homogènes que possible.

Le plus souvent, on veut, en général, obtenir des sections à l'intérieur des groupes principaux, puis des subdivisions plus petites de ces sections, et ainsi de suite. En bref, on désire avoir une hiérarchie, c'est à dire une suite de partitions "emboîtées", de plus en plus fines, sur l'ensemble d'observations initial.

On parle ainsi de classification automatique : les classes seront obtenues au moyen d'algorithmes formalisés et non par des méthodes subjectives ou visuelles faisant appel à L'initiative du praticien.

On distingue deux grandes familles de techniques de classification :

- La classification hiérarchique : pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents.
- La classification non hiérarchique ou partitionnement, aboutissant à la décomposition de l'ensemble de tous les individus en ***m*** ensembles disjoints ou classes d'équivalence; le nombre ***m*** de classes est ***fixé***.

VII.3 Les éléments d'une classification

Les problèmes de classification automatique diffèrent selon le type d'information recherché: une hiérarchie, une partition, une recouvrement ...

VII.3.1 Les partitions

Une partition de l'ensemble des observations Ω est un ensemble de parties non vides $P = (P_1,\ldots,P_k)$ d'intersection vides deux à deux et dont la réunion forme Ω avec :

1) $\forall j \in \{1,2,\ldots,k\}\ Pj \neq \varnothing$ ➔ Les parties ne sont pas vides

2) $\forall i,j \in \{1,2,\ldots,k\}\ i \neq j\ ;\ Pi \cap Pj = \varnothing$ ➔ les parties sont d'intersections vides deux à deux

3) $\bigcup_{i=1}^{k} P_j = \Omega$ ➔ la réunion de toutes les partitions est l'ensemble Ω

Ainsi avec les sept points suivants:

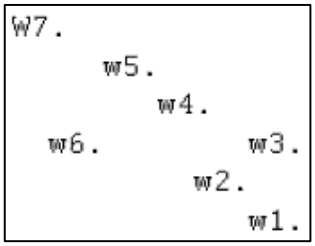


Fig. VII.2 Ensemble des observations

on peut, par exemple, construire une partition en trois classes:
$P=(P_1, P_2, P_3)$ représentée par
$P_1 = \{ w_7\}$,
$P_2 = \{ w_5 , w_4, w_6\}$
et $P_3 = \{ w_1 , w_2, w_3\}$.

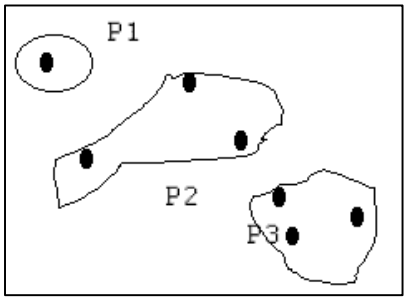


Fig. VII.3 Partition en classes

VII.3.2 Les recouvrements :

Un recouvrement de Ω est un ensemble de parties non vides $P = (P_1, \ldots, P_k)$ dont la réunion forme Ω.

1) $\forall j \in \{1,2,\ldots,k\}\ Pj \neq \varnothing$ ➔ Les parties ne sont pas vides

2) $\bigcup_{i=1}^{k} P_j = \Omega$ ➔ la réunion de toutes les partitions est l'ensemble Ω

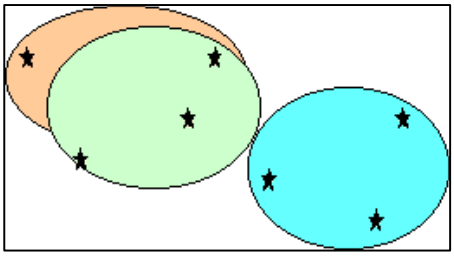


Fig. VII.4 Recouvrement en classes

Avec les sept points précédents, on peut aussi construire un recouvrement à trois classes $P=(P_1, P_2,P_3)$: $P_1 =\{ w_7 , w_5,w_4\}$; $P_2 =\{ w_5 , w_4,w_6\}$; et $P_3 =\{ w_1 , w_2,w_3\}$.

➔ **Donc une partition est un cas particulier de recouvrement.**

VII.3.3 Les Hiérarchies :

On cherche à représenter Ω par un ensemble de partitions emboîtées. Soit Ω un ensemble fini, H un ensemble de parties (appelées paliers) non vides de Ω. H est une hiérarchie sur Ω si :

1) $\Omega \in H$ (Le palier le plus haut contient tous les individus)
2) $\forall w \in \Omega, \{ w \} \in H$ *(Les points terminaux)*
3) $\forall h, h' \in H$ on a : $h \cap h' \neq \varnothing \Rightarrow h \subset h'$ ou $h' \subset h$

Nous utilisons encore l'ensemble Ω formé des sept points précédents; une hiérarchie associée H associée peut être:

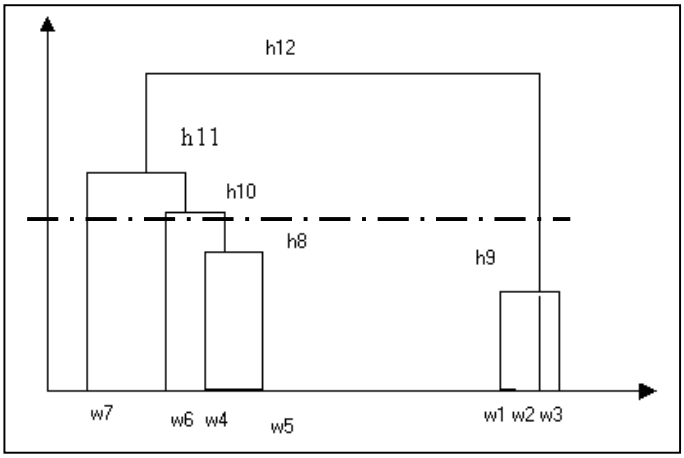


Fig. VII.5 Hiérarchie

On a bien $H=\bigcup_{i=1,12} h_i$ avec $h_i=\{w_i\}$ pour i=1,7

h8={w4,w5} ;
h9={w1,w2,w3} ;
h10= {w6}∪ h8
h_{11} = {w7} ∪h_{10}
et $h_{12}= h_{11}\cup h_9$.

On vérifie facilement que H satisfait bien aux trois axiomes de la définition d'une hiérarchie H={h1,h2,h3,h4,h5,h6,h7,h8,h9,h10,h11,h12}

Le trait horizontal mixte indique un niveau de troncature définissant une partition en quatre classes :
w7,h8,w6 et h9

Une telle hiérarchie peut être résumée par un arbre hiérarchique [**figure ci-dessus**] dont les noeuds (h8,..,h12) symbolisent les diverses subdivisions de l'échantillon ; les éléments de ces subdivisions étant les objets (w1,…,w7), placés à l'extrémité inférieure des branches qui leur sont reliées.
En faisant varier ce niveau de troncature on obtient les diverses partitions constituant la hiérarchie.

VII.4 La critère de ressemblance : une notion de distance

Comme nous l'avons dit en introduction, pour assembler des individus proches et séparer des individus éloignés il faut tout d'abord définir la notion de "distance", que l'on appelle aussi ***dissimilarité***. Le critère de ressemblance entre deux individus (ou deux classes) est une ***dissimilarité*** calculée en fonction des caractéristiques retenues pour chaque individu.

VII.4.1 Distances entre deux individus

Il existe plusieurs distances possibles entre deux individus. Le choix de la distance dépend de la problématique que l'on se pose.

- **Distance euclidienne :**

On utilise en général la distance euclidienne qui est la distance classique entre deux points. On rappelle ici la définition de la distance euclidienne.
Soit deux points ***(M1,M2)*** qui ont 2 variables uniquement : (x1, y1) et (x2, y2).
La distance euclidienne (distance classique) : $d(M_1,M_2)=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$
Cette distance est généralement utilisée dans le contexte des données quantitatives.

- **Distance du χ^2 :**

Cette distance est couramment utilisée en analyse factorielle des correspondances.
Cette distance est généralement utilisée dans le contexte des données qualitatives.

VII.4.2 Matrice des distances

Pour un nuage d'individus, on peut résumer l'ensemble des distances entre individus au sein d'une matrice des distances que l'on note $\Re$. Chaque coefficient d_{ij} représente la distance entre l'individu M_i et l'individu M_j . Par exemple, si l'on choisit comme critère de ressemblance la distance euclidienne, on a $d_{ij} = d_2(M_i,M_j)$.
Voici les propriétés de ce type de matrice. Une matrice de distances est :

- Une matrice carrée.
- Une matrice symétrique ($d_{ij} = d_{ji}$).
- De coefficients positifs ($d_{ij} > 0$).
- De coefficients nuls sur la diagonale ($d_{ii} = d(M_i,M_i) = 0$).

VII.5. La classification automatique hiérarchique :

Il s'agit d'effectuer une partition de classes de plus en plus vaste (classification hiérarchique ascendante) ou de moins en moins vaste (classification automatique descendante).

Nous développerons l'algorithme de la classification hiérarchique ascendante.

VII.5.1 La classification hiérarchique ascendante (CHA)

But : Trouver la meilleure partition pour chaque nombre de classes k (k=1,..,n) selon un critère de «qualité défini».

A ce niveau, on peut relever deux problèmes : une limite théorique et une limite numérique.

- La limite théorique est que la notion de meilleure partition est abstraite. Il faut à nouveau définir un critère de qualité pour trouver la meilleure partition selon ce critère.
- La seconde limite est numérique. Le nombre de partition possible à partir d'une famille d'individus explose rapidement et le temps de calcul devient alors problématique même pour des ordinateurs modernes. Par exemple, il y a plus de 4 millions de possibilités de regrouper 100 individus en 5 classes distinctes. $\left(\frac{n^k}{k!}\right) = \left(\frac{100^4}{4!}\right) =$ **4.166.667**

Donc l'objectif est de trouver parmi l'ensemble de ces partitions la meilleure partition selon le couple [***critère de qualité ; nombre de classes de la partition***].

Hypothèse :

Avant d'effectuer un algorithme, il faut choisir le critère de ressemblance entre deux individus, i.e. comment mesurer la distance entre deux individus.

Nous choisissons la distance entre deux individus par la distance euclidienne.

Notations : On note Γ(i) la partition obtenue à la I^{ème} itération. Dans l'algorithme CHA, Γ(i) est composée de n − i classes.

Les 4 étapes de la méthode

1. Explication du choix du critère de qualité et le critère de ressemblance qui en découle.
2. Description de l'algorithme.
3. Construction du ***dendrogramme*** (graphique récapitulatif).
4. Présentation du critère de décision associé au ***dendrogramme***.

1ère étape : Critère de ressemblance et critère de qualité entre deux classes : Notion d'inertie et **Ecart de Ward :**

- **Critère de qualité et notion d'inertie :**

  Une partition pour être bonne doit satisfaire les deux critères suivants :

  1. Les individus proches doivent être regroupés.
  2. Les individus éloignés doivent être séparé.

  Pour évaluer l'homogénéité d'un nuage de points où la distance entre deux points est la distance euclidienne, une mesure classique est l'inertie de ce nuage pour la distance euclidienne.
- **Inertie interclasse et inertie intraclasse**

On appelle inertie totale d'un nuage $\Gamma = \{M_i, i = 1, \ldots, n\}$ la moyenne des carrés des distances de ses points au centre de gravité du nuage. Donc, si G désigne le centre de gravité de Γ, l'inertie totale de Γ est :

$$\boxed{\mathcal{L}(\Gamma) = \sum_{i=1}^{n} p_i d_2(M_i, G)^2}$$

Si tous les points du nuage sont de même poids égal à $1/n$

$$\boxed{\mathcal{L}(\Gamma) = \frac{1}{n}\left(d_2(M_1, G)^2 + d_2(M_2, G)^2 + \ldots + d_2(M_n, G)^2\right)}$$

L'inertie totale d'une classe est :

$$\boxed{\mathcal{L}(\gamma_j) = \sum_{i \mid M_i \in \gamma_j}^{n} p_i d_2(M_i, G_j)^2}$$

L'inertie est une mesure de l'homogénéité d'un ensemble de points (nuage ou classe). Une classe (ou un nuage) sera d'autant plus homogène que son inertie totale sera faible.

Ce qui nous intéresse n'est pas l'homogénéité d'une classe mais l'homogénéité de l'ensemble des classes d'une partition. Pour cela, on introduit la notion d'inertie intraclasse d'une partition :

$$\boxed{\mathcal{L}_{intra}(\Gamma) = \sum_{j=1}^{k} \mathcal{L}(\gamma_j)}$$

L'inertie intraclasse d'une partition est la somme des inerties totales de chaque classe de la partition. L'inertie intraclasse mesure l'homogénéité de l'ensemble des classes. Plus l'inertie intraclasse est faible, plus la partition est composée de classes homogènes. On peut noter deux cas particuliers :
Si chaque classe de la partition représente un seul individu

$$\boxed{\Gamma = (\gamma_1 = M_1, \gamma_2 = M_2, \ldots, \gamma_n = M_n)}$$

l'inertie intraclasse de cette partition est nulle :

$$\boxed{\mathcal{L}_{intra}(\Gamma) = 0}$$

En effet, comme le centre de gravité d'une classe composée d'un seul point est le point lui même, l'inertie totale de chaque classe est nulle :

$$\boxed{\mathcal{L}(\gamma_j) = d_2(M_i, M_i)^2 = 0}$$

Si la partition n'est composée que d'une seule classe

$$\boxed{\Gamma = (\gamma_1) \quad où \quad \gamma_1 = (M_1, M_2, \ldots, M_n)}$$

l'inertie intraclasse de la partition est égale à l'inertie totale du nuage :

$$\boxed{\mathcal{L}_{intra}(\Gamma) = \mathcal{L}(\Gamma)}$$

On introduit maintenant une mesure du niveau de séparation entre les classes : l'inertie interclasse.
On suppose qu'une bonne mesure du niveau de séparation des classes est la somme pondérée des distances entre le centre de gravité de chaque classe et le centre de gravité du nuage :

$$\boxed{\mathcal{L}_{inter}(\Gamma) = \sum_{j=1}^{k} \mu_j d_2(G_j, G)^2}$$

L'inertie interclasse mesure la séparation entre les classes d'une partition. Plus l'inertie interclasse est grande plus les classes sont séparées. On peut noter deux cas particuliers :
Si chaque classe de la partition représente un seul individu

$$\boxed{\Gamma = (\gamma_1 = M_1, \gamma_2 = M_2, ..., \gamma_n = M_n)}$$

l'inertie interclasse de cette partition est l'inertie totale de Γ :

$$\boxed{\mathcal{L}_{inter}(\Gamma) = \mathcal{L}(\Gamma)}$$

Si la partition n'est composée que d'une seule classe

$$\boxed{\Gamma = (\gamma_1) \quad où \quad \gamma_1 = (M_1, M_2, ..., M_n)}$$

l'inertie interclasse de la partition est égale à l'inertie totale du nuage :

$$\boxed{\mathcal{L}_{intra}(\Gamma) = 0}$$

Notre critère de qualité pour assurer une partition de classes homogènes et distinctes est alors de minimiser l'inertie intraclasse des partitions et de maximiser l'inertie interclasse.

Le théorème suivant nous donne une relation entre inertie intraclasse et inertie interclasse.

Décomposition de Huygens :

$$\boxed{\mathcal{L}(\Gamma) = \mathcal{I}_{intra}(\Gamma) + \mathcal{I}_{inter}(\Gamma)}$$

Pour chaque partition, l'inertie intraclasse plus l'intertie interclasse est égale à l'inertie totale de notre nuage de points.

Cette relation permet de faire l'interprétation suivante :

Minimiser l'inertie intraclasse est équivalent à maximiser l'inertie interclasse. Notre critère de qualité est alors soit minimiser l'inertie intraclasse, soit maximiser l'inertie interclasse. On choisit ici de minimiser l'inertie intraclasse.
Dans l'approche CHA, le critère de qualité d'une partition est d'avoir ***une inertie intraclasse la plus petite possible*** (ou, de manière équivalente, ***une inertie interclasse la plus grande possible***) en fonction du nombre de classes de chaque partition. Car le but initial de la classification est d'avoir un nombre de classes restreint par rapport au nombre d'individus.

Critère de ressemblance : écart de Ward

Le but de l'algorithme CHA est de passer de la meilleure partition de k classes à la meilleure partition de k − 1 classes selon le critère de qualité (i.e. sélectionner la partition de k − 1 classes qui

a l'inertie intraclasse minimale). Le critère de ressemblance est une conséquence du critère de qualité :
La proximité de deux classes est donnée par le gain d'inertie intraclasse engendré par le regroupement de ces deux classes.

On choisit un critère de ressemblance qui satisfait cette propriété. Ce critère permet de plus d'obtenir directement le niveau d'inertie intraclasse des partitions.
Pour passer d'une partition $\Gamma(i)$ de k classes à une partition $\Gamma(i+1)$ de k−1 classes, on choisira de regrouper les deux classes qui génèrent le gain d'inertie intraclasse minimum. Ce gain d'inertie intraclasse est calculé grâce à l'écart de Ward :

Théorème : Ecart de Ward :

Soit la partition $\Gamma(i+1)$ de k−1 classes composée de la manière suivante :

- On part de la partition $\Gamma(i)$ de k classes
- On enlève les classes $(\gamma_m, \gamma_l) \in \Gamma(i)$
- On rajoute la classe $\gamma_{n+i} = (\gamma_m, \gamma_l)$

Le gain d'inertie intraclasse (ou de manière équivalente la perte d'inertie interclasse) entre la partition $\Gamma(i+1)$ et la partition $\Gamma(i)$ est noté $d(\gamma_m, \gamma_l)$:

$$\boxed{d(\gamma_m, \gamma_l) = \mathcal{I}_{intra}(\Gamma(i-1)) + \mathcal{I}_{intra}(\Gamma(i))}$$

Ce gain d'inertie intraclasse, appelé écart de Ward, se calcule grâce à la formule suivante :

$$\boxed{d(\gamma_m, \gamma_l) = \frac{\mu_m \mu_l}{\mu_m + \mu_l} d_2(G_m, G_l)^2}$$

- μ_m et μ_l sont les poids des classes γ_m et γ_l
- G_m et G_l sont les centres de gravité des classes γ_m et γ_l

L'écart de Ward peut être vu comme une distance entre deux classes si la distance entre deux individus est la distance euclidienne.

2ème étape : L'algorithme pas à pas

1. Initialisation : Chaque individu représente une classe

   $\gamma_1 = \{M1\}, \gamma_2 = \{M2\}, \ldots, \gamma_n = \{Mn\})$

   La partition initiale est alors : $\Gamma(0) = (\gamma_1, \gamma_2, \ldots, \gamma_n)$
2. Itération i :

- Calculer la matrice des distances de la partition $\Gamma(i-1)$ selon l'écart de Ward.
- Repérer l'écart de Ward minimal dans la matrice des distances et les deux classes correspondantes.

Soit $d_{m,l} = d(\gamma_m, \gamma_l)$ la distance minimale et γ_m et γ_l ces deux classes correspondantes.

– Créer une classe γ_{n+i} composée des deux classes d'écart de Ward minimal :
$\gamma_{n+i} = \{\gamma_m, \gamma_l\}$
Composer la nouvelle partition Γ(i) de la manière suivante :

- Prendre la partition Γ(i−1).
- Enlever les deux classes γ_m et γ_l.
- Ajouter la classe γ_{n+i}.
- Calculer l'inertie intraclasse de la partition Γ(i) grâce à la formule suivante :

$$\boxed{I_{intra}(\Gamma(i)) = I_{intra}(\Gamma(i-1)) + d(\gamma_m, \gamma_l)}$$

3. Stop : on arrête les itérations lorsque la partition Γ(i) n'est composée que d'une seule classe.
Cet algorithme comporte donc (n−1) itérations.
Cet algorithme permet de trouver les partitions ayant l'inertie intraclasse minimale pour un nombre de classes allant de 1 à n. Nous avons donc rempli le premier but de la CHA. Le second objectif consiste à trouver le meilleur couple (qualité de la partition ; nombre de classe de la partition) parmi toutes ces partitions. Pour cela, nous allons transcrire les résultats de l'algorithme sur un graphique que l'on appelle le dendrogramme.

3ème étape : Le dendrogramme
On peut représenter les résultats de l'algorithme CHA à l'aide d'un dendrogramme. Ce graphique se dessine de la manière suivante :

1. Initialisation : On place en abscisse tous les individus.

2. Itération i :

- On joint les deux classes sélectionnées (écart de Ward minimal) pour former un palier.
- La hauteur de ce palier à la ième itération est la valeur du ratio d'homogénéité :

$$\boxed{\frac{I_{intra}(\Gamma(i))}{I(\Omega)}}$$

On préfère exprimer l'inertie intraclasse en proportion de l'inertie totale du nuage (i.e. sous la forme de ce ratio d'homogénéité) pour plus de lisibilité.

3. Stop : on s'arrête lorsque toute les individus sont rassemblés dans une seule classe. Le niveau (en ordonnée) de ce palier est 1.
D'après l'algorithme, on connait à chaque étape les deux classes à rassembler ainsi que l'inertie intraclasse de la partition ainsi composée. Pour tracer le dendrogramme, il nous suffit de repérer l'inertie totale du nuage. Soit on l'a calculée, soit on remarque que $I(\Omega) = I_{intra}(\Gamma_{(n-1)})$, donnée que l'on a obtenue dans l'algorithme.

Le ratio d'homogénéité mesure en pourcentage :

- L'inertie intraclasse de la partition.
- Le gain d'inertie intraclasse entre la partition $\Gamma(i + 1)$ et la partition $\Gamma(i)$. Le dendrogramme permet :
- De lire toutes les étapes successives de l'algorithme.
- De repérer à l'oeil l'étape (ou les étapes) où le gain d'inertie intraclasse a été le plus grand.

4ème étape : Choix de la meilleure partition
Une fois le dendrogramme constitué entièrement, comment l'utilise-t-on ?

La problématique est de trouver la partition qui a le « meilleur » couple [inertie intraclasse, nombre de classes], contraintes opposées (à chaque fois que l'on diminue le nombre de classes, on augmente l'inertie intraclasse). Pour cela, on repère sur le dendrogramme le seuil qui nous convient selon deux critères :

- Le premier seuil où le gain d'inertie intraclasse est grand.
- S'il existe plusieurs seuils pertinents, on choisit le nombre de classes qui nous convient le mieux en fonction de la problématique.

Une fois ce seuil déterminé, on trace une coupure à ce niveau. La partition optimale ainsi choisie est composée des classes se situant sous la coupure.

Commentaires : Lors de la lecture de permettre la classification de populations très nombreuses (plusieurs bien repérer où lire le gain d'inertie intraclasse pour trouver le seuil idéal. Une astuce pour bien repérer quand le gain d'inertie intraclasse est grand est de noter sur l'axe des ordonnées le ratio d'homogénéité pour chaque partition $\Gamma(i)$.

VII.6 Méthodes de partitionnement

Les méthodes de partitionnement permettent de traiter rapidement un grand nombre de données en optimisant localement un critère tel que l'inertie intraclasse.
Ces méthodes sont itératives : à partir d'une partition initiale les individus sont affectés à la classe qui leur est la plus proche. Les difficultés dans l'implémentation de ces méthodes résident dans la manière de choisir une partition initiale et de définir la proximité entre un individu et une classe (choix du mode de représentation des classes et de la distance employée). Les méthodes de partitionnement se divisent en trois catégories :

1. **Méthode *K-Means* (MacQueen)** :
   Les centres des classes sont calculés après chaque affectation d'un point (individu) dans une classe ; les centres sont des points de l'espace ; les distances entre les individus et les classes correspondent aux distances entre les individus et les centres.
   L'algorithme des K-means s'efforce de trouver les centroïdes les plus représentatifs de nuages de points.

   Algorithme :
   - ✓ **Initialisation :** choisir les centres initiaux des classes (choix aléatoire)
   - ✓ **Affectation :** affecter chaque point du nuage au groupe dont le centre est le plus proche
   - ✓ **Mise à jour des centres :** calculer les nouveaux centres (la moyenne des nouveaux points de la classe)
   - ✓ **Test de convergence :**
     - soit le nombre d'itérations
     - soit les centroïdes sont inchangés

2. **Méthode des nuées dynamiques (Migrating Means) : ISODATA. (Diday) :**

Les centres des classes sont calculés après chaque itération et les classes sont représentées par un ensemble de ***centroïdes*** et non par des points de l'espace; les distances entre les individus et les classes correspondent aux distances entre les individus et les représentants de la classe ;

Algorithmes simplifié :

- C'est le même algorithme que le ***K-means***.
- Mais cherche à équilibrer les classes.
- Fusion de deux classes (diminution du nombre de classes) si la distance inter centre est trop faible

3. **Méthode des centres mobiles (Forgy)** :

Cette méthode s'applique lorsque l'on connaît à l'avance combien de classes on veut obtenir. Appelons ***k*** ce nombre de classes.

Hypothèses : La distance entre deux individus est la distance euclidienne.

Commentaires : L'algorithme est le même pour une autre distance. Il suffit de remplacer la distance euclidienne par cette nouvelle distance.

Algorithme

Notation : On note $\Gamma(i)$ la partition obtenue à la $i^{ème}$ itération et $C_j(i)$ le centre de la classe γ_j à la $i^{ème}$ itération.

1. **Initialisation** : On choisit au hasard ***k*** individus dans la population Ω. Ce sont les ***k*** centres initiaux **(noyaux)** : $(C_1(1), \ldots, C_k(1))$.
2. **Itération i** :
   - ✓ On regroupe les individus autours de ***k*** centres de manière à former une partition $\Gamma(i)$ de ***k*** classes γ_k. Le critère de regroupement est le suivant : chaque classe γ_j est constituée des points plus proches du centre $C_j(i)$ que des autres centres $C_m(i)$, $m \neq j$ au sens de la distance euclidienne :
     Pour tout $m \neq j$, si $d_2(M_l,C_j(i)) \leq d_2(M_l,C_m(i)) \; \forall$ alors $M_1 \in \gamma_j$.
   - ✓ On calcule les centres de gravité de chaque classe γ_j que l'on note G_j. On désigne ces points comme nouveaux centres **(noyaux)** :
     $$(C_1(i+1) = G_1, \ldots, C_k(i+1) = G_k)$$
     On peut alors effectuer une itération supplémentaire à partir de ces nouveaux centres.
3. **Stop** : Deux itérations successives génèrent la même partition : $\Gamma(i) = \Gamma(i+1)$.

Commentaires :

Dans cet algorithme, nous regardons la distance entre deux individus et non entre deux classes.

Tous les calculs se font avec la distance euclidienne. Nous n'avons pas besoin d'introduire une distance entre classes.

Le principe de l'algorithme fait que l'inertie intraclasse décroît à chaque itération. Le critère de qualité est à nouveau de minimiser l'inertie intraclasse de la partition.

Inconvénient de la méthode : Cet algorithme est rapide mais on n'est pas sûr d'obtenir la meilleure partition possible pour un nombre de classes déterminé.

Inconvénient de ces méthodes : Le découpage Final dépend de l'initialisation.

Institut de sciences exactes
Département d'Informatique
Module : Analyse de données

Fiche de TD n° 05 : Classification automatique

Exercice n° 01:

On a relevé, pour quatre individus, les valeurs de quatre variables quantitatives. Toutes ces variables sont définie en référence à une même unité de mesure. Tous les individus sont munis d'un poids unitaire
Les données sont rassemblées dans le tableau X suivant :

	VAR.1	VAR.2	VAR.3	VAR.4
A	5	10	13	4
B	16	7	6	1
C	8	15	14	3
D	9	2	11	12

Les individus sont (notés A, B, C et D)

Réaliser une classification ascendante hiérarchique indicée (CAHI) de cet ensemble d'individus en utilisant le critère d'agrégation de Ward (maximisation de l'inertie inter classe de la partition construite à chaque étape)

Exercice n° 02 :

Soit un ensemble de six individus {1, 2, 3, 4, 5, 6} sur lesquels nous avons effectués une classification hiérarchique ascendante avec le critère de la minimisation de l'inertie de la réunion de deux classes.
Les résultats obtenus (perte d'inerties inter classes) sont les suivants : $a,b \in \Re$.
$V(\{i\}) = 0$ pour i=1 à 6 ; $V(\{1, 2\}) = a + 2b$; $V(\{3, 4\}) = 2a$; $V(\{1, 2, 5\}) = a – b$; $V(\{1, 2, 5, 6\}) = 2a – b$; $V(\{1, 2, 3, 4, 5, 6\}) = 18$.

1. Représenter le dendrogramme correspondant
2. Que doivent vérifier les paramètres a et b ?
3. Quelle est la valeur de l'inertie totale ?

Institut de sciences exactes
Département d'Informatique
Module : Analyse de données

Fiche de TD n° 05 : Classification automatique

Exercice n° 01:

On a relevé, pour quatre individus, les valeurs de quatre variables quantitatives. Toutes ces variables sont définie en référence à une même unité de mesure. Tous les individus sont munis d'un poids unitaire

Les données sont rassemblées dans le tableau X suivant :

	VAR.1	VAR.2	VAR.3	VAR.4
A	5	10	13	4
B	16	7	6	1
C	8	15	14	3
D	9	2	11	12

Les individus sont (notés A, B, C et D)

Réaliser une classification ascendante hiérarchique indicée (CAHI) de cet ensemble d'individus en utilisant le critère d'agrégation de Ward (maximisation de l'inertie inter classe de la partition construite à chaque étape)

Exercice n° 02 :

Soit un ensemble de six individus {1, 2, 3, 4, 5, 6} sur lesquels nous avons effectués une classification hiérarchique ascendante avec le critère de la minimisation de l'inertie de la réunion de deux classes.

Les résultats obtenus (perte d'inerties inter classes) sont les suivants : $a,b \in \Re$.

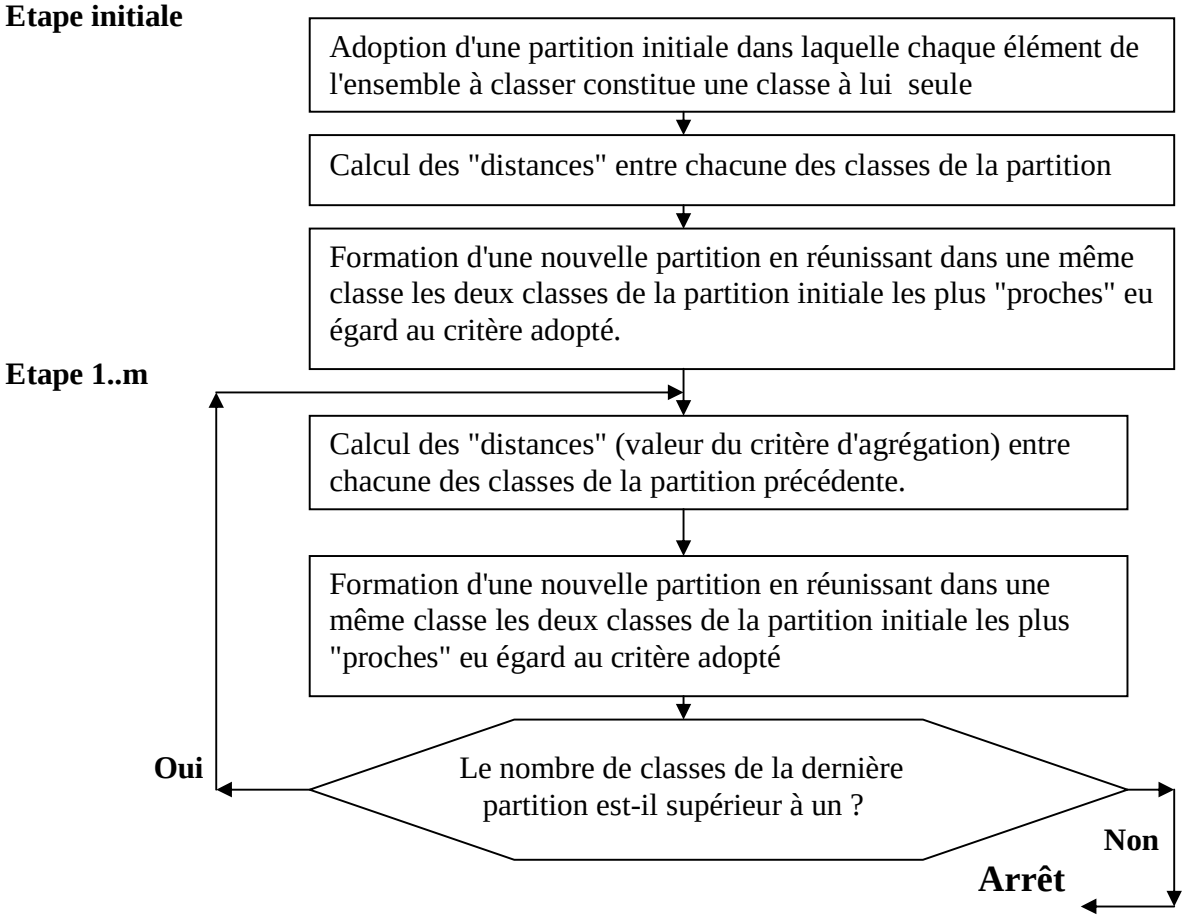
$V(\{i\}) = 0$ pour i=1 à 6 ; $V(\{1, 2\}) = a + 2b$; $V(\{3, 4\}) = 2a$; $V(\{1, 2, 5\}) = a - b$; $V(\{1, 2, 5, 6\}) = 2a - b$; $V(\{1, 2, 3, 4, 5, 6\}) = 18$.

1. Représenter le dendrogramme correspondant
2. Que doivent vérifier les paramètres a et b ?
3. Quelle est la valeur de l'inertie totale ?

Solution d'exercice n° 01:

Voici une l'organigramme equivalent à l'algorithme de la classification de CAHI :

- Options préalables :
- choix d'un indice de distance entre éléments de base de l'ensemble à classer
- choix d'un critère d'agrégation (indice de "distance" entre les classes d'éléments)
- Procédure :



A chacun des éléments de l'ensemble à classer on associe un point de l'espace R^4 dont les coordonnées correspondent aux valeurs des quatre variables.

A chaque point on associe en plus le poids attaché à l'élément concerné.

Exemple : au premier individu correspond un point A de coordonnées (5,10,13,4), munis d'un pods m_A égal à 1.

Notons : n : le nombre d'individus actifs (ici n=4),

i : indice désignant les individus

m_i : le poids associé à l'individus i (ici m_i=1 quel que soit i)

p : le nombre de variable actives (ici p=4)

j : indice désignant les variables

G_k : centre de gravité de la classe k

l_k : ensemble des individus appartenant à la classe k

on définit également un indice de distance entre les points correspondant aux éléments de base de l'ensemble à classer.

Les données prenant ici la forme d'un tableau de mesures on utilise la distance euclidienne usuelle.

$$d^2(i,i') = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Sur cette base on détermine les distances entre chacun des points du nuage de Rp associé à l'ensemble à classer.
Ces distances sont calculées ici sans "réduction" préalable des données puisque toutes les mesures sont exprimées au travers de la même unité.

On obtient le tableau suivant :

	A	B	C	D
A	0	188	36	148
B	188	0	196	220
C	36	196	0	260
D	148	220	260	0

Tableau des distances entre les éléments de base

Exemple de calcul : $d^2(A,B)=(5-16)^2+(10-7)^2+(13-6)^2+(4-1)^2=188$

Procédure de classification :

Etape initiale :

On commence par définir une partition initiale dans laquelle chaque point forme à lui seul une classe. Pour simplifier la présentation on désigne ici chacune de ces classes par la lettre identifiant le point qui la constitue. La partition de départ est composée des **quatre classes** : A, B, C et D.
Dans le cadre de cette partition initiale **le centre de gravité de chaque classe** est confondu avec l'unique point qui la compose. La distance entre les centres de gravités de deux classes quelconques est alors égale à la distance entre les deux points qui définissent ces classes.
Exemple : $d^2(G_A,G_B)=d^2(A,B)=188$
La phase suivante consiste à construire une nouvelle partition en fusionnant deux classes de la partition initiale. On associe à la nouvelle classe définie par cete fusion un poids égal à la somme des poids des classes qu'elle réunit.
Lorsqu'on utilise le critère d'agrégation de Ward, les deux classes agrégées (réunies) sont celles dont la réunion provoque la perte d'inertie inter classe la plus faible. En procédant ainsi on retient, parmi toutes les partitions qui peuvent être créées en fusionnant deux classes de la partition initiale, celle qui est doté de la plus grande inertie inter classe. On sait qu'une partition n'est peinement satisfaisante que lorsque qu'elle possède une inertie inter classe élevée, les classes sont alors fortement différenciées et, à l'intérieur de chacune d'elles, les éléments rassemblés tendent à être très homogènes.
La perte d'inertie inter classe P que provoque la réunion de deux classes quelconques k et k' est définie par la relation suivante :

$$P = \frac{m_k m_{k'}}{m_k + m_{k'}} d^2(G_k, G_{k'}) \qquad [I] \qquad \text{(critère d'agrégation)}$$

Dans laquelle on note : G_k et $G_{k'}$ le centre de gravité de chacun des deux classes considérées.
m_k et $m_{k'}$ désignant le poids qui leur est assicié.
En application des règles qui viennent d'être indiquées, à ce stade de l'algorithme, on utilise la relation [I] pour calculer la perte d'inertie inter classe qu'engendrerait chacun des regroupements possibles de deux classes de la partition initiale.
Le résultat des calculs est produit dans le tableau suivant :

	A	B	C	D
A	0	94	18	74
B	94	0	98	110
C	18	98	0	130
D	74	110	130	0

Tableau de perte d'inertie inter classe

Exemple de calcul : $\frac{m_A m_B}{m_A + m_B} d^2(G_A, G_B) = \frac{(1x1)}{(1+1)} x188 = 94$

L'étude de ce tableau permet de déterminer quelles sont les deux classes qu'il faut réunir pour minimiser la valeur du critère d'agrégation.
On voit ici que c'est en réunissant les classes A et C que l'on minimise la perte d'inertie inter classe.
On agrége donc ces deux classes A et C pour former une nouvelle classe que l'on nomme E. le poids asocié à cette nouvelle classe est $m_E=m_A+m_C=2$
Au terme de cette étape initiale l'ensemble à classer est patitionné en trois classes : B, D et E

	B	D	E
Poids	1	1	2

Tableau des poids

Première étape :

- **calcul du tableau de distances entre les centres de gravités des classes :**

avec la formule suivante :

$$d^2(G_r, G_l) \frac{1}{m_k + m_{k'}} \left[m_k d^2(G_k, G_l) + m_{k'} d^2(G_{k'}, G_l) - \frac{m_k m_{k'}}{m_k + m_{k'}} d^2(G_k, G_{k'}) \right]$$

	B	D	E
B	0	220	183
D	220	0	195
E	183	195	0

Exemple de calcul :

$$d^2(G_E, G_B) = \frac{1}{m_A + m_C} \left[m_A d^2(G_A, G_B) + m_C d^2(G_C, G_B) - \frac{m_A m_C}{m_A + m_C} d^2(G_A, G_C) \right]$$

$$= \frac{1}{1+1} \left[1x188 + 1x196 - \frac{1x1}{1+1} x36 \right] = 183$$

- **tableau de perte d'inertie inter classe**

	B	D	E
B	0	110	122
D	110	0	130
E	122	130	0

Exemple de calcul :

$$\frac{m_B m_D}{m_B + m_D} d^2(G_B, G_D) = \frac{(1x1)}{(1+1)} x220 = 110$$

- **Tableau des poids associés aux classes**

Réunion des classes B et D : on agrégé ces deux classes pour former une nouvelle classe que l'on nomme **F**

De poids $m_F=m_B+m_D=2$

Donc l'ensemble à classer est partitionné en deux classes E et F de poids

	E	F
Poids	2	2

Seconde étape :

- distance entre E et F :

$$d^2(G_F,G_E)=\frac{1}{m_B+m_D}\left[m_B d^2(G_B,G_E)+m_D d^2(G_D,G_E)-\frac{m_B m_D}{m_B+m_D}d^2(G_B,G_D)\right]$$

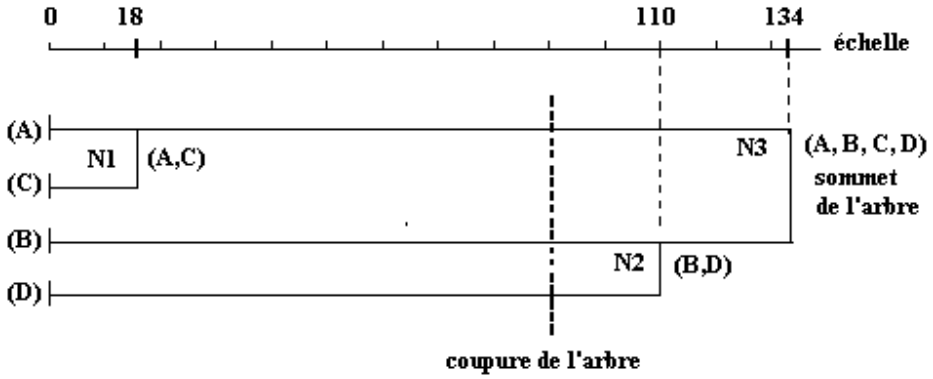
$$=\frac{1}{(1+1)}\left[(1x183)+(1x195)-\frac{1x1}{(1+1)}220\right]=134$$

- perte d'inertie inter classe engendré par la fusion des classes E et F :

$$\frac{m_F m_E}{m_F+m_E}d^2(G_F,G_E)=\frac{(2x2)}{(2+2)}x134=134$$

la partition obtenue à l'issue de cette ultime étape ne comporte d'une classe qui correspond à l'ensemble à classer (A, B, C, D)

Elaboration de l'arbre :

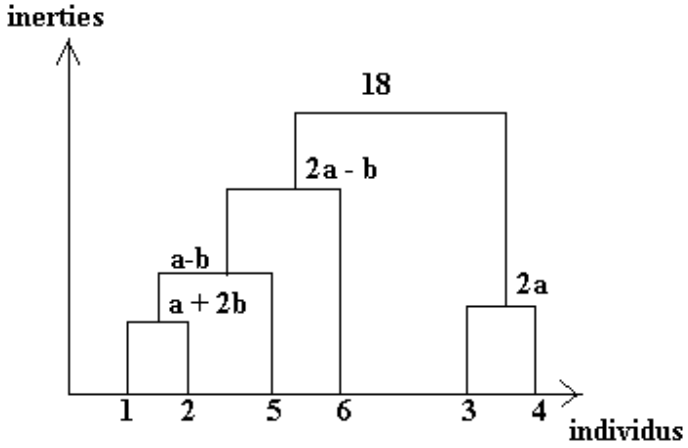


Solution exercice n° 02 :

Les pertes d'inerties inter-classes à la reunion des classes sont :

V({i}) = 0 pour i=1 à 6 ; V({1, 2}) = a + 2b ; V({3, 4}) = 2a ; V({1, 2, 5}) = a – b ; V({1, 2, 5, 6}) = 2a – b ; V({1, 2, 3, 4, 5, 6}) = 18.

1. Le dendogramme est le suivant :



2. Représentation des paramètres a et b : il faut avoir ces inégalités :
$0 < a + 2b < a - b < 2a - b < 18$ et $2a < 18$

$a + 2b < a - b \rightarrow 3b < 0 \rightarrow b < 0$
$a + 2b > 0 \rightarrow b > -a/2$
$2a > 0 \rightarrow a > 0$
$2a - b < 18 \rightarrow a < 9 + b/2$

$\rightarrow$

$a \in]\, 0\, ,\ 9+b/2\, [$

$b \in]\, -a/2\, ,\ 0\, [$

3. La valeur de l'inertie totale : est la somme de touts les inerties :

$\sum V_i = a + 2b + 2a + a - b + 2a - b + 18 = 6a + 18$