ESIAI
ECOLE SUPÉRIEURE D'INGÉNIERIE
APPLIQUÉE ET INNOVATION

ANALYSE DES DONNEES MULTIVARIEES

Par R. El Mehdi

Support de cours
2024/2025

Introduction

L'analyse des données est une technique relativement récente, qui s'est constituée dans la décennie 1960-1970. Elle permet de décrire plus sûrement de grands gisements de données, et constitue un outil précieux pour le chercheur qui veut extraire le maximum de résultats des données qu'il a collectées.

Nombreuses sont les disciplines dans divers domaines qui font appel à des outils statistiques pour traiter des centaines et des milliers de données, mais dans un univers aléatoire, il n'est absolument pas prouvé qu'on puisse connaître avec certitude les lois et les distributions auxquelles obéissent les phénomènes observés. Il est donc indispensable d'utiliser les méthodes de l'analyse des données car elles s'appliquent à des faits bruts, et le recours à des hypothèses probabilistes contestables est pratiquement absent de l'analyse des données.

Cette technique est une branche de la statistique descriptive perfectionnée. Son propre est de raisonner sur un nombre quelconque de données concernant un nombre quelconque de variables, d'où le nom d'analyse multivariée qu'on lui donne souvent. Pour effectuer ce raisonnement, l'analyse des données a fait appel aux espaces mathématiques comportant un nombre quelconque de dimensions et aux outils informatiques.

Liée à l'informatique, l'analyse multidimensionnelle n'a pu être développée qu'après la relance de l'informatique, car elle nécessite la réalisation des calculs matriciels infaisables en l'absence de l'ordinateur. Ces calculs automatiques ont permis le développement des deux grands groupes de l'analyse des données, qui sont les méthodes d'analyse factorielle et les méthodes de classification automatique.

L'analyse factorielle porte sur des nuages de points dont on cherche à trouver les directions d'allongement maximal. Elle traite des tableaux de nombres et remplace un tableau difficile à lire par un tableau plus simple à lire qui soit une bonne approximation de celui-ci. Chaque méthode correspond à un procédé particulier pour construire le nuage et mesurer son allongement. Parmi les méthodes d'analyse factorielle on cite, l'Analyse en Composantes Principales (ACP), l'Analyse Factorielle des Correspondances (AFC), l'Analyse des Correspondances Multiples (ACM), l'Analyse Canonique (AC), ...

La classification automatique porte sur des ensembles d'individus qu'il faut regrouper en catégories jugées homogènes au regard de tel ou tel critère. La nature des variables observées et le calcul de l'homogénéité des catégories varient d'une méthode à l'autre. Parmi les méthodes de classification on trouve les méthodes ascendantes et les méthodes descendantes. L'usage des méthodes ascendantes est plus fréquent, car les méthodes descendantes manquent de précision.

Le principe de la technique de classification ascendante est de construire à partir des éléments de l'ensemble I une suite finie des partitions emboîtées. Au niveau le plus bas de cette hiérarchie sont placées les classes à un élément, appelées classes terminales ou minimales. Les autres classes sont appelées noeuds de la hiérarchie, et l'ensemble I

constitue le noeud le plus haut. On note ici qu'un noeud est une réunion de deux classes qui se trouve au-dessous de lui.

Chapitre1 **Statistique descriptive élémentaire**

1. Rappel sur l'analyse descriptive simple

1.1. Mesures de tendance centrale et de dispersion

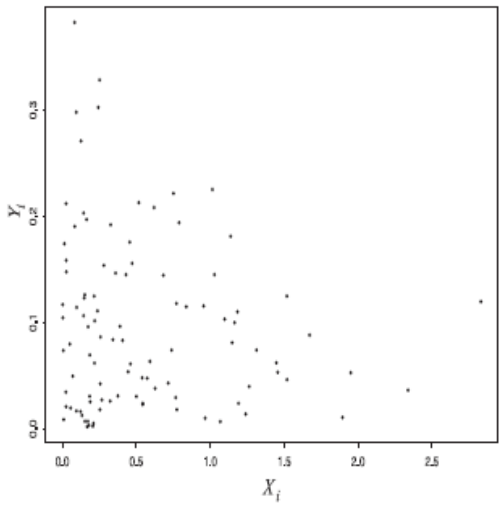
(Voir le support du cours Proba / Stat – CP2).

1.2. Graphisme

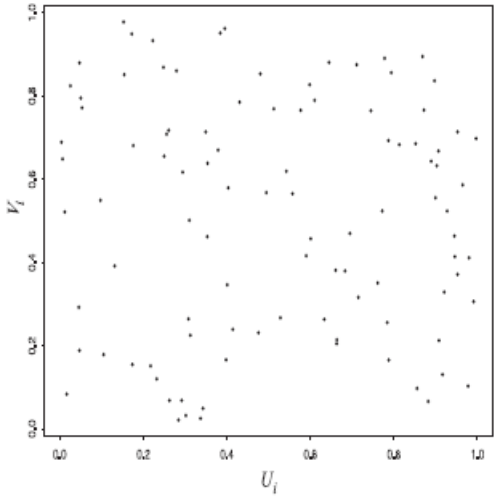
Le graphique est un élément clef pour communiquer des résultats d'une analyse statistique simple ou multivariée. La plupart des observations que l'on peut faire sur des séries de données peuvent en général être illustrées sur la base des graphiques et les utilisateurs de la statistique sont de plus en plus demandeurs de cet outil. C'est un outil souvent simple à lire et à interpréter surtout s'il est représenté dans un espace de dimension 2 ou 3. Parmi les graphes usuels on cite :

- **Graphe X-Y (Scatter plot)**

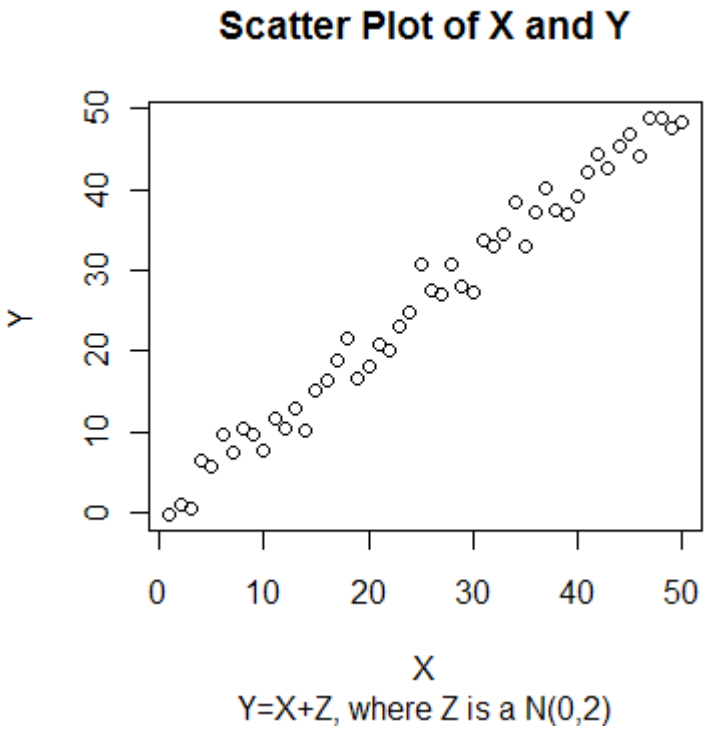
  Le scatter plot est une méthode standard pour visualiser les données. Il représente le nuage de points (x_i, y_i) pour tout $i = 1,...,n$ et il permet entre autre de détecter une probable relation entre deux variables si le nuage a une tendance particulière.



Are the variables X and Y independent?



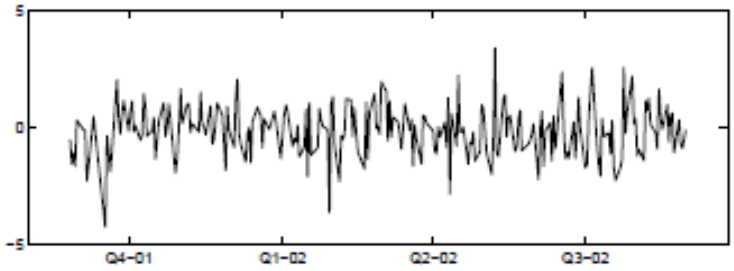
Are $U = F(X)$ and $V = G(Y)$ independent?



• **Graphique temporel**

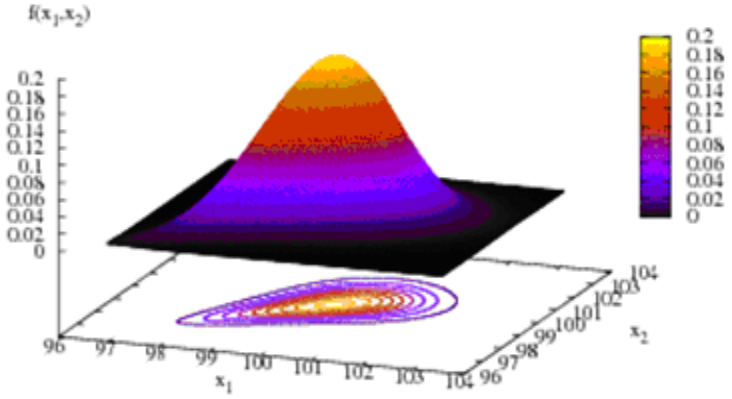
Le graphique temporel est une représentation graphique de l'évolution d'une série dans le temps. Il est parfois appelé le chronogramme.

Le codage d'un signal électrique dans le temps, à titre d'exemple, ou de l'effet d'un appareil sur la santé peut être représenté par la série suivante :



• **Surfaces de réponse**

La surface de réponse est une courbe représentée dans un espace de dimension 3 (3D). Les deux axes du plan (x_1, x_2) représentent les variables et le dernier axe représente la densité $f(x_1, x_2)$. Ce dernier axe dresse en couleurs les niveaux de la fonction pour faciliter la lecture du graphique. La projection de la courbe de f sur le plan donne un graphique appelé Contour.



• **Graphe d'autocorrélation**

Autocorrélation

L'autocorrélation entre deux variables X_i et X_{i-k} mesure la dépendance d'une variable et son passé. L'intensité de la dépendance dans ce cas est définie par le coefficient d'autocorrélation d'ordre k

$$\rho_k = \rho(X_i, X_{i-k}) = \frac{Cov(X_i, X_{i-k})}{\sqrt{V(X_i).V(X_{i-k})}}$$

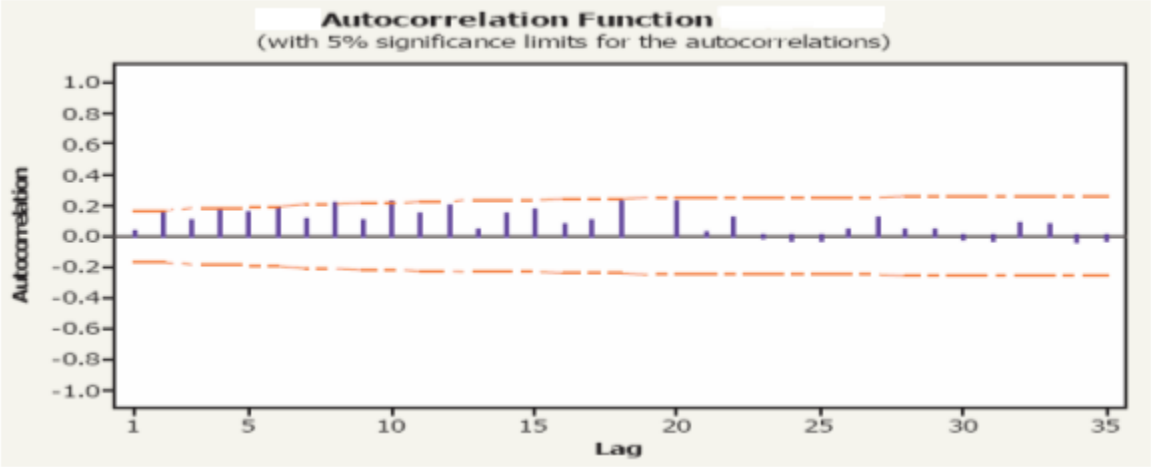
Il est estimé par $$r_k = \frac{\sum_{i=k+1}^{N}(X_i - \bar{X})(X_{i-k} - \bar{X})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} .$$

A titre indicatif, si les données sont décrites par un modèle Autorégressif d'ordre 1 :

$$X_i = \mu + \phi(X_{i-1} - \mu) + \varepsilon_i \quad , \quad \varepsilon_i \sim iN(0, \sigma_\varepsilon^2) \quad \text{et} \quad -1 < \phi < 1 \quad .$$

on a $$\rho_k = \frac{\text{cov}(X_i, X_{i-k})}{\sqrt{V(X_i).V(X_{i-k})}} \quad \Rightarrow \quad r_k = \phi^k .$$

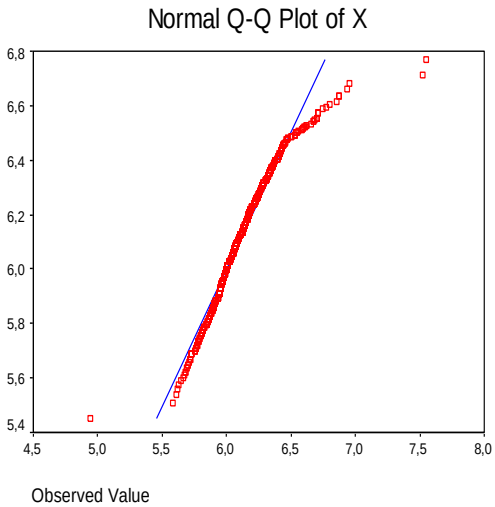
L'autocorrélogramme est un graphique sur lequel sont présentées les r_1, r_2, r_3, … sous forme de bâtonnets. Sont présentées également sur le graphique la ligne $y = 0$ et les deux bornes de l'intervalle de confiance des autocorrélations placé souvent à $\pm \frac{2}{\sqrt{N}}$. La variable est autoccorélée s'il existe des bâtonnets d'autocorrélation qui sortent de l'intervalle, par conséquent l'indépendance n'est pas remplie.

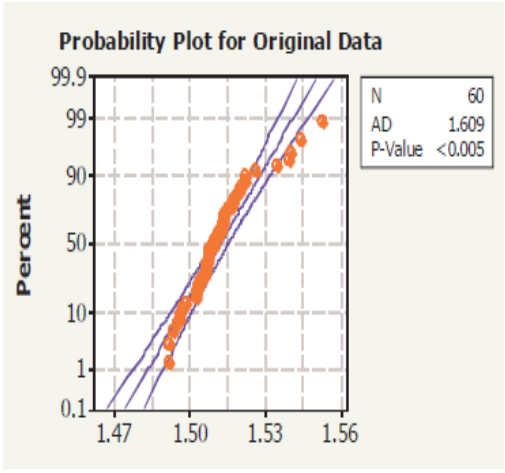


L'indépendance est assurée pour les données du graphique ci-dessus car aucune pique ne sort de l'intervalle.

• **QQPlot** (Quantile-Quantile Plot)

Si la variable X pour laquelle on teste la normalité est gaussienne, les points de coordonnées $\left(x_{(i)}, x^*_{(i)}\right)$ sont alignés sur la droite d'équation $x^*_{(i)} = \sigma.z^*_{(i)} + \bar{x}$ appelée la droite d'Henri, où $z^*_{(i)}$ sont les quantiles d'ordre $F_i = \dfrac{i-0.375}{n+0.25}$ calculés en utilisant la loi normale centrée réduite. On compare donc les valeurs des quantiles de la loi empirique $\left(x_i\right)$ au quantiles de la loi normale centrée réduite $\left(x^*_i\right)$. Cette méthode peut également se généraliser à d'autres distributions en comparant là encore les quantiles théoriques aux quantiles empiriques.





• **Histogramme**

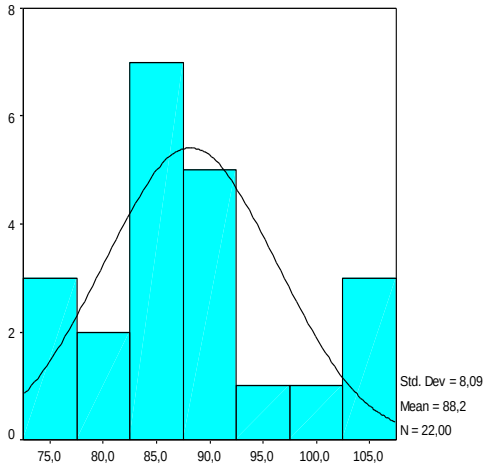
L'histogramme est un graphique qui permet de visualiser la distribution de la variable quantitative. A la différence du diagramme en barre, l'histogramme est constitué d'un certain ensemble de classes $[a_i, a_{i+1}[$ d'amplitudes égales, à chaque classe on associe un effectif n_i. $([a_i, a_{i+1}[, n_i)$ sur les axes des abscisses et des ordonnées respectivement sont les rectangles de l'histogramme.

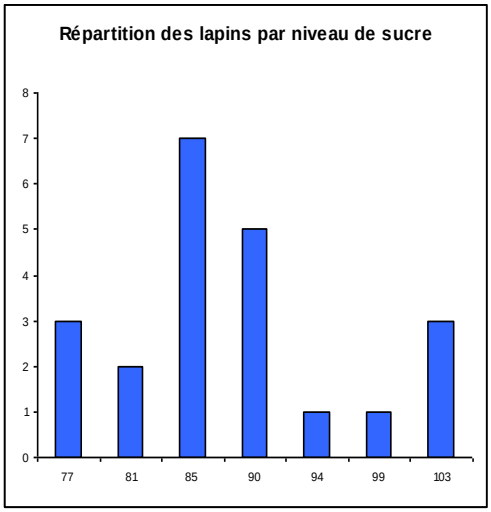
Si les amplitudes des classes $(a_{i+1} - a_i)$, $\forall i$ ne sont pas égales, la largeur du rectangle restera $(a_{i+1} - a_i)$ et la hauteur devient $\dfrac{n_i}{a_{i+1} - a_i}$.

Le nombre de classes K : Il n'y a pas une méthode standard pour calculer le nombre de classes dans un histogramme, mais généralement on utilise :

$$K = 1 + \frac{10 \log(n)}{3} \quad \text{ou} \quad K = \sqrt{n}$$

Il est souvent préférable de faire varier le nombre de classes afin de voir la meilleure façon de représenter l'histogramme de la variable et d'avoir une vision clair sur sa distribution. Cependant, le recours à des logiciels dédiés à cette fin facilitera la tâche de la formation des classes.





• **Box plot**

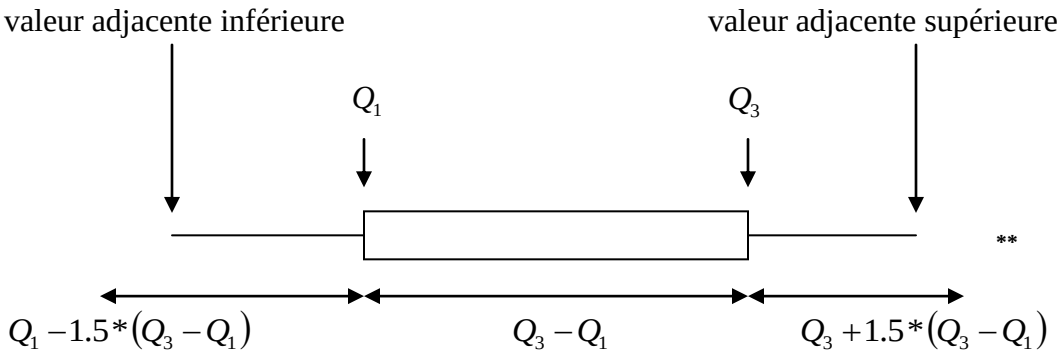
Le box plot est un graphe sous forme de boîtes réalisé par Tukey (1977). Il représente la valeur adjacente inférieure, les trois quartiles (Q_1, $Q_2 = Me$, Q_3), la valeur adjacente supérieure et les outliers (les valeurs aberrantes) des données de la variable.

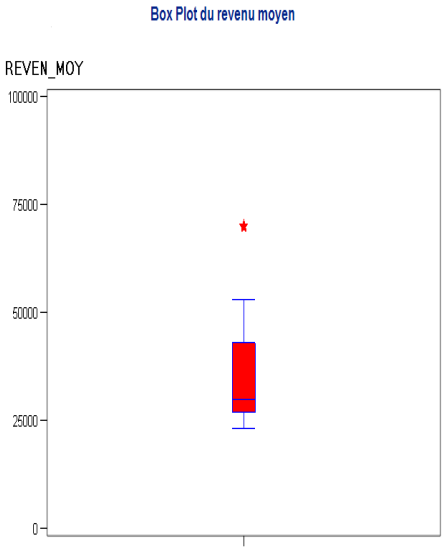
Les quartiles Q_1, Q_2 et Q_3 répartissent les données en quatre parties égales.

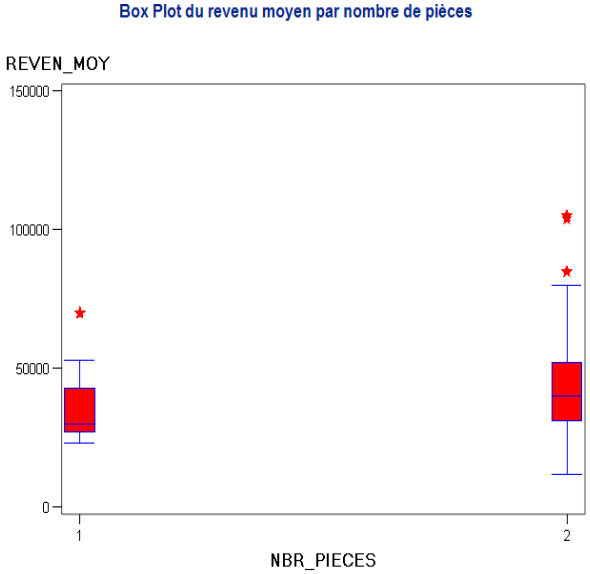
La valeur adjacente inférieure est la valeur minimum dans les données qui est supérieure à la valeur frontière basse $Q_1 - 1.5*(Q_3 - Q_1)$

La valeur adjacente supérieure est la valeur maximum dans les données qui est inférieure à la valeur frontière haute $Q_3 + 1.5*(Q_3 - Q_1)$

Les outliers ou les observations aberrantes sont des observations qui paraissent étrangères aux valeurs de la variable. Leur détection ne sera pas traitée dans ce cours.

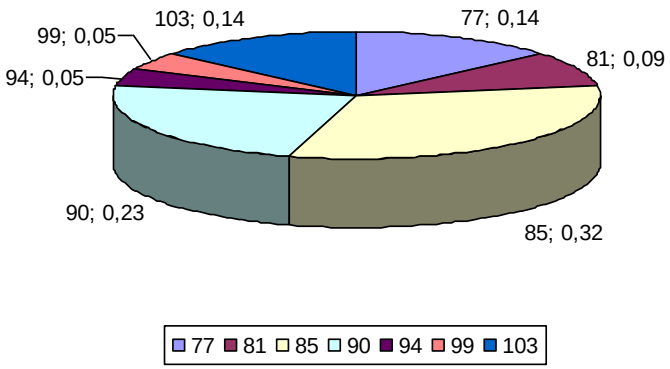






• **Diagramme**

C'est un graphe sous forme d'un cercle réparti en segments (même représentation pour les effectifs absolus et relatifs).



2. Notations et notions importantes

2.1. Moyenne et variance d'un vecteur

Dans le cadre de la statistique univariée (une seule variable à analyser), il est souvent utile de résumer les informations de la variable dans certaines grandeurs statistiques. Les plus utilisées sont la moyenne et la variance.

Une variable X est une série d'observations $(x_1, x_2, ..., x_i, ..., x_n)$, elle peut être exprimée par un vecteur (Colonne ou Ligne)

$$X_{(1,n)} = (x_1, x_2, ..., x_i, ..., x_n)^t$$

La moyenne d'un vecteur X, qui est une mesure de tendance centrale notée $\bar{x}$, et la variance de X qui est une mesure de dispersion notée σ^2 sont définies respectivement par

$$E(X) = \bar{x} = \frac{1}{n}\sum_{i=1}^{I} n_i x_i \ ,$$

$$Var(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{I} n_i (x_i - \bar{x})^2 \ .$$

où n_i est l'effectif de x_i et I est le nombre de modalités. La variance empirique S^2 est définie par $S^2 = \frac{n}{n-1}\sigma^2$. L'écart-type, quant à lui, est $\sigma = +\sqrt{Var(X)}$.

2.2. Vecteur de Moyennes

Dans le cas de l'analyse multivariée (p vecteurs), les données sont décrites par une matrice X de format (n, p) qui est en réalité un tableau à double entrée.

$$X_{(n,p)} = \{x_{ij}\} = \begin{array}{c} \begin{matrix} X_1 & \cdots & X_j & \cdots & X_p \end{matrix} \\ \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & S_{nj} & \cdots & x_{np} \end{pmatrix} \end{array}$$

Un vecteur de moyennes, noté $\overline{X}_{(p,1)}$, est un vecteur pour lequel chaque composante est la moyenne $\bar{x}$ de la variable correspondante.

$$\overline{X}_{(p,1)} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_j \\ \vdots \\ \bar{x}_p \end{pmatrix},$$

où $$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\ .$$

$\overline{X}_{(p,1)}$ peut être exprimée par $\overline{X} = \frac{1}{n} X^t I_n$ où $I_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ de format $(n,1)$ et X^t est de format (p,n).

2.3. Matrice de Covariance

Généralement, la Covariance, notée S_{ij}, mesure l'association entre deux variables X_i et X_j de dimension n. La matrice de Covariance, appelée aussi la matrice de Variance-Covariance, est une matrice symétrique qui englobe l'information de la covariance pour tout $i = 1,..., p$ et $j = 1,..., p$. D'où

$$S = \begin{pmatrix} S_{11} & \cdots & S_{1j} & \cdots & S_{1p} \\ \vdots & & \vdots & & \vdots \\ S_{i1} & \cdots & S_{ij} & \cdots & S_{ip} \\ \vdots & & \vdots & & \vdots \\ S_{p1} & \cdots & S_{pj} & \cdots & S_{pp} \end{pmatrix}$$

et $$Cov(X_i, X_j) = S_{ij} = \frac{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{n}$$

Il est à noter que:

i) Les éléments de la diagonale de la matrice de Covariance sont les variances des i.

Par conséquent $S_{ii} = S_i^2 = \dfrac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2}{n}$

ii) S est symétrique, alors $S_{ij} = S_{ji}$ et $S = S^t$

2.4. Matrice de corrélation

La corrélation r_{ij} entre deux vecteurs x_i et x_j est une covariance standardisée (étant donné que la covariance est sensible au choix de l'unité de mesure). La matrice de corrélation, notée $R_{(p,p)}$ est constituée des paires de corrélation r_{ij}, $i = 1,..., p$ et $j = 1,..., p$.

$$R_{(p,p)} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

où $$r_{ij} = \frac{S_{ij}}{S_i . S_j} = \frac{S_{ij}}{\sqrt{S_{ii} . S_{jj}}}$$

La matrice R peut être obtenue par simple transformation de la matrice de covariance S.

$$R = D_s^{-1/2} . S . D_s^{-1/2}$$

où $$D_{s_{(p,p)}} = diag\left(S_1^2,..., S_j^2,..., S_p^2\right) = \begin{pmatrix} S_1^2 & & & & \\ & \ddots & & 0 & \\ & & S_j^2 & & \\ & 0 & & \ddots & \\ & & & & S_p^2 \end{pmatrix}$$

et $$D_s^{-1/2} = \begin{pmatrix} 1/S_1 & & & & \\ & \ddots & & 0 & \\ & & 1/S_j & & \\ & 0 & & \ddots & \\ & & & & 1/S_p \end{pmatrix}$$

Chapitre2 **L'Analyse en Composantes Principales (A.C.P.)**

1. Généralités

L'A.C.P. est une méthode mathématique d'analyse des données qui consiste à rechercher les directions de l'espace qui représentent le mieux les corrélations entre p variables aléatoires. Elle vise à représenter graphiquement les relations entre des variables quantitatives (ou assimilées à des variables quantitatives) et également leurs relations avec les individus décrits par ces variables. A partir du graphique présenté par l'A.C.P., on analyse les axes factoriels, le plan, la proximité entre les variables et (ou) les individus.

Lorsqu'on a un espace de dimension 100 (le nombre de vecteurs ou de variables indépendantes), il y aura 100 axes à déterminer qui expliquent le mieux la dispersion du nuage des points. L'A.C.P. va les ordonner par 'l'inertie expliquée' et les axes expliquant les plus grandes inertie sont les axes qui sont interprétés. Si on décide de ne retenir que les deux premiers axes de l'A.C.P. selon le critère de l'inertie, on pourra alors projeter et visualiser le nuage de dimension 100 sur un plan.

L'A.C.P. est généralement utilisée pour visualiser des données, mais elle est encore un moyen de décorréler et de débruiter (supprimer le bruit) les données. En effet, ces dernières sont décorrélées car dans la nouvelle base, constituée des nouveaux axes, les points ont une corrélation nulle ; elles sont débuiter car les axes que l'on décide d'ignorés sont considérés comme des axes bruités (sont un bruit).

On utilise l'A.C.P. quand on est face à un tableau ou une table de données quantitatives, les lignes et les colonnes représentent respectivement les individus et les variables. De ce fait, l'A.C.P. touche, alors, plusieurs domaines socio-économiques tels le commerce, l'industrie, l'agriculture, les services, la santé, la finance, …

Soit T la matrice décrivant la base de données quantitatives,

$$T_{(n\ ,\ p)} = \begin{array}{c} \\ I_1 \\ I_2 \\ \vdots \\ I_i \\ \vdots \\ \\ I_n \\ \downarrow \\ I \end{array} \begin{array}{c} \begin{array}{cccccccc} V_1 & V_2 & \ldots & V_j & \ldots & & V_p & \longrightarrow\ J \end{array} \\ \begin{pmatrix} & & & \cdot & & & & \\ & & & \cdot & & & & \\ & & & \cdot & & & & \\ \cdot & \cdot & \cdot & t_{ij} & \cdot & \cdot & \cdot & \\ & & & \cdot & & & & \\ & & & \cdot & & & & \\ & & & \cdot & & & & \end{pmatrix} \end{array}$$

$$T_{(n\ ,\ p)} = \left\{ \quad t_{ij} \,/\, i \in I \quad et \quad j \in J \quad \right\}$$

$$\left(T^i\right)^t\ ,\ \left(T^1\right)^t = I_1 \longrightarrow \left(t_{11},\ t_{12},\ \ldots,\ t_{1j},\ \ldots,\ t_{1p}\right) \in IR^p\ ;$$

$$\left(T^j\right),\ \left(T^1\right) = V_1 \longrightarrow \left(t_{11},\ t_{21},\ \ldots,\ t_{i1},\ \ldots,\ t_{n1}\right)^t \in IR^n\ ;$$

donc, les individus sont représentés dans l'espace IR^p et les variables dans l'espace IR^n

L'ensemble des points représentant les individus est représenté par le Nuage I, noté $N(I)$, et l'ensemble des points représentant les variables est représenté par le Nuage J, noté $N(J)$. En général, le nombre d'individus est supérieur au nombre de variables $(n > p)$.

Dans une première étape et avant d'appliquer l'A.C.P., il faut s'assurer de :

- Les variables sont homogènes (même unité de mesure ou moyennes et (ou) écart types très proches). Dans ce cas, on applique l'A.C.P. sur le tableau initial qui est la matrice T.
- Les variables sont hétérogènes quant à leurs moyennes et leurs dispersions. Il faut centrer et réduire les données, ceci veut dire qu'on obtient des données de moyennes nulles ($\bar{t}_j = \mathbf{O}$) et de variances égale à l'unité ($S_j^{\,2} = 1$).

On analyse donc le tableau T' noté pour simplification T également tel que

$$T'_{(n\ ,\ p)} = \left\{ \quad t'_{ij} = \frac{t_{ij} - \bar{t}_j}{S_j \sqrt{n}}, \quad i \in I \quad et \quad j \in J \quad \right\},$$

noté

$$T_{(n\ ,\ p)} = \left\{ \quad t_{ij} = \frac{t_{ij} - \bar{t}_j}{S_j \sqrt{n}}, \quad i \in I \quad et \quad j \in J \quad \right\}.$$

$$E(t_{ij}) = E\left(\frac{t_{ij} - \bar{t}_j}{S_j\sqrt{n}}\right) = \frac{1}{S_j\sqrt{n}} E(t_{ij} - \bar{t}_j) = \frac{1}{S_j\sqrt{n}} (E(t_{ij}) - \bar{t}_j) = 0, \quad \forall j \in J$$

$$V(t_{ij}) = V\left(\frac{t_{ij} - \bar{t}_j}{S_j\sqrt{n}}\right) = \frac{1}{n.S_j^2} V(t_{ij}) = \frac{n.S_j^2}{n.S_j^2} = 1, \quad \forall j \in J$$

On divise par $S_j\sqrt{n}$ et non pas par S_j pour des besoins de calcul. Ceci ne change rien dans les positions relatives de individus et (ou) des variables. Donc, l'écart-type de T, étant une matrice centrée réduite, n'est plus 1 mais $1/\sqrt{n}$.
L'implication géométrique de la standardisation se résume dans ce qui suit :

- Les nouvelles données ne dépendent plus des unités de mesures.
- Le produit scalaire entre deux variables j et k, notées T^j et T^k, est donné par

$$(T^j)^t \; T^k = \sum_{i=1}^{n}\left(\frac{t_{ij} - \bar{t}_j}{S_j\sqrt{n}}\right).\left(\frac{t_{ik} - \bar{t}_k}{S_k\sqrt{n}}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t_{ij} - \bar{t}_j}{S_j}\right).\left(\frac{t_{ik} - \bar{t}_k}{S_k}\right) = \frac{S_{jk}}{S_j.S_k} = r_{jk}$$

  Ce produit scalaire n'est autre que la corrélation entre les deux variables T^j et T^k qui est égale à la corrélation entre T^k et T^j.

- Si on remplace T^k par T^j dans ce produit scalaire on trouve que le carré de la norme de T^j est égale à 1.

$$\left\|T^j\right\|^2 = (T^j)^t.T^j = 1$$

$$\left\|T^j\right\|^2 = (T^j)^t.T^j = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t_{ij} - \bar{t}_j}{S_j}\right).\left(\frac{t_{ij} - \bar{t}_j}{S_j}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t_{ij} - \bar{t}_j}{S_j}\right)^2 = \frac{S_{jj}}{S_j.S_j} = \frac{S_j^2}{S_j.S_j} = 1$$

En conséquence de la standardisation, dans un plan $x \times y$, les variables sont représentés dans un cercle de centre 0 et de rayon 1. De plus, plus les points sont proches du cercle meilleure est la représentation car meilleure est la projection.

2. Description mathématique de l'A.C.P .

2.1. La matrice à diagonaliser

La matrice R à diagonaliser est définie par

$$R = {}^tT.T$$

$$(p,p) \quad (p,n) \quad (n,p)$$

de terme général

$$r_{jj'} = \sum_{i=1}^{n}\left(\frac{t_{ij}-\bar{t}_j}{s_j.\sqrt{n}}\right).\left(\frac{t_{ij'}-\bar{t}_{j'}}{s_{j'}.\sqrt{n}}\right),$$

c'est la matrice des corrélations linéaires entre les variables. Elle est une matrice symétrique qui contient des valeurs réelles. R est donc diagonalisable[1] $\left(P^{-1}.R.P = D\right)$ et admet des valeurs propres λ_1, ..., λ_p distinctes (Corollaire du Théorème1 et Théorème2). Ces valeurs propres sont classées dans un ordre décroissant. Pour chaque valeur propre, on cherche le vecteur propre associé. Les vecteurs propres donnent des axes propres indépendants qui passent tous par l'origine. Ces axes sont dits composantes, facteurs ou bien axes factoriels (on dit facteur même pour une ACP).

$\lambda_1 > \lambda_2 > \ldots > \lambda_p$

$\downarrow \quad \downarrow \quad \quad \downarrow$

$V_1 \quad V_2 \quad \quad V_p$

$\downarrow \quad \downarrow \quad \quad \downarrow$

1er axe propre | 2ème axe propre | pème axe propre

Ces composantes ou facteurs regroupent, dans une certaine mesure, un certain nombre d'individus ou de variables corrélées dans le but d'expliquer un phénomène par un nombre plus restreint d'éléments ou de variables.

Une fois les λ_k et les V_k déterminés, les vecteurs propres U_k de la matrice $T.\left({}^tT\right)$ sont déterminés par

$$U_k = \frac{1}{\sqrt{\lambda_k}}.T.V_k$$

---

[1] P est la matrice de passage qui est une matrice inversible constituée des vecteurs propres de R.
D est la matrice diagonale constituée des valeurs propres de R.

Théorème1: A est une matrice diagonalisable si et seulement si elle admet p vecteurs propres linéairement indépendants (les p vecteurs propres formant une base).

Théorème2: Soient V_1, …, V_p des vecteurs propres associés respectivement à des valeurs propres λ_1,.., λ_p distinctes. Alors la famille $\{V_1,..,V_p\}$ est une famille libre.

$$\left(\sum_{j=1}^{p}\alpha_j.V_j = 0 \quad \Rightarrow \quad \alpha_j. = 0, \quad \forall j = 1,..., p\right)$$

2.2. La distance entre deux points

1/ La distance entre deux points individus

La distance entre deux points individus i et i' de $N(I)$ est définie par

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

avec t_{ij} transformée on trouve

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

car

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{t_{ij} - \bar{t}_j}{S_j.\sqrt{n}} - \frac{t_{i'j} - \bar{t}_j}{S_j.\sqrt{n}} \right)^2$$

$$= \sum_{j=1}^{p} \frac{\left(t_{ij} - \bar{t}_j - t_{i'j} + \bar{t}_j\right)^2}{n.S_j^{\,2}}$$

$$= \sum_{j=1}^{p} \frac{\left(t_{ij} - t_{i'j}\right)^2}{n.S_j^{\,2}} \quad ,$$

alors le poids de chaque variable est égal à $1/n$, ce qui implique qu'on donne la même importance à toutes les variables.

2/ La distance entre deux points variables

La distance entre deux points variables j et j' de $N(J)$ est définie par

$$d^2(j,j') = 2.(1 - r_{jj'})$$

où $r_{jj'}$: le coefficient de corrélation linéaire entre j et j'.

$$d^2(j,j') = \sum_{i=1}^{n} \left(\frac{t_{ij} - \bar{t}_j}{S_j.\sqrt{n}} - \frac{t_{ij'} - \bar{t}_{j'}}{S_{j'}.\sqrt{n}} \right)^2$$

$$= \sum_{i=1}^{n} \left[\left(\frac{t_{ij} - \bar{t}_j}{S_j.\sqrt{n}} \right)^2 + \left(\frac{t_{ij'} - \bar{t}_{j'}}{S_{j'}.\sqrt{n}} \right)^2 - 2.\frac{\left(t_{ij} - \bar{t}_j\right).\left(t_{ij'} - \bar{t}_{j'}\right)}{n.S_j.S_{j'}} \right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\left(t_{ij}-\bar{t}_j\right)^2}{S_j^{\,2}} + \frac{1}{n}\sum_{i=1}^{n}\frac{\left(t_{ij'}-\bar{t}_{j'}\right)^2}{S_{j'}^{\,2}} - 2.\frac{1}{n}\sum_{i=1}^{n}\frac{\left(t_{ij}-\bar{t}_j\right).\left(t_{ij'}-\bar{t}_{j'}\right)}{S_j.S_{j'}}$$

$$= \frac{S_j^{\,2}}{S_j^{\,2}} + \frac{S_{j'}^{\,2}}{S_{j'}^{\,2}} - 2.r_{jj'} = 1+1-2.r_{jj'}$$

$$d^2\left(j,j'\right) = 2-2.r_{jj'} = 2.\left(1-r_{jj'}\right)$$

Si $r_{jj'} = 1 \quad \Rightarrow$ les deux points j et j' sont confondus ;

Si $r_{jj'} = -1 \quad \Rightarrow$ les deux points j et j' sont opposés.

Il est possible d'utiliser le résultat d'une ACP pour construire une classification statistique des variables aléatoires en utilisant la distance suivante (où $r_{jj'}$ n'est autre que la corrélation entre j et j'):

$$d\left(j,j'\right) = \sqrt{2.\left(1-r_{jj'}\right)}$$

2.3. Les coordonnées

1/ La coordonnée de l'individu i de $N(I) \subset IR^p$ sur le 1er axe

La coordonnée de l'individu i de $N(I)$ sur le 1er axe est définie par

$$F_1(i) = < T^i \quad , \quad V_1 >$$
$$= T^i . \quad V_1$$
$$(1,p) \quad (p,1)$$

Donc, la coordonnée de i sur l'axe k est

$$F_k(i) = T^i . \quad V_k$$

et les coordonnées de tous les points i sur l'axe k sont définies par les lignes (éléments) du vecteur colonne

$$T. \quad V_k \qquad , \ \forall i$$
$$(n,p) \quad (p,1)$$

2/ La coordonnée de la variable j de $N(J) \subset IR^n$ sur le 1er axe

Elle est définie par

$$G_1(j) = < T^j \quad , \quad U_1 > \quad = \quad {}^t\left(T^j\right). \quad U_1$$
$$(1,n) \qquad (n,1)$$

alors sur l'axe k on a

$$G_k(j) \;=\; {}^t\left(T^j\right). \quad U_k$$

Pour tous les points variables de $N(J) \subset IR^n$

$$G_k \;=\; {}^tT. \quad U_k \qquad , \; \forall j$$

Les coordonnées des points variables sur l'axe k sont les coefficients de corrélation entre les variables et l'axe k .

La projection orthogonale des points individus (ou variables) sur le plan 1x2 peut être schématisée par la figure 1.

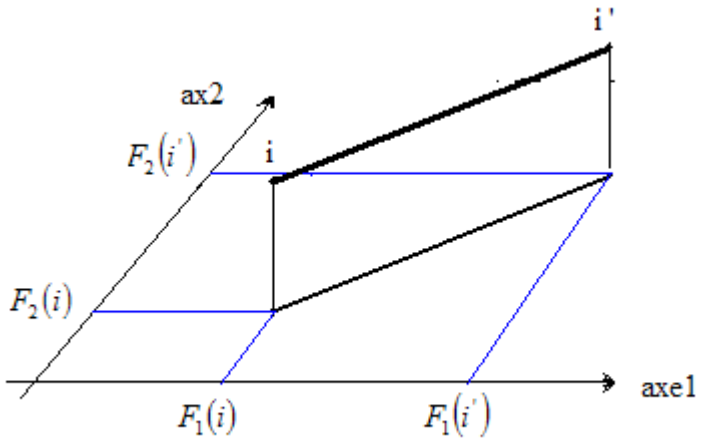


Figure 1 : La projection orthogonale des points individus

2.4. Les contributions dans l'inertie de l'axe

1/ La contribution de l'individu i dans la variance de l'axe k

Par définition le coefficient de corrélation est

$$r = \frac{Cov(X,Y)}{\sqrt{V(X)}.\sqrt{V(Y)}}$$

Soit la matrice R constituée de ces coefficients de corrélation linéaires définie par

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{pmatrix}$$

On a $D_\lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$ est semblable à R.

où λ_k , la variance expliquée par l'axe factoriel k , est définie par

$$\lambda_k = \sum_{i=1}^{n} F_k^2(i) \quad , \quad k = 1,...., p$$

De plus, on a : $trace(R) = trace(D) = \sum_{k=1}^{p} \lambda_k = I$ (Inertie) >0

I est dite la variance totale. Elle est donc, égale à la somme des variances expliquées par les p axes.

La contribution de l'individu i dans la variance expliquée par l'axe k est

$$CTR_k(i) = \tau_k(i) = \frac{F_k^2(i)}{\sum_{i=1}^{n} F_k^2(i)} = \frac{F_k^2(i)}{\lambda_k}$$

avec $\sum_{i=1}^{n} \tau_k(i) = 1 \quad = \frac{100}{100}$ (en %)

$CTR_k(i)$ est exprimée en (%). Si par exemple on a $\tau_1(i) + \tau_2(i) \cong 80\%$, l'individu i contribue à hauteur de 80% dans la variance expliquée par le plan1x2.

2/ <u>La contribution de la variable j dans la variance de l'axe k</u>

La contribution de la variable j à l'inertie de l'axe k est définie comme pour les individus par

$$CTR_k(j) = \tau_k(j) = \frac{G_k^2(j)}{\sum_{j=1}^{p} G_k^2(j)} = \frac{G_k^2(j)}{\lambda_k}$$

Elle est exprimée également en (%) et $\sum_{j=1}^{p}\tau_k(j)=1$. Enfin, on remarque que cette contribution est une contribution relative.

2.5. La contribution relative de l'axe k à l'excentricité de l'individu i ou de la variable j

1/Pour l'individu $i \in N(I)$

La contribution relative de l'axe k à l'excentricité de l'individu i est

$$Cor_k{}^2(i) = Cos_k{}^2(i) = \frac{F_k^2(i)}{\left\|T^i - G\right\|^2} = \frac{F_k^2(i)}{\sum_k F_k^2(i)} = \frac{F_k^2(i)}{\rho^2(i)}$$

où G : le centre de gravité. Cette grandeur mesure la qualité de la représentation de l'individu i. Plus le $Cos_k{}^2(i)$ est grand, meilleure est la représentation.

2/ Pour la variable $j \in N(J)$

La contribution relative de l'axe k à l'excentricité de la variable j est

$$Cor_k{}^2(j) = Cos_k{}^2(j) = G_k{}^2(j)$$

Normalement $$Cor_k{}^2(j) = Cos_k{}^2(j) = \frac{G_k{}^2(j)}{\rho^2(j)} \quad ,$$

mais on sait dans ce cas que $\rho^2(j) = \left\|T^j - \bar{T}^j\right\|^2 = 1,$

car

- on cherche alors la moyenne de $\frac{t_{ij} - \bar{t}_j}{S_j.\sqrt{n}}$ pour le vecteur T^j (le centre de gravité est une moyenne) :

$$\frac{1}{n}\sum_{i=1}^{n}\frac{t_{ij} - \bar{t}_j}{S_j.\sqrt{n}} = \frac{1}{n.S_j.\sqrt{n}}\sum_{i=1}^{n}\left(t_{ij} - \bar{t}_j\right) = \frac{1}{n.S_j.\sqrt{n}}\left[\sum_{i=1}^{n} t_{ij} - n\bar{t}_j\right] = 0$$

- Pour la $\left\|T^j\right\|^2 = 1$: Voir p15.

Comme pour les individus, une variable sera d'autant mieux représentée sur un axe, un plan ou un sous-espace que sa corrélation avec la composante principale correspondante est en valeur absolue proche de 1.

Une variable sera bien représentée sur un plan si elle est proche du bord du cercle des corrélations, car cela signifie que le cosinus de l'angle du vecteur joignant l'origine au point représentant la variable avec le plan est, en valeur absolue, proche de 1.

2.6. Le poids

En A.C.P. la même importance est attribuée à chaque individu, de ce fait chaque individu a une probabilité égale à $\frac{1}{n}$ de se réaliser. $\frac{1}{n}$, qui est donc le poids affecté à chaque individu, nous permet de définir une matrice diagonale de poids notée $P_{1/n}$

$$P_{1/n} = \begin{pmatrix} 1/n & 0 & \cdots & 0 \\ 0 & 1/n & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \vdots & 0 & 1/n \end{pmatrix} = 1/n \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \vdots & 0 & 1 \end{pmatrix} = 1/n . Id_n$$

2.7. Eléments supplémentaires

Si on craint que l'influence de certains individus soit excessive pour la détermination des axes principaux, il est possible de les placer en ***éléments supplémentaires***, c'est à dire de les considérer comme s'ils ne font pas partie du nuage dont on cherche les directions principales, mais on peut, par la suite, représenter leurs positions sur les plans principaux obtenus.

On traite de la même manière des variables en éléments supplémentaires, elles ne font pas partie de l'ensemble des variables de base mais on peut examiner leurs corrélations avec les composantes principales obtenues.

Il est parfois recommandé, après une première ACP des données étudiées, d'éprouver la stabilité des configurations observées en effectuant de nouvelles analyses laissant en éléments supplémentaires les individus ou variables d'importance trop marquée, ou encore les données douteuses.

2.8. Résultats

Dans le cas général (la possibilité d'existence de variables dépendantes), l'ACP remplace les p variables de départ par q nouvelles composantes $q \leq p$:

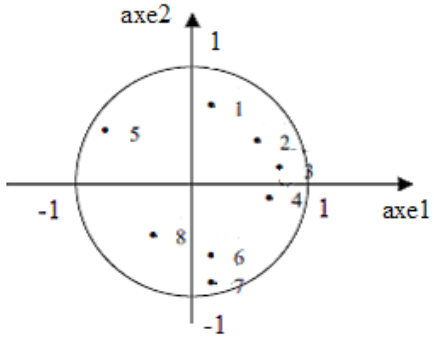
- Orthogonales 2 à 2, c-à-d $\text{cov}(V_j, V_{j'}) = 0$ pour tout $j \neq j'$;
- De variances maximales telle que $\sigma^2_{V_1} \geq \sigma^2_{V_2} \geq ... \geq \sigma^2_{V_j} \geq ... \geq \sigma^2_{V_q}$
- Le nombre maximum de composantes principales $q \leq p$ avec $q < p$ dès que l'une des variables d'origine est une combinaison linéaire d'autres variables.

• Choix des r premiers axes principaux (composantes principales) :
Un nombre $r << p$ d'axes est retenu afin de réduire la dimension de l'espace tout en gardant un maximum d'information des données initiales. La mesure appropriée de cette information est le % de variance expliquée définie par $\tau_k = \frac{\lambda_k}{\sum_{k=1}^{q} \lambda_k} \times 100$.

Il est à signaler également que si les variables originales sont fortement corrélées entre elles, un nombre réduit d'axes (composantes) permet d'expliquer 80% à 90% de variance, et la perte d'information dans ce cas est minime.

Géométriquement : Projeter les données dans un sous-espace de dimension r, centré sur g, revient à aplatir le nuage sur le sous-espace de r axes (exemple d'un ballon). Par ailleurs le % de variance expliquée par les r axes mesure d'aplatissement du nuage sur le sous-espace. En outre, Plus ce % est grand, meilleure est la représentation des données dans le sous-espace. Mais il faut faire attention car
Proximité de j et j' sur le plan par exemple $\not\Rightarrow$ proximité de j et j' dans l'espace initial.

• La représentation des points variables sur les deux premiers axes (composantes), c.à.d. sur un plan vectoriel se fait dans un cercle de rayon 1. De plus, plus les points sont proches du cercle meilleure est la représentation (car dans ce cas l'effet des autres axes est minime étant donné que tous les axes passent par le centre).



2.9. Simple exemple
Soient les données représentant les notes de $n = 9$ étudiants dans $p = 4$ disciplines.

	Ind	MATH	PHYS	FRAN	ANGL
a	1	6.00	6.00	5.00	5.50
b	2	8.00	8.00	8.00	8.00
c	3	6.00	7.00	11.00	9.50
d	4	14.50	14.50	15.50	15.00
e	5	14.00	14.00	12.00	12.50
f	6	11.00	10.00	5.50	7.00
g	7	5.50	7.00	14.00	11.50
h	8	13.00	12.50	8.50	9.50
i	9	9.00	9.50	12.50	12.0

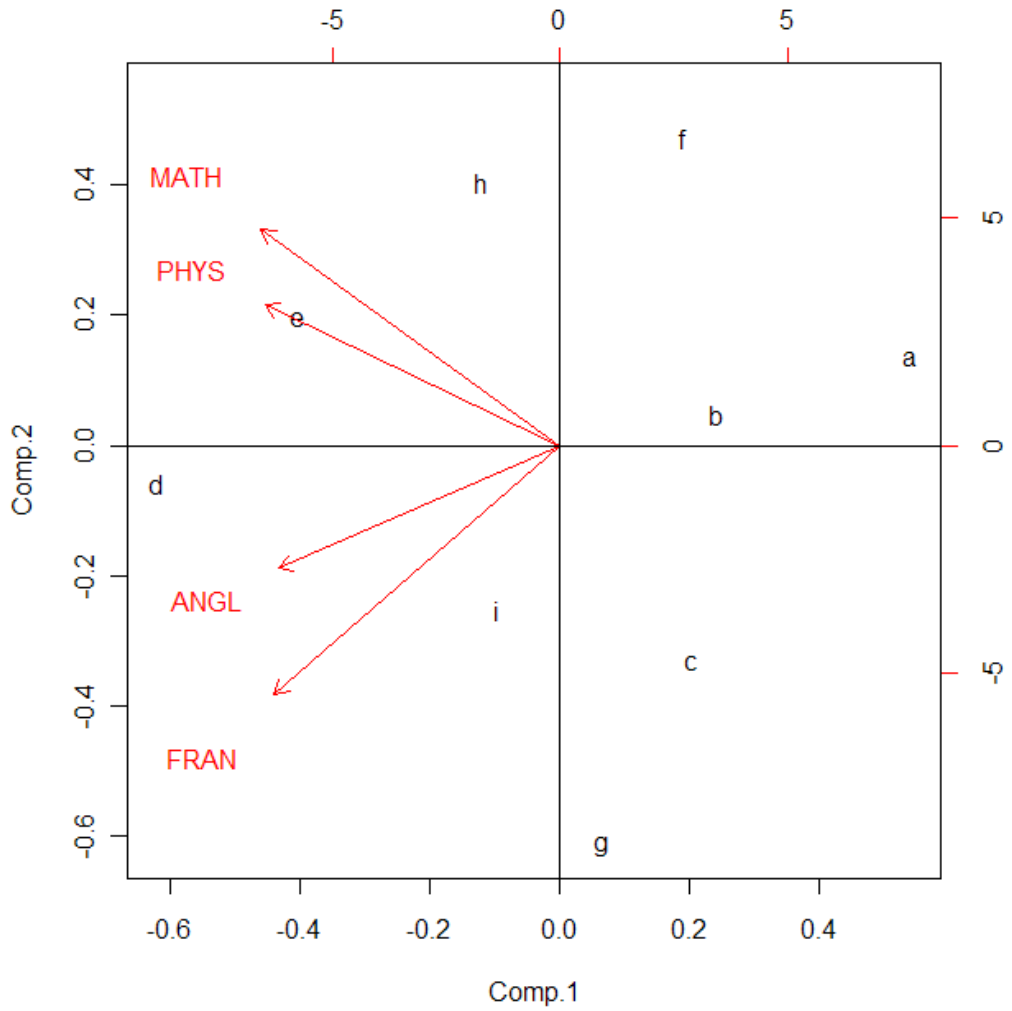
Results

Proportion of Variance	0.700467	0.2984607	0.0008095538	0.0002627896
Cumulative Proportion	0.700467	0.9989277	0.9997372104	1.0000000000

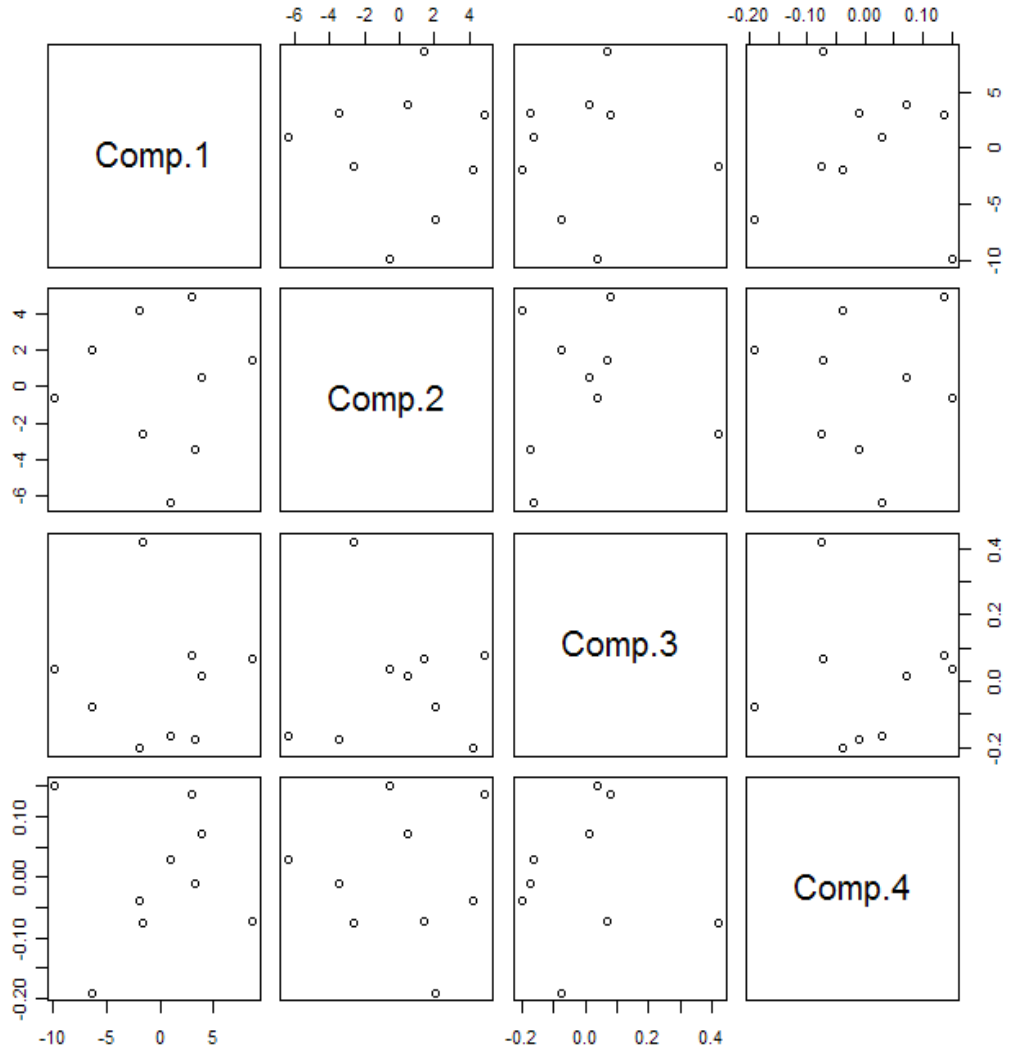
scores

	Comp.1	Comp.2	Comp.3	Comp.4
a	8.612059	1.4093727	0.06752404	-0.07158969
b	3.878793	0.5022279	0.01309446	0.07093634
c	3.213388	-3.4683149	-0.17497150	-0.01065973
d	-9.851807	-0.5995132	0.03680819	0.14998275
e	-6.406574	2.0465857	-0.07561885	-0.19044801
f	3.033102	4.9211080	0.07749344	0.13542301
g	1.025444	-6.3771179	-0.16386970	0.02986136
h	-1.953971	4.1995965	-0.20192835	-0.03907002
i	-1.550436	-2.6339447	0.42146828	-0.07443601

Les individus sont représentés, puis les variables sont ajoutées sur le plan.



En général en ACP le plan vectoriel des deux premières composantes est représenté car il représente le plus d'informations, mais parfois il est souhaitable de voir si les autres plans apportent d'informations supplémentaires. Pour notre exemple les plans possibles pour les individus sont



2.10. Exemple

2.10.1. Présentation des données

Le tableau ci-dessous dresse le comportement de consommation de 12 ménages concernant 7 biens. Les 7 biens sont : bread, vegetables, fruits, meat, poultry, milk et water and drinks. Les individus sont : ′w′=manual worker, ′e′=employee et ′m′=manager ;

les lettres représentants les individus sont suivies d'un chiffre qui indique le nombre des personnes dans le ménage (les parents + les enfants).

	bread	veget.	fruits	meat	poul.	milk	water
w2	332	428	354	1437	526	247	427
e2	293	559	388	1527	567	239	258
m2	372	767	562	1948	927	235	433
w3	406	563	341	1507	544	324	407
e3	386	608	396	1501	558	319	363
m3	438	843	689	2345	1148	243	341
w4	534	660	367	1620	638	414	407
e4	460	699	484	1856	762	400	416
m4	385	789	621	2366	1149	304	282
w5	655	776	423	1848	759	495	486
e5	584	995	548	2056	893	518	319
m5	515	1097	887	2630	1167	561	284
mean	446,67	732,00	505,00	1886,75	803,17	358,25	368,58
sd	107,15	189,18	165,09	395,75	249,56	117,13	71.78

2.10.2. Résultats de l'ACP

- La matrice de corrélation

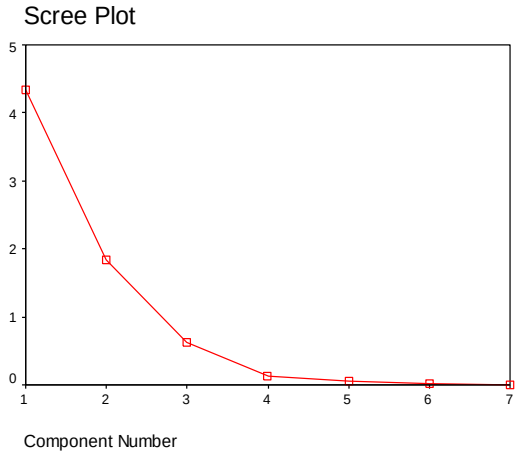
Correlation Matrix

		BREAD	FRUITS	MEAT	MILK	POUL	VEGET	WATER
Correlation	BREAD	1,000	,196	,321	,856	,248	,593	,304
	FRUITS	,196	1,000	,959	,332	,926	,856	-,486
	MEAT	,321	,959	1,000	,375	,982	,881	-,437
	MILK	,856	,332	,375	1,000	,233	,663	,007
	POUL	,248	,926	,982	,233	1,000	,827	-,400
	VEGET	,593	,856	,881	,663	,827	1,000	-,356
	WATER	,304	-,486	-,437	,007	-,400	-,356	1,000

- Les valeurs propres

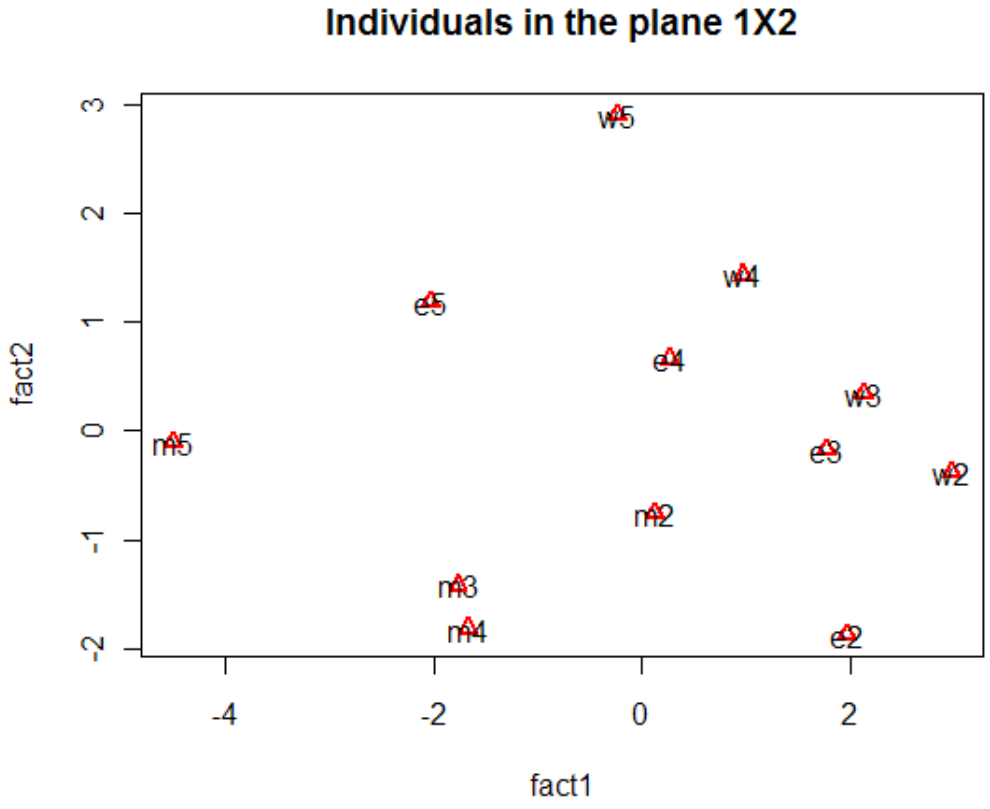
Total Variance Explained

Component	Total	% of Variance	Cumulative %
1	4,333	61,903	61,903
2	1,830	26,147	88,050
3	,631	9,012	97,062
4	,128	1,833	98,896
5	5,756E-02	,822	99,718
6	1,885E-02	,269	99,987
7	9,038E-04	1,291E-02	100,000



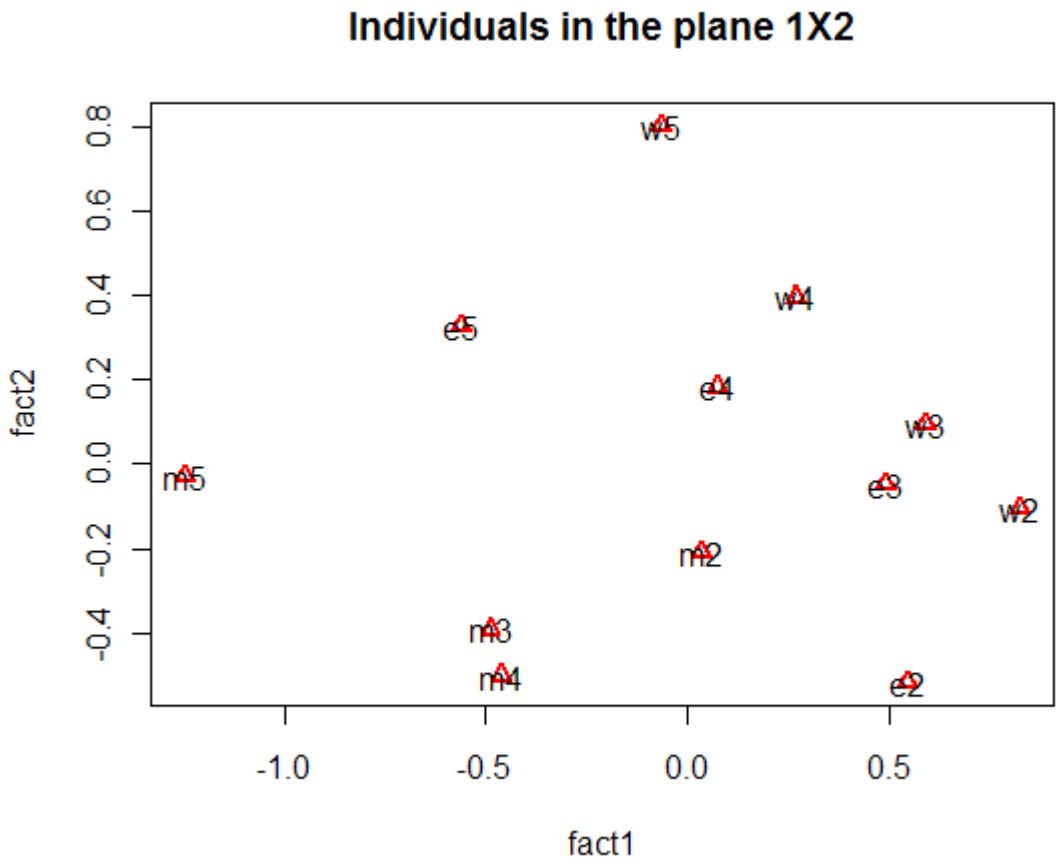
- Les scores des individus sur les 5 premiers axes

		Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
1	w2	2,99	-0,38	-0,42	0,38	-0,24
2	e2	1,97	-1,87	1,36	-0,17	0,10
3	m2	0,12	-0,76	-1,48	0,20	0,46
4	w3	2,13	0,34	0,11	0,11	-0,01
5	e3	1,77	-0,17	0,54	0,16	0,18
6	m3	-1,77	-1,42	-1,04	-0,45	0,08
7	w4	0,97	1,43	0,29	-0,28	-0,10
8	e4	0,26	0,66	-0,29	0,30	-0,17
9	m4	-1,67	-1,81	-0,10	-0,42	-0,44
10	w5	-0,23	2,90	-0,59	-0,26	-0,13
11	e5	-2,04	1,18	1,03	-0,34	0,34
12	m5	-4,51	-0,11	0,59	0,75	-0,08



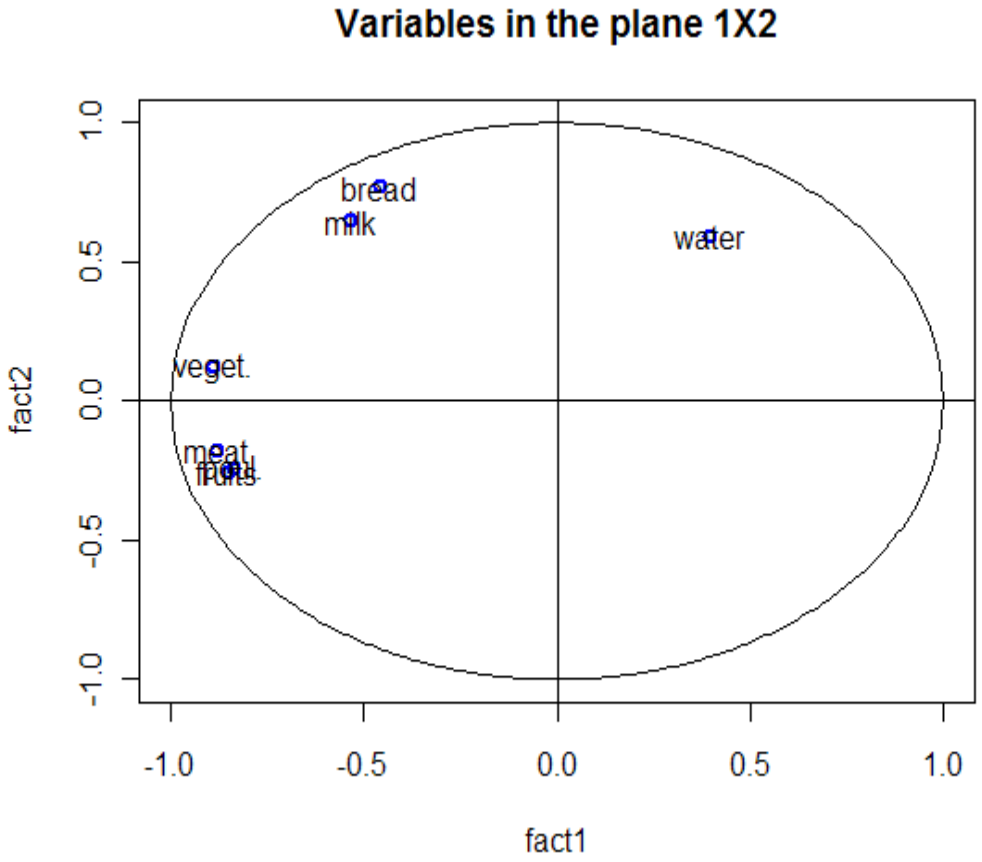
- Les projections des individus sur les 5 premiers axes (F)

		comp1	comp2	comp3	comp4	comp5
1	w2	0,83	-0,10	-0,12	0,10	-0,07
2	e2	0,55	-0,52	0,38	-0,05	0,03
3	m2	0,03	-0,21	-0,41	0,06	0,13
4	w3	0,59	0,09	0,03	0,03	0,00
5	e3	0,49	-0,05	0,15	0,05	0,05
6	m3	-0,49	-0,39	-0,29	-0,12	0,02
7	w4	0,27	0,40	0,08	-0,08	-0,03
8	e4	0,07	0,18	-0,08	0,08	-0,05
9	m4	-0,46	-0,50	-0,03	-0,12	-0,12
10	w5	-0,06	0,80	-0,16	-0,07	-0,04
11	e5	-0,56	0,33	0,29	-0,09	0,09
12	m5	-1,25	-0,03	0,16	0,21	-0,02



- Les projections des variables sur les 5 premiers axes (G)

	comp1	comp2	comp3	comp4	comp5
bread	-0,46	0,77	0,01	-0,18	-0,01
veget.	-0,89	0,12	0,05	-0,01	0,18
fruits	-0,85	-0,25	-0,11	0,18	0,01
meat	-0,88	-0,18	-0,15	-0,02	-0,09
poul.	-0,84	-0,24	-0,26	-0,11	-0,05
milk	-0,54	0,65	0,32	0,15	-0,08
water	0,39	0,59	-0,57	0,1	0,02



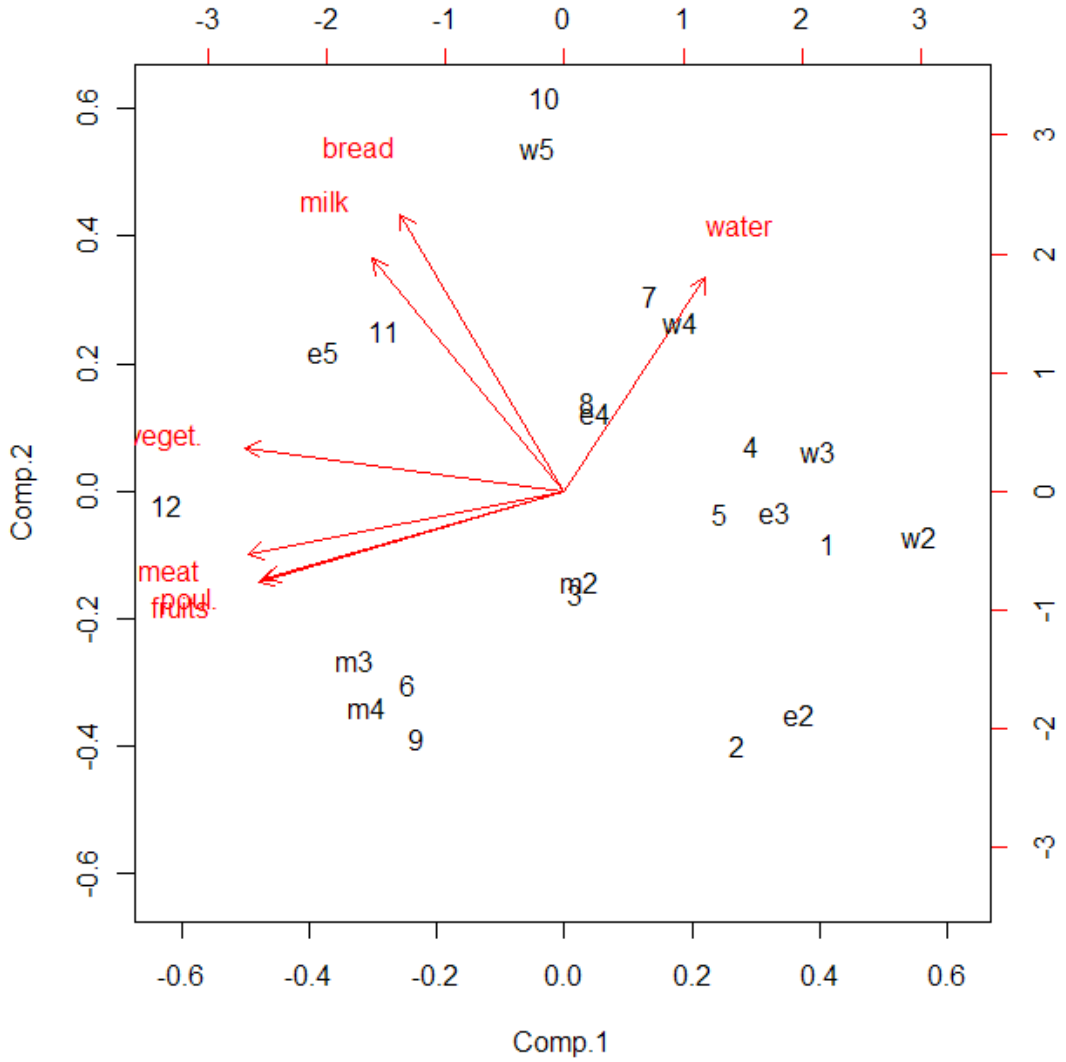
- Les cos2 des individus sur les 4 premiers axes

	comp1	comp2	comp3	comp4
w2	0,944	0,015	0,019	0,015
e2	0,419	0,377	0,200	0,003
m2	0,005	0,187	0,716	0,014
w3	0,969	0,024	0,003	0,002
e3	0,892	0,008	0,082	0,008
m3	0,482	0,308	0,166	0,031
w4	0,299	0,646	0,026	0,024
e4	0,097	0,607	0,113	0,125
m4	0,429	0,505	0,002	0,027
w5	0,006	0,945	0,039	0,007
e5	0,605	0,203	0,156	0,017
m5	0,956	0,001	0,017	0,027

- Les cos2 des variables sur les 4 premiers axes

	comp1	comp2	comp3	comp4
bread	0,25	0,71	0,00	3,79E+04
veget.	0,94	0,02	0,00	6,79E+01
fruits	0,86	0,08	0,01	3,85E+04
meat	0,93	0,04	0,03	3,64E+02
poul.	0,83	0,07	0,08	1,35E+04
milk	0,34	0,50	0,12	2,59E+04
water	0,18	0,42	0,38	1,20E+04

• Le bi-plot



Remarque1
Les signes des colonnes des vecteurs propres et des scores (coordonnées) sont arbitraires et peuvent changés selon les programmes de l'ACP et selon les logiciels utilisés

Remarque2
Dans la pratique utilisez $n \times Var(F_k)$ pour calculer les valeurs propres et non pas $\sum_{i=1}^{n} F_k^2(i)$.

2.11. Formulation mathématique de l'ACP (directions successives d'inertie maximale du nuage)

Soit la matrice T centrée réduite si nécessaire,
et soit le vecteur v_k de IR^p pour lequel la norme est égale à 1,

$T.v_k$: Le vecteur $T.v_k$ de IR^n a pour composantes les produits scalaires des observations (centrée réduite si nécessaire) avec v_k. Il représente *les distances à l'origine des projections* des observations selon la direction de v_k ;

$v_k^{'}T^{'}T.v_k$: Le produit matriciel $v_k^{'}T^{'}T.v_k$ représente *l'inertie totale du nuage* dans cette direction v_k ;

$T^{'}T$: Est une matrice symétrique représentant la *matrice d'inertie* du nuage.
Elle est simplement, au facteur (1/n) près, *la matrice des corrélations* entre les variables colonnes initiales.

La recherche des directions principales, c'est à dire des directions successives d'inertie maximale du nuage, se traduit donc par le problème de maximisation sous contrainte suivant :

$$\underset{v_k}{Max} \ \left(v_k^{'}T^{'}Tv_k\right)$$
$$s.c. \quad v_k^{'}v_k = 1$$

les vecteurs v_k successifs doivent être orthogonaux $v_k v_{k^{'}} = 0$.

L'algèbre linéaire enseigne que les vecteurs propres normés v_k, associés à la suite décroissante des valeurs propres (positives) λ_k de $T^{'}T$, apportent la solution du problème. La valeur propre λ_k mesurant l'inertie dans la $k^{ième}$ direction principale v_k :

$$v_k^{'}T^{'}T.v_k = \lambda_k v_k^{'}v_k = \lambda_k$$

Les vecteurs $F_k = T.v_k$ de IR^n sont les *composantes principales* successives du nuage, centrées, de variances respectives $\frac{\lambda_k}{n}$ et non corrélées (de covariances nulles). Ce sont les nouvelles variables dont les composantes donnent les coordonnées des points du nuage sur les axes factoriels.

Chapitre 3 L'analyse Factorielle des Correspondances (A .F.C.)

L'analyse factorielle des correspondances est un mode de présentation graphique d'un tableau de contingence. Elle vise à rassembler en un ou plusieurs graphes (très souvent un seul), la plus grande partie possible de l'information contenue dans le tableau, en prenant en considération, non pas les valeurs absolues, mais les correspondances entre les caractères, c'est à dire les valeurs relatives. Cette méthode de présentation est d'autant plus utile que la dimension du tableau est grande, car un petit tableau n'a pas besoin d'A.F.C. pour être interprété.

Un tableau de contingence, fréquent en statistique, est un tableau qui donne la ventilation d'une population ou d'une quantité selon deux critères qualitatifs que l'on croise. On le reconnaît si on obtient des quantités qui ont un sens en calculant les sommes en lignes ou en colonnes.

Mais, vu les avantages que représente l'A.F.C., son utilisation peut être étendue à certains cas où l'on ne dispose pas de tableaux de contingence. On cite parmi ces cas celui du tableau de notes, très souvent utilisé pour les enquêtes d'opinion auprès du public ; et celui du tableau logique qui ne contient que des 0 et des 1. Il est clair que la somme des notes obtenues par un individu a un sens, et que la somme des 0 et des 1 d'une ligne a également un sens. Ainsi, On peut les considérer comme des faux tableaux de contingence.

SECTION 1 : LECTURE DES DONNEES

1.1. Le tableau de données

Le tableau des données met en correspondance deux ensembles que l'on a l'habitude de noter I (lignes) et J (colonnes). Ce tableau est noté K_{IJ} ,

$$K_{IJ} = \{k(i, j) / i \in I, j \in J\}.$$

I et J sont deux ensembles d'éléments finis, soit $CardI = n$ et $CardJ = p$. Le terme général du tableau K_{IJ} est $k(i, j)$.

Généralement, les éléments mis en ligne sont nommés individus, ceux mis en colonne sont nommés variables. Les deux ensembles I et J jouent des rôles symétriques, en sorte qu'on ne changerait rien aux résultats en changeant les lignes par les colonnes. Donc, la présentation du tableau est indifférente.

1.2. Les marges et leurs profils

Du tableau K_{IJ} on peut définir deux marges :

- une colonne de marge dont le i^{e} terme est la somme des nombres inscrits dans la i^{e} ligne.

$$k(i) = \sum_{j=1}^{p} k(i, j) \, . \qquad (1)$$

- une ligne de marge dont le j^{e} terme est la somme des nombres inscrits dans la j^{e} colonne.

$$k(j) = \sum_{i=1}^{n} k(i, j) \, . \qquad (2)$$

La ligne et la colonne de marge ont le même total, noté k , qui est égal à la somme de tous les éléments $k(i, j)$ du tableau K_{IJ} .

$$k = \sum_{i=1}^{n} \sum_{j=1}^{p} k(i, j) = \sum_{i=1}^{n} k(i) = \sum_{j=1}^{p} k(j) \qquad (3)$$

	j	Colonne de marge
i	. . . . . $k(i, j)$	$k(i)$
ligne de marge	$k(j)$	k

Pour comparer deux lignes i et i', il faut utiliser les valeurs relatives car les sommes par lignes sont différentes. Nous définissant alors, le tableau des fréquences relatives.

En rapportant les $k(i, j)$ au total général du tableau K_{IJ} , on définit un tableau de fréquence, noté f_{IJ} :

$$f_{IJ} = \left\{ f_{ij} / i \in I, j \in J \right\},$$
$$f_{ij} = k(i, j) / k \qquad (4)$$

f_{IJ} est la loi conjointe du couple (i, j) sur l'ensemble fini IxJ .

En rapportant les k(i) à leur total k, on obtient le profil de la colonne de marge:

$$f_I = \{ f_i = k(i)/k \mid i \in I \} = (f_1, f_2, ..., f_i, ..., f_n)' \quad (5)$$

On définit de la même façon le profil de la ligne de marge en divisant les k(j) par le total général k :

$$f_J = \{ f_j = k(j)/k \mid j \in J \} = (f_1, f_2, ..., f_j, ..., f_p) \quad (6)$$

$f_i = k(i)/k$ est appelé masse ou poids de i, il représente la loi marginale de i,
$f_j = k(j)/k$ est appelé masse ou poids de j, il représente la loi marginale de j.

Cette masse mesure l'importance relative de l'élément i ou j au sein de I ou J respectivement.

La colonne $\{ f_i / i \in I \}$ et la ligne $\{ f_j / j \in J \}$ ont le même total qui est égal à 1.

$$\sum_{i=1}^{n} f_i = \sum_{j=1}^{p} f_j = 1. \quad (7)$$

On obtient ainsi le tableau f_{IJ}, appelé tableau de fréquences :

	j	f_I
i	. . . . . f_{ij}	f_i
f_J	f_j	1

f_{IJ} est représenté par la matrice F =$\{ f_{ij} \}$qui est une matrice de format (n, p).

1.3. Définition des profils des lignes et des colonnes du tableau K_{IJ}

Dans notre analyse, l'élément i de I ne sera pas caractérisé par sa ligne brute $\{ k(i, j)/ j \in J \}$ et par son total $k(i)$, mais par son total $k(i)$ et son profil $\{ k(i, j) / k(i) \mid j \in J \}$ de total égal à 1. Ceci nous permet de comparer deux éléments i et i' de I.

On est amené alors, à construire le tableau des profils des lignes. Ce tableau peut être obtenu en divisant chaque ligne i par son total $k(i)$.

Soit $f_j^i = k(i,j) / k(i)$ la part relative ou la proportion de j dans la i[e] ligne, le profil de la ligne i, dit profil de i sur J, est :

$$f_J^i = \{ \; f_j^i \; / \; j \in J \} = (f_1^i, f_2^i, ..., f_j^i, ..., f_p^i) \qquad (8)$$

f_J^i définit la loi conditionnelle de j pour i donnée.

De même tout élément j de J est caractérisé par son total $k(j)$ et son profil $\{ k(i,j)/ k(j) , i \in I \}$. Le tableau des profils des colonnes est obtenu en divisant chaque colonne j par son total $k(j)$:

$$f_i^j = k(i,j) / k(j) , \qquad (9)$$

$$f_I^j = \{ f_i^j / i \in I \} = (f_1^j, f_2^j, ..., f_i^j, ..., f_n^j) \qquad (10)$$

f_I^j définit la loi conditionnelle de i pour j donnée.

Au tableau des profils des lignes, on adjoint une colonne poids f_I, et au tableau des profils des colonnes, on adjoint une ligne poids f_J.

1.4. Notations

Afin de faciliter et d'alléger l'écriture, une notation matricielle est adoptée pour décrire les différents tableaux définis précédemment. Ces notations sont dressées dans ce point 4 de la section1 et seront utilisées dans la suite du document.

- On note $r_{ij} = f_j^i$ (i donné),
  et on définit une matrice R de format (n , p) par $R = \{ r_{ij} \}$
- On note $c_{ij} = f_i^j$ (j donné),
  et on définit une matrice C de format (n , p) par $C = \{ c_{ij} \}$
- On note $H = f_I = (f_1, f_2, ..., f_i, ..., f_n)'$
- On note $G = f_J = (f_1, f_2, ..., f_j, ..., f_p)'$
- Soient deux matrices D_n de format (n , n) et D_p de format (p , p) tel que

$D_n = \text{diag}(f_1, f_2, ..., f_i, ..., f_n)$ (11)

$D_p = \text{diag}(f_1, f_2, ..., f_j, ..., f_p)$ (12)

On peut alors écrire

$R = D_n^{-1}.F$ (13)

$C = F.D_p^{-1}$ (14)

$G = D_p.i_p$ (15)

$H = D_n.i_n$ (16)

$i_p = (1, 1, \ldots, 1)'$ de dimension p et $i_n = (1, 1, \ldots, 1)'$ de dimension n.

SECTION 2 : LES NUAGES DE PROFILS ET LEURS CENTRES DE GRAVITE

L'ensemble I est représenté dans l'espace des profils sur J de dimension (p-1), et l'ensemble J dans celui des profils sur I de dimension (n-1).

2.1. Le nuage N(I)

Dans l'espace des profils sur l'ensemble des variables J, chaque ligne (individu) i du tableau brut est représentée par son profil f_J^i affecté à sa masse f_i. On appelle nuage de l'ensemble I, l'ensemble des profils des diverses lignes i, chacun menu de la masse de la ligne qu'il représente.

$$N(I) = \{ (f_J^i , f_i) / i \in I \} \subset \mathrm{R}^p.$$

Le centre de gravité du nuage $N(I)$ est une sorte de moyenne spatiale, où chaque point joue un rôle proportionnel à sa masse.

$$g_J = \sum_{i=1}^{n} f_i f_J^i / \sum_{i=1}^{n} f_i \qquad (17)$$

et comme $\sum_{i=1}^{n} f_i = 1$, la formule s'écrit simplement:

$$\begin{aligned} g_J &= \sum_{i=1}^{n} f_i f_J^i \\ &= \sum_{i=1}^{n} f_i (f_1^i, f_2^i, ..., f_j^i, ..., f_p^i) \\ &= \sum_{i=1}^{n} (f_i f_1^i, f_i f_2^i, ..., f_i f_j^i, ..., f_i f_p^i) \\ &= (\sum_{i=1}^{n} f_i f_1^i, \sum_{i=1}^{n} f_i f_2^i, ..., \sum_{i=1}^{n} f_i f_j^i, ..., \sum_{i=1}^{n} f_i f_p^i) \\ &= (g_1, g_2, ..., g_j, ..., g_p) \end{aligned}$$

$$\begin{aligned} g_j &= \sum_{i=1}^{n} f_i f_j^i \\ &= \sum_{i=1}^{n} (k(i)/k).(k(i,j)/k(i)) \\ &= (1/k) \sum k(i,j) \\ &= (1/k).k(j) = f_j \end{aligned}$$

On en déduit que le centre de gravité du nuage $N(I)$ est la ligne f_J :

$$g_J = f_J = (f_1, f_2, ..., f_j, ..., f_p) \in \mathrm{R}^p. \quad (18)$$

2.2. Le nuage N(J)

De la même manière, on constitue le nuage de J :

$$N(J) = \{ (f_I^j , f_j) / j \in J \} \subset \mathrm{R}^n.$$

Le centre de gravité du nuage $N(J)$ est la colonne f_I :

$$g_I^{'} = f_I^{'} = (f_1, f_2, ..., f_i, ..., f_n) \in \mathrm{R}^n. \quad (19)$$

Si le tableau K_{IJ} comporte des éléments supplémentaires (qui vont être expliqués dans un point ultérieur), on se restreint pour la recherche du centre de gravité au tableau $K_{I'J'}$ tel que :

$I^{'}$: est le sous ensemble des individus non supplémentaires.

$J^{'}$: est le sous ensemble des variables non supplémentaires.

SECTION 3 : LA DISTANCE DISTRIBUTIONNELLE

3.1. La distance distributionnelle sur l'espace des profils sur J (où se trouve N(I))

Etant donné qu'un nuage sans métrique n'a pas de forme car il n'a pas de directions principales d'allongement, il est indispensable de définir une métrique sur les deux nuages $N(I)$ et $N(J)$.

Pour comparer deux lignes i et i' du nuage $N(I)$, il faut définir une distance mettant en jeu toutes les dimensions de l'espace. On utilise la distance distributionnelle entre les profils, qui est une distance propre à l'analyse des correspondances.

Partons de la formule de distance Euclidienne en échelle quelconque dans l'espace des profils sur J. Le carré de la distance entre les deux lignes i et i' de I, pondérée par α_J est

$$d^2(i,i^{'}) = d^2(f_J^i, f_J^{i'}) = \sum_{j=1}^{p} \alpha_j (f_j^i - f_j^{i'})^2 \quad (20)$$

$\alpha_J = (\alpha_1, ..., \alpha_j, ..., \alpha_p)$ est un vecteur où les coefficients α_j, strictement positifs, pondèrent l'influence de la $j^{ème}$ variable.

En prenant $\alpha_j = 1/f_j$, on trouve la formule de distance distributionnelle appelée distance de Chi-2 :

$$d^2(f_J^i, f_J^{i'}) = \sum_{j=1}^{p} (1/f_j)(f_j^i - f_j^{i'})^2 \qquad (21)$$

En fait, le choix de cette métrique de χ^2 est lié au choix de l'indépendance statistique exprimé par $f_{IJ} = f_I . f_J$ qui signifie l'hypothèse du point de vue probabiliste que la loi f_{IJ} est le produit des deux lois marginales f_I et f_J ; ou encore les couples aléatoires (i, j) apparaissent comme si i et j étaient indépendants l'un de l'autre.

L'égalité $f_{IJ} = f_I . f_J$ signifie encore que toutes les lignes ont pour profil f_J et toutes les colonnes ont pour profil f_I. L'A.F.C. a précisément pour objet de découvrir dans quelles directions principales les données s'écartent de cette hypothèse nulle.

La matrice de distance est symétrique car la distance entre f_J^i et $f_J^{i'}$ est exactement la même que celle entre $f_J^{i''}$ et f_J^i. Elle est de diagonale nulle et de valeur maximale généralement non limitée.

3.2. La distance distributionnelle sur l'espace des profils sur I (où se trouve N(J))

La formule de distance entre deux colonnes j et j' sur l'espace des profils sur I est :

$$d^2(f_I^j, f_I^{j'}) = \sum_{i=1}^{n} (1/f_i).(f_i^j - f_i^{j'})^2 \qquad (22)$$

Les deux formules (21) et (22) sont compatibles avec le principe d'équivalence distributionnelle :
Dans $N(I)$, si deux points f_J^i et $f_J^{i'}$ coïncident et reçoivent respectivement les masses f_i et $f_{i'}$, on peut les considérer comme un seul point i'' affecté de la masse (nous traitons d'une manière similaire deux points confondus j et j') :

$$f_{i''} = f_i + f_{i'} \qquad (f_{j''} = f_j + f_{j'}) \qquad (23)$$

Le nuage $N(I)$ ne change pas quand le tableau est modifié en cumulant deux lignes en une seule. Le nuage $N(J)$ ne change pas lui aussi, car la distance entre deux points de $N(J)$ n'est pas modifiée. Néanmoins, cette modification remplace l'ensemble I par l'ensemble $I' = I - i - i' + i''$.

SECTION 4 : LES AXES FACTORIELS ET LES FACTEURS

Le but de l'analyse factorielle des correspondances est de représenter géométriquement, dans un espace Euclidien de faible dimension les diverses informations. Il s'agit d'un algorithme de traitement des données qui fournit des images simplifiées de la réalité multidimensionnelle.

La réduction de la dimension de l'espace où figure le nuage est effectuée par la définition dans l'espace de nouveaux axes, appelés axes principaux d'inertie ou axes factoriels, sur lesquels sont définies de nouvelles coordonnées, appelées facteurs.

4.1. Pour le nuage N(I)

Les individus sont représentés par les lignes de la matrice X de format (n,p) définie par

$$X = R.D_p^{-1/2} = D_n^{-1}.F.D_p^{-1/2} \qquad (24)$$

où chaque ligne a une masse ou un poids f_i. D_n, D_p et R sont définies respectivement par les formules (11), (12) et (13).

L'élément x_{ij} de la matrice X est écrit $\dfrac{f_{ij}}{f_i\sqrt{f_j}}$.

Dans ce cas, on doit prendre en considération ce poids, exprimé par la matrice D_n, dans le calcul des axes factoriels, notés U_α, α = 1,…, r , et des valeurs propres (inerties), notées λ_α , α = 1,…,r .

Les axes factoriels U_α sont déterminés par les vecteurs propres de la matrice T de format (p,p) , associés aux valeurs propres $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r$, tel que

$$T.U_\alpha = \lambda_\alpha.U_\alpha \quad , \quad \alpha = 1,...,r .$$

La matrice symétrique T est définie par

$$T = X^{'}.D_n.X \qquad (25)$$

ou par $T = D_P^{-1/2}.F^{'}.D_n^{-1}.D_n.D_n^{-1}.F.D_P^{-1/2}$

$$T = D_P^{-1/2}.F^{'}.D_n^{-1}.F.D_P^{-1/2} \qquad (26)$$

Les axes factoriels, ou axes principaux d'inertie, correspondent aux directions principales dans lesquelles s'allonge le plus le nuage autour de son centre de gravité. Ils sont rangés de 1 à r dans l'ordre décroissant des λ_α , et les axes successifs sont orthogonaux deux à deux :

$$\sum_{j=1}^{p} U_{\alpha}^{j}.U_{\beta}^{j}.f_j \;=\; \begin{cases} 0 & si \quad \alpha \neq \beta \\ 1 & si \quad \alpha = \beta \end{cases} \tag{27}$$

Le premier axe factoriel est obtenu en maximisant $U_1^{'}TU_1$ sachant que $\|U_1\| = 1$,

c.à.d. $$U_1 \leftarrow \arg\max_{U_1^{'}U_1=1} U_1^{'}TU_1 \quad \Rightarrow \quad U_1 \leftarrow \arg\max_{U_1^{'}U_1=1} U_1^{'}(X^{'}D_nX)U_1 \tag{28}$$

Nous cherchons le deuxième axe factoriel en maximisant le programme suivant :

$$U_2 \leftarrow \arg\max_{\substack{U_2^{'}U_2=1 \\ U_1^{'}U_2=0}} U_2^{'}TU_2 \tag{29}$$

et ainsi de suite, jusqu'à ce que nous obtenons les r axes factoriels.

Les facteurs F_α sont donnés par la projection des lignes de X sur l'axe factoriel U_α.

$$F_\alpha \;=\; X.U_\alpha \tag{30}$$

alors $$F_\alpha(i) = \sum_{j=1}^{p} U_{\alpha j}.f_{ij} / f_i\sqrt{f_j} \;=\; \sum_{j=1}^{p} U_{\alpha j}.(f_j^i.f_j^{-1/2}) \tag{31}$$

$F_\alpha(i)$, qui est appelé facteur, mesure la distance du profil f_J^i au profil moyen f_J, en projection sur l'axe α :

$$|\,F_\alpha(i)| \;=\; d(pr_\alpha(f_J^i), f_J)\,. \tag{32}$$

F_α est une fonction de moyenne 0 et de variance λ_α :

$$\sum_{i=1}^{n} f_i.F_\alpha(i) = 0 \quad ; \quad \sum_{i=1}^{n} f_i.(F_\alpha(i))^2 = \lambda_\alpha \tag{33}$$

$\lambda_\alpha \;=\; \sum_{i=1}^{n} f_i.(F_\alpha(i))^2$ mesure la dispersion globale ou l'inertie du nuage le long de l'axe de rang α .

4.2. Pour le nuage N(J)

De même, à chaque axe factoriel du nuage $N(J)$ est associé un triplé $(V_\alpha, G_\alpha, \mu_\alpha)$ avec des propriétés analogues, où V_α sont les axes factoriels ; G_α les facteurs ; μ_α les valeurs propres.

Les variables sont représentées par les colonnes de la matrice Y de format (n,p) définie par

$$Y = D_n^{-1/2} C = D_n^{-1/2} F D_p^{-1} = \left\{ \frac{f_{ij}}{\sqrt{f_i}\, f_j} \right\} \tag{34}$$

et la matrice à diagonaliser dans ce cas est une matrice de format (n,n) définie par

$$W = Y D_p Y' = D_n^{-1/2} F D_p^{-1} F' D_n^{-1/2} \tag{35}$$

Les facteurs des variables G_α ,qui sont les projections des colonnes de Y sur V_α , sont donnés par

$$G_\alpha = Y'.V_\alpha \quad , \quad \alpha = 1,...,r \tag{36}$$

Deux remarques particulières sont à mentionner :

1 - Les nuages $N(I)$ et $N(J)$ ont les mêmes valeurs propres non nulles ; et les axes principaux d'inertie de $N(J)$ se déduisent de ceux de $N(I)$ et réciproquement.

Démontrons que les deux nuages ont les mêmes valeurs propres non nulles. Afin de faciliter la démonstration nous allons noter

$T = X^{*'} X^*$ et $W = X^* X^{*'}$ tel que $X^* = D_n^{1/2} X$

Soient T la matrice à diagonaliser associée au nuage $N(I)$;

W la matrice à diagonaliser associée au nuage $N(J)$;

$\lambda_1, \lambda_2, ..., \lambda_r$ les valeurs propres non nulles de la matrice T ;

$U_1, U_2, ..., U_r$ les vecteurs propres de la matrice T ;

$\mu_1, \mu_2, ..., \mu_r$ les valeurs propres non nulles de la matrice W ;

$V_1, V_2, ..., V_r$ les vecteurs propres de la matrice W ;

D'une part

$\lambda_1, \lambda_2, ..., \lambda_r$ les valeurs propres non nulles de la matrice T

$\Rightarrow \quad (T - \lambda_\alpha I) U_\alpha = \underline{0}$, α = 1, …, r

$\Rightarrow \quad T U_\alpha = \lambda_\alpha U_\alpha$

$\Rightarrow \quad X^{*'} X^* U_\alpha = \lambda_\alpha U_\alpha$

$\Rightarrow \quad X^* (X^{*'} X^*) U_\alpha = X^* (\lambda_\alpha U_\alpha)$

$\Rightarrow \quad (X^* X^{*'}) X^* U_\alpha = \lambda_\alpha (X^* U_\alpha)$

$\Rightarrow \quad X^* U_\alpha$ est un vecteur propre de $X^* X^{*'}$

or V_α est un vecteur propre de $X^* X^{*'}$

$\Rightarrow \quad X^* U_\alpha$ et V_α sont colinéaires

$$\Rightarrow \begin{cases} V_\alpha = a_\alpha (X^* U_\alpha) \quad , \quad a_\alpha \in IR \\ \{\lambda_\alpha\} \subset \{\mu_\alpha\} \end{cases} \qquad \text{(a)}$$

L'ensemble des valeurs propres λ_α est un sous-ensemble des valeurs propres μ_α ;

D'une autre part

$\mu_1, \mu_2, ..., \mu_r$ les valeurs propres non nulles de la matrice W

$\Rightarrow (W - \mu_\alpha I) V_\alpha = \underline{0}$

$\Rightarrow W V_\alpha = \mu_\alpha V_\alpha$

$\Rightarrow X^* X^{*'} V_\alpha = \mu_\alpha V_\alpha$

$\Rightarrow X^{*'} (X^* X^{*'}) V_\alpha = X^{*'} (\mu_\alpha V_\alpha)$

$\Rightarrow (X^{*'} X^*) X^{*'} V_\alpha = \mu_\alpha (X^{*'} V_\alpha)$

$\Rightarrow X^{*'} V_\alpha$ est un vecteur propre de $X^{*'} X^*$

or U_α est un vecteur propre de $X^{*'} X^*$

$\Rightarrow X^{*'} V_\alpha$ et U_α sont colinéaires

$$\Rightarrow \begin{cases} U_\alpha = b_\alpha (X^{*'} V_\alpha) \quad , \quad b_\alpha \in IR \\ \{\mu_\alpha\} \subset \{\lambda_\alpha\} \end{cases} \qquad \text{(b)}$$

L'ensemble des valeurs propres μ_α est un sous-ensemble des valeurs propres λ_α ;

De (a) et (b) nous déduisons que $\lambda_\alpha = \mu_\alpha$

Les axes principaux d'inertie d'un nuage se déduisent de ceux de l'autre par

$$V_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} (X^* U_\alpha), \; U_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} (X^{*'} V_\alpha) \quad , \quad \alpha = 1, \ldots, r \qquad (37)$$

Effectivement, sachant que $V_\alpha = a_\alpha (X^* U_\alpha)$, $V_\alpha' V_\alpha = 1$ et que $X^{*'} X^* U_\alpha = \lambda_\alpha U_\alpha$,

nous pouvons démontrer que $a_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}$

$V_\alpha' V_\alpha = 1 \Rightarrow (a_\alpha X^* U_\alpha)' (a_\alpha X^* U_\alpha) = 1$

$\Rightarrow a_\alpha^2 U_\alpha' X^{*'} X^* U_\alpha = 1$

$\Rightarrow a_\alpha^2 \lambda_\alpha = 1$

$\Rightarrow a_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} = \frac{1}{\sqrt{\mu_\alpha}}$

$$\Rightarrow \quad V_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}(X^* U_\alpha)$$

D'une manière analogue nous pouvons démontrer que $U_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}(X^{*'} V_\alpha)$

2 - En analyse des correspondances, les valeurs propres sont comprises entre 0 et 1

$$(\ 0 \leq \ \lambda_\alpha \ \leq 1\ ). \tag{38}$$

SECTION 5 : FORMULES DE TRANSITION ET DE RECONSTITUTION

5.1. Les formules de transition

La formule de transition, dite aussi formule barycentrique, évite de faire l'analyse du nuage N(J) si celle du nuage N(I) est déjà faite, et réciproquement. Il suffit de déterminer les axes principaux d'inertie et les facteurs relatifs à l'un des nuages N(I) ou N(J) , les facteurs relatifs à l'autre nuage se calculent par les formules de transition.

Soient F_α les facteurs extraits de N(I) et G_α les facteurs extraits de N(J), on a les deux formules suivantes, dites de transition :

$$F_\alpha(i) \ = \ (\lambda_\alpha)^{-1/2} \sum_{j=1}^{p} f_j^{\,i}.G_\alpha(j), \tag{39}$$

$$G_\alpha(j) \ = \ (\lambda_\alpha)^{-1/2} \sum_{i=1}^{n} f_i^{\,j}.F_\alpha(i) \tag{40}$$

ou sous forme matricielle par :

$$F_\alpha \ = \ (\lambda_\alpha)^{-1/2}.R.G_\alpha \tag{41}$$

$$G_\alpha \ = \ (\lambda_\alpha)^{-1/2}.C^{'}.F_\alpha \tag{42}$$

$F_\alpha(i)$ apparaît, au coefficient $(\lambda_\alpha)^{-1/2}$ prés, comme la moyenne des $G_\alpha(j)$, pour j parcourant J , pondérée par les $f_j^{\,i}$. De même, $G_\alpha(j)$ apparaît, au coefficient $(\lambda_\alpha)^{-1/2}$ prés, comme la moyenne des $F_\alpha(i)$) , pour i parcourant I, pondérée par les $f_i^{\,j}$.

Ces deux expressions permettent le passage des facteurs sur I aux facteurs sur J et réciproquement, et manifestent la parfaite symétrie des rôles que jouent les deux ensembles I et J mis en correspondance par le tableau $K_{I,J}$. Ces deux relations de dualité sont la clé de la représentation graphique des deux nuages N(I) et N(J) sur le même graphique.

5.2. La formule de reconstitution

A partir du tableau $K_{I,J}$, on a pu construire les nuages N(I) et N(J). Ces nuages sont rapportés à leurs axes principaux d'inertie, et les éléments i de I et j de J ont comme coordonnées les facteurs $F_\alpha(i)$ et $G_\alpha(j)$.

Inversement, à partir des facteurs sur les ensembles I et J , des valeurs propres λ_α et des profils moyen f_I et f_J , nous pouvons reconstituer exactement le tableau initial en utilisant la formule suivante, dite formule de reconstitution des données en fonction des facteurs :

$$k(i,j)/k = f_{ij} = f_i f_j (1 + \sum_{\alpha=1}^{r} (\lambda_\alpha)^{-1/2} F_\alpha(i).G_\alpha(j)) \ , \qquad (43)$$

où r est le nombre de facteurs ou tout simplement le nombre de valeurs propres non nulles.

SECTION 6 : LES ELEMENTS SUPPLEMENTAIRES

6.1. Définition

Certains éléments de I et de J peuvent être mis en éléments supplémentaires pour une raison ou une autre : l'élément a par exemple perturbé une analyse antérieure ou il comporte des erreurs. Ces éléments dits éléments supplémentaires, figurent au tableau $K_{I,J}$ comme les autres éléments, dits éléments principaux, mais on les exclut des calculs des $k(i)$, des $k(j)$ et du total général k .

Si l'élément j_s de J est mis en élément supplémentaire, le total de la ligne i n'est plus k(i) mais k'(i) = k(i) - k(i , j_s) ; de même si l'élément i_s de I est mis en élément supplémentaire, le total de la colonne j n'est plus k(j) mais k'(j) = k(j) - k(i_s , j) ; le total général dans ce cas est k' défini par :

$$k' = \sum_{i \in I - \{i_s\}} \sum_{j \in J - \{j_s\}} k(i,j) \ . \qquad (44)$$

6.2. Les facteurs des éléments supplémentaires

Ecartés du calcul, les éléments supplémentaires ne servent pas à construire les axes factoriels. Mais il est utile de savoir leurs places par rapport aux autres éléments de l'ensemble, chose possible en projetant leurs profils sur ces axes.
Les coordonnées ou les facteurs relatifs à ces éléments supplémentaires sont calculés par la formule de transition.

Soit s un élément supplémentaire, $F_\alpha(s)$ et $G_\alpha(s)$ sont :

$$F_\alpha(s) = (\lambda_\alpha)^{-1/2} \sum_{j=1}^{p} f_j^s . G_\alpha(j), \quad (45)$$

$$G_\alpha(s) = (\lambda_\alpha)^{-1/2} \sum_{i=1}^{n} f_i^s . F_\alpha(i). \quad (46)$$

Les nouveaux individus introduits après avoir fait l'analyse sont considérés comme éléments supplémentaires. Pour situer ces individus par rapport aux autres déjà étudiés, nous calculons leurs facteurs sans refaire l'analyse.

Les nouvelles variables introduites peuvent également être placées en éléments supplémentaires pour être situées par rapport aux autres variables principales de J.

6.3. La reconstitution d'un élément supplémentaire

On peut utiliser la formule de reconstitution pour reconstituer un élément (ligne) supplémentaire s. Soit $k^h(s , j)$ la reconstitution de k(s , j) à l'ordre h .
$k^h(s , j)$ est définie par :

$$k^h(s, j) = k^{'} . f_s f_j (1 + \sum_{\alpha=1}^{h} (\lambda_\alpha)^{-1/2} F_\alpha(s).G_\alpha(j)) \quad (47)$$

où k' désigne la total général du tableau, calculé sans la ligne supplémentaire ; et $h \leq r$. Si $h = r$, la reconstitution tient compte de tous les facteurs existants.

Une formule analogue peut être donnée pour la reconstitution d'une colonne supplémentaire.

SECTION 7 : CALCUL DES CONTRIBUTIONS (CTR)

Avant de donner les formules de calcul qui servent à l'interprétation des résultats de l'analyse, on définit quelques notations statistiques indispensables au calcul.

7.1. Définitions

- Contribution absolue et contribution relative :

Si une somme S de nombres positifs t_u est donnée par l'expression : $S = \sum_{u \in U} t_u$; on dit que t_u , la part revenant à l'élément u , est la contribution absolue

de l'élément u à la somme S, et que le quotient $\frac{t_u}{S}$ est la contribution relative de l'élément u au total S , $\frac{t_u}{S}$ est compris entre 0 et 1 .

- Rayon polaire :

On appelle rayon polaire ρ(i) la distance au sens de χ^2 d'un élément du nuage N(I) au centre de gravité du nuage :

$$\rho^2(i) = d^2(f_J^i, f_J) = \| f_J^i - f_J \|^2_{f_J} = \sum_{\alpha \in A} F_\alpha^2(i), \qquad (48)$$

où A est l'ensemble des indices des valeurs propres.
Pour les éléments du nuage N(J) on a :

$$\rho^2(j) = d^2(f_I^j, f_I) = \| f_I^j - f_I \|^2_{f_I} = \sum_{\alpha \in A} G_\alpha^2(j). \qquad (49)$$

- Trace et taux d'inertie :

L'inertie totale par rapport à f_J du nuage N(I), appelée aussi trace, s'écrit inert $_{fJ}$ (N(I)), et est égale à la somme des valeurs propres.

$$\text{inert}_{fJ}(N(I)) = \text{Trace} = \lambda_1 + \lambda_2 + \cdots + \lambda_\alpha + \cdots + \lambda_r . \qquad (50)$$

La part relative de l'axe α à l'inertie totale du nuage est définie par le rapport (λ_α / Trace) . Ce rapport est appelé taux d'inertie de l'axe α , noté τ_α , et exprimé généralement en pourcentage.

$$\tau_\alpha = \lambda_\alpha / \text{Trace} . \qquad (51)$$

On peut aussi calculer la part relative ou la contribution relative d'un groupe d'axes à l'inertie totale. Par exemple, pour les deux axes α et β on a :

$$(\lambda_\alpha + \lambda_\beta) / \text{Trace} = \lambda_\alpha / \text{Trace} + \lambda_\beta / \text{Trace} = \tau_\alpha + \tau_\beta . \qquad (52)$$

Comme les axes sont rangés dans l'ordre des valeurs propres décroissantes :

$$\lambda_r < \lambda_{r-1} < \cdots < \lambda_\alpha < \cdots < \lambda_1 , \qquad (53)$$

alors $$\tau_r < \tau_{r-1} < \cdots < \tau_\alpha < \cdots < \tau_1 . \qquad (54)$$

7.2. La contribution relative d'un point à un axe (CTR)

La contribution relative d'un point i du nuage N(I) à l'axe α , notée $CTR_\alpha(i)$, donne la part de ce point dans l'inertie λ_α du nuage.

$$CTR_\alpha(i) = f_i.F_\alpha^2(i) / \lambda_\alpha \ , \qquad (55)$$

avec, comme on le sait selon la formule 33, $\lambda_\alpha = \sum_{i=1}^{n} f_i.(F_\alpha(i))^2$. Le terme $f_i.F_\alpha^2(i)$ est appelé la contribution absolue de i à l'axe α.

D'une façon analogue on définit la contribution relative d'un point j du nuage N(J) à l'axe α :

$$CTR_\alpha(j) = f_j.G_\alpha^2(j) / \lambda_\alpha \ , \qquad (56)$$

et la contribution absolue de j à l'axe α est $f_j.G_\alpha^2(j)$.

7.3. La contribution relative d'un axe à l'écart d'un point au centre (COR)

Cette contribution nous permet de savoir de quels axes un point de N(I) ou de N(J) est plus voisin. En d'autres termes, quels facteurs expliquent la position du point relativement au point moyen. Le point moyen est représenté par le centre de gravité du nuage qui est l'origine des axes factoriels.

La formule de définition de COR pour un point i du nuage N(I) est :

$$COR_\alpha(i) = F_\alpha^2(i) \ / \ \rho^2(i) \ . \qquad (57)$$

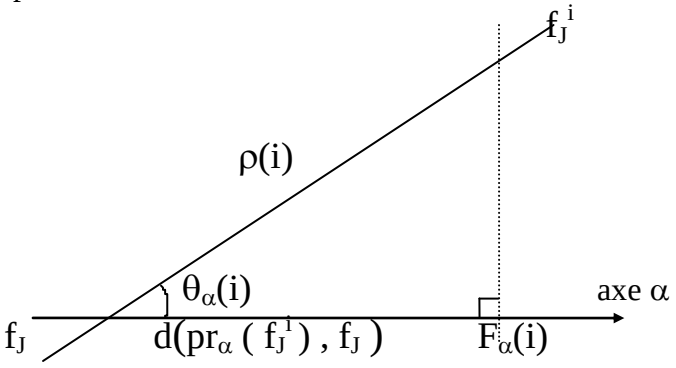
De même, pour un point j de N(J) , la formule est :

$$COR_\alpha(j) = G_\alpha^2(j) \ / \ \rho^2(j) \qquad (58)$$

Pour que les $\{ COR_\alpha(i) \ , \alpha \in A \}$ soient définis, il faut que les facteurs ne soient pas tous nuls. Il est à signalé aussi que $COR_\alpha(i)$ s'interprète comme le carré d'un coefficient de corrélation, et qu'il est compris entre 0 et 1

$$0 \leq COR_\alpha(i) \leq 1 \ . \qquad (59)$$

Géométriquement, $COR_\alpha(i)$ est un cosinus carré. Ceci peut être visualisé en considérant le plan suivant :



$$\cos(\theta_\alpha(i)) = |F_\alpha(i)| / d(f_J^i, f_J)$$

$$\cos^2(\theta_\alpha(i)) = F_\alpha^2(i) / d^2(f_J^i, f_J)$$
$$= F_\alpha^2(i) / (F_1^2(i) + ... + F_\alpha^2(i) + ... + F_r^2(i))$$
$$= F_\alpha^2(i) / \rho^2(i)$$
$$= COR_\alpha(i) .$$

si l'angle $\theta_\alpha(i) = 0$, c.à.d. si le point f_J^i est sur l'axe α , $COR_\alpha(i) = 1$. Dans ce cas on dit que l'axe α explique à lui seul l'écart de i au centre.

si $\theta_\alpha(i) = 90°$, alors $COR_\alpha(i) = 0$. On dit alors que l'axe est étranger à l'écart du point i au centre.

7.4. La qualité de la représentation d'un point sur un sous-espace

Pour tout élément i de I, les facteurs sont les coordonnées du profil f_J^i sur les r axes principaux d'inertie. Ils décrivent la projection de f_J^i sur le sous-espace engendré par ces axes. Notons cette projection $pr_{1,...r}(f_J^i)$.

La qualité de la représentation du point f_J^i par sa projection sur le sous- espace engendré par les axes factoriels α , β et γ est mesurée par le rapport :

$$QLT(i) = d^2(pr_{\alpha,\beta,\gamma}(f_J^i), f_J) / d^2(f_J^i, f_J) \leq 1 . \quad (60)$$

or $$d^2(pr_{\alpha,\beta,\gamma}(f_J^i), f_J) = F_\alpha^2(i) + F_\beta^2(i) + F_\gamma^2(i) ,$$

donc $$QLT(i) = F_\alpha^2(i)/d^2(f_J^i,f_J) + F_\beta^2(i)/d^2(f_J^i,f_J) + F_\gamma^2(i)/d^2(f_J^i,f_J)$$

$$QLT(i) = COR_\alpha(i) + COR_\beta(i) + COR_\gamma(i) . \quad (61)$$

La représentation sera d'autant meilleure que f_J^i sera plus proche de sa projection et QLT plus voisin de 1, elle est parfaite quand QLT(i) = 1. QLT(i) = 1 quand le sous-espace est engendré par les r axes principaux d'inertie, car

$$d^2(pr_{1,...,r}(f_J^i), f_J) = F_1^2(i) + F_2^2(i) + \cdots + F_r^2(i) = d^2(f_J^i, f_J). \quad (62)$$

De même, pour tout élément j de J on a :

$$QLT(j) = d^2(pr_{\alpha,\beta,\gamma}(f_I^j), f_I) / d^2(f_I^j, f_I). \quad (63)$$
$$QLT(j) = COR_\alpha(j) + COR_\beta(j) + ... + COR_\gamma(j) . \quad (64)$$

En conclusion, on note qu'il existe aujourd'hui des programmes rapides et pas très complexes qui permettent d'effectuer tous les développements mathématiques de ce chapitre. Les valeurs propres, les facteurs, les axes factoriels, les contributions, les corrélations, les qualités et tout autre calcul indispensable peuvent être donnés

par l'ordinateur, le chercheur intervient en dernier lieu pour interpréter les résultats.

SECTION 8 : APPLICATION SUR LES RESSOURCES FINANCIERES DES COMMUNES RURALES ORIENTALES MAROCAINES POUR L'EXERCICE 1998/1999

8.1. Le tableau des données

Le tableau analysé dans ce point est un tableau $K_{91,10}$, qui comprend 91 lignes et 10 colonnes. Les lignes représentent les noms des communes, alors que les colonnes représentent les catégories de ressources locales, regroupées en dix catégories. On a ainsi:

Les individus : 91 communes, numérotées de 1 à 91.
Les variables : 10 catégories qui sont la taxe urbaine, la taxe d'édilité, la patente, le produit du domaine forestier, les impôts et taxes assimilées, le produit des services, le produit et revenu des biens, les concessions, les subventions et concours et enfin les recettes d'ordre.

L'intersection d'une ligne et d'une colonne du tableau donne le nombre k(i,j) qui représente le montant, en dirhams, de la catégorie j correspondant à la commune i. Par exemple, à l'intersection de la ligne 8 et de la colonne 3 , on lit la valeur 225157 qui représente, en dirhams, le montant de la patente récolté par la commune GAFAÏT durant l'exercice1998/1999.

8.2. L'analyse

A l'aide du logiciel SPAD et du programme que nous avons implémenté sur R Gui, nous avons obtenu le tableau des valeurs propres, et c'est à partir de ce tableau que commence l'analyse. La table1 et la figure1 nous guident dans le choix du nombre de facteurs et des plans susceptibles d'être interprétés. Nous voyons clairement que la première valeur propre est nettement supérieure aux autres et que le plan 1x2 résume une part importante d'informations.

DIMENSION	VALEUR PROPRE	POURCENTAGE	POURCENTAGE CUMULE
1	.45303	.573	.573
2	.11687	.148	.720
3	.07821	.099	.819
4	.04910	.062	.881
5	.03138	.040	.921
6	.01940	.025	.945
7	.01796	.023	.968
8	.01708	.022	.990
9	.00821	.010	1.000
Total	.79124	1.000	1.000

Table 1

La figure 3 présente l'histogramme des valeurs propres

Figure 1

Comme on a 9 valeurs propres, l'espace des profils sur J où se trouve le nuage N(I) est de dimension 9 qui est égal à (card (I) - 1) .

$$\begin{aligned} 9 &= \text{Min (card (I) - 1 , card (J) - 1)} \\ &= \text{Min (91 - 1 , 10 - 1)} \\ &= \text{Min (90 , 9).} \end{aligned}$$

Les valeurs propres sont toutes inférieures à 1 et sont classées dans un ordre décroissant. En général, le nombre de chiffre après la virgule est grand pour distinguer les petites valeurs propres les unes des autres. Dans notre cas le chiffre est égal à 5.

Le premier axe factoriel associé à la première valeur propre $\lambda_1 = 0.45303$ explique 57.3% de l'inertie totale. Le deuxième axe factoriel explique 14.8% de cette inertie. Le plan 1x2 explique 72% de l'inertie totale, c.à.d. presque 72 % des informations se trouvent dans le plan 1x2. Il faut à présent expliquer ces axes.

L'axe 1:

Les individus qui contribuent le plus à l'inertie de l'axe 1 sont décrits dans la table2 qui résume cinq types d'informations :

Individu	C.P	Commune	$CTR_1(i)$ en %	$F_1(i)$	$COR_1(i)$
54	27	AIN LEHJER	42.2	-1.12	0.963
84	56	SELOUANE	13.5	-0.960	0.840
16	26	LAAOUINATE	7.3	-0.750	0.751
85	56	AREKMANE	1.7	0.580	0.870
89	25	BNI-CHIKER	1.3	0.550	0.662
62	49	TENDRARA	0.9	0.680	0.148
91	56	DRIOUCH	0.9	0.440	0.291
-	-	Total	67.8	-	-

Table 2

où

Individu : L'ordre de la commune parmi 91 communes de la région orientale classées par ordre croissant de la population.

C.P. : Le code de la préfecture ou la province à laquelle appartient la commune.

Commune : Le non de la commune.

$CTR_1(i)$: La contribution relative d'un point i du nuage N(I) à l'axe 1, donnée par la formule (55).

$F_1(i)$: La projection de la ligne i de X sur le premier axe factoriel, donnée par la formule (39).

$COR_1(i)$: La contribution qui permet de savoir si le point i est proche ou non du premier axe, donnée par la formule (57).

Le premier axe qui est expliqué par six individus, oppose les communes d'AIN LEHJER(54), SELOUNE(84) et LAAOUINATE(16) aux communes d'AREKMANE(85), BNI-CHIKER(89), TENDRARA(62) et DRIOUCH(91). Les communes d'AIN LEHJER, SELOUNE et LAAOUINATE, ont de fortes corrélations, elles sont par conséquent bien représentées. Généralement, les individus contribuant bien à l'axe, sont bien représentés sur celui-ci.

Nous remarquons que les trois communes d'AIN LEHJER(54), SELOUNE(84) et LAAOUINATE(16) ont dégagé un montant de recettes d'origine fiscal beaucoup plus important que les autres (selon la base de données). En même temps, se sont les communes qui ont eu les montants les plus élevés de subventions et concours.

Les variables qui contribuent le plus à l'inertie de l'axe 1 sont dressées dans la table3.

Variable	$CTR_1(j)$ en %	$G_1(j)$	$COR_1(j)$
Subv.concours	24	0.41	0.928
Imp.tax.ass	23.3	-1.00	0.767
Tax.urbaine	21.7	-1.49	0.845
Patente	15.4	-0.80	0.461
Tax.édilité	11.3	-1.25	0.784
Total	95.7	-	-

Table 3

où les colonnes représentent :

Variable	:	Le sigle de la catégorie de recettes.
$CTR_1(j)$	:	La contribution relative d'un point j du nuage N(J) à l'axe 1, donnée par la formule (56).
$G_1(j)$	:	La projection de la colonne j de X sur le premier axe factoriel, donnée par la formule (40)
$COR_1(j)$	:	La contribution qui permet de savoir si le point j est proche ou non du premier axe, donnée par la formule (58).

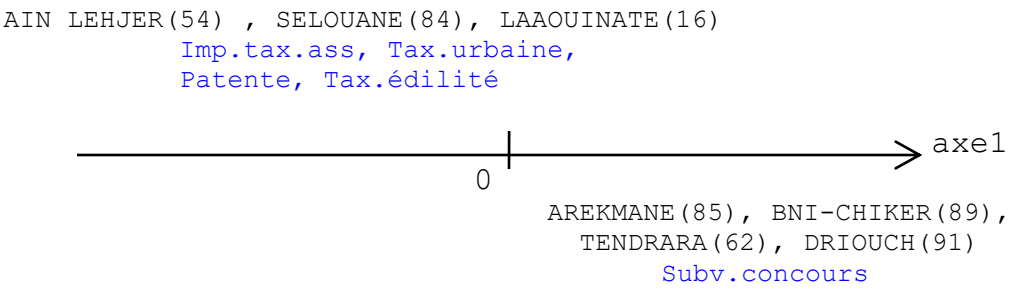
Le premier axe factoriel est expliqué par les cinq variables Subv.concours, Imp.tax.ass, Tax.urbaine, Patente et Tax.édilité. Il oppose les subventions et concours à la fiscalité locale. Par conséquent, l'axe peut être appelé "axe de subvention" , "axe de fiscalité" ou "axe d'autonomie financière". Puisque, plus une variable est corrélée avec un axe, plus elle est importante pour le décrire, à part la patente qui a une corrélation moyenne, ces variables jouent un rôle primordial dans la description de l'axe.

L'opposition de ces deux catégories de ressources est évidente, car moins les communes ont de ressources propres plus elles ont besoin de subventions de fonctionnement pour faire face à leurs dépenses de fonctionnement et de subventions d'équipement pour financer leurs projets d'investissement. Autrement exprimé, la commune qui souffre de la faiblesse de ses ressources propres compte énormément sur la subvention de l'Etat. Mais conscient de cette dépendance du financement de la commune des ressources étatiques et dans le but de la minimiser, l'Etat a décidé depuis 1997 d'introduire les ressources propres comme critère d'octroi de la subvention. Bien entendu, un important effort fiscal attire un montant de subventions important.

Certes, la préoccupation majeure d'une collectivité qui souffre de la faiblesse de ses ressources propres est d'équilibrer son budget de fonctionnement qui absorbe une bonne part de la subvention de l'Etat. C'est la raison pour laquelle les projets d'équipement sont presque absents dans les collectivités pauvres. Or, une collectivité riche peut facilement équilibrer son budget de fonctionnement et réserver une bonne part de la subvention de l'Etat au budget d'équipement.

D'une façon générale, le premier axe factoriel indique que la fiscalité occupe une part importante dans les recettes d'AIN LEHJER(54), SELOUANE(84) et de LAAOUINATE(16) contrairement aux communes d'AREKMANE(85), BNI-CHIKER(89), TENDRARA(62) et DRIOUCH(91). Cet axe peut être appelé, comme déjà cité, "axe d'autonomie financière" car il oppose les communes d'AIN LEHJER(54) avec un taux d'autonomie de 1040%, SELOUANE(84) avec un taux de 108% et LAAOUINATE(16) avec un taux de 449% qui ont réalisées leurs autonomie financière à celles qui n'ont pas pu la réalisée (moins de 100%). En outre, les communes autonomes ont reçus également un montant de subventions très élevé car leurs ressources propres sont intéressantes.

Nous déduisons alors, que le nouveau système de répartition de la T.V.A. sous forme de subventions a accentué l'écart entre les communes riches et les communes pauvres et a détérioré la situation de ces dernières. Le problème qui se pose est que les projets d'équipement sont presque absents dans plusieurs communes autonomes.
L'axe 1 peut être schématisé par



L'axe 2:

Le deuxième axe associé à la deuxième valeur propre λ_2 = 0.11687 est dominé par les commune IKSANE(50), TIOULI(39), SELOUANE(84), LAAOUINATE(16) et RISLANE(27) qui sont toutes des communes qui ont réalisé leurs autonomies financières. La table4 résume les contributions, les composantes et les corrélations de ces principales communes avec cet axe.

Individu	C.P	Commune	$CTR_2(i)$ en %	$F_2(i)$	$COR_2(i)$
50	56	IKSANE	49.2	-1.14	0.932
39	26	TIOULI	14.9	-0.99	0.885
84	56	SELOUANE	06.8	0.35	0.110
16	26	LAAOUINATE	05.5	-0.33	0.145
27	25	RISLANE	05.2	0.59	0.236
-	-	Total	81.6	-	-

Table 4

L'axe oppose principalement les commune d'IKSANE(50), TIOULI(39) et LAAOUINATE(16) à celle de SELOUANE(84) et RISLANE(27). Les communes de IKSANE(50) et TIOULI(39) qui ont une grande contribution à l'inertie de l'axe ont également une grande corrélation avec celui-ci. Les points correspondant sont par conséquent très voisins de l'axe

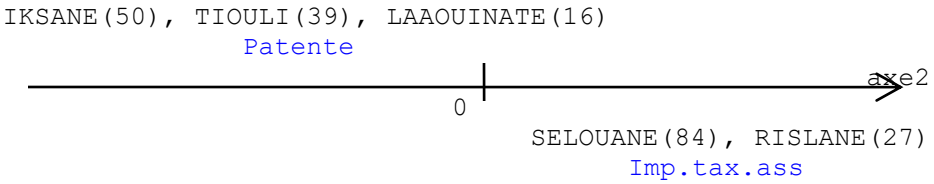
Les variables qui contribuent le plus à l'inertie du deuxième axe factoriel et qui dominent l'axe sont la Patente et les Imp.tax.ass.

Variables	$CTR_2(j)$ en %	$G_2(j)$	$COR_2(j)$
Patente	64.4	-0.84	0.497
Imp.tax.ass	16.7	0.43	0.141
Total	81.1	-	-

Table 5

Les communes d'IKSANE(50), TIOULI(39) et LAAOUINATE(16) ont dégagé un important montant de la patente évalué respectivement à 45.7%, 29% et 42% et un faible montant des impôts et taxes assimilées évalué respectivement à près de 0.02%, 19% et 2% des recettes de fonctionnement. Ceci justifie l'association de ces communes avec la patente. La part importante de la patente a permis la réalisation des taux d'autonomies respectives de 332%, 449% et 132%. En revanche, SELOUANE(84) et RISLANE(27) ont atteint des taux d'autonomie de 108% et 101% grâce surtout à la part de 28% et 47% des impôts et taxes assimilées dans les recettes de fonctionnement. De ces dernières, la patente ne constitue que 15% et 0.02% .

Nous schématisons l'axe2 par



Le plan 1x2 :

Le plan est représenté par la figure2 ci-dessous qui dresse les points des nuages N(I) et N(J). Nous allons décrire le résultat selon les quatre quadrants du plan.

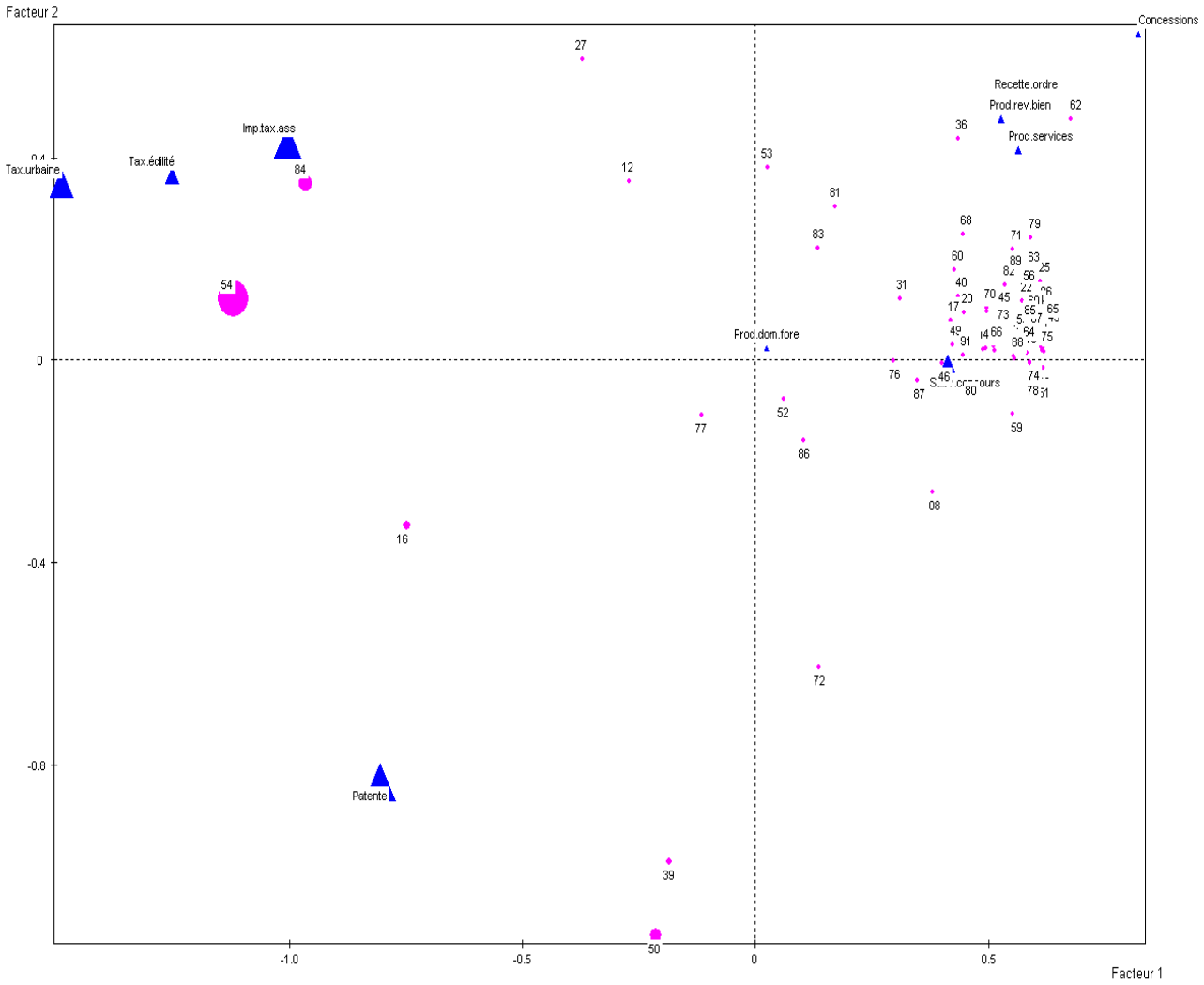
Le quadrant (- , -) marque l'association des communes autonomes d'IKSANE(50), TIOULI(39) et LAAOUINATE(16) avec la patente. L'association est due, comme déjà expliqué, au rôle que joue cette catégorie dans les ressources de fonctionnement de ces communes. Ce rôle est important mais aucune des communes ne dépend intégralement de la patente, c'est pourquoi leurs points représentatifs se trouvent un peu loin du point représentant la patente.

Le quadrant (- , +) nous indique que les impôts et taxes assimilées, la taxe urbaine et la taxe d'édilité sont déterminantes dans les recettes de fonctionnement d'AIN LEHJER(54), SELOUANE(84), RISLANE(27) et SIDI BOUBKER(12). Plus le montant de ce type de recettes fiscales est élevé plus les ressources propres sont élevées et plus la communes dispose du financement pour faire face aux dépenses de fonctionnement. En effet, à part la commune de SIDI BOUBKER(12), cette recette fiscale a aidé ces communes à atteindre l'autonomie financière.

Dans le quadrant (+ , -) et (+ , +) nous notons l'existence d'un nuage de points représentants les communes concentré autour du point représentant la subvention. Ce sont des communes qui souffrent de problèmes de financement et qui dépendent fortement du concours de l'Etat. Le produit du domaine forestier est mal représenté car il très proche de zéro.

Globalement, L'analyse a marqué l'association des communes autonomes d'IKSANE(50), TIOULI(39) et LAAOUINATE(16) avec la patente et celle d'AIN LEHJER(54), SELOUANE(84), RISLANE(27) avec les impôts et taxes assimilées, la taxe urbaine et la taxe d'édilité. En fait, le rôle de ces catégories est important mais aucune des communes n'en dépend intégralement. En revanche, la plupart des points représentants les communes rurales orientales se concentrent autour du point représentant la subvention. Ce sont des entités qui souffrent de problèmes de financement et qui dépendent fortement du concours de l'Etat

Figure 2 : Représentation graphique des communes rurales orientales marocaines et des différentes catégories de recettes de fonctionnement

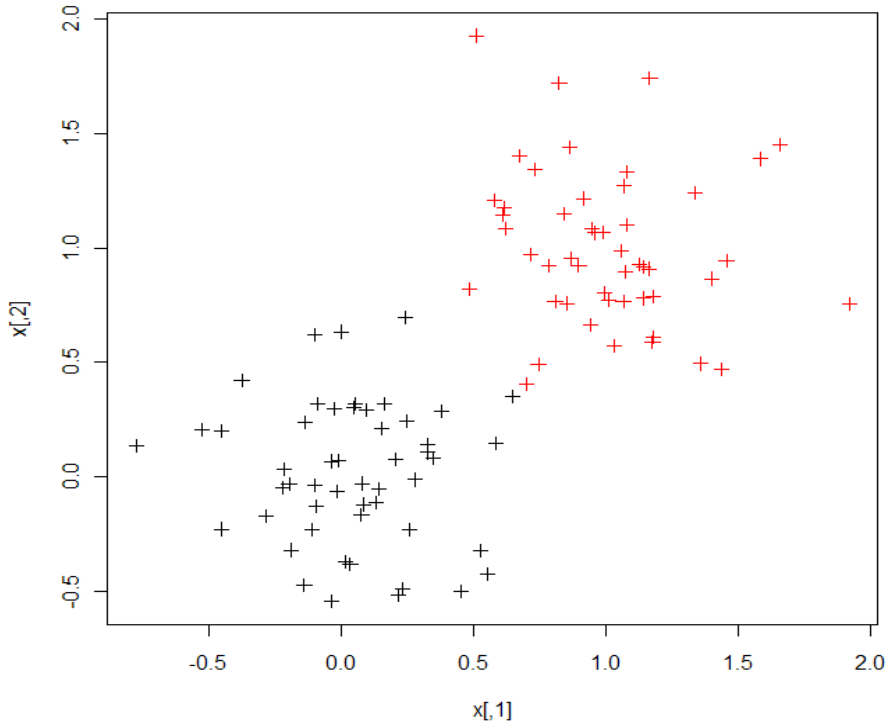


Chapitre 4 **Classification hiérarchique (C.H.)**

1. Le principe général

La classification est un problème de construction des groupes (clusters) à partir de données multivariées. Son but est de former des groupes homogènes selon un certain critère (la distance à titre d'exemple) telle que la différence entre les groupes est la plus grande possible. Il faut donc, classer des objets dans des classes homogènes. Ainsi l'analyse de classification peut être effectuée dans différents domaines notamment la psychologie, la médecine, la biologie, l'industrie et la finance.

Pour la base de données représentée par le graphique ci-dessous, on remarque bien qu'on peut distinguer entre deux groupes, mais le problème qui se pose c'est comment placer les points dans telle ou telle classe ?



La classification des objets en classes dépend de deux choix :

- Le choix de la distance entre les objets.
- Le choix de l'algorithme de construction des groupes. L'idée de base est de construire, en se basant sur le critère de la distance, les groupes de telle sorte que la différence dans le groupe est petite et la différence entre groupes est grande.

Lorsque le nombre de classes k n'est pas spécifié d'avance, un grand problème de spécification se présente c'est celui de la détermination de k. Si l'on connaît le nombre de classes à constituer, la classification est dite une classification supervisée, sinon elle dite non-supervisée. Généralement, le dernier cas est plus utilisé dans le cadre de l'analyse des données.

Par ailleurs, plusieurs méthodes de clustering sont disponibles dans la littérature et la méthode qui va être présentée dans ce chapitre est la méthode de classification ascendante.

2. Notions de Classifications Ascendantes Hiérarchique (C.A.H.)

a. Classification

Selon les représentations géométriques propres à l'analyse factorielles des correspondances par exemple, l'ensemble I est identifié par un nuage $N(I)$ qui est l'ensemble des points munis des masses dans l'espace euclidien des profils sur J où la distance est celle de Chi-2 de centre f_J.

Faire une classification sur I c'est édifier un système de classes ou parties sur I d'après la représentation géométrique.

b. Hiérarchie et partition

La partition est la forme la plus simple de la classification. On partage I en un système de classes non vides, de telle sorte que tout individu i appartienne à une classe et une seule. La classification sert aussi à désigner un système emboîté ou une hiérarchie de classes. A titre d'exemple, en sciences naturelles les êtres vivants sont partagés en deux grandes règnes, animal et végétal, chacun est divisé en embranchement, ainsi les animaux sont partagés en vertébrés, mollusques, arthropodes, … ; les vertébrés sont à leur tour subdivisés en classes (mammifères, oiseaux, reptiles, …). Cette classification est appelée classification hiérarchique ou hiérarchie de classes.

c. Classification descendante et classification ascendante

Contrairement à la classification ascendante, la classification descendante part du sommet jusqu'à la constitution de classes avec un seul élément. Le nœud qui prend le numéro r, où $r = 2.Card(I) - 1$, se scinde en deux descendants immédiats A(r) et B(r), le nœud A(r) se scinde à son tour en deux et ainsi de suite jusqu'à la formation de classes d'un seul élément.

La classification ascendante, quant à elle, elle part de la base jusqu'à la constitution de la dernière classe. Si on dispose de n individus à titre d'exemple, les deux individus i et i' s'agrège pour former la classe $n+1$, et on itère jusqu'à former le dernier nœud.

En résumé

- Un algorithme descendant part du tout qu'il scinde en deux classes ; puis il scinde chacune de ces deux classes en deux et ainsi de suite jusqu'à isoler les individus.
- Un algorithme ascendant part des individus et d'un critère de ressemblance des individus qui s'étend aux classes, agrège les individus qui se ressemblent le plus ;

puis il agrège soit deux autre individus soit un individu et une classe, puis des classes entre elles, créant ainsi des nœuds.

3. Critères d'agrégation et algorithme de CAH

3.1. Le critère d'agrégation

Il existe plusieurs critères simples d'agrégation dont la définition part d'une distance entre points. Ces critères sont : le critère de saut minimum $D_{saut}(C,C')$; le critère du diamètre $D_{diam}(C,C')$; le critère de la distance moyenne $D_{moy}(C,C')$ et le critère de l'inertie $D_{inert}(C,C')$, avec C et C' deux partie finies quelconques de l'ensemble I à classer.

Il est à noter que parmi les 4 critères cités seul D_{moy} est une véritable distance[2].

a/ Critère du saut minimum

$D_{saut}(C,C')$ est la distance minimale entre un point de la classe C et un point de la classe C'. Autrement exprimé, c'est la distance entre deux points i et i' appartenant l'un à C et l'autre à C' qui sont les plus proches possibles.

$$D_{saut}(C,C') = \min_{\substack{i\in C \\ i'\in C'}} D(i,i')$$

si par exemple on trouve que $\min D$ est entre les éléments 4 et 17, alors ces deux éléments forment une classe.

b/ Critère du diamètre

Le critère du diamètre est la distance maximale entre $i \in C$ et $i' \in C'$

$$D_{diam}(C,C') = \max_{\substack{i\in C \\ i'\in C'}} D(i,i')$$

c/ critère de la distance moyenne

$D_{moy}(C,C')$ est la moyenne des distances séparant $i \in C$ et $i' \in C'$, chaque $D(i,i')$ a pour poids le produit $f_i \cdot f_{i'}$.

$$D_{moy}(C,C') = \frac{1}{f_C \cdot f_{C'}} \sum_{\substack{i\in C \\ i'\in C'}} f_i \cdot f_{i'} D(i,i')$$

---

[2] La distance métrique doit verifier les axioms suivantes:

1/ La symétrie : $D(i,i') = D(i',i')$

2/ La positivité stricte : $D(i,i') > 0$, si $i \neq i'$

$D(i,i') = 0$ si $i = i'$

3/ L'inégalité du triangle : quels que soient les trois points i, i' et i''

$$D(i,i'') \leq D(i,i') + D(i',i'').$$

où f_C et $f_{C'}$ sont les masses totales respectivement des classes C et C' tel que

$$f_C = \sum_{i \in C} f_i \quad \text{et} \quad f_{C'} = \sum_{i' \in C'} f_{i'} ,$$

D_{moy} prend en considération les distances minimales et les distances maximales entre les points de C et de C'. Elle tient compte alors de D_{saut} et D_{diam}.

d/ Critère de l'inertie

On associe C et C' à leurs centres de gravités qui sont notes, simplement, C et C'

$$D_{inert}(C, C') = \frac{f_C \cdot f_{C'}}{f_C + f_{C'}} \left\| C - C' \right\|^2$$

avec f_C et $f_{C'}$ sont les masses totales de C et C' définies précédemment,

$\left\| C - C' \right\|^2$ est le carré de la distance euclidienne entre les centres de gravité des classes C et C'.

$$\left\| C - C' \right\|^2 = \sum_{j}^{p} \left(C_j - C'_j \right)^2$$

$$\begin{array}{cc} & \\ C & \left(\begin{array}{ccc} \vdots & \vdots & \vdots \end{array} \right) \} f_C \\ C' & \left(\begin{array}{ccc} \vdots & \vdots & \vdots \end{array} \right) \} f_{C'} \end{array}$$

On agrège les classes C et C' qui donnent le $\min D_{inert}$.

Il est à noter que le critère le plus utilisé parmi les critères cités est celui de l'inertie.

3.2. Algorithme de la CAH

Etape1 : On part de la partition la plus fine de I dont chaque classe est formée d'un seul individu i, qu'on numérote de 1 à $Card(I)$, et on calcul $D(i, i')$ pour tout couple $\{i\}$ et $\{i'\}$. On a donc $C^2_{Card(I)} = \frac{Card(I).(Card(I)-1)}{2}$ distances à calculer.

On agrège, alors, le couple (i, i') qui donne le $\min D(i, i')$ pour créer le nœud 'n' constitué de i et i' avec $A(n) = i$ et $B(n) = i'$. Ce nœud va porter le numéro $Card(I)+1$.

On a, donc, une nouvelle partition de I formée de $(Card(I)-1)$ classes.

Généralement de chaque nœud partent deux branches. Ainsi, du nœud $(Card(I)+1)$ part deux branches, l'une est $A(Card(I)+1) = i$ disposée à gauche (A est l'initiale d'aîné), et l'autre est $B(Card(I)+1) = i'$ disposée à

droite (B est l'initiale de benjamin). $A(Card(I)+1)$ et $B(Card(I)+1)$ sont les deux descendants immédiats du nœud $(Card(I)+1)$.

Etape2 : On calcule les écarts $D(n,i^{"})$ pour tout $i^{"} \neq i$ et $i^{"} \neq i^{'}$.

On a $(Card(I)-2)$ distances à calculer car les autres distances entre les classes d'un seul individu sont déjà calculées.

On agrège la paire réalisant le $\min D$, qui donne naissance à un nouveau nœud ; il reçoit le numéro $(Card(I)+2)$.

La nouvelle partition de I est formée de $(Card(I)-2)$ classes.

Etape3 : On itère jusqu'à l'obtention d'un seul sommet qui va prendre le N° $(2.Card(I)-1)$ et qui n'est autre que l'ensemble I tout entier.

Remarque

Il existe d'autres algorithmes accélérés qui agrègent dans la même étape plusieurs paires à la fois et ils ne donnent que rarement le même résultat.

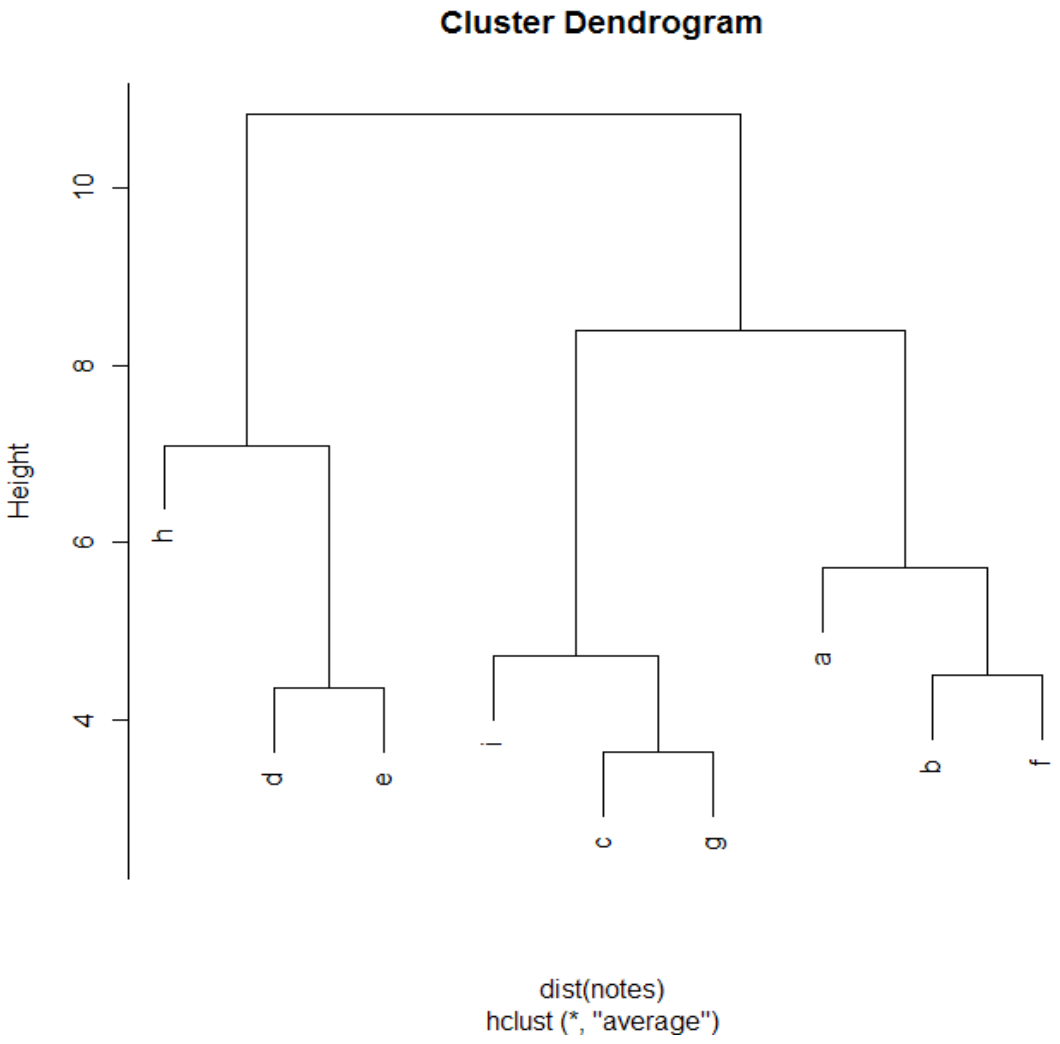
3.3. Le dondrogramme

L'idée du dondrogramme est de construire des séquences de partitions à partir des partitions les plus fines, où on a n classes contenant chacune un seul élément. La procédure, qui est celle décrite par l'algorithme précédent, part, étape par étape, de n classes vers $n-1$, $n-2$, … jusqu'à avoir un sommet d'une seule classe. Le dondrogramme est, donc, la représentation graphique de ces séquences, il est dit également 'l'arbre de calssification'.

Par ailleurs, sur le dondrogramme, la hiérarchie peut figurer, sur un des axes, la valeur de la distance minimale entre deux classes. Cette valeur (indice) indique le 'niveau d'agrégatiion' de l'étape. Bien évidemment, un niveau bas de l'indice indique qu'à ce niveau des goupes plus homogènes sont regroupés, un niveau élevé de l'indice indique que des goupes hétérogènes sont regroupés.

L'arbre est coupé à un certain niveau δ et le nombre de groupes à former est défini à ce niveau. Il est clair que, une fois que le niveau change, le nombre de groupes peut changer. Globalement, la statistique descriptive des sous-groupes aide à valider le choix de δ, mais il n'y a pas de méthodes précises qui aident à décider où couper l'arbre.

Pour l'exemple des notes des étudiants cité précédemment, le dondrogramme se présente comme suit :



Chapitre 5 **Analyse canonique (AC)**

Le principe de l'analyse canonique est de mettre en évidence des proximités entre deux ensembles de données et de décrire ces proximités entre les variables de ces deux ensembles. Cette description nécessite la détermination des composantes canoniques.

1. La base de données

Dans l'analyse canonique deux tableaux de données sont traités simultanément. Ces données sont dressées de telle sorte que

- Le tableau 1, noté X_1, comporte n lignes et m_1 colonnes où chaque ligne i représente l'individu i et chaque colonne j représente une variable quantitative centrée ou une modalité d'une variable qualitative.
- Le tableau 2, noté X_2, comporte n lignes et m_2 colonnes où chaque ligne i représente l'individu i et chaque colonne j représente une variable quantitative centrée ou une modalité d'une variable qualitative.
  Pour chaque tableau, les colonnes sont supposées linéairement indépendantes.

2. L'algorithme de l'analyse canonique et les composantes canoniques

Etape1 : Déterminer un couple de variables canoniques $\left(z_1^1, z_2^1\right)$ tel que

$$Arg\max R^2\left(z_1^1, z_2^1\right)$$

$$sc \quad \begin{cases} Var\left(z_1^1\right)=1 \\ Var\left(z_2^1\right)=1 \end{cases}$$

où R est le coefficient de détermination défini par

$$R^2\left(z_1^1, z_2^1\right)=\frac{Cov^2\left(z_1^1, z_2^1\right)}{Var\left(z_1^1\right)Var\left(z_2^1\right)}$$

Si on définit P_1 (respectivement P_2) comme étant la projection orthogonale de des points sur l'espace engendré par la colonnes de X_1 (respectivement X_2), on peut dire donc que

z_1^1 (z_2^1) est le premier vecteur propre de P_1P_2 (P_2P_1)

Par conséquent

- z_1^1 est une combinaison linéaire des variables du tableau X_1,
- z_2^1 est une combinaison linéaire des variables du tableau X_2.

Selon l'algèbre linéaire, les projections orthogonales P_1 et P_2 sont définies par

$$P_i = X_i\left(X_i{}' X_i\right)^{-1} X_i{}'$$

Les vecteurs propres z_1^1 et z_2^1 sont associés à la même valeur propre qui est égale au coefficient de détermination $R^2\left(z_1^1, z_2^1\right)$

- z_1^1 est la première composante canonique du tableau X_1,
- z_2^1 est la première composante canonique du tableau X_2,

Etape2 : On itère jusqu'à la détermination des $k^{ème}$ composante canonique du tableau X_1 et de X_2 qui sont notées respectivement z_1^k et z_2^k par

$$Arg \max R^2\left(z_1^k, z_2^k\right)$$

$$sc \quad \begin{cases} Var\left(z_1^k\right) = Var\left(z_2^k\right) = 1 \\ R^2\left(z_1^r, z_1^k\right) = R^2\left(z_2^r, z_2^k\right) = 0, \quad \forall r < k \end{cases}$$

Deux remarques importantes sont à signaler

a. Les composantes canoniques d'un même tableau sont donc deux à deux non corrélées.

b. La composante canonique d'ordre k d'un tableau est non corrélée avec les composantes canoniques d'ordre différent de k de l'autre tableau.

3. Les facteurs

A la $k^{ème}$ étape : z_i^k est une combinaison linéaire des variables du tableau i (i = 1; 2), doù

$$z_i^k = X_i a_i^k \quad , \text{i=1,2}$$

où a_1^k et a_2^k sont les facteurs d'ordre k.

Les facteurs a_1^k sont solutions de

$$V_{11}^{-1} V_{12} V_{22}^{-1} V_{21} . a_1^k = R^2\left(z_1^k, z_2^k\right) a_1^k$$

$$\text{où } V_{ij} = \frac{1}{n}\left(X_i^{'}\right) X_j$$

- Dans le cas où les deux variables sont quantitatives, V_{ij} est la matrice des covariances entre les variables du tableau i et celles du tableau j.
- Dans le cas où les deux variables sont qualitatives, V_{ij} est la matrice des fréquences relatives des variables du tableau i et celles du tableau j.

Les a_1^k sont, donc, les vecteurs propres de $V_{11}^{-1} V_{12} V_{22}^{-1} V_{21}$ et $R^2\left(z_1^k, z_2^k\right)$ les valeurs propres de la même matrice

De la même façon on définit les facteurs a_2^k sont solutions de

$$V_{22}^{-1} V_{21} V_{11}^{-1} V_{12} . a_2^k = R^2\left(z_1^k, z_2^k\right) a_2^k$$

Les a_2^k sont, donc, les vecteurs propres de $V_{22}{}^{-1}V_{21}V_{11}{}^{-1}V_{12}$ et $R^2\left(z_1^k, z_2^k\right)$ les valeurs propres de la même matrice

4. La relation entre a_1^k et a_2^k

$$V_{11}{}^{-1}V_{12}.a_2^k = R\left(z_1^k, z_2^k\right)a_1^k$$
$$V_{22}{}^{-1}V_{21}.a_1^k = R\left(z_1^k, z_2^k\right)a_2^k$$

5. Comment procède-t-on en pratique ?

- On calcule la matrice des corrélations
- On extrait de la matrice des corrélations les matrices dont nous allons avoir besoin par la suite qui sont V_{11} , V_{22} , V_{12} et V_{21}.
- On calcule alors les facteurs a_1^k et a_2^k qui sont respectivement les vecteurs propres associés aux valeurs propres de la matrice $V_{11}{}^{-1}V_{12}V_{22}{}^{-1}V_{21}$ et de matrice $V_{22}{}^{-1}V_{21}V_{11}{}^{-1}V_{12}$ en les diagonalisant.
- Enfin on calcule les premières composantes canoniques z_1^k et z_2^k .

6. Les proximités entre les individus

L'AC détermine z_1^1 et z_2^1 telles qu'en moyenne les 2 variables soient le plus proches possibles pour les n individus, c-à-d de telle sorte que l'expression

$$\frac{1}{n}\sum_{i=1}^{n}\left(z_{1i}^k - z_{2i}^k\right)^2$$

soit la plus petite possible, sous les mêmes contraintes que dans l'espace des variables.

7. Les représentations graphiques

Puisque le but de l'analyse canonique est de mettre en évidence des proximités entre deux ensembles de données, la représentation graphique a pour objectif de décrire ces proximités, aussi bien pour les variables que pour les individus.

- **La représentation des variables**

  Dans cette représentation il faut expliquer pourquoi la corrélation z_1^k et z_2^k est élevée. Ceci implique qu'il faut expliquer pourquoi la corrélation entre une combinaison linéaire de variables de X_1 et une combinaison linéaire de variables de X_2 est élevée. Il est donc nécessaire de représenter sur un même graphique l'ensemble des variables de départ ($m_1 + m_2$). Cette représentation des variables se fait comme en ACP à l'aide d'un cercle des corrélations.

L'axe correspondant à la $j^{ème}$ étape est un compromis (une moyenne) entre z_1^j et z_2^j

Tel que

$$z^j = \frac{z_1^j + z_2^j}{2}$$

- **La représentation des individus**

Chacun des 2 tableaux de données décrit un nuage pour les mêmes n individus. Donc, la représentation des individus en AC permet de cerner ce qui caractérise le mieux ces nuages d'individus dans les directions pour lesquelles ces nuages sont les plus ressemblants possibles.

• A la $j^{ème}$ étape, il s'agit de comparer la description des individus donnée par la variable canonique z_1^j à la description des individus donnée par la variable canonique z_2^j.

• Enfin la proximité plus ou moins importante entre les deux descriptions des individus peut aussi être mise en évidence en calculant l'écart résiduel qui est défini auparavant par $\left|z_{1i}^j - z_{2i}^j\right|$,

Si l'écart résiduel de l'individu i pour la $j^{ème}$ étape est élevé, cet individu joue un rôle particulier dans le phénomène mis en évidence à la $j^{ème}$ étape. Ce rôle est à identifier.

8. Les cas particuliers de l'AC

L'AC présente un grand intérêt d'un point de vue théorique car plusieurs techniques statistiques très utilisées en sont des cas particuliers pour lesquels le problème est de maximiser le coefficient de corrélation entre une variable quantitative X_1 et un ensemble de variables X_2. On cite dans le cas où X_1 décrit une seule variable quantitative, l'AC se ramène à :

- La Régression Linéaire Simple (RLS) si X_2 est constitué d'une seule variable quantitative,
- La Régression Linéaire Multiple (RLM) si X_2 est constitué par plusieurs variables quantitatives,
- L'analyse de la variance si X_2 est une ou plusieurs variables qualitatives,
- L'analyse de la covariance si X_2 est un mélange de variables quantitatives et qualitatives.

Par ailleurs, l'analyse factorielle des correspondances, est le cas particulier de l'AC pour lequel les tableaux X_1 et X_2 décrivent chacun les modalités d'une variable qualitative. L'analyse factorielle discriminante, qui ne sera pas présentée dans ce cours, est le cas particulier de l'AC pour lequel X_1 décrit un ensemble de variables quantitatives et X_2 une variable qualitative.