Mohammed Dennaï

Leçons de statistique matricielle

EDILIVRE

Leçons de statistique matricielle.

Mohammed DENNAÏ

Préface

Ce cours d'analyse en composantes principales et analyse factorielle est destiné aux étudiants en statistique, des grandes écoles, des universités, aux agrégatifs, aux capétiens, aux étudiants en psychologie, aux laboratoires de toutes sortes, afin de palier à la statistique descriptive.

Un tableau de dimension $n \times p$, formé de nombres soient entiers, réelles donc quantitatifs ou alors qualitatifs, dans ce dernier le choix est porté sur un codage de ces modalités, une dimension inférieure $n \times p$, s'impose dès que celle-ci dépasse 3×3, pour visualiser le nuage de points que le tableau nous impose, on abaisse la dimension, et on essaie de déformer le moins possible l'inertie que représente ce nuage. Les corrélations entre coordonnées nous donnent une idée de la linéarité ou de la causalité qui lie ces coordonnées.

Ce cours se compose des leçons suivantes :

- Statistique descriptive unidimensionnelle et multidimensionnelle
- Analyse en composantes principales
- Analyse factorielle
- Analyse canonique
- Régression linéaire.

Des exemples traités sous le langage R, complètent celui-ci.

Mes remerciements vont à Mme Rezzouk, à Mr Cellier tous deux professeurs de statistique à l'université de Rouen. Aux éditions Edilivre pour leurs encouragements, et à ma famille.

1

[1] Ce cours a été réalisé sous word de microsoft©, graph-easy, FactoMineR le pack A.C.P de R.

Chapitre 1 Statistique descriptive

Population

Définition 1.1 On appelle population une collection d'objets ayant des propriétés communes. Les objets sont appelés individus.

Exemple 1.1 Dans l'exemple du blé les individus peuvent être les différentes parcelles répertoriées.

Exemple 1.2 La population d'une ville : ensemble de ses habitants.

Il est essentiel, dans tous les cas, de définir avec précision et sans ambiguïté les ensembles que l'on étudie.

Un malade de longue durée ou un malade occasionnel.

Il est souhaitable que la délimitation précise de la population étudiée soit non seulement connue du statisticien lui-même, mais diffusée en même temps que les résultats de l'étude effectuée. Ce point de déontologie est bien respecté.

Caractère

Chacun des individus de la population peut être considéré du point de vue d'un ou plusieurs caractères.

Le personnel enseignant peut être décrit selon divers caractères : sexe, âge, qualification, ancienneté dans l'enseignement, salaire mensuel…etc.

Les lézards d'Afrique du sud, la longueur, le nombre d'écailles,…etc.

Modalité

Chaque caractère peut présenter une ou plusieurs modalités ; marié, célibataire, divorcé ou non et le fait d'avoir moins de 40 ans, les modalités d'un même caractère sont à la fois incompatibles et exhaustives ; chaque individu de la popuation présente une et une seule des modalités du caractère.

Caractère quantitatif, qualitatif

- Qualitatif : le sexe, la nationalité,…etc.
- Quantitatif : la longueur, le nombre soit discret ou continu donc un intervalle.

Caractère statistique

Soit C un caractère dont l'ensemble des k modalité est M.

Définition 1.1 On appelle un caractère statistique l'application $C: P \to M$, P une population, qui à chaque individu de la population associe la modalité du caractère qui lui est associée.

Définition 2.1 Codage numérique, soit un caractère $C: P \to M$, toute application $\varphi: M \to M'$ définit un autre caractère $C' = \varphi(C)$, l'intérêt de ceci-ci a lieu si M' est numérique et M non, car alors φ permet d'associer au caractère C un caractère numérique. φ porte le nom de codage numérique. Le codage de la sécurité sociale : 1 pour un homme, é pour une femme.

Tableaux

Soient $C: P \to M$ un caractère et la suite $x = \{x_i\}_{i=1\dots n}$, n modalités associées à P, on note$u = \{u_i\}_{i=1\dots p}$, les p modalités différentes qui constituent $x = \{x_i\}_{i=1\dots n}$.

Pour $i = 1 \dots p$, on note n_i, l'effectif (ou le nombre d'occurrences, ou la fréquence absolue) de la modalité u_i, $n_i = Card\{a \in P, C(a) = u_i\}$, on a évidemment : $\sum_{i=1}^{p} n_i = CardP$.

Exemple 1.2 $\{x_i\}_{i=1\dots 6}$.=$\{1,2,1,1,3,2\}$ alors : $u = \{u_i\}_{i=1\dots 3} = \{1,2,3\}$ et $n_1 = 3, n_2 = 2,\ n_3 = 1$.

On dispose ces résultats dans un tableau :

u_i	n_i

Chapitre 2 Statistique descriptive Caractéristiques de position et de dispersion

Série statistique

Moyenne arithmétique

Soit une population P et soit une pondération définie sur P (i.e : une mesure positive définie sur P), $\{p_i\}_{i=1\dots n}$et soit un caractère quantitatif $x = \{x_i\}_{i=1\dots n}$ défini sur P à valeurs dans M où $M \subset \mathbb{R}$.

On note : $P = \{a_i\}_{i=1\dots n}$, $a_i \to p_i$ et $x_i = C(a_i)$.

Définition 2.1 On appelle série statistique l'ensemble : $\{(x_i, p_i)\}_{i=1\dots n}$.

Définition 2.2 On appelle moyenne arithmétique de la série statistique $\{(x_i, p_i)\}_{i=1\dots n}$ ou de la variable x le nombre réel défini par :

$$\bar{x} = \frac{\sum_{i=1..n} p_i x_i}{\sum_{i=1..n} p_i},$$

Si la population est équiprobable alors : $\bar{x} = \frac{\sum_{i=1..n} x_i}{n}$.

<u>Lien avec le centre de gravité :</u>

Soient n points alignés $(M_i)_{i=1..n}$ sur un axe d'origine O et M_i étant d'abscisse x_i affecté de la masse p_i alors le centre de gravité G de ces n points pondérés a pour abscisse : $\bar{x}$.

Propriétés 2.1

- $min(x_1, \dots, x_n) \leq : \bar{x} \leq max(x_1, \dots, x_n)$.
- Soient deux nuages de points : $X = (x_1, \dots, x_n)$, $X' = (x'_1, \dots, x_m)$ liés par l'application affine : φ: $\begin{cases} X \to X' \\ X \to aX + b \end{cases}$, alors :

$$\bar{x}' = \frac{\sum_{i=1..n} p_i x'_i}{\sum_{i=1..n} p_i} = \frac{\sum_{i=1..n} p_i (a x_i + b)}{\sum_{i=1..n} p_i} = a \frac{\sum_{i=1..n} p_i x_i}{\sum_{i=1..n} p_i} + b.$$

- $\sum_{i=1..n} p_i (x_i - \bar{x}) = 0$.

Définition 2.3 La série statistique $\{((x_i - \bar{x}), p_i)\}_{i=1\dots n}$, est dite centre, ou encore la variable x est centrée.

Comme ci-dessus, soient n points alignés sur un axe d'origine O et M_i étant d'abscisse x_i, affecté de la masse p_i.

Nuage de points alignés, inertie d'un nuage de points

Définition 2.4 L'ensemble $N = (M_i)_{i=1..n}$ s'appelle un nuage de points.

La précédente propriété s'écrit alors : $\sum_{i=1..n} p_i \overrightarrow{GM_i}$=0, G étant le centre de gravité.

Soit A un point de l'axe sur lequel se trouve les M_i, on note la distance d euclidienne, si a est l'abscisse de A alors :

$$d^2(A, M_i) = (x_i - a)^2.$$

Définition 2.5 On appelle inertie au point A du nuage de points N la quantité :

$$I_A(N) = \sum_{i=1\dots n} p_i \, d^2(A, M_i).$$

Théorème 2.1(Théorème de Huyghens)

L'inertie est donnée par, si G est le centre de gravité du nuage:

$$I_A(N) = I_G(N) + \sum_{i=1\dots n} p_i \, d^2(A, G). \qquad (2.1)$$

Preuve 2.1

$$I_A(N) = \sum_{i=1\dots n} p_i\,(x_i - a)^2 = \sum_{i=1\dots n} p_i\,(x_i - \bar{x} + \bar{x} + a)^2 = I_G(N) + \sum_{i=1\dots n} p_i\, d^2(A,G). \quad \bullet$$

On constate que : $I_A(N) = I_G(N)$ +une quantité positive, donc le point de l'axe où cette inertie est minimale est le centre de gravité. La valeur a qui minimise l'expression : $I_A(N) = \sum_{i=1\dots n} p_i\,(x_i - a)^2$ est $a = \bar{x}$.

Proposition 2.1 L'inertie en un point d'un nuage de points alignés est minimale au centre de gravité du nuage.

Variance, écart type

Définition 2.6 On appelle variance de x et on note $var\ x$ l'expression :

$$varx = \frac{I_A(N)}{\sum_{i=1\dots n} p_i}. \qquad (2.2)$$

Qui s'écrit :

$$varx = \frac{\sum_{i=1\dots n} p_i (x_i - \bar{x})^2}{\sum_{i=1\dots n} p_i} = \frac{\sum_{i=1\dots n} p_i {x_i}^2}{\sum_{i=1\dots n} p_i} - \bar{x}^2. \qquad (2.3) \quad .$$

Définition 2.7 On appelle écart type de x et on note σ_x la racine carrée positive de la variance de x, c'est une caractéristique de dispersion.

Médiane

Soit la série statistique $\{(x_i, p_i)\}_{i=1\dots n}$. Des notes des étudiants d'une grande école, on affecte le même poids à chaque étudiant (donc à sa note), plutôt que de s'intéresser à la note moyenne des étudiants, on se pose la question suivante :

- Existe –t-il une note N_m qui partage les étudiants en deux groupes d'effectifs égaux, ceux qui ont eu une note supérieure à N_m et ceux qui ont obtenu une note inférieure à N_m.
- Et pour une note N quelconque, on s'intéresse à la quantité : $\sum_{i=1\dots n} |x_i - N|$, existe-t-il une note ${N_m}'$ pour laquelle cette somme est minimale ?
  La réponse est : $N_m = {N_m}'$, qui est la note médiane.
  Elle n'est pas unique, en effet il faut réordonner les notes par ordre croissant puis :
- Si le nombre d'étudiants est impair il faut prendre la note milieu.
- Si le nombre d'étudiants est pair il faut prendre l'une quelconque des valeurs médianes comprise dans un intervalle médian : entre les deux notes de la liste.

Définition 2.8 On appelle médiane de la série statistique : $\{(x_i, p_i)\}_{i=1\dots n}$. Toute valeur a qui rend minimum l'expression :

$$\sum_{i=1\dots n} p_i |x_i - a|.$$

On trouve aussi pour définition de la médiane la valeur a si elle existe telle que :

$$\sum_{i=1\dots n} p_i (x_i \le a) = \sum_{i=1\dots n} p_i (x_i \ge a).$$

Cette valeur centrale est peu sensible aux valeurs aberrantes, une noe très faible avoisinante du 1 ou du zéro, ou alors aux notes extrêmes.

Et il est difficile de déterminer la sensibilité aux fluctuations d'échantillonnage, et le comportement d'une telle note pour des réunions statistiques.

Une approche matricielle de la notion de moyenne arithmétique

Soit $\mathbb{R}^n$ muni de sa base canonique : $(e_i)_{i=1\dots n}$ et x un point de $\mathbb{R}^n, x = \sum_{i=1\dots n} e_i x_i$, on peut munir $\mathbb{R}^n$ de la métrique définie à partir de la matrice :

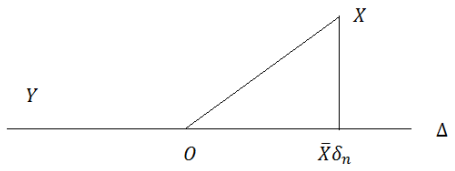
$$D_p = \begin{pmatrix} p_1 & 0 \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 \cdots & p_n \end{pmatrix},$$

- Soit la droite Δ la droite engendrée par : $\delta_n = \sum_{i=1\dots n} e_i$, la propriété : $\sum_{i=1..n} p_i (x_i - \bar{x}) = 0$, s'écrit :

$$(x - \bar{x}\delta_n)^t D_p \delta_n = 0. \qquad (2.4)$$

Ou encore $\bar{x}\delta_n$ est la projection D_p −orthogonale de x sur Δ, illustrée sur la figure 0-1.

Figure 2-Error! Use the Home tab to apply 0 to the text that you want to appear here.**-1 Projection orthogonale**



Soit $y \in \Delta$, $d(x, y)$ est minimale pour : $y = \bar{x}$.

Exprimons $varx$:

$$varx = \frac{\|x - \bar{x}\delta_n\|_{D_p}}{\|\delta_n\|_{D_p}}, \qquad (2,5)$$

Si p est une mesure de probabilité alors : $\sum_{i=1..n} p_i = 1$; on peut écrire :

$$\bar{x} = x^t D_p \delta_n \qquad (2.6)$$

Et $\sigma_x \|x - \bar{x}\delta_n\|_{D_p}$ (2.7)

Chapitre 3 Statistique descriptive Description bidimensionnelle

Table de contingence

Soit une population $P = \{a_1, \dots, a_n\}$ sur laquelle on observe simultanément deux caractères A, B à respectivement r, s modalités, les différentes modalités sont incompatibles et exhaustives.

Soit l'application $\varphi: \begin{cases} A \times B \to M \times M' \\ a = (a_1, a_2) \to m = (m^1, m^2) \end{cases}, m^1 = (m_1, \dots, m_r), m^2 = (m'_1, \dots, m'_s), cardM = r, cardM' = s.$

On note $n_{i,j}$ le nombre d'individus de P qui présentent à la fois la modalité m_i de A et m'_j de B.

$$n_{i,j} = card\{a \in P; \varphi(a) = (m^1{}_i, m^2{}_j)\} \qquad (3.1)$$

On a bien évidemment :

$$n = \sum_{i=1\dots r} \sum_{j=1\dots s} n_{i,j} \qquad (3.2)$$

Fréquences

On note :

$n_{i\bullet} = card\{a = (a_1, a_2) \in P,\ \ \varphi\,(a_1, B) = m_i\}\,, n_{\bullet,j} = card\{a = (a_1, a_2) \in P,\ \ \varphi\,(A, a_2) = m_j{}'\}\,.$

On a :

$\sum_{i=1\dots r} n_{i\bullet} = \sum_{j=1\dots s} n_{\bullet,j} = n.$ (3.2)

Définition 3.1

On appelle table de contingence le tableau à double entrée :

$$(n_{i,j})_{i=1..r, j=1..s}.$$

Définition 3.2

On appelle fréquence relative du couple $(\boldsymbol{m^1{}_i, m^2{}_j})$ le nombre défini par :

$$f_{i,j} = \frac{n_{i,j}}{n}. \qquad (3.3)$$

Définition 3.3

On appelle fréquence marginale de la modalité $\boldsymbol{m^1{}_i}$ (resp de la modalité $\boldsymbol{m^2{}_j}$) le nombre défini par :

$$f_{i,\bullet} = \frac{n_{i,\bullet}}{n} \quad (\text{resp} : f_{\bullet j,} = \frac{n_{\bullet,j}}{n} \quad)$$

On a ainsi :

$$\sum_{i=1..r,} \sum_{j=1..s} f_{i,j} = 1 = \sum_{i=1..r} f_{i,\bullet} = \sum_{j=1.s} f_{\bullet j,}. \qquad (3.4)$$

On peut alors construire deux tableaux unidimensionnels : les tableaux de marge.

Fréquence conditionnelle-indépendance

On considère les $n_{\bullet,j}$, individus qui présentent la modalité $m^2{}_j$, de B, la j-ème colonne de la table de contingence décrit cette sous-population, suivant le caractère A : $n_{i,j}$ individus sur $n_{\bullet,j}$, présentent la modalité $m^1{}_i$.

Définition 3.4

On appelle fréquence conditionnelle de la modalité $m^1{}_i$. le nombre défini par :

$f_i^j = \frac{n_{i,j}}{n_{\bullet,j}}.$

ceci se lit $f, i\ si\ j$.

f_i^j, représente la proportion d'individus qui, parmi les $n_{\bullet,j}$, présentant la modalité $m^2{}_j$, de B, présentent la modalité $m^1{}_i$. de A.

Remarque 3.1

$$f_i^j = \frac{n_{i,j}}{n_{\bullet,j}} = \frac{\frac{n_{i,j}}{n}}{\frac{n_{\bullet,j}}{n}} = \frac{f_{i,j}}{f_{\bullet,j}}.$$

Remarque 3.2 On définit de même :

$$f_j^i = \frac{n_{i,j}}{n_{i,\bullet}} = \frac{\frac{n_{i,j}}{n}}{\frac{n_{i,\bullet}}{n}} = \frac{f_{i,j}}{f_{i,\bullet}}.$$

Définition 3.5 On dit que les caractères A, B sont indépendants ssi :

$$f_j^i = f_{i,\bullet} \cdot f_{\bullet,j}. \qquad (3.5)$$

Il est équivalent de dire que :

$$n_{i,j} = \frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}. \qquad (3.6)$$

Dans ce cas on écrit :

$$f_i^j = f_{i,\bullet}, \quad f_j^i = f_{\bullet,j}. \qquad (3.7)$$

Ceci signifie, d'une part que, pour tout couple (i, j) , la fréquence de $m^1{}_i$ liée par $m^2{}_j$, est indépendante de $m^2{}_j$, et d'autre part la fréquence $m^2{}_j$, liée par $m^1{}_i$ est indépendante de $m^1{}_i$.

Explicitons ces notions sur un exemple :

Exemple 3.1

Soit A l'ensemble des facultés de médecine, et B l'ensemble des formations qui y sont proposées.

La population est l'ensemble des étudiants en médecine qui ont été reçus à au moins une des formations en une date donnée.

- $n_{i,j}$ représente le nombre d'étudiants de la fac i, reçus dans la formation j.
- $n_{i,\bullet}$ est le nombre d'étudiants reçus dans la fac i, toutes formations confondues.
- $n_{\bullet,j}$ est le nombre d'étudiants reçus dans la formation j.
- $n_{\bullet,\bullet}$ est le total de la population estudiantine reçue à au moins un examen.
- $f_{i,j}$ est la proportion d'étudiants reçus dans la fac i pour la formation j dans la population totale.
- $f_{i,\bullet}$ est la proportion d'étudiants reçus dans la fac par rapport à la population estudiantine totale.
- $f_{\bullet,j}$ est la proportion d'étudiants de la formation j dans la population totale.
- f_j^i représente la proportion d'étudiants reçus dans la formation j sachant qu'il s'agit d'étudiants de la fac i.

Supposons que la proportion d'étudiants reçus dans les différentes formations qui y sont proposées soit la même, quelle que soit la formation et supposons maintenant que ceci soit vrai pour toutes les fac d'un pays comme la France, alors on dira qu'il y'a indépendance entre la variable 'fac' et la variable 'formation'. Si l'on se trouve dans ce cas, alors on peut dire également que pour une formation (et ce quelle que soit cette formation), la proportion d étudiants reçus dans telle ou telle fac est la même quelle que soit la fac.

Centre de gravité-covariance- coefficient de corrélation

Soient les deux caractères A, B quantitatifs et $\varphi: A \times B \in P \to \mathbb{R}^2, a_i \mapsto (x_i, y_i)$, on peut représenter les n points $((M_i(x_i, y_i))_{i=1..n}$, dans $\mathbb{R}^2, M_i$ étant de coordonnées (x_i, y_i), on suppose que chaque individu muni d'un poids et on munit M_i, de ce poids : on a ainsi un nuage de points munis de masses dans $\mathbb{R}^2$, $N = \{(M_i(x_i, y_i))_{i=1..n}\}$, on peut définir le centre de gravité G de ces n points pondérés :

Définition 3.6

Le centre de gravité du nuage $N = \{(M_i(x_i, y_i))_{i=1..n}\}$, est le point de coordonnées :

$$\bar{x} = \frac{\sum_{i=1..n} p_i x_i}{\sum_{i=1..n} p_i}, \qquad \bar{y} = \frac{\sum_{i=1..n} p_i y_i}{\sum_{i=1..n} p_i}.$$

$\bar{x}$(resp $\bar{y}$)

Etant la moyenne de la variable $x: P \to \mathbb{R}, a_i \mapsto x_i$, (resp $y: P \to \mathbb{R}, a_i \mapsto y$), on rappelle que :

$$varx = \frac{\sum_{i=1..n} p_i (x_i - \bar{x})^2}{\sum_{i=1..n} p_i}, \quad vary = \frac{\sum_{i=1..n} p_i (y_i - y)^2}{\sum_{i=1..n} p_i}.$$

Définition 3.7

On appelle covariance de x et y la quantité :

$$cov(x, y) = \frac{\sum_{i=1..n} p_i (x_i - \bar{x})(y_i - y)}{\sum_{i=1..n} p_i}.$$

Définition 3.8

On appelle coefficient de corrélation de x et y la quantité :

$$r(x, y) = \frac{cov(x,y)}{\sqrt{varx.vary}}.$$

Définition 3.9

On appelle matrice de covariance de x et y la matrice :

$$V = \begin{pmatrix} varx & cov(x, y) \\ cox(x, y) & vary \end{pmatrix}.$$

Définition 3.10

On appelle matrice de corrélation de x et y la matrice :

$$R = \begin{pmatrix} 1 & r(x, y) \\ r(x, y) & 1 \end{pmatrix}.$$

Proposition 3.1

La covariance et le coefficient de corrélation sont sensibles aux changements d'échelle :
Soient $(a, b, c, d) \in \mathbb{R}^4$ alors :

$$cov(ax, +b, cy + d) = accov(x, y)$$
$$r(ax, +b, cy + d) = \frac{ac}{|ac|} r(x, y).$$

- Si ac est positif $r(x, y)$ est inchangé.
- Si ac est négatif $r(x, y)$ est changé en -$r(x, y)$.
- $cov(x, x) = varx, r(x, x) = 1$.
- On pose :

$$D_{\frac{1}{\sigma}} = \begin{pmatrix} \frac{1}{\sigma_x} & 1 \\ 1 & \frac{1}{\sigma_y} \end{pmatrix}, \text{alors : } R = D_{\frac{1}{\sigma}} V D_{\frac{1}{\sigma}}.$$

Approche géométrique du coefficient de corrélation

La statistique $\{((x_i, y_i), p_i)_{i=1..n}\}$, a pour image dans $\mathbb{R}^n$, le couple de points :

- $x = \sum_{i=1..n} x_i e_i, y = \sum_{i=1..n} y_i e_i,$

On considère $\mathbb{R}^n$, muni de la métrique D_p. Il vient que :

$\bar{x} = \frac{x^t D_p \delta_n}{\|\delta_n\|_{D_p}}, \quad \bar{y} = \frac{y^t D_p \delta_n}{\|\delta_n\|_{D_p}}$, où $\delta_n = \sum_{i=1..n} e_i$, $varx = \frac{\|x - \bar{x}\delta_n\|^2_{D_p}}{\|\delta_n\|_{D_p}}, \quad vary = \frac{\|y - \bar{y}\delta_n\|^2_{D_p}}{\|\delta_n\|_{D_p}}, cov(x, y) = \frac{(x - \bar{x}\delta_n)^t D_p (y - \bar{y}\delta_n)}{\|\delta_n\|_{D_p}}$.

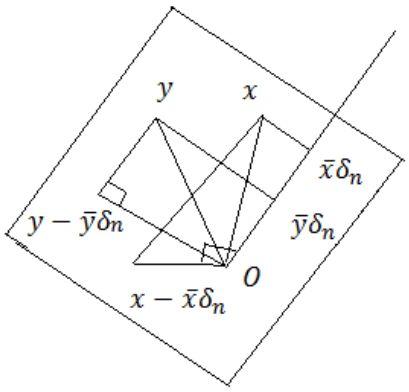
Δ est le sous-espace vectoriel engendré par δ_n, ($\Delta = \mathbb{R}\delta_n$), on considère la décomposition en somme directe de $\mathbb{R}^n = \Delta \oplus \Delta^{\perp}$.

$\Delta^{\perp}$, désigne l'orthogonal de Δ au sens de la métrique D_p; $\Delta^{\perp} = \{x \in \mathbb{R}^n;\ x^t D_p \delta_n = 0\}$, on sait que :

$$x - \bar{x}\delta_n \in \Delta^{\perp} \text{ et } y - \bar{y}\delta_n \in \Delta^{\perp}.$$

$x - \bar{x}\delta_n, y - \bar{y}\delta_n$ sont les caractères centrés de x, y et ce sont aussi les projections orthogonales de x, y sur $\Delta^{\perp}$. Centrer les variables x, y revient à choisir le centre de gravité pour origine.

Figure 3-1 Projections orthogonales



Définition 3.11

- Notion de M-angle.
  Etant donné un espace vectoriel réel E, sur lequel est définie une forme bilinéaire symétrique, définie positive M, on a le résultat suivant plus connu sous le nom d'inégalité de Cauchy-Schwarz :

$$\forall (x, y) \in E^2, \{(x, y) \neq (0,0)\}, M^2(x, y) \leq M(x, x)M(y, y).$$

Ou encore :

$$-1 \leq \frac{M(x,y)}{\sqrt{M(x,x)M(y,y)}} \leq 1. \qquad (3.8)$$

$l'égalité$ à 1 ou -1 ayant lieu ssi x, y sont liés.

- Puisque M est positive, on peut définir la norme associée au produit scalaire, soit : $\|x\|_M = \sqrt{M(x, x)}$.
- $\frac{M(x,y)}{\sqrt{M(x,x)M(y,y)}}$ est considéré comme le cosinus de l'angle formé par x, y pour la métrique M.
- **<u>Application à $\mathbb{R}^n$ muni de la métrique D_p</u>**:

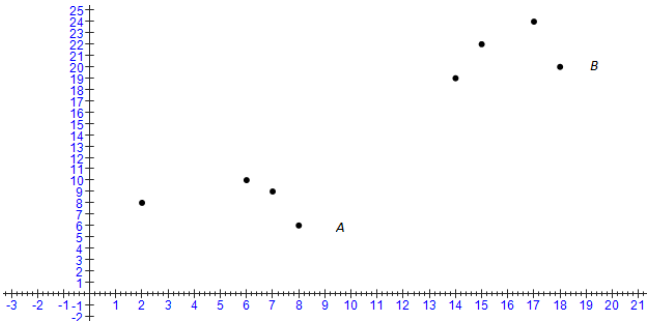
  - $$r(x, y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{D_p(x - \bar{x}\delta_n, y - \bar{y}\delta_n)}{\|x - \bar{x}\delta_n\|_{D_p} \|y - y\delta_n\|_{D_p}}. \qquad (3.9)$$

- On peut considéré $r(x, y)$ comme le cosinus de l'angle formé par x, y pour la métrique D_p.

- **<u>Interprétation des coefficients de corrélation :</u>**
  Il faut être extrêmement prudent quand à la manipulation de ces coefficients de corrélation, une valeur ne pose pas de problème, il s'agit de $r(x, y) = 1$, cela veut dire que les n points sont alignés, le reste tout est permis, alors par blocs de nuage de points, si $r(x, y)$ est proche de 1, les blocs de points sont rapprochés et éloignés entre eux quand $r(x, y)$ est proche de 0. Se référer aux exemples publiés par Saporta, voir bibliographie.

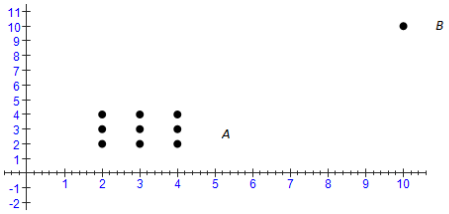
Du bon usage du coefficient de corrélation $r(x, y)$:

Figure 3-2 Coefficients de corrélation ; deux nuages A, B



$$r_A = 0{,}02 \quad , r_B = 0.43 \quad , r = 0{,}024$$

Figure 3-3 Coefficients de corrélation un nuage A et un point B



$$\bar{x}_A = 2, \bar{y}_A = 3, r_A = 0, \bar{x}_B = 10, \bar{y}_B = 10, r_B = indeterminé,$$

$$\bar{x} = 1{,}9, \qquad \bar{y} = 3{,}7, \quad r = -0{,}01$$

Figure 3-4 Coefficients de corrélation un nuage A sous forme de parabole

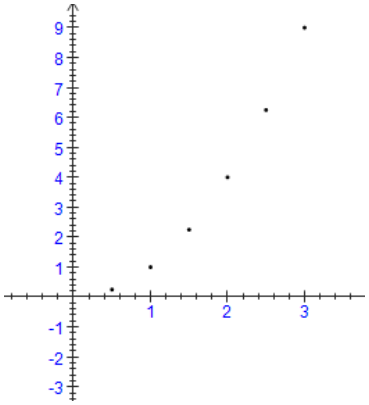


Figure 3-5 Coefficients de corrélation un nuage A sous forme de racine de x

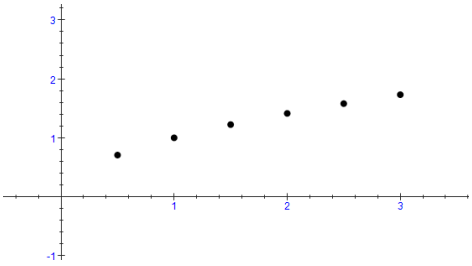


Figure 3-5 Quatre blocs de nuages

	Abscisse	Ordonnée	
1	.2	.1	
2	0.3	.3	A
3	.5	.4	
4	0.6	0.3	
5	0.7	.6	
6	1.2	1.4	
7	1.3	1.5	B
8	1.7	1.6	
9	1.5	1.1	
10	1.8	1.7	
11	2.1	2.5	
12	2.4	2.6	C
13	2.9	2.4	
14	3.4	3.1	
15	3	3.2	
16	2.9	3.7	D
17	3.2	3.6	

D'après Figure 3-5

$\bar{x}_A = 0{,}474, \bar{y}_A = 0{,}34, r_A = 0{,}84.$

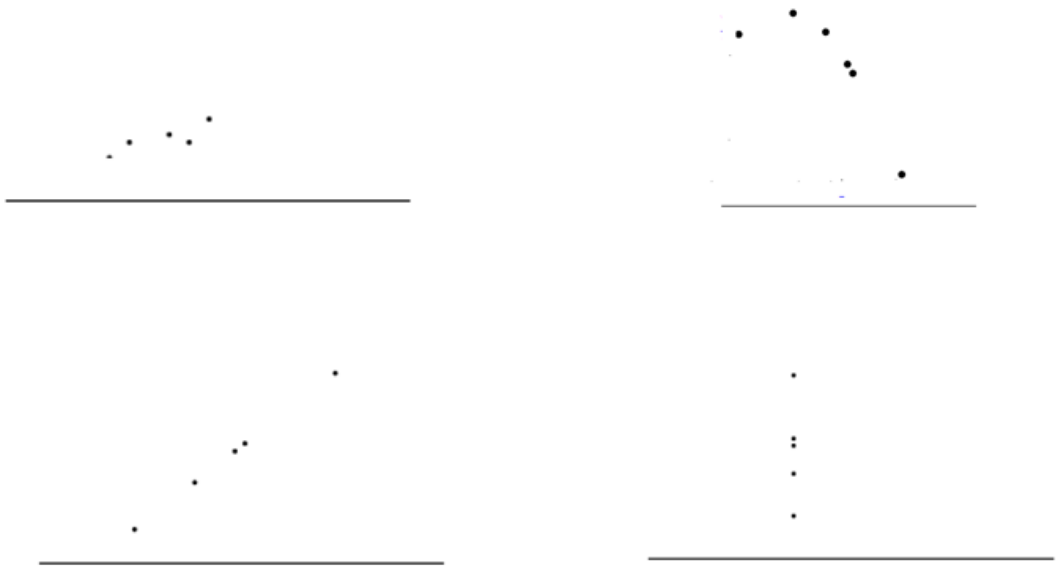
$\bar{x}_B = 1{,}5, \bar{y}_B = 1{,}46, r_B = 0{,}46.$

$\bar{x}_C = 2{,}46, \bar{y}_C = 2{,}5, r_C = -0{,}618.$

$\bar{x}_D = 3{,}125, \bar{y}_D = 3{,}4, r_D = -0{,}56.$

$\bar{x} = 1{,}747, \bar{y} = 1{,}77, r = 0{,}96$

Figure 3-6 Les quatre nuages ont mêmes données statistiques



Les quatre nuages de la figure 3-6 ont mêmes moyennes, mêmes variances et mêmes coefficients de corrélation :

$$\bar{x} = 9, \bar{y} = 7{,}5, s_x^2 = 10{,}0, s_y^2 = 3{,}75, r = 0{,}82.$$

Seul le premier nuage justifie l'usage de r, car les autres nuages de points sont corrélés.

Le fait que deux variables soient « fortement corrélées » ne démontre pas qu'il y ait une relation de causalité entre l'une et l'autre. Le contre-exemple le plus typique est celui où elles sont en fait liées par une causalité commune. Cette confusion est connue sous l'expression *Cum hoc ergo propter hoc*[2].

Remarque 3.3 La corrélation n'est pas transitive, x corrélé avec y, y avec z, cela ne veut pas dire que x est corrélée avec z.

[2] ***Cum hoc ergo propter hoc*** (latin signifiant *avec ceci, donc à cause de ceci*[1]) est un sophisme qui consiste à prétendre que si deux événements sont corrélés alors il y a un lien de cause à effet entre les deux. La confusion entre corrélation et causalité est appelée **effet cigogne** en zététique (corrélation trompeuse entre le nombre de nids de cigognes et celui des naissances humaines) ; en science et particulièrement en statistique cette erreur est rappelée par la phrase « la corrélation n'implique pas la causalité ». Le contraire est *Cum hoc sed non propter hoc*.(Extrait de Wikipédia).

Chapitre 4 Statistique descriptive multidimensionnelle ; analyse en composantes principales

Dans les chapitres suivants, nous allons présenter un ensemble de méthodes d'analyse d'un tableau de données. Sur un tableau dans $\mathbb{R}^{mn}$ étudier l'inertie globale nous nous informe pas ce les lignes et les colonnes de celui-ci, pour pouvoir le cerner et étudier l'apport de chaque ligne ou de chaque colonne il faut le projeter sur un sous-espace de dimension inférieure à n et m. Si ce tableau illustre des parcelles agricoles, où le blé a été semé,

- En fonction de paramètres climatiques et de composition du sol, on effectuera une analyse des correspondances après avoir codé les différents paramètres.
- Ou bien en fonction des quantités d'engrais, d'insecticides, de désherbants : on fera alors une analyse en composantes principales.

En face d'un tableau à double entrée, constitué de nombres réels, comment rendre compte de l'information contenue dans ce tableau ?

On peut utiliser les outils de la statistique descriptive uni- ou bidimensionnelle très satisfaisant. On désire utiliser une méthode synthétique qui rende compte du tableau dans son ensemble. On voudrait trouver des associations entre des lignes ou des colonnes du tableau, déterminer quels caractères contribuent à telle ou à telle association…etc. L'analyse factorielle a pour but de répondre à ce propos.

Soit un tableau croisant individus et caractères ; conventionnellement, les lignes du tableau sont appelées individus et les colonnes caractères.

- On peut considérer un tableau issu d'un recensement, croisant par exemple des régions de France, et des tranches d'âges de la population. A la rencontre de la ligne région i de la colonne tranche d'âge j, on notera le nombre de personnes résidantes dans la région i et dont l'âge est dans la tranche j. on a affaire ici à une table de contingence, ie : un tableau de nombres entiers.
- On peut disposer d'un tableau de mesures de caractéristiques sur une population animale. Le tableau à étudier est alors un tableau de nombres réels.
- On peut avoir affaire à des caractères qualitatifs, comme dans l'exemple croisant parcelles de blé et paramètres climatiques et géophysiques. A la suite d'un codage des modalités de ces paramètres, on aura à traiter un tableau de nombres réels.

Notations en A.C.P

Analyse en composantes principales

On note un tableau X par :

$$X = \left(x_j^i\right)_{1\le i\le n, 1\le j\le p}, x_j^i \in \mathbb{R}.$$

x_j^i est la valeur prise par le caractère j pour l'individu i.

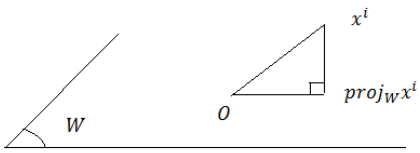
L'individu i es représenté par le vecteur : $x^i = \sum_{j=1..p} x_j^i e_j$ de $\mathbb{R}^p$, muni de sa base canonique : $E = \left\{(e_j)_{1\le j\le p}\right\}$. Au tableau X, on associe n individus :$(x^i)_{1\le i\le n}$ de $\mathbb{R}^p$, qui sont les n lignes de X.

Présentation informelle de l'A.C.P :

Considérons les n points $(x^i)_{1\le i\le n}$ de $\mathbb{R}^p$, et munissons $\mathbb{R}^p$ d'une distance euclidienne d(associée à un produit scalaire).

Dès que p est supérieur à 3 on ne peut visualiser cet ensemble de points. D'où l'idée de projeter celui-ci sur un sous espace W de dimension inférieure à p et d'approcher le nuage de points par sa projection sur W.

4-1 Projection sur W



Dès Le choix de W est le suivant : minimiser

$$\sum_{i=1..n} d^2(x^i, proj_W\, x^i). \qquad (4.1)$$

D'après le théorème de Pythagore on peut écrire :

$$d^2(O, x^i) = d^2(x^i, proj_W x^i) + d^2(O, proj_W x^i). \qquad (4.2)$$

$d^2(O, proj_W x^i)$, étant une donnée du nuage, le critère de choix de W est de maximiser l'expression par antagonisme:

$$\sum_{i=1..n} d^2(O, proj_W\, x^i). \qquad (4.3)$$

i.e. le nuage projeté sur W est le moins déformé possible.

On peut introduire en plus une pondération $(p_i)_{I+1..n}$ sur chacun des points, et le critère (4.1) devient minimiser :

$$\sum_{i=1..n} p_i d^2(x^i, proj_W\, x^i). \qquad (4.4)$$

i.e : maximiser :

$$\sum_{i=1..n} p_i d^2(O, proj_W\, x^i). \qquad (4.5)$$

Le but de l'A.C.P est la détermination de tels espaces W.

Proposition 4.1 Le caractère j est représenté par le vecteur :

$x_j = \sum_{i=1..n} x_j^i f_i$ de $\mathbb{R}^n$, muni de sa base canonique $F = (f_i)_{1i\leq n}$, au tableau X on associe p caractères $(x_j)_{1\leq j\leq p}$ de $\mathbb{R}^n$ qui sont les colonnes de X.

Preuve 4.1 Soit $\mathbb{R}^{p^*}$, muni de sa base duale E^*, et $\mathbb{R}^{n^*}$, muni de sa base duale F^*, x_j^i apparait comme la j-ème composante de x^i:

$$x_j^i =< e_j^*, x^i > \qquad (4.6)$$

De même, on peut dire que x_j^i apparait comme la i-ème composante de x_j:

$$x_j^i =< f_i^*, x_j > \qquad (4.7) \quad \blacksquare$$

Considérons alors les décompositions suivantes, où Δe_j, (resp Δf_i) désigne l'axe de $\mathbb{R}^p$ (resp $\mathbb{R}^n$) engendré par e_j($(resp\ f_i)$:

$$\mathbb{R}^p = \oplus_{j=1..p} \Delta e_j, \quad \mathbb{R}^n = \oplus_{i=1..n} \Delta f_i. \qquad (4.8)$$

<u>A l'individu i est associé :</u>

- Le vecteur x^i de $\mathbb{R}^n$,
- L'axe Δe_j,
- La forme linéaire f_i^* de $\mathbb{R}^{n^*}$.

<u>De même au caractère j est associé :</u>

- Le vecteur x_j de $\mathbb{R}^n$,
- L'axe Δf_i,
- La forme linéaire e_j^* de $\mathbb{R}^{p^*}$.

Notations

- Si on considère les lignes de X alors :

$$X = \begin{pmatrix} x^{1^t} \\ \dots \\ x^{n^t} \end{pmatrix},$$

- Si on considère les colonnes de X alors :

$$X = (x_1, \dots, x_p)$$

Soit l'application linéaire x définie sur $\mathbb{R}^{p^*}$ à valeurs dans $\mathbb{R}^{p^*}$par $x(e_j^*) =$; $x: \mathbb{R}^{p^*} \to \mathbb{R}^{p^*}$, $e_j^* \to x_j$.

Proposition 4.1 La matrice associée à l'application linéaire x: $\mathbb{R}^{p^*} \rightarrow \mathbb{R}^n$, e_j^* de $\mathbb{R}^{p^*}$ étant muni de la base E^*, $\mathbb{R}^n$ muni de la base F est le tableau X.

Proposition 4.2 La matrice associée à l'application linéaire x^i: $\mathbb{R}^{n^*} \rightarrow \mathbb{R}^p$, $\mathbb{R}^{n^*}$ étant muni de la base F^*, $\mathbb{R}^p$ muni de la base Eest le tableau X.

Preuve 4.2 La preuve de l'une des deux propositions implique l'autre par dualité.

On a pour tout $k \in \{1, \dots, p\}$, $<e_j^*, x^t(f_i^*)> = < x(e_j^*), f_i^* > = < f_i^*, x_k > = x_k^i$.■

Pour pouvoir parler de proximité entre deux points x^i, x^k, il est nécessaire d'avoir défini une distance sur l'espace sur lequel se trouve ces points. En analyse factorielle, les proximités sont mesurées à l'aide de distances euclidiennes. On considère dans la suite du chapitre, la métrique M définie sur $\mathbb{R}^p$, Il est logique de chercher à définir sur $\mathbb{R}^{n^*}$ une métrique M' qui conserve le produit scalaire, ie telle que :

$$\forall i, k \quad M(x^i, x^k) = M'(f_i^*, f_k^*),$$

On a : $M'(f_i^*, f_k^*) = M(x^i, x^k) = < M(x^i), x^k > = < M \circ x^t(f_i^*), x^t(f_i^*) > = < x^t \circ M \circ x^t(f_i^*), f_i^* >$,

Soit $M' = xMx^t$, M' ainsi définie est une application linéaire de $\mathbb{R}^{n^*}$ dans $\mathbb{R}^n$, symétrique, positive, mais pas définie en général. En effet :

$$\forall a \in \mathbb{R}^{n^*}, M'(a, a) = M(x^t(a), x^t(a) \geq 0$$

$$M'(a, a) = 0 \Leftrightarrow x^t(a) = 0 \Leftrightarrow a \in kerx^t.$$

On a : $\dim kerx^t = n - rgx^t$, si $p < n$ alors $rgx^t \leq p$ et $dimkerx^t \geq n - p$, par suite x^t n'est pas injective et M' est non définie. Par conséquent, en général, M' n'est pas un produit scalaire.

Schéma de dualité

$$\begin{tikzcd} \mathbb{R}^{p^*} \arrow[r, "x"] & \mathbb{R}^n \arrow[d, "M'"] \\ \mathbb{R}^p \arrow[u, "M"] \arrow[r, "x^t"] & \mathbb{R}^{n^*} \end{tikzcd}$$

Inertie et moments d'inertie

Définition 4.1 Soit le nuage $N = \{(x^i, p_i), i = 1 \dots n\}$ et $a \in \mathbb{R}^p$, on appelle inertie du nuage $N(I)$ au point a et on note :

$I_a(N(I))$ l'expression suivante : $I_a(N(I)) = \sum_{i=1..n} p_i \|x^i - a\|_M$.

Dans la suite on supposera : $\forall i, p_i > 0, \sum_{i=1..n} p_i = 1$.

Définition 4.2 on appelle centre de gravité du nuage $N(I)$ le point g de $\mathbb{R}^p$ défini par :

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

Soit g_j la j-ième composante de g alors : $g_j = \sum_{i=1..n} p_i {x_j}^i = \overline{x_j}$, attention la j-ième composante du centre de gravité du nuage $N(I)$ est la moyenne du caractère.

Théorème 4.1(Théorème de Huyghens-Koenig)

$$I_a(N(I)) = I_g(N(I)) + \|g - a\|_M^2. \qquad (4.9)$$

Preuve 4.1

On a : $x^i - a = x^i - g + g - a$, par suite on peut écrire :

$$\|x^i - a\|^2{}_M = \|x^i - g\|^2{}_M + \|g - a\|_M^2 + 2M(x^i - g, g - a)$$

$$\sum_{i=1..n} p_i M(x^i - g, g - a) = M(\sum_{i=1..n} p_i(x^i - g), g - a) = 0. \blacksquare$$

Proposition 4.1

L'inertie du nuage $N(I)$ est minimale au point g centre de gravité du nuage.

Définition 4.3 on appelle variable centrée associée à la variable x_j , la variable : $x'_j = x_j - \overline{x_j}\delta_n$.

Propriété 4.1

- $\bar{x}'_j = \sum_{i=1..} p_i(x_j^i - \overline{x_j}) = 0.$
- Etant donné $X = (x_1, \dots, x_p)$, $X' = (x'_1, \dots, x'_p)$, on munit $\mathbb{R}^n$, espace des caractères, de la métrique euclidienne des poids ; D_p, est définie par :
  - $D_p(f_i, f_{i'}) = p_i \delta_{ii'}$, où $\delta_{ii'}$ est le symbole de Kronecker.

Rappels :

A x_j on associe la série statistique : $\{(x_j^i, p_i), i = 1..n\}$, on a :

$$var\ x_j = \|x_j - \overline{x_j}\delta_n\|_{D_p}^2 = D_p(x_j - \overline{x_j}\delta_n, x_{j'} - \overline{x_{j'}}\delta_n) = D_p(x'_j, x'_j) = D_p(x'_j - \overline{x_j}'\delta_n, x'_j - \overline{x'}_j\delta_n) = var\ x'_j.$$

De même : $cov(x_j, x_{j'}) = D_p(x_j - \overline{x_j}\delta_n, x_{j'} - \overline{x_{j'}}\delta_n) = cov(x'_j, x'_{j'})$.

Définition 4.4 on appelle matrice de covariance des caractères $(x_j)_{1 \le j \le p}$ de $\mathbb{R}^n$, la matrice :

$$V = X'^t D_p X' = \begin{pmatrix} x'^t_1 \\ \dots \\ x'^t_p \end{pmatrix} D_p(x'_1, \dots, x'_p) = \{D_p(x'_j, x'_k)\}_{j=1..p, k=1..p} = \{cov(x'_j, x'_k)\}_{j=1..p, k=1..p} = \{cov(x_j, x_k)\}_{j=1..p, k=1..p}. \quad (4.10)$$

Il vient :

$$x^i - g = \begin{pmatrix} x_1^i - \bar{x}_1 \\ \dots \\ x_p^i - \bar{x}_p \end{pmatrix} \text{ et } (x^i - g)^t(x^i - g) = \begin{pmatrix} (x_1^i - \bar{x}_1)^2 & (x_1^i - \bar{x}_1)(x_2^i - \bar{x}_2) \cdots & (x_1^i - \bar{x}_1)(x_p^i - \bar{x}_p) \\ \vdots & \ddots & \vdots \\ (x_1^i - \bar{x}_1)(x_p^i - \bar{x}_p) & (x_2^i - \bar{x}_2)(x_p^i - \bar{x}_p) \cdots & (x_p^i - \bar{x}_p)^2 \end{pmatrix}$$

D'où :

$$\sum_{i=1..n} p_i(x^i - g)^t(x^i - g) = \begin{pmatrix} var x_1 & cov(x_1, x_2) \cdots & cov(x_1, x_p) \\ \vdots & \ddots & \vdots \\ cov(x_1, x_p) & cov(x_2, x_p) \cdots & var x_p \end{pmatrix} = V. \quad (4.11)$$

Définition 4.5

V est aussi appelée matrice d'inertie du nuage $N(I)$.

<u>Autre expression de $I_g(N(I))$:</u>

$$I_g(N(I)) = \sum_{i=1..n} p_i \|x^i - g\|_M^2 = \sum_{i=1..n} p_i(x^i - g)^t M(x^i - g), \quad (4.12)$$

$(x^i - g)^t M(x^i - g)$ est un nombre réel, et $tr\left(\left((x^i - g)^t M(x^i - g)\right)\right) = tr\left((x^i - g)((x^i - g)^t M)\right) = tr\left(((x^i - g)^t(x^i - g)M)\right)$,

$$I_g(N(I)) = \sum_{i=1..n} p_i\left((x^i - g)^t M(x^i - g)\right) = \sum_{i=1..n} p_i\, tr\left((x^i - g)((x^i - g)^t M)\right) = \sum_{i=1..n} p_i\, tr\left(((x^i - g)^t(x^i - g)M)\right) = trVM.$$

D'après le théorème de Koenig, on a : $I_0 = I_g + \|g\|^2$.

Cas de variables centrées

On dit que les variables sont centrées ssi pour tout j, $\bar{x}_j = 0$. par suite $g = 0$. on dit aussi que le nuage $N(I)$ est centré, l'origine du nuage est en g. Alors on a : $V = X^t D_p X = \sum_{i=1\ldots n} x^i\,(x^i)^t$ et $I_g = I_0 = tr(VM)$.

Schéma de dualité

A l'isomorphisme D_p de $\mathbb{R}^n \to \mathbb{R}^{n*}$ est associée l'application linéaire V de $\mathbb{R}^{p*} \to \mathbb{R}^p$ définie par : $V = x^t D_p x$.

V définit sur $\mathbb{R}^{p*}$ une forme bilinéaire symétrique semi-définie positive telle que :

$$\begin{tikzcd} \mathbb{R}^{p*} \arrow[r, "x"] \arrow[d, "V"] & \mathbb{R}^n \arrow[d, "D_p"] \\ \mathbb{R}^p \arrow[r, "x^t"] & \mathbb{R}^{n*} \end{tikzcd}$$

La forme bilinéaire symétrique semi-définie positive V sur $\mathbb{R}^{p*}$ vérifie :

$$V(e_j^*, e_k^*) = D_p(x_j, x_k) = cov(x_j, x_k), \quad V(e_j^*, e_j^*) = var x_j.$$

La matrice associée à la forme bilinéaire V par rapport à la base E^* est la matrice de covariance des caractères défie précédemment.

Dans toute la suite de ce chapitre, on supposera le nuage $N(I)$ centré.

Moment d'inertie par rapport à un sous-espace

Soit W un sous-espace vectoriel de $\mathbb{R}^p$, et la décomposition de $\mathbb{R}^p$ en somme directe :

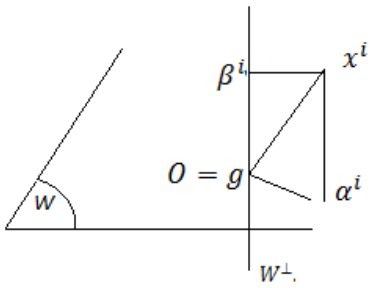
$$\mathbb{R}^p = W \oplus W^{\perp}.$$

Définition 4.6

On appelle inertie du nuage N(I) par rapport à W l'expression :

$$I_W(N(I)) = \sum_{i=1..n} I_W(x^i) = \sum_{i=1..n} p_i d^2(x^i, W),$$

Figure 4-1



On peut écrire : $x^i = proj_W x^i + proj_{W^{\perp}} x^i = \alpha^i + \beta^i$, l'inertie N(I) par rapport à W s'écrit :

$$I_W(N(I)) = \sum_{i=1..n} p_i \|x^i - proj_W x^i\|^2 = \sum_{i=1..n} p_i \|proj_{W^{\perp}} x^i\|^2. \qquad (4.13)$$

D'après le théorème de Pythagore :

$$\|x^i\|^2 = \|proj_W x^i\|^2 + \|proj_{W^{\perp}} x^i\|^2 \text{ il vient alors :}$$

Proposition 4.2

$$I_W(N(I)) = I_W(N(I)) + I_{W^\perp}(N(I)). \tag{4.14}$$

Moment d'inertie par rapport à un sous-espace affine W_1 de W

Soit a un élément de $W^\perp$. On définit : $W_1 = \{x + a, x \in W\} = W + a$.

Soit : $x^i = \alpha^i + \beta^i = \alpha^i + a + \beta^i - a, \alpha^i \in W, \beta^i \in W^\perp$.

Définition 4.7

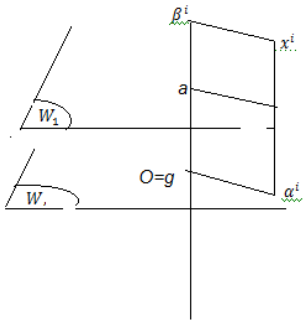
On appelle inertie du nuage $N(I)$ par rapport à W_1 l'expression :

$$I_{W_1}(N(I)) = \sum_{i=1..n} p_i \|\beta^i - a\|^2. \tag{4.15}$$

Théorème 4.1(Théorème de Huyghens-Koenig)

$$I_{W_1}(N(I)) = I_W(N(I)) + \|a\|^2, \|a\|^2 = d(W, W_1). \tag{4.16}$$

Figure 4-2



On a :

$$\sum_{i=1..n} p_i \|\beta^i - a\|^2 = \sum_{i=1..n} p_i \|\beta^i\|^2 + \|a\|^2 - 2M(p_i\beta^i, a) = I_W + \|a\|^2.$$

Car : $g = 0 = \sum_{i=1..n} p_i x^i = \sum_{i=1..n} p_i \alpha^i + \sum_{i=1..n} p_i \beta^i$, toutes ces sommes sont donc nulles par conséquent : $I_{W_1}(N(I)) \geq I_W(N(I))$, le sous-espace affine de dimension k le plus proche de $N(I)$ passe par g, c'est donc un sous-espace vectoriel de $\mathbb{R}^p$, puisque $g = 0$.

Conséquence du théorème de Pythagore

$$W = W_1 \oplus W_2 \; et \; W_1 \perp W_2 \Rightarrow I_{W^\perp} = I_{W_1{}^\perp} + I_{W_2{}^\perp}.$$

C'est l'application du théorème de Pythagore.

Soit $\mathbb{R}^p = W \oplus W^\perp$. On a : $x^i = \alpha^i + \beta^i$, où α^i est dans W et β^i est dans $W^\perp$. on peut écrire : $x^i = \alpha_1{}^i + \alpha_2{}^i$, avec $\alpha_1{}^i \in W_1, \alpha_2{}^i \in W_2$.

$$I_{W^\perp} = \sum_{i=1..n} p_i \|\alpha_1{}^i\|^2 + \sum_{i=1..n} p_i \|\alpha_2{}^i\|^2 = I_{W_1{}^\perp} + I_{W_2{}^\perp}.$$

Remarque 4.1

$$I_W = I_g - I_{W^\perp} = I_g - (I_{W_1{}^\perp} + I_{W_2{}^\perp}). \tag{4.17}$$

Définition 4.8

Un sous-espace vectoriel W_k de dimension k est (s'il) à inertie minimale ssi, parmi tous les sous-espaces E_k de dimension k, il réalise le minimum de I_{E_k}.

Ce sous-espace n'est pas nécessairement unique.*

Puisque I_g, est une donnée du nuage, dire que W_k est à inertie minimale revient à dire que $I_{W_1{}^{\perp}}$ est maximale.

Théorème 4.2

Soit W_k un sous-espace vectoriel de $\mathbb{R}^p$ de dimension k, à inertie minimale. Alors le sous-espace vectoriel de dimension (k+1) à inertie maximale est la somme directe de W_k et du sous espace de dimension 1 M-orthogonal à W_k à inertie minimale dans $W_k{}^{\perp}$, : les solutions sont emboîtées.

Preuve 4.2

Soit W_k un sous-espace vectoriel de dimension k à l'inertie minimale et soit E_{k+1} un sous-espace vectoriel de dimension k+1 quelconque :

$$\dim(E_{k+1} \cap I_{W^{\perp}}) = \dim(E_{k+1}) + \dim(W_k^{\perp}) + dim(E_{k+1} + W_k^{\perp})$$

soit dim $(E_{k+1} \cap I_{W^{\perp}}) \geq k$ +1+$(p-k)$-$p \geq 1$

donc il existe une vecteur v non nul appartenant à $(E_{k+1} \cap W_k^{\perp}))$:

on pose Δv le sous-espace engendré par v.

$W_k \oplus v = W_{k+1}$.

Et $I_{W_{k+1}{}^{\perp}} = I_{k^{\perp}} + I_{v^{\perp}}$

$E_{k+1} = \Delta v \oplus R$, où R est le sous-espace vectoriel supplémentaire M-orthogonal de Δv dans E_{k+1} est maximal, par suite : $I_{v^{\perp}} \geq I_{k^{\perp}}$.

$I_{W_{k+1}{}^{\perp}} \geq I_{E_{k+1}{}^{\perp}}$ et $I_{W_{k+1}} \geq I_{E_{k+1}}$.

Le moment d'inertie de N(I) par rapport au sous-espace vectoriel W_{k+1} construit à partir de W_k est inférieur au moment d'inertie par rapport au sous-espace vectoriel E_{k+1}, il est donc minimum.

A partir d'un sous-espace vectoriel W_k de dimension k à inertie minimale, on peut construire, en lui adjoignant une doite M-orthogonale, un sous-espace vectoriel de dimension k+1 à inertie minimale.

But de l' A.C.P

Au tableau de données X, on associe le nuage pondéré $N(I)$, composé des vecteurs lignes de X, la i-ème ligne x^i, de X étant pondérée par p_i, on se donne une métrique euclidienne M de $\mathbb{R}^p$, ce qui permet de définir une distance entre deux vecteurs lignes de X.

L' A.C.P a pour but de trouver un sous-espace de dimension k, de $\mathbb{R}^p$, qui approche au mieux le nuage $N(I)$, alors on considère la projection de $N(I)$, sur ce sous-espace comme une bonne représentation de ce nuage.

Le critère retenu sur ce sous-espace est que l'inertie soit minimale.

La recherche de ce sous-espace ; consiste en la recherche :

- D'une droite Δu_1 la plus proche de $N(I)$.
- La droite Δu_2 $M-$ orthogonale à Δu_1, la plus proche du nuage $N(I)$.
- Et ainsi de suite.
  L'étude consistera à offrir les meilleurs plans qui donneront la meilleure 'prise de vue' du nuage.

Moment d'inertie par rapport à Δu et $\Delta u^{\perp}$

Soit Δu une droite engendrée par le vecteur M-normé u de $\mathbb{R}^p$.

Le problème consiste en la recherche de u tel que : Δu soit à inertie minimale, soit encore : $I_{\Delta u^{\perp}}(I)$ soit maximale.

Par définition, on a :

$$I_{\Delta u^{\perp}}(I) = \sum_{i=\cdots 1} p_i \left\| x^i - proj_{\Delta u^{\perp}} x^i \right\|^2 = \sum_{i=\cdots 1} p_i \left\| proj_{\Delta u} x^i \right\|^2 \qquad (4.18)$$

Or : $proj_{\Delta u} x^i = M(u, x^i)u = ((x^i)^t Mu)u$.

$$I_{\Delta u^{\perp}}(I) = \sum_{i=\cdots 1} p_i ((x^i)^t Mu)^2 = \sum_{i=\cdots 1} p_i ((x^i)^t Mu)((x^i)^t Mu) = u^t M(\sum_{i=\cdots 1} p_i ((x^i)^t (x^i))Mu = u^t MVMu = V(Mu, Mu) \qquad (4.19)$$

Si on pose : $proj_{\Delta u} x^i = c_i u$, alors : $c_i = M(u, x^i)$ et $I_{\Delta u^{\perp}}(I) = \sum_{i=\cdots 1} p_i {c_i}^2 = \|c\|_{D_p}$.

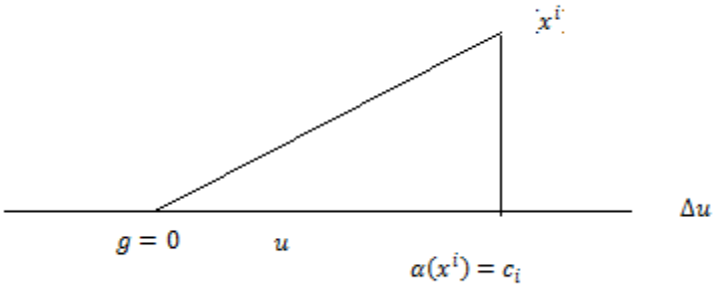
$\mathbb{R}^p \to^M \mathbb{R}^{p*} \to^x \mathbb{R}^n$;$u \mapsto M(u) \mapsto xoM(u) = c$; $xMu = \begin{pmatrix} (x^1)^t \\ \dots \\ (x^n)^t \end{pmatrix} Mu = \begin{pmatrix} (x^1)^t Mu \\ \dots \\ (x^n)^t Mu \end{pmatrix} = \begin{pmatrix} c_1 \\ \dots \\ c_n \end{pmatrix}$.

Le vecteur c représente les coordonnées des individus sur l'axe Δu, par suite c est centré, de plus la solution u au problème posé, si elle existe, impliquera que c est de variance maximale.

Si on pose $\alpha = Mu$, on a : $I_{\Delta u^\perp}(I) = V(\alpha, \alpha)$, le problème d'analyse est le suivant : maximiser $u^t MVMu$ sous la contrainte $\|u\|_M = 1$,

On peut transporter le problème dans $\mathbb{R}^{p*}$, au moyen de l'isomorphisme M, car il induit sur $\mathbb{R}^{p*}$, une métrique M^{-1}, donc dans $\mathbb{R}^{p*}$, il s'agit de maximiser $\alpha^t VM\alpha$, sous la contrainte $\|\alpha\|_{M^{-1}} = 1$.

Figure 4-3



$$\mathbb{R}^p \to^\alpha \mathbb{R}$$

$$x^i \to \alpha(x^i)$$

$$\alpha(x^i) =< x^i, \alpha >=<x^i, Mu >= M(u, x^i) = c_i$$

α est la forme linéaire coordonnée sur l'axe factoriel Δu.

Propriété des vecteurs propres et des valeurs propres de VM

Propriété 4.1

- VM est M-symétrique : $(MVM)^t = MVM$.

Propriété 4.2

- Les valeurs propres de VM sont toutes réelles.
- $\mathbb{R}^p = \oplus_{i=1..k} E_i$, E_i est le sous-espace propre associé à la valeur propre λ_i.
- Les E_i sont deux-à-deux orthogonaux.
- $\boldsymbol{dim} E_i$ = ordre de multiplicité de la valeur propre λ_i.

Propriété 4.3

- Les valeurs propres de VM sont positives ou nulles.
- Si u est vecteur propre de VM alors $VMu = \lambda$.
- $I_{\Delta u^\perp}(I) = u^t M(\lambda u) = u^t MVMu = \lambda^t uMu = V(Mu, Mu) = \lambda \|u\|_M^2$ d'après l'égalité (2.19)..

On peut donc construire une base M-orthonormée de $\mathbb{R}^p$ à l'aide de vecteurs propres de VM.

Résolution du problème d'optimisation, axes et facteurs principaux- Composantes principales

Il s'agit de maximiser la fonction f définie par : $f(\alpha) = \alpha^t V\alpha$ sous la contrainte : $g(\alpha) = \alpha^t M^{-1}\alpha - 1 = 0$.

Si α^0 réalise un maximum alors il existe un réel λ appelé multiplicateur de Lagrange tel que :

$$\left(\frac{df}{d\alpha} - \lambda\frac{dg}{d\alpha}\right)(\alpha^0) = 0. \qquad (4.20)$$

Pour tout k ; $\left(\frac{df}{d\alpha_k} - \lambda\frac{dg}{d\alpha_k}\right)(\alpha^0) = 0.$

$$2V\alpha^0 - 2\lambda M^{-1}\,\alpha^0 = 0$$

$$MV\alpha^0 = \lambda\alpha^0,$$

α^0 doit être un vecteur propre de MV associé à la valeur propre λ.

De plus :

$$I_{\Delta u^\perp}(I) = {\alpha^0}^t V\alpha^0 = {\alpha^0}^t(\lambda M^{-1}\alpha^0) = \lambda\left({\alpha^0}^t M^{-1}\alpha^0\right) = \lambda.$$

Si u^0 réalise le maximum de $I_{\Delta u^\perp}(I)$ les valeurs de VM (ou de MV) étant toutes réelles alors λ doit être la plus grande valeur propre.

La solution α^0 est un vecteur propre M^{-1} normé de MV associé à la plus grande valeur propre. Il est unique si cette valeur propre est simple.

Axes et facteurs principaux- composantes principales

Définition 4.9

On appelle premier axe principale d'inertie l'axe engendré par u_1 où u_1 est un vecteur propre M- Normé, de VM associé à la plus grande valeur propre λ.

u_1 est un vecteur de $\mathbb{R}^p$, on a :

$$VMu_1 = \lambda_1 u_1\,,\quad I_{\Delta {u_1}^\perp}(I) = \lambda_1.$$

Définition 4.10

On appelle premier facteur principal d'inertie α_1 un vecteur propre de MV associé à la plus grande valeur propre λ_1.

α_1 est un vecteur propre de $\mathbb{R}^{p*}$, c'est donc une forme linéaire sur $\mathbb{R}^p$, et :

$\alpha_1 = Mu_1$.

Définition 4.11

On appelle première composante principale d'inertie c^1, le vecteur $\mathbb{R}^n$, dont les coordonnées sur F sont les composantes des x^i, sur l'axe Δu_1,

$c^1 = xM(u_1)$, ou encore : $c^1 = ({c^1}_i)_{i=1..n}$, avec ${c^1}_i = M(u_1, x^i)$.

Remarque 4.2

- c^1 es centré
  En effet : $\sum_{i=1..n} p_i\,{c^1}_i = \sum_{i=1..n} p_i\,M(u_1, x^i)M(\sum_{i=1..n} p_i\,x^i, u_1) = 0.$■
- c^1 est de variance λ_1,
  En effet :
  $\sum_{i=1..n} p_i\,({c^1}_i)^2 = \lambda_1.$ ■

.

Définition 4.12

On appelle part d'inertie expliquée par l'axe Δu_1 la quantité :

$$\frac{\lambda_1}{trVM} = \frac{\lambda_1}{\lambda_1 + \cdots + \lambda_k} \qquad (4.21)$$

Elle exprime l'importance relative du prmier axe factoriel.

L'axe Δu_1, est le ous-espace vectoriel de dimension 1 de $\mathbb{R}^p$, le plus proche du nuage N(I), au sens donné ici, i.e : au sens de l'inertie minimale. La recherche du sous-espace vectoriel de dimension 2 de $\mathbb{R}^n$, le plus proche du nuage N(I) revient à rechercher un axe Δu_2, engendré par le vecteur M-unitaire u_2, de $\mathbb{R}^p$, M-orthognal à u_1, à inertie minimale. Ce sous-espace sera alors : $\Delta u_1 \oplus \Delta u_2$, si λ_1 est d'ordre 2 au moins, il suffit de choisir un vecteur orthogonal à u_1 dans le sous-espace propre associé à λ_1 si la première valeur propre est simple, le problème est de maximiser $u^t MVMu$, sous la contrainte : $u^t Mu = 1$, et $u^t Mu = 0$. soit encore en transposant dans $\mathbb{R}^{p*}$, au moyen de l'ismorphisme, M : maximiser $\alpha^t V\alpha$. sous la contrainte, $\alpha^t M\alpha = 1$, et $\alpha^t M\alpha = 0$.

<u>On cherche à maximiser :</u>

$$\begin{cases} f(\alpha) = \alpha^t V\alpha, \\ g_1(\alpha) = \alpha^t M\alpha - 1 \ . \\ g_2(\alpha) = \alpha^t M\alpha_1 = 0 \end{cases}$$

Si α^0, réalise un maximum, alors il existe $\lambda, \mu \in \mathbb{R}$, tels que α^0, soit la solution de :

$$\begin{cases} \left(\frac{df}{d\alpha} - \lambda\frac{dg_1}{d\alpha} - 2\mu\frac{dg_2}{d\alpha}\right)(\alpha^0) = 0. \\ 2V\ \alpha^0 - 2\lambda M^{-1}\ \alpha^0 - 2\mu M^{-1}\ \alpha_1 = 0 \\ \alpha_1^t V\alpha^0 = \lambda^t \alpha_1 M^{-1}\ \alpha^0 + \mu^t \alpha_1 M^{-1}\ \alpha_1 = \mu \end{cases}$$

$$\alpha_1^t V\alpha^0 = \alpha^{0^t} V\alpha_1 = \alpha^{0^t}(\lambda_1 M^{-1}\ \alpha_1) = \lambda_1{}^t(\alpha^0 M^{-1}\ \alpha_1) \Rightarrow \mu = 0.$$

$$\Rightarrow V\alpha^0 = \lambda M^{-1}\ \alpha^0.$$

Et α^0, doit être vecteur propre de MV pour la valeur propre λ.

Comme : $I_{\Delta u_0^\perp} = \alpha^0 M^{-1}\ \alpha^0 = \lambda \Rightarrow$ le maximum aura lie en $\lambda = \lambda_2$. λ_2 est la deuxième plus grande valeur propre de MV et α_2 un vecteur propre M^{-1} -normé de MV associé à λ_2.

En conclusion la solution $u_2, (resp\ \alpha_2)$ est vecteur propre de $VM\ (resp\ MV)$ associé à la deuxième plus grande valeur propre. ■

Définition 4.12

On appelle deuxième axe principal d'inertie l'axe engendré par u_2 où u_2 est vecteur propre M-normé de VM associé à la deuxième plus grande valeur propre λ_2.

u_2 est vecteur de $\mathbb{R}^p$, on a : $VMu_2 = \lambda_2 u_2$, $I_{\Delta u_2^\perp} = \alpha^0 M^{-1}\ \alpha^0 = \lambda_2$.

Définition 4.13

On appelle deuxième facteur principal d'inertie l'axe engendré par α_2 où α_2 est un vecteur propre de MV associé à la deuxième plus grande valeur propre λ_2.

α_2 est vecteur de $\mathbb{R}^{p*}$, c'est donc une forme linéaire sur $\mathbb{R}^p$; on a :$MVu_2 = \alpha_2$, $I_{\Delta u_2^\perp} = \alpha^0 M^{-1}\ \alpha^0 = \lambda_2$.

Définition 4.14

On appelle deuxième composante principale c^2 le vecteur c^2 de $\mathbb{R}^n$ dont les coordonnées sur F sont des composantes des x^i, sur l'axe Δu_2.

$$\begin{cases} c^2 = xM(u_2) \\ c^2 = (c_i^2)_{I=1..n} \\ c_i^2 = M(u_2, x^i) \end{cases}$$

Remarque 4.3

- c^2 est centré.
  En effet : $\sum_{i=1..n} p_i\, c^2{}_i = \sum_{i=1..n} p_i\, M(u_2, x^i) = M(\sum_{i=1..n} p_i\, x^i, u_2) = 0$.■
- c^2 est de variance λ_2,
  En effet :
  $\sum_{i=1..n} p_i\, (c^2{}_i)^2 = \lambda_2$. ■

Définition 4.15

On appelle part d'inertie expliquée par le deuxième axe :

$$\frac{\lambda_2}{trVM} = \frac{\lambda_2}{\lambda_1+\cdots+\lambda_k} \qquad (4.22)$$

Définition 4.16

On appelle plan principal le plan engendré par u_1, u_2

On a : $I_{(\Delta u_1,\Delta u_2)} = I_g - I_{\Delta u_1{}^{\perp}} - I_{\Delta u_2{}^{\perp}} = trVM - \lambda_1 - \lambda_2$.

Définition 4.17

On appelle part d'inertie expliquée par le : plan engendré par u_1, u_2 la quantité :

$$\frac{\lambda_1+\lambda_2}{trVM} = \frac{\lambda_1+\lambda_2}{\lambda_1+\cdots+\lambda_k} \qquad (4.23)$$

$\mathbb{R}^p$ admettant une base orthonormé constituée de vecteurs propres de VM, on met ainsi en place, un système d'axes principaux d'inertie. De plus on sait que le sous espace W_k, de dimension k, qui approche au mieux le nuage au sens de l'inertie sera le sous-espace vectoriel, engendré, par les k premiers axes principaux.

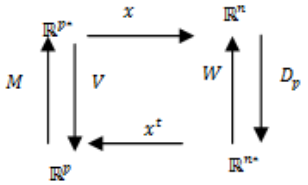
Les valeurs propres sont rangées par ordre décroissant il vient :

$$I_{W_k} = I_g - \sum_{i=1..k} \lambda_k \qquad (4.24)$$

La part d'inertie expliquée par W_k est égale à :

$$\frac{\sum_{i=1..k} \lambda_k}{trVM} \qquad (4.25)$$

Dualité



$V = x^t D_p x$
$W = xMx^t$

On met en place ainsi, une M-base orthonormé de $\mathbb{R}^p$, constituée de vecteurs propres de VM qui engendre les k axes principaux ($k \leq p$) associés aux valeurs propres rangées par ordre décroissant.

Si toutes les valeurs propres de VM sont strictement positives alors il y a p axes factoriels.

Si VM admet la valeur propre 0, soit ($dimkerVM = r > 0$), l'ordre de multiplicité de cette valeur propre nulle est r.

Soit $(v_i)_{i=1\ldots r}$ un système orthonormé de vecteurs propres de VM associés à la valeur propre 0 alors : $I_{\Delta v_i^{\perp}}(N(I)) = 0$.

Par ailleurs : $kerVM = \oplus_{i=1\ldots r} v_i$, on a donc :

: $I_{(kerVM)^{\perp}}(N(I)) = \sum_{i=1..} I_{\Delta v_i^{\perp}}(N(I)) = 0$, par suite $I_{(kerVM)^{\perp}}(N(I)) = \sum_{i=1..r} p_i d^2(x^i, (kerVM)^{\perp}) = 0$, et pour tout i, $x^i \in (kerVM)^{\perp}$), il y' a alors p-r axes factoriels.

Axes principaux d'inertie

Dans $\mathbb{R}^p$, on a :

$$\forall i,\ \ VMu_i = \lambda_i u_i, \text{ avec : } \forall i,\ \ M(u_i, u_i) = 1,\ \ M(u_i u_{i'}) = 0, i \neq i' \ .$$

Facteurs principaux d'inertie

De même $\mathbb{R}^{p*}$, on a une base M^{-1} −orthonormée constituée de vecteurs propres de MV, $(\alpha_i)_{i=1\ldots k}$ qui sont les facteurs principaux :

$$\forall i,\ \ Mu_i = \alpha_i,$$

Dans $\mathbb{R}^{p*}$;

$$\forall i,\quad MV\alpha_i = \lambda_i\alpha_i,\ \forall i,\ \ M^{-1}(\alpha_i,\alpha_i) = 1,\ M(\alpha_i\alpha_{i'}) = 0, i \neq i'\ ,\ \ \forall i,\ \ V\ (\alpha_i,\alpha_i) = \lambda_i,$$

Et on a le résultat :

Proposition 4.3

Les $(\alpha_i)_{i=1\ldots k}$ sont V-orthogonaux.

Preuve 4.3

$$\forall i,\quad MV\alpha_i = \lambda_i\alpha_i \Rightarrow V\alpha_i = \lambda_i M^{-1}\alpha_i \Rightarrow\ {\alpha_i}^t V\alpha_{i'} = \lambda_i {\alpha_i}^t M^{-1}\alpha_i' = \begin{cases} 1\ si\ i = i' \\ 0\ si\ i \neq i \end{cases}.\blacksquare$$

Composantes principales

Dans $\mathbb{R}^n$ on obtient les composantes principales :

$$c^i = xMu_i = x\alpha_i$$

Proposition 4.4

Les composantes principales c^i sont centrées, de variance la valeur propre correspondante λ_i

- $\forall i, D_p(c^i, c^i) = \lambda_i$.
- $\forall i \neq i',\ D_p(c^i, c^{i'}) = {c^i}^t D_p c^{i'} = \alpha_i^t x^t D_p x\alpha_{i'} = \alpha_i^t V\alpha_{i'} = cov(\alpha_i, \alpha_{i'}) = 0$.
- Les c^i sont deux-deux D_p −orthogonales, leur corrélation est nulle, $r(c_i, c_{i'}) = 0$, elles sont donc non corrélées.
- $\forall i,\ \ WD_p c^i = xMx^t D_p x\alpha_i = xMV\alpha_i = \lambda_i x\alpha_i = \lambda_i {c^i}_i$.

Proposition 4.5

c^i est donc vecteur propre de WD_p par rapport à la valeur propre λ_i.

Autre présentation de la notion de composante principale

Soit α un élément de $\mathbb{R}^{p*}$, dans la base duale $({e^*}_i)_{i=1..p}$, celui-ci s'écrit : $\alpha = \sum_{i=1..p} \alpha_i\, {e^*}_i$, on a donc :

$$c = x\alpha = x\left(\sum_{i=1..p} \alpha_i\, {e^*}_i\ \right) = \sum_{i=1..p} \alpha_i\, x({e^*}_i) = \sum_{i=1..p} \alpha_i\, x_i,$$

L'A.C.P consiste à remplacer les p-variables $(x_1, \ldots, x_p)$ corrélées par de nouvelles variables fictives, non corrélées entre elles, de variances maximales, combinaisons linéaires des p variables initiales.

Propriété fondamentale d'une composante principale

Soient données les deux variables x^i, x^k, on note : $d(x^i, x^k) = d_{ik}$ et ${d_{ik}}^2 = (x^i - x^k)^t M(x^i - x^k)$, on pose alors :

$$\begin{aligned} {d_{i.}}^2 &= \sum_{k=1..n} p_k\, {d_{ik}}^2 \\ {d_{..}}^2 &= \sum_{i=1..n} p_i\, {d_i}^2 \end{aligned}.$$

D'après le théorème de Huyghens : $I_{x^i} = I_g + \|x^i\|_M{}^2$ or

$I_{x^i} = \sum_{k=1..n} p_k\, d_{ik}{}^2 = d_{i.}{}^2 \Rightarrow d_{..}{}^2 = \sum_{i=1..n} p_i\, d_i{}^2 = (\sum_{i=1..n} p_i) I_g + \sum_{i=1..n} p_i \|x^i\|_M{}^2 = 2I_g.$

$$\Rightarrow \|x^i\|_M{}^2 = d_{i.}{}^2 - \frac{1}{2d_{..}{}^2}$$

$$\Rightarrow d_{ik}{}^2 = \|x^i\|_M{}^2 + \|x^k\|_M{}^2 - 2M(x^i, x^k)$$

$$\Rightarrow M(x^i, x^k) = \frac{1}{2(d_{i.}{}^2 + d_{k.}{}^2 - d_{..}{}^2 - d_{ik}{}^2)} \qquad (4.26)$$

Par définition :

$$W(f_i^*, f_k^*) = M(x^i, x^k) = \frac{1}{2(d_{i.}{}^2 + d_{k.}{}^2 - d_{..}{}^2 - d_{ik}{}^2)} \qquad (4.27)$$

Sachant qu'une composante principale est vecteur propre de WD_p donc celle-ci ne dépend des individus que par leurs distances mutuelles d_{ik} et des paramètres de pondérations p_i.

Choix de la métrique M

Deux métriques M, M' sont choisies.

On fait souvent le choix de la métrique : $M = I$, c'est-à-dire on travaille sur la matrice d'inertie V. Si on change linéairement d'unité de mesure sur une variable, cela revient à multiplier les covariances entre cette variable et les autres par un coefficient μ et la variance de la variable en question par μ^2 . les résultats de l'A.C.P sont fondamentalement changés.

Par exemple, supposons qu'une des variables relevées sur un ensemble d'individus comporte la variable taille et supposons qu'elle soit exprimée en mètre, soit X le tableau de données :

Choisissons d'exprimer cette taille en centimètre ; le nouveau tableau des données Y ne diffère de X que par la colonne taille, celle-ci étant multipliée par 100. Comparons les matrices d'inertie V_X, V_Y, leurs éléments sont les mêmes, excepté la ligne et la colonne correspondant à la variable taille. Dans , V_Y les éléments non diagonaux correspondant à cette variable sont par rapport à ceux de V_X multipliés par 100, l'élément diagonal par 10 000. Il est clair que les vecteurs propres de V_X et de V_Y sont très différents et donc les résultats de l'A.C.P alors que l'information dans les deux tableaux est la même. Pour pallier à ce problème on fait souvent le choix $M = D_{1/\sigma^2}$, Explicitons ce cas :

Soit l'isomorphisme sur $\mathbb{R}^{p*}$ défini par : $D_{1/\sigma}$, et posons $Z = X\, D_{1/\sigma}$. Ceci revient à faire le changement de variables dans l'espace des caractères : $\forall j, \quad z_j = \frac{x_j}{\sigma_j}$, alors : $var z_j = 1$, on dit que l'on a réduit les variables, X étant centré, alors Z l'est aussi.

Définition 4.18

X étant centré, le tableau Z défini par $Z = X\, D_{1/\sigma}$ porte le nom de tableau de variables centrées réduites.

Soit V_X, la matrice de covariance des variables $(z_j)_{j=1..p}$ et R la matrice de corrélation des variables $(x_j)_{j=1..p}$ alors :

$$V_X = Z^t D_p Z = D_{1/\sigma} X^t D_p D_{1/\sigma} X = D_{1/\sigma} V_X D_{1/\sigma} = R,$$

Considérons le diagramme commutatif suivant :

$$\begin{tikzcd}
\mathbb{R}^p \arrow[r, "D_{1/\sigma^2}"] \arrow[d, "D_{1/\sigma}"'] & \mathbb{R}^{p*} \\
\mathbb{R}^p \arrow[r, "I_p"'] & \mathbb{R}^{p*} \arrow[u, "D_{1/\sigma}"']
\end{tikzcd}$$

Soit D_{1/σ^2},= $D_{1/\sigma}{}^t I_p D_{1/\sigma}$.

L'isomorphisme $D_{1/\sigma}$ défini sur $\mathbb{R}^p$ conserve la norme, à savoir :

$$\forall u, v \in \mathbb{R}^p, \quad \|u - v\|_{D_{1/\sigma^2}} = \left\|D_{1/\sigma}u - D_{1/\sigma}v\right\|_{I_p}$$

Soit $N_1(I)$ l'image de $N(I)$ par $D_{1/\sigma}$ alors : $N_1(I) = \{(D_{1/\sigma}x^i), i = 1 \dots n\}$, la matrice d'inertie de $N_1(I)$ s'écrit : $\sum_{i=1\dots n} p_i \left(D_{1/\sigma}x^i\right)^t \left(D_{1/\sigma}x^i\right) = D_{1/\sigma}(\sum_{i=1\dots n} p_i\,(x^i)^t(x^i))D_{1/\sigma}$=$D_{1/\sigma}V_X D_{1/\sigma} = R$, soit c une composante principale dans l'A.C.P sur $N_1(I)$ dans l'espace : ($\mathbb{R}^p$, I_p) alors : c=Zα où α est vecteur propre I_p-normé de R.

Etant donné que :

- L'isomorphisme $D_{1/\sigma}$ défini sur $\mathbb{R}^p$ conserve les distances.
- Une composante principale ne dépend que des distances entre individus et des poids.
- c est composante principale de l'A.C.P sur $N(I)$ dans l'espace euclidien ($\mathbb{R}^{p*}$, D_{1/σ^2}), Z étant le tableau de variables centrées réduites, un changement d'unité sur une variable n'affecte pas les composantes principales.

Proposition 4.6

On choisit la métrique : $M = D_{1/\sigma^2}$, alors la prmière composante principale c^1, est la plus liée aux $(x_j)_{j=1..p}$, au sens suivant : c^1, rend l'expression : $\sum_{i=1\dots p} r^2(c, x_i)$ maximale, où $r(c, x_i)$ désigne le coefficient de corrélation de c, x_i.

Preuve 4.6

Soit c un élément de $\mathbb{R}^n$**,** soit Z le tableau de données centrées réduites alors :

$$Z = (z_1, \dots, z_p) = XD_{1/\sigma}$$

$$r(c, z_j) = r(c, x_j) = \frac{cov(c, z_j)}{\sqrt{varc}\sqrt{varz_j}} = \frac{cov(c, z_j)}{\sqrt{varc}}$$

$$r^2(c, z_j) = \frac{\left(z_j^t D_p z_j\right)^2}{c^t D_p c} = \frac{z_j^t D_p z_j z_j^t D_p z_j}{c^t D_p c}$$

$$f(c) = \sum_{j=1\dots p} r^2(c, z_j) = \frac{c^t D_p \sum_{j=1\dots p} z_j^t\, z_j D_p c}{c^t D_p c} = \frac{c^t D_p Z^t Z D_p c}{c^t D_p c}$$

Une condition nécessaire pour que f soit maximale elle doit vérifier $\frac{df}{dc} = 0$,

Ceci implique : $(D_p Z^t Z D_p c)(c^t D_p c) - (c^t D_p Z^t Z D_p c) D_p c = 0$,

Ou encore : $Z^t Z D_p c = \frac{c^t D_p Z^t Z D_p c}{c^t D_p c} = f(c)c$,

c est donc vecteur propre de $Z^t Z D_p$ pour valeur propre $f(c)$, soit la valeur propre maximale, de $Z^t Z D_p$ or $Z^t Z = XD_{1/\sigma^2}X^t$donc c est vecteur propre de WD_p associé à la plus grande valeur propre.■

Description du nuage des individus

L'A.C.P construit de nouvelles variables fictives permettant des représentations graphiques entre variables d'origine et un changement d'axes permettant d'approcher aux mieux les individus de manière à rendre compte de leurs regroupements et de leurs oppositions éventuelles.

Si la part d'inertie expliquée par le plan des deux premiers axes factoriels est assez proche de 1 on possède à l'aide de projections des individus x^i sur ce plan une assez bonne représentation du nuage N(I).

Soit la décomposition de $\mathbb{R}^p$, suivants les axes factoriels :

$\mathbb{R}^p = \oplus_{j=1..p} \Delta u_j$,

x^i s'écrit : $x^i = \sum_{j=1..p} c_j^i u_j$, avec $c_j^i = x^i\ {}^t M u_j$ la i-ème composante de la j-ème composante principale, on peut aussi écrire :

$$\|x^i\|_M^2 = \sum_{j=1..p} (c_j^i)^2.$$

Qualité de la représentation d'un individu sur un axe

La qualité de la représentation d'un individu sur l'axe Δu_j, est donnée par le carré du cosinus de l'angle que forme le vecteur x^i avec l'axe Δu_j,

Définition 4.19

On appelle corrélation de l'individu vecteur x^i avec l'axe Δu_j, l'expression :

$$COR_j(x^i) = \frac{\left(x^{i^t} M u_j\right)}{\|x^i\|_M^2} = \frac{\left(c_i^j\right)^2}{\sum_{j=1..p}\left(c_i^j\right)^2}, \qquad (4.28)$$

On rappelle que c_i^j est la composante de x^i sur l'axe Δu_j,

Plus la valeur de l'égalité (4.28) est proche de 1 plus le point x^i est proche de l'axe Δu_j, meilleur est sa représentation par sa projection sur cet axe.

La qualité de la représentation de x^i sur le plan engendré par les deux axes principaux $\Delta u_j, \Delta u_k$ est donnée par :

$$QLT_{j,k}(x^i) = COR_j(x^i) + COR_k(x^i) \qquad (4.29)$$

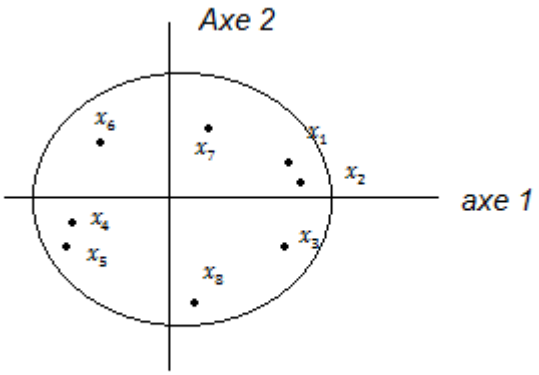
Corrélation entre composantes et variables initiales

On choisit la métrique : $M = D_{1/\sigma^2}$, pour faire l'A.C.P du tableau X on est ramené à déterminer les valeurs propres et les vecteurs propres de la matrice de corrélation R. Soit v un vecteur propre de $R\ I_p$ −normé associé à la valeur propre λ on a :

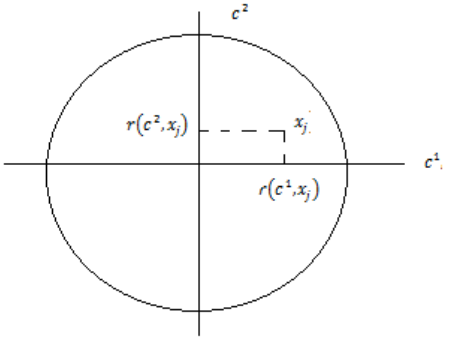
$$r(c, x_j) = r(c, z_j) = \frac{z_j{}^t D_p c}{\sqrt{varc}} = \frac{z_j{}^t D_p c}{\sqrt{\lambda}} = \frac{z_j{}^t D_p Z v}{\sqrt{\lambda}} = \sqrt{\lambda} v_j.$$

Pour un couple de composantes principales c^1, c^2 on synthétise les corrélations sur une figure appelée « cercle des corrélations », où chaque variable x_j, est repérée par un point de coordonnées : $\left(r(c^1, x_j), r(c^2, x_j)\right)$

Figure 4-4



Error! Use the Home tab to apply 0 to the text that you want to appear here.**Figure 4-5**

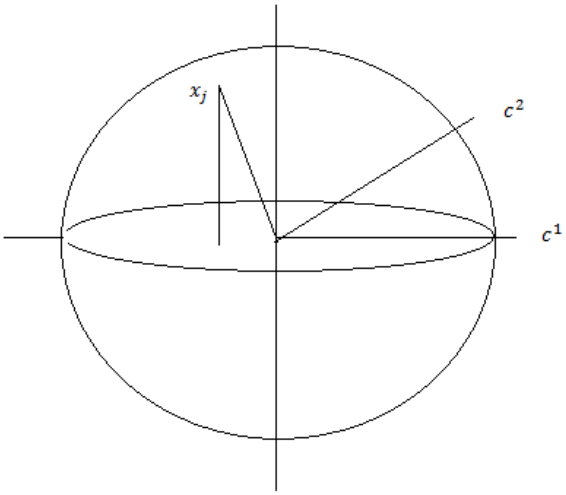


On se gardera d'interprétera des proximités entre variables si les projections de celles-ci ne sont pas proches de la circonférence.

Si la métrique choisie est $M = D_{1/\sigma^2}$, le cercle des corrélations représente la projection de l'ensemble des variables centrées réduites sur le sous-espace engendré par c^1, c^2, soit cet espace A. z_j est de variance 1, donc z_j est sur la sphère unité de $\mathbb{R}^n$, si on projette z_j sur A l'extrémité de cette projection se trouve à l'intérieur du grand cercle intersection de A et de la sphère unité. L'abscisse de z_j est

$\cos(c, z_j) = r(c, x_j) = r(c, z_j)$, le cercle des corrélations est le pendant de la projection des individus sur le plan principal.

Figure 4-6



Importance d'un individu dans la détermination d'un axe

On a :

$$I_{\Delta u}{}^{\perp}{}_{k}\big(N(I)\big) = \sum_{i=1..n} I_{\Delta u}{}^{\perp}{}_{k}(x^i) = \sum_{i=1..n} p_i\big(c_i^k\big)^2 \qquad (4.30)$$

Définition 4.20

On appelle contribution de l'individu x^i à l'axe Δu_k (notée CT$R_k(x^i)$) le rapport :

$$\frac{p_i\big(c_i^k\big)^2}{\sum_{i=1..n} p_i\big(c_i^k\big)^2} \qquad (4.31)$$

Comme rapport d'un élément d'une somme de nombres positifs à la somme totale.

On peut aussi écrire :

$$\mathrm{CT}R_k(x^i) = \frac{p_i\big(c_i^k\big)^2}{\lambda_k} = \frac{I_{\Delta u}{}^{\perp}{}_{k}(x^i)}{I_{\Delta u}{}^{\perp}{}_{k}\big(N(I)\big)} \qquad (4.32)$$

Ce rapport est d'autant plus proche de 1 que l'individu x^i intervient dans le calcul de la variance de c^k.

Définition 4.21

On appelle taux d'inertie expliquée par l'axe Δu_k la quantité :

$$\tau_k = 100 \times \frac{\lambda_k}{\sum_{i=1..p} \lambda_i} = 100 \times \frac{I_{\Delta u}{}^{\perp}{}_{k}\big(N(I)\big)}{I_O\big(N(I)\big)} \qquad (4.33)$$

Cette expression mesure l'importance relative de l'axe Δu_k les valeurs propres étant rangées par ordre décroissant, les taux d'inertie expliquée par les axes vont eux aussi en décroissant.

Formule de reconstitution du tableau X

Connaissant les résultats de l'A.C.P d'un tableau X, soient les facteurs $(\alpha_k)_{k=1\ldots p}$ et les composantes principales $(c^k)_{k=1\ldots p}$, on peut « reconstituer » le tableau d'origine X en posant : $c^k = X\alpha_k$ par suite : : $c^k{\alpha_k}^t M^{-1}$=$X\alpha_k{\alpha_k}^t\ M^{-1}$ en sommant sur k il vient :

$$\sum_{k=1\ldots p} c^k{\alpha_k}^t M^{-1} = X\sum_{k=1..p}\alpha_k{\alpha_k}^t\ M^{-1} = X.$$

Proposition 4.6

Le tableau d'origine X s'exprime en fonction des résultats de l'A.C.P par la formule :

$$\sum_{k=1\ldots p} c^k{\alpha_k}^t M^{-1} = X\sum_{k=1..p}\alpha_k{\alpha_k}^t\ M^{-1} = X.$$

Preuve 4.6

$\left(\sum_{k=1..p}\alpha_k{\alpha_k}^t\ M^{-1}\right)\alpha_j = \sum_{k=1..p}\alpha_k{\alpha_k}^t\ M^{-1}\,\alpha_j = \sum\ \ \alpha_k\delta_{kj} = \alpha_j$ donc : $\sum_{k=1..p}\alpha_k{\alpha_k}^t\ M^{-1} = I_p$. ■

Exemple 4.1

<u>Analyse en composantes principales :</u>

Cet exemple est tiré de http://factominer.free.fr/classical-methods/analyse-en-composantes-principales.html.

On utise ici l'exemple du tableau de données 'decathlon' qui contient les performences réalisées par des athlètes lors de deux compétitions, vous pouvez charger le jeu de données ici.

	row.names	X100m	Long.jump	Shot.put	High.jump	X400m	X110m.hurdle	Discus	Pole.vault	Javeline	X1500m	Rank	Points	Competition
1	SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.7	1	8217	Decastar
2	CLAY	10.76	7.4	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.5	2	8122	Decastar
3	KARPOV	11.02	7.3	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.2	3	8099	Decastar
4	BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.1	4	8067	Decastar
5	YURKOV	11.34	7.09	15.19	2.1	50.42	15.31	46.26	4.72	63.44	276.4	5	8036	Decastar
6	WARNERS	11.11	7.6	14.31	1.98	48.68	14.23	41.1	4.92	51.77	278.1	6	8030	Decastar
7	ZSIVOCZKY	11.13	7.3	13.48	2.01	48.62	14.17	45.67	4.42	55.37	268	7	8004	Decastar
8	McMULLEN	10.83	7.31	13.76	2.13	49.91	14.38	44.41	4.42	56.37	285.1	8	7995	Decastar
9	MARTINEAU	11.64	6.81	14.57	1.95	50.14	14.93	47.6	4.92	52.33	262.1	9	7802	Decastar
10	HERNU	11.37	7.56	14.41	1.86	51.1	15.06	44.99	4.82	57.19	285.1	10	7733	Decastar
11	BARRAS	11.33	6.97	14.09	1.95	49.48	14.48	42.1	4.72	55.4	282	11	7708	Decastar
12	NOOL	11.33	7.27	12.68	1.98	49.2	15.29	37.92	4.62	57.44	266.6	12	7651	Decastar
13	BOURGUIGNON	11.36	6.8	13.46	1.86	51.16	15.67	40.49	5.02	54.68	291.7	13	7313	Decastar
14	Sebrle	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5	70.52	280.01	1	8893	OlympicG
15	Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.9	69.71	282	2	8820	OlympicG
16	Karpov	10.5	7.81	15.93	2.09	46.81	13.97	51.65	4.6	55.54	278.11	3	8725	OlympicG
17	Macey	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.4	58.46	265.42	4	8414	OlympicG
18	Warners	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.9	55.39	278.05	5	8343	OlympicG
19	Zsivoczky	10.91	7.14	15.31	2.12	49.4	14.95	45.62	4.7	63.45	269.54	6	8287	OlympicG
20	Hernu	10.97	7.19	14.65	2.03	48.73	14.25	44.72	4.8	57.76	264.35	7	8237	OlympicG
21	Nool	10.8	7.53	14.26	1.88	48.81	14.8	42.05	5.4	61.33	276.33	8	8235	OlympicG
22	Bernard	10.69	7.48	14.8	2.12	49.13	14.17	44.75	4.4	55.27	276.31	9	8225	OlympicG
23	Schwarzl	10.98	7.49	14.01	1.94	49.76	14.25	42.43	5.1	56.32	273.56	10	8102	OlympicG
24	Pogorelov	10.95	7.31	15.1	2.06	50.79	14.21	44.6	5	53.45	287.63	11	8084	OlympicG
25	Schoenbeck	10.9	7.3	14.77	1.88	50.3	14.34	44.41	5	60.89	278.82	12	8077	OlympicG
26	Barras	11.14	6.99	14.91	1.94	49.41	14.37	44.83	4.6	64.55	267.09	13	8067	OlympicG
27	Smith	10.85	6.81	15.24	1.91	49.27	14.01	49.02	4.2	61.52	272.74	14	8023	OlympicG
28	Averyanov	10.55	7.34	14.44	1.94	49.72	14.39	39.88	4.8	54.51	271.02	15	8021	OlympicG
29	Ojaniemi	10.68	7.5	14.97	1.94	49.12	15.01	40.35	4.6	59.26	275.71	16	8006	OlympicG
30	Smirnov	10.89	7.07	13.88	1.94	49.11	14.77	42.47	4.7	60.88	263.31	17	7993	OlympicG
31	Qi	11.06	7.34	13.55	1.97	49.65	14.78	45.13	4.5	60.79	272.63	18	7934	OlympicG
32	Drews	10.87	7.38	13.07	1.88	48.51	14.01	40.11	5	51.53	274.21	19	7926	OlympicG
33	Parkhomenko	11.14	6.61	15.69	2.03	51.04	14.88	41.9	4.8	65.82	277.94	20	7918	OlympicG
34	Terek	10.92	6.94	15.15	1.94	49.56	15.12	45.62	5.3	50.62	290.36	21	7893	OlympicG
35	Gomez	11.08	7.26	14.57	1.85	48.61	14.41	40.95	4.4	60.71	269.7	22	7865	OlympicG
36	Turi	11.08	6.91	13.62	2.03	51.67	14.26	39.83	4.8	59.34	290.01	23	7708	OlympicG
37	Lorenzo	11.1	7.03	13.22	1.85	49.34	15.38	40.22	4.5	58.36	263.08	24	7592	OlympicG
38	Karlivans	11.33	7.26	13.3	1.97	50.54	14.98	43.34	4.5	52.92	278.67	25	7583	OlympicG
39	Korkizoglou	10.86	7.07	14.81	1.94	51.16	14.96	46.07	4.7	53.05	317	26	7573	OlympicG

Présentation des données :

Le tableau de données contient 41 lignes et 13 colonnes.

Les colonnes de 1 à 12 sont des variables continues, les dix premières épreuves correspondent aux performences des athlètes pour les dix épreuves du décathlon et les colonnes 11 et 12 correspondent au rang et au nombre de points obtenus. La dernière colonne est une variable quantitative correspondant au nom de la compétition (jeux olympiques ou decastar 2004).

Télécharger R

Charger le package **FactomiineR** et le tableau de données en tapant les lignes de code suivantes :

```
library(FactoMineR)
data(decathlon)
```

Charger :

```
library(Rcmdr)
```

On arrive à une interface Gui, dans cette interface cocher le package **FactoMineR** et dans options cocher la case console R, charger le jeu de données 'decathlon' dans la rubrique 'données dans le package' .

Objectifs :

L' A.C.P permet de décrire un jeu de données, de le résumer, d'en réduire la dimensionnalité.

L' A.C.P réalise sur les individus du tableau de données répond à différentes questions :

1- Etude des individus(ie : des athlètes) : deux athlètes sont proches s'ils ont des résultats similaires. On s'intéresse la variabilité entre individus. Y a-t - il des profiles d'athlètes ? Peut-on opposer un groupe d'individus à un autre ?
2- Etude des variables (ie : des performances) : on étudie les liaisons linéaires entre les variables. Les objectifs sont de résumer la matrice des corrélations et de chercher des variables synthétiques : peut-on résumer les performences des athlètes par un petit nombre de variables ?
3- Lien entre les deux études : peut-on caractériser des groupes d'individus par des variables ?

A.C.P :

On étudie les profils d'athlètes uniquement en fonction de leur performance. Les variables actives ne seront donc que celles qui concernent les dix épreuves du décathlon. Les autres variables ('Rank', 'points' et 'competition') n'appartiennent pas aux profils d'athlètes et utilisent une information déjà donnée par les autres variables (dans le cas de 'Rank', 'Points') mais il intéressant de les confronter aux composantes principales. Nous les utiliserons comme variables illustratives. Dans ce tableau de données, les variables mesurées dans les mêmes unités. On doit les réduire de façon à donner la même influence à chacune.

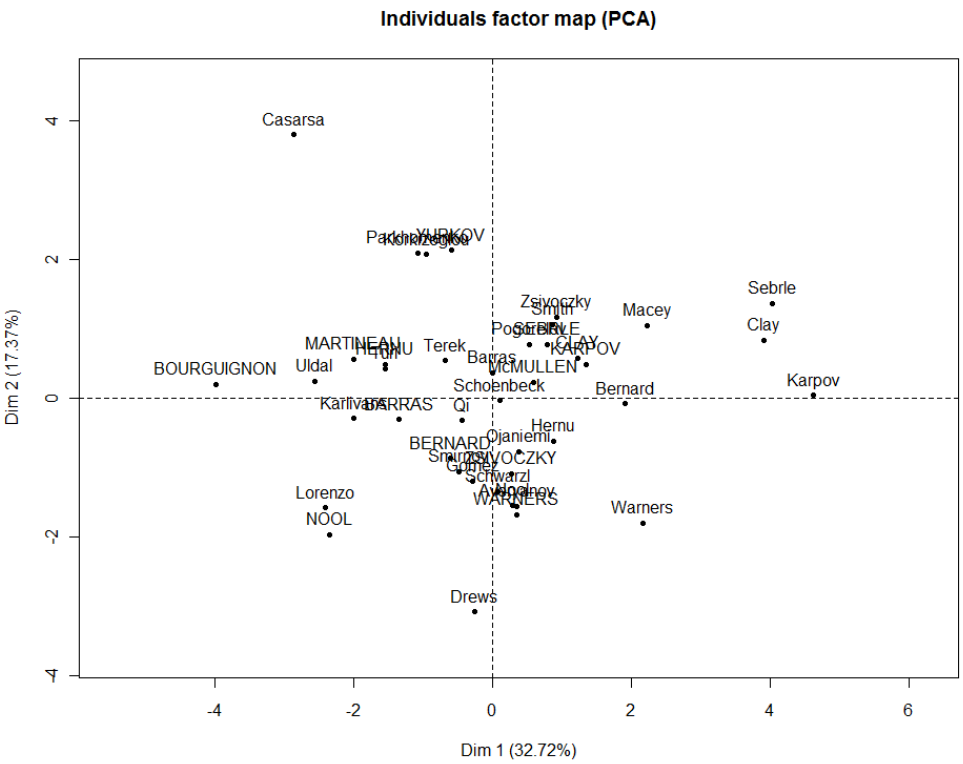
Individus et variables actives :

Réaliser l'A.C.P sur tous les individus et seulement les variables actives (ie : les dix premières) en tapant la ligne :

```
res.pca = PCA(decathlon[,1:10], scale.unit=TRUE, ncp=5, graph=T)
```

```
#decathlon: le tableau de données utilisé
#scale.unit: pour choisir de réduire ou non les variables
#ncp: le nombre de dimensions à garder dans les résultats
#graph: pour choisir de faire apparaître les graph
```

Figure 4-8



Les deux premières dimensions contiennent 50% de l'inertie totale(l'inertie est la variance totale du tableau e données, ie : la trace de la matrice des corrélations).

La variable *"X100m"* est négativement corrélée à la variable *"long.jump"*. Quand un ahtlète réalise un temps faible au 100m, il peut sauter loin. Il faut faire attention ici qu'une petite valeur pour les variables *"X100m"*, *"X400m"*, *"X110m.hurdle"* et*"X1500m"* correspond à un score élevé : plus un athlète court rapidement, plus il gagne de points.

Le premier axe oppose les athlètes qui sont "bons partout" comme Karpov pendant les Jeux Olympiques à ceux qui sont "mauvais partout" comme Bourguignon pendant le Décastar. Cette dimension est particulièrement liée aux variables de vitesse et de saut en longueur qui constituent un groupe homogène.

Le deuxième axe oppse les athlètes qui sont forts (variables *"Discus"* et *"Shot.put"*) à ceuw qui ne le sont pas.
Les variables *"Discus"*, *"Shot.put"* et *"High.jump"* ne sont pas très corrélées aux variables *"X100m"*, *"X400m"*, *"X110m.hurdle"* et *"Long.jump"*. Cela signifie que force et vitesse ne sont pas très corrélées.

A l'issue de cette première approche, on peut diviser le premier plan factoriel en quatre parties : les athlètes rapides et puissants (comme Sebrle), les athlètes lents (comme Casarsa), les athlètes rapides mais faibles (comme Warners) et les ahtlètes ni forts ni rapides, relativement parlant (comme Lorenzo).

Variables illustratives :

Les variables illustratives n'influencent pas la construction des composantes principales de l'analyse. Elles aident à l'interprétation des dimensions de variabilité.

On peut ajouter deux types de variables : continues et qualitatives.

On ajoute les variables *"Rank"* and *"Points"* comme variables continues illustratives. Tapez la ligne de code suivante :

```
res.pca = PCA(decathlon, scale.unit=TRUE, ncp=5,graph=T)
```

```
#decathlon: le tableau de données utilisé
#scale.unit: pour choisir de réduire ou non les variables
#ncp: le nombre de dimensions à garder dans les résultats
#quanti.sup: vecteur des index des variables continues illustratives
```

#graph: pour choisir de faire apparaître les graphiques ou non

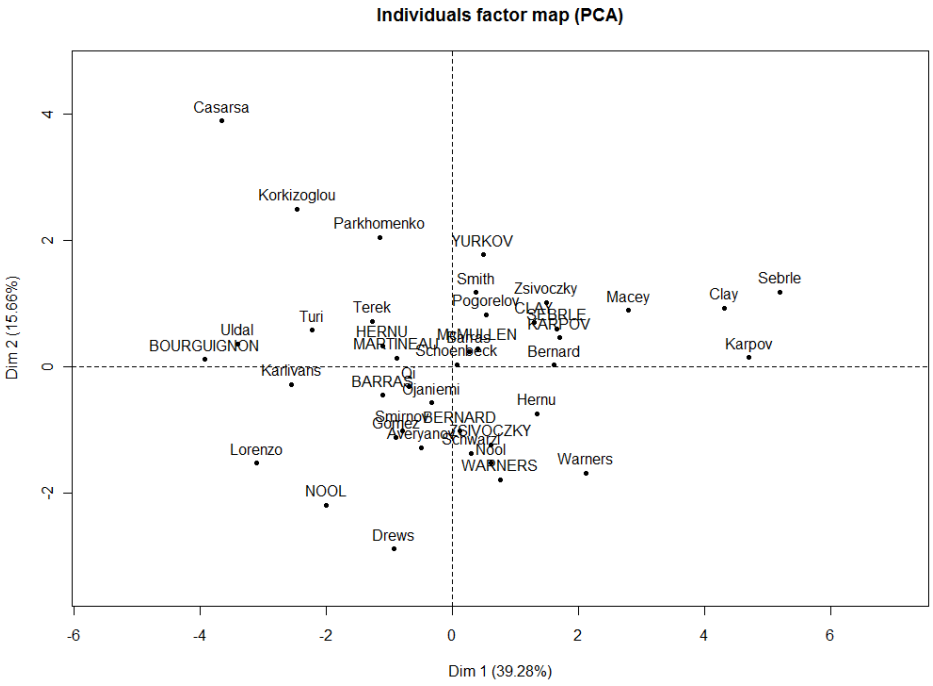


Figure 4-9

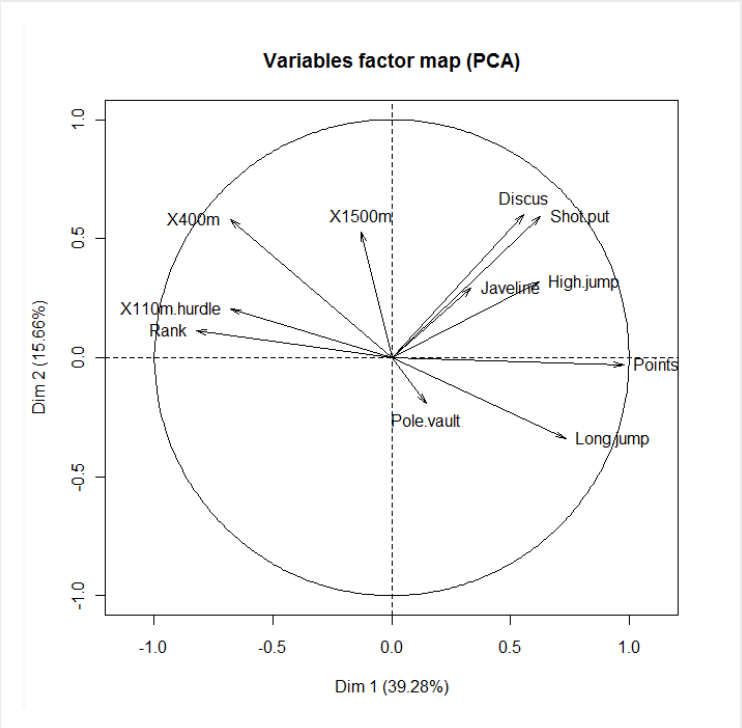


Figure 4-10

Les gagnants du décathlon sont ceux qui marquent le plus de points (ou ceux dont le rang est faible). Les variables les plus liées au nombre de points sont les variables qui réfèrent à la vitesse (*"X100m", "X110m.hurdle", "X400m"*) et au

saut en longueur. Au contraire,*"Pole-vault"* et *"X1500m"* n'ont pas une grande influence sur le nombre de points. Les athlètes qui sont bons à ces deux épreuves ne sont pas favorisés.

On ajoute la variable "Competition" comme variable qualitative illustrative :

```
res.pca = PCA(decathlon, scale.unit=TRUE, ncp=5, quanti.sup=c(11: 12), quali.sup=13, graph=T)
#decathlon: le tableau de données utilisé
#scale.unit: pour choisir de réduire ou non les variables
#ncp: le nombre de dimensions à garder dans les résultats
#quanti.sup: vecteur des index des variables continues illustratives
#quali.sup: vecteur des index des variables qualitatives illustratives
#graph: pour choisir de faire apparaître les graphiques ou non
```

Les centres de gravité des modalités de cette variable supplémentaire apparaissent sur le graphe des individus. Ils sont localisés au barycentre des individus qui les possèdent et représentent un individu moyen.

On peut également colorier les individus selon la couleur des centres de gravité des modalités :

```
plot.PCA(res.pca, axes=c(1, 2), choix="ind")
#res.pca: le résultat d'une ACP
#axes: les axes à représenter
#choix: le graph à représenter ("ind" pour les individus, "var" pour les variables)
```

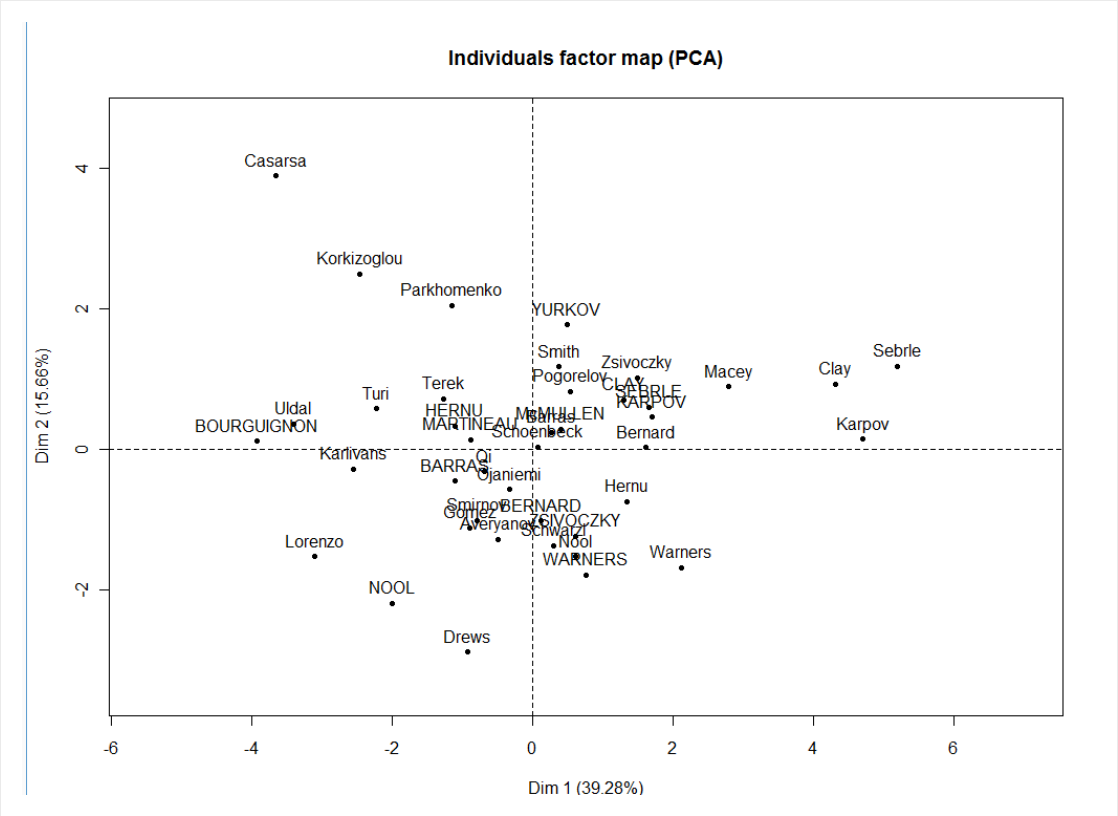


Figure 4-11

En regardant les points qui représentent *"Decastar"* et *"Olympic Games"*, on voit que*"Olympic Games"* a une coordonnée plus élevée sur le premier axe que *"Decastar"*. Ceci montre une évolution des performances des athlètes. Tous les athlètes qui ont participé aux deux compétitions ont obtenu des résultats légèrement meilleurs aux jeux Olympiques.
Cependant, il n'y a aucune différence entre les points *"Decastar"* et *"Olympic Games"*sur le deuxième axe. Cela signifie que les athlètes ont amélioré leurs performances mais n'ont pas changé de profil (à l'exception de Zsivoczky qui est passé de lent et fort pendant le Décastar à rapide et faible pendant les Jeux Olympiques).

Les points qui représentent un même individu vont dans le même direction. Par exemple, Sebrle a obtenu de bons résultats aux deux compétitions mais le point qui représente sa performance aux J.O. est plus extrême. Sebrle a obtenu plus de points pendant les J.O. que pendant le Décastar..

On peut envisager deux interprétations :

- Les athlètes qui participent aux J.O. sont meilleurs que ceux qui participent au Décastar
- Les athlètes font de leur mieux aux J.O. (plus motivés, plus entraînés)

La fonction *dimdesc()* permet de décrire les dimensions.
On peut décrire les dimensions données par les variables en écrivant :

```
dimdesc(res.pca, axes=c(1,2))#res.pca: le résultat d'une PCA
#axes: les axes choisi
```

```
$Dim.1$quanti
              correlation      p.value
Points          0.9765361 1.277562e-27
Long.jump       0.7332187 4.981554e-08
Shot.put        0.6264887 1.177321e-05
High.jump       0.6192263 1.587077e-05
Discus          0.5566053 1.573425e-04
Javeline        0.3344155 3.259493e-02
X110m.hurdle   -0.6775081 1.143028e-06
X400m          -0.6796006 1.028661e-06
Rank           -0.8206733 5.070567e-11


$Dim.2
$Dim.2$quanti
          correlation      p.value
Discus      0.6010837 3.242251e-05
Shot.put    0.5935892 4.300945e-05
X400m       0.5805188 6.925042e-05
X1500m      0.5279666 3.886354e-04
High.jump   0.3166238 4.370533e-02
Long.jump  -0.3403452 2.945501e-02
```

La fonction *dimdesc()* calcule le coefficient de corrélation entre une variable et une dimension et réalise un test de significativité. Ces tableaux donnent le coefficient de corrélation et la probabilité critique des variables qui sont significativement corrélées aux dimensions principales. Les variables actives et illustratives dont le probabilité critique est inférieure à 0.05 apparaissent.
Les tableaux de la description des deux axes principaux montrent que les variables*"Points"* et *"Long.jump"* sont les plus corrélées à la première dimension et que *"Discus"*est la variable la plus corrélée à la deuxième dimension. Ceci confirme la premièer interprétation.

Cette fonction est très utile lorsqu'on est en présence d'un nombre élevé de variables. Elle facilite l'interprétation.

Si on ne veut pas qu'un (ou plusieurs) individu participe à l'analyse, il est possible de l'ajouter en tant qu'individu illustratif. Ainsi, il ne sera pas actif dans l'analyse mais apportera de l'information supplémentaire.
Pour ajouter des individus illustratifs, utilisez l'argument suivant de la fonction A.C.P :

```
ind.sup
```

Tous les résultats détaillés peuvent être vus dans l'objet *res.pca*. On peut récupérer les valeurs propres, les résultats des individus actifs et illustratifs, les résultats des variables actives et les résultats des variables continues et qualitatives illustratives en tapant :

```
res.pcanames(res.pca)res.pca$eig, res.pca$ind, res.pca$ind.sup, res.pca$var,
```

```
res.pca$quali.sup, res.pca$quanti.sup
```

<u>Pour Aller Plus Loin :</u>

Dans la plupart des ACPs, le poids des individus est égal à 1/(nombre d'individus). Néanmoins, il est parfois nécessaire de donner un poids spécifique à certains individuss.
Pour cela, l'utilisateur doit utiliser l'argument :

```
row.w
```

Il peut également être intéressant de donner un poids spécifique à certaines variables.
L'utilisateur doit alors utiliser l'argument :

```
col.w
```

Pour voir si les modalités d'une variable illustrative sont significativement différentes les unes des autres, on peut tracer des ellipses de confiance autour d'elle.
Pour ce faire, écrivez :

```
concat = cbind.data.frame(decathlon[,13],res.pca$ind$coord)
```

```
ellipse.coord = coord.ellipse(concat,bary=T)
```

```
plot.PCA(res.pca,habillage=13,ellipse=ellipse.coord,cex=0.8)
```

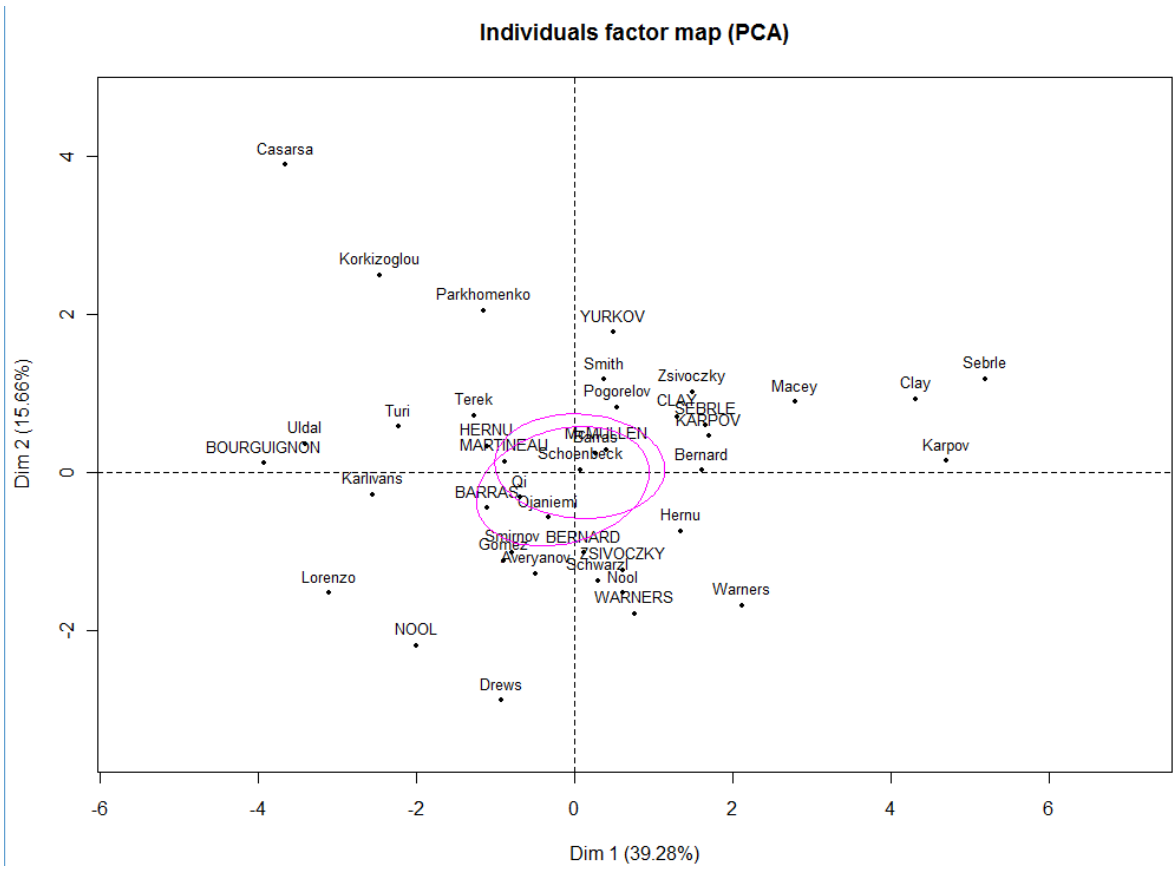
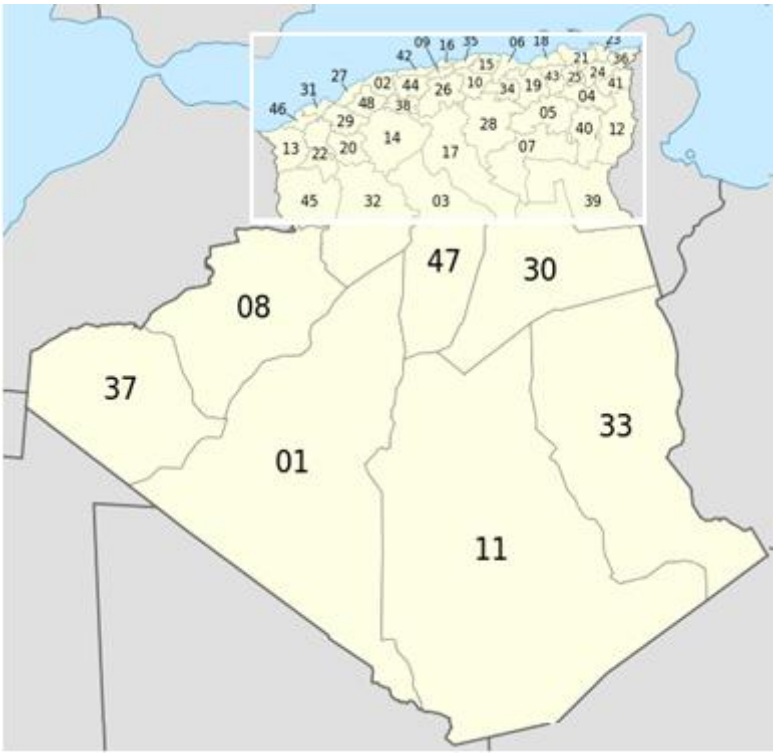


Figure 4-11

Exemple : Une étude sur le réseau téléphonique des départements d'Algérie :



	IGQS	IZAA	EZAA	TSI	VR2	TCR	TCOM
Alger	61.4	58.6	56.6	43.3	53	31.6	13.4
Oran	89.1	66.3	66.2	28	83.9	68	18.7
Tlemcen	95.7	69.5	68	19	87.5	68.3	15.2
Constantir	89.3	58.6	66	29.2	84.6	70.6	15.9
Annaba	88	66.3	66.1	31.5	92.9	74.1	15.6
Bel Abbès	95.8	69.5	66.6	22.5	93.4	78.9	16.6
Adrar	96	67.2	69.3	22.5	87.5	75.6	16.1
Skikda	93.4	67.1	68.3	23.2	86.7	74.6	16.3
Tebessa	91.8	67.9	67.1	28.5	90.2	58.5	15.5
Tizi-Ouzou	83.4	70.8	64.7	38.6	78.9	66.4	16.3
Gardhaia	93.5	68.1	68.4	21.2	86.1	67.8	15.5
Tindouf	87.1	67.1	63.9	29.7	86.3	54.4	17.3
	89.3	66.7	64.7	27.6	85.9	60.7	16.2
	93.8	67.4	68.9	25.4	89.7	73.1	15.5
	91.8	64	67.8	28.9	88.3	74.3	16.6
	92.7	66.3	67	25.7	89	70.2	16.7
	90.2	67.6	66.8	28.5	84.3	74.2	16.3
	91.5	67.8	67.9	28.1	85.5	66.2	17.3
	87.4	67.6	67.8	31.5	7.7	71.3	15.1
	97.8	67.7	70.1	18.8	93.5	82	15.8
	88	67	66.9	30.5	80.3	70.2	16.2
	98.2	66.6	68.4	23.2	89.8	54.1	17.1

Dans le but d'expliquer les indicateurs de qualité de service, on totalise les indicateurs technico-économiques suivants :

1- ACLP. Accroissement brut du parc de lignes principales
2- DLAI délai moyen de raccordement
3- QRAC indice de qualité de raccordement
4- SDCM demande en instance inférieure à 1 an
5- UNAN demande en instance inférieure à 15 jours
6- DEM1 delai de réalisation inférieur à 15 jours
7- DEM2 delai de réalisation inférieur à 30 jours
8- TFPB trafic financier des postes publiés

Qualité de service du réseau téléphonique

La gestion de la qualité de service du réseau téléphonique est fondée sur des indicateurs. Les données présentées concernée les indicateurs clients qui servent à mesurer de façon uniforme la qualité de service offerte aux usagers.

Les données de base sont publiées par région et pour huit indicateurs élémentaires et un indice global combinaison des six indicateurs élémentaires.

Voici la signification des indicateurs avec leurs règles :

1 IGQS – indice global de qualité de service. Combinaison des six autres indicateurs (n° 2 à 7) ; intervalle de variation [0,100]

2 ATON- pourcentage d'appels ayant attendu la tonalité d'invitation à transmettre plus de 3 '' par rapport au nombre total d'appels effectués.

3 IZAA- pourcentage d'appels efficaces par rapport au nombre des appels correctement effectués par les usagers, en ne tenant compte que des appels intérieurs à la zone autonome d'acheminement (ZAA) où à la zone urbaine (ZU)

4 EZAA- définition identique au cas précédent en remplaçant 'intérieur' par 'extérieurs' (appels à destination de l'étranger non pris en considération)

5 TSI – taux de signalisation des dérangements reçus et traités durant une période donnée par rapport au nombre de lignes principales en service.

6 VR2- vitesse de relevé des dérangements dans les 2 jours. C'est le pourcentage de dérangements relevés le jour même et le lendemain de la signalisation par rapport au nombre de dérangements qui ont donné lieu à déplacent d'agents.

7 VR8- vitesse de relevé des dérangements dans les 8 jours : définition identique à la précédente en remplacement le jour même et le lendemain par le jour même et les sept jours suivants.

8 TCR- taux d'efficacité à l'arrivée des centres de renseignement pourcentage d'appels efficaces par rapport au nombre des appels au service des renseignements.

9 TCOM- temps d'établissement des communications : délai exprimé en secondes de mise en relation de deux abonnés.

Les unités d'observations sont constituées des directions régionales des télécommunications : 22 régions ont été retenues. Chacune d'entre elles est identifiée par le nom de la ville représentative de la région.

Remarque. Deux indicateurs n'ont pas été pris en compte dans les tableaux des données : ATON, variable 'attente de tonalité' n'est pas assez précise et fiable dans les tableaux statistiques publiés pour devoir être en considération : VR8 'vitesse de relevé dans les 8 jours' a des valeurs presque constantes (=100)

Voici donc d'une part la carte des unités statistique, d'autre part des tableaux statistiques de base.

Dans le but d'expliquer les indicateurs de qualité de service, on totalise les indicateurs technico-économiques suivants :

1 ACLP accroissement brut du parc de lignes principales

2 DLAL délai moyen de raccordement

QRAC indice de qualité de raccordement

SIXM demande en instance inférieure à 1 an

UNAN demande en instance inférieur à 15 jours

DEM1 délai de réalisation inférieur à 15

DEM2 délai de réalisation inférieur à 30

TFPB trafic financier des postes publiés.

Tableau de données de base(rappel)

Indicateurs de qualité

I / J	1	izaa	ezaa	tsi	Vr2	tcr	tcom
1	61.4	58.6	55.6	43.3	53	31.6	13.4
2	89.1	66.3	66.2	28	83.9	68	18.7
3	95.7	69.5	68	19	87.5	68.3	19.2
4	89.3	67.2	66.1	29.2	84.6	70.6	15.9
5	88	67.1	66.6	31.5	82.9	74.1	15.6
6	95.8	67.9	68.1	22.5	93.4	78.9	166
7	96	70.8	69.3	22.5	87.5	75.6	15.2
8	93.4	68.1	68.3	23.2	86.7	74.6	15.8
9	91.8	67.1	67.7	28.5	9.2	58.5	15.6
10	83.4	66.7	64.7	38.6	78.9	66.4	16.6
11	93.5	67.4	68.4	21.2	86.1	67.8	16.6
12	87.1	66.3	63.9	29.7	86.3	54.4	16.1
13	89.3	67.6	64.7	27.6	85.9	60.7	16.3

Tiaret	93.8	67.8	68.9	25.4	89.7	73.1	15.5
15	91.8	67.6	67.8	25.4	88.3	74.3	16.8
16	92.7	67.7	67	28.9	89	70.2	16.5
17	90.2	68.3	66.8	25.7	84.3	74.2	16.3
18	91.5	68.1	67.9	28.5	85.5	66.2	17.3
19	87.4	67.9	67.8	28.1	78.7	71.3	5.1
20	97.8	67.ç	70.1	31.5	93.5	82	15.8
21	88	67 .6	66.9	30.5	80.3	70.2	16.2
22	98.2	66.6	68.4	23.2	89.8	54.1	17.1

Tableau des données : qualité de service pour l'année 1984 par région

Valeurs propres de la matrice des corrélations

numéro	itération	Valeur propre	pourcentage	cumul
1	0	3.82	54.61	54.6
2	1	1.52	21.78	76.3
3	1	.82	11.71	88.1
4	3	.46	6.7	94.8
5	5	1.18	2.5	97.4
6	1	.16	2.2	99.68
7	2	.02	.3	100

TABLEAU DES VALEURS PROPRES DE LA MATRICE DES CORRELATIONS ;

Numéro=numéro d'ordre des valeurs propres classées dans l'ordre décroissant

Itération= nombre d'itérations du programme de calcul de diagonalisation pour obtenir le résultat numérique.

Valeur propre= valeur propre de la matrice des corrélations

Pourcentage= pourcentage de variance de chaque axe factoriel.

Cumul= pourcentage de variance cumulé en ne considérant que 1 puis 2pour axes factoriels.

Matrice des corrélations des variables

Tous les coefficients sont multipliés par 1000

	IGQS	IZAA	EZAA	TSI	VR2	TCR	TCOM
igqs	1000						
izaa	489	1000					
ezaa	808	740	1000				
tsi	-918	-435	-705	1000			
Vr2	829	131	503	-744	1000		
tcr	230	597	547	-251	131	1000	
tccm	-224	-94	-114	-179	-261	-71	1000

P(tcom,vr2)=0.261

=coefficient de corrélation linéaire entre les variables tcom et vr2 ;

	C^1	C^2
IGQS	-944	252
IZAA	-674	-600
EZAA	-909	-221
TSI,	892	-233

VR2	-748	494
TCR	-505	-646
TCOM	-202	580

Interprétation du premier axe factoriel ($r_1 = (54.6.\lambda_1 = 3;8)$.

Le premier axe caractérise une opposition entre les indices suivants : TSI

D'une part et l'IGQS, L'EZAA, L'IZAA et VR2 d'autre part : c'est une opposition de codage des données dans la mesure où quand la qualité générale augmente (IGQS, IZAA, EZAA,VR2) la variable TSI diminue.

Le premier facture est donc un facteur de qualité globale générale de la qualité de service caractérisé par la position de l'IGQS.

Interprétation du deuxième axe factoriel ($r_2 = 21.6 : \lambda_2 = 1.52$).

L'importance deux fois moindre que le premier, il nuance le facteur global, la qualité de service en introduisant une dimension qui ne suit pas le même sens de variation des variables. Ainsi, voit-on une opposition sur cette axe entre les variables IZAA(-600) et TCR(--646) d'une part, et les variables VR2(494) et TCCOM(580), d'autre part. Par cet axe, on introduit ce qu'on peut appeler un facteur de qualité de deuxième espèce(non corrélé avec le premier) Ceci montre que la qualité de service n'est pas réductible à l'indice global de la qualité de IGQS qui, synthétique, ne prend pas en compte toutes les dimensions du problème.

Interprétation du plan factoriel composé des deux premiers axes ($+, r_2 = 76.3\%$).

Ce plan représente 76.3% de la variance expliquée. La qualité de service semble bien représentée par ces deux premiers axes factoriels. Il aurait été imprudent de ne considérer qu'un axe factoriel (54%) pour caractériser la qualité de service.

Chapitre 5 Analyse canonique

Soit une population de n individus sur lesquels on a relevé deux ensembles de caractères numériques, en nombre respectivement p et q.

On se pose la question de savoir si ces deux ensembles mesurent ou non les mêmes propriétés. On suppose p et q nettement inférieurs à n.

Exemple 5.1

Comparer deux ensembles de notes dans disciplines littéraires et scientifiques dans une population scolaire.

Exemple 5.2

Comparer les résultats d'analyses médicales faites par deux laboratoires différents.

Le tableau de données se présente sous la forme suivante :

$X = (X_1/X_2)$, où X_1 est le tableau à n lignes et p colonnes des mesures du premier caractère et X_2 est le tableau à n lignes et q colonnes des mesures du second.

Soient les deux sous-espaces de $\mathbb{R}^n$, engendrés par les colonnes de X_1, X_2 respectivement.

$$W_1 = \{x \in \mathbb{R}^n, x = X_1 b, b \in \mathbb{R}^{p*}\},$$
$$W_2 = \{x \in \mathbb{R}^n, x = X_2 b, b \in \mathbb{R}^{q*}\}.$$

Un élément de W_1 (resp W_2) est une combinaison des p premières variables (resp q dernières variables).

Définition 5.1

On appelle potentiels de prévision les sous-espaces W_1 , W_2.

<u>Schéma de dualité :</u>

Soient les applications x_1 (resp ,x_2) définies comme en A.C.P dont la matrice par rapport aux bases canoniques de $\mathbb{R}^n$, et de $\mathbb{R}^{p*}$ $(resp$ $\mathbb{R}^{q*})$ est X_1, $(resp\ X_2)$

$$\begin{tikzcd}
\mathbb{R}^p \arrow[r, "x_1^t"] & \mathbb{R}^{n*} \arrow[r, "x_2^t"] & \mathbb{R}^q \\
\mathbb{R}^p \arrow[r, "x_1"] & \mathbb{R}^{n*} \arrow[u, "M"] & \mathbb{R}^q \arrow[l, "x_2"']
\end{tikzcd}$$

Soient $\left(\mathbb{R}^{p*}, \left(e_j^*\right)_{j=1..p}\right)$, $(\mathbb{R}^n, (f_i)_{i=1..n})$, $X_1 = \left(x_j^i\right)_{i=1..n,,j=1..p}$, on note $x_j$le j-ème vecteur colonne de X_1, on a : $X_1\left(e_j^*\right) = x_j$. Soit $x \in W_1$, x un caractère fictif, combinaison linéaire des p caractères x_j, $x = \sum_{j=1..p} a_j x_j$, on peut écrire : $W_1 = x_1(\mathbb{R}^{p*})$, $W_2 = x_2(\mathbb{R}^{q*})$.

Le problème qui se pose quels sont les liens entre ces deux sous-espace ?

Si $W_1 = W_2$ alors les deux ensembles de variables ont même pouvoir de description.

Si $W_1 = {W_2}^{\perp}$ pour une m étrique M donnée sur $\mathbb{R}^n$, les deux ensembles représetent des caractéristiques complétement différentes.

Entre ces deux positions extrêmes on peut aussi s'intéresser aux espaces : $W_1 \cap W_2$, $W_1 \cap {W_2}^{\perp}$ et ${W_1}^{\perp} \cap W_2$.

Recherche des variables canoniques

<u>Position du problème :</u>

On considère $\mathbb{R}^n$ muni de la métrique M. On cherche un couple $(\xi, \eta) \in \mathrm{W}_1 \times \mathrm{W}_2$, M −normés tel que le M-angle entre ξ, η soit le plus petit possible, soit encore tel que la corrélation entre ξ, η soit maximale.

Supposons que ce problème ait une solution notée : $\left(\xi_1, \eta_1\right)$. Dans ce cas on cherche $(\xi, \eta) \in \mathrm{W}_1 \times \mathrm{W}_2$, M −normés avec la contrainte :

$\xi \perp \xi_1, \eta \perp \eta_1$tel que le M-angle entre ξ, η soit le plus petit possible. Si ce problème admet une solution on la note : $\left(\xi_2, \eta_2\right)$.

On cherche ainsi à mettre en place un ensemble de variables $\left(\xi_k, \eta_k\right)_{k=1..\min(p,q)}$, dites variables canoniques. Géométriquement cela revient à déterminer les lignes de plus grande pente des deux sous-espaces.

<u>Opérateur de projection orthogonale sur un sous-espace W_1 :</u>

Soit A_1 l'opérateur de projection sur W_1 . Soit u $\in \mathbb{R}^n$ et soit $u_1$sa projection sur W_1, on suppose que : $dim\mathrm{W}_1 = rgx_1$=p.

Soit $a \in \mathbb{R}^{p*}$, on a :

$$x_1^t M x_1(a, a) =< x_1^t M x_1 a, a >=< M x_1 x_1 a >$$

Par suite : $x_1^t M x_1(a,a) = 0 \Leftrightarrow < x_1^t M x_1 a, a >= 0 \Leftrightarrow < M x_1 x_1 a \geq 0 \Leftrightarrow x_1 a = 0$.

Mais x_1 est injective donc $a = 0$. Par suite $x_1^t M x_1$ est une forme bilinéaire symétrique définie positive sur $\in \mathbb{R}^{p*}$donc elle est inversible et on a : $u - u_1 \perp \mathrm{W}_1 = x_1(\mathbb{R}^{p*})$.

Par suite on peut écrire :

$\forall i \in \{1, \dots, p\}, M\big(u - u_1, x_1(e_i^*)\big) =< M(u - u_1), x_1(e_i^*) >$, d'où : $< {x_1}^t M(u - u_1), e_i^* >= 0 \Rightarrow {x_1}^t M(u - u_1) = 0$, comme $u_1 = A_1 u = x_1(a)$ avec $a \in \mathbb{R}^{p*}$ alors :

$$x_1^t M u = x_1^t M u_1 = x_1^t M x_1(a) \Rightarrow a = (x_1^t M x_1)^{-1} x_1^t M u \text{ et } A_1 u = x_1(a) = x_1(x_1^t M x_1)^{-1} x_1^t M u \Rightarrow A_1 = x_1(x_1^t M x_1)^{-1} x_1^t M. \blacksquare$$

Proposition 5.1

Si $dim W_1 = p$, l'opérateur de projection M-orthogonale sur W_1, A_1 est donné par :

$$A_1 = x_1(x_1^t M x_1)^{-1} x_1^t M \tag{5.1}$$

<u>Recherche de la ligne de plus grande pente de W_1 et W_2 :</u>

Soit A_1 (resp A_2) l'opérateur de projection M-orthogonale sur W_1(resp W_2), et soit $\xi \in W_1$, le vecteur de W_2 le plus proche de ξ est $A_2\xi$,, le vecteur de W_1 le plus proche de $A_2\xi$,, est $A_1A_2\xi$,, , le vecteur le plus proche de $A_1A_2\xi$, est $A_2A_1A_2\xi$, si on projette $A_2A_1A_2\xi$ sur W_1 on obtient $A_1A_2A_1A_2\xi$ qui est égal à $A_1A_2\xi$.

Proposition 5.2

Pour obtenir le vecteur ξ dans W_1 le plus proche de W_2, il faudra que l'angle de ξ et $A_1A_2\xi$ soit nul, donc que ξ soit colinéaire à $A_1A_2\xi$, i. e: ξ soit vecteur propre de A_1A_2.

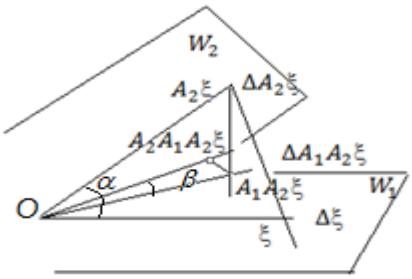


Figure 5-1

Preuve 5.2

On pose :

$cos\beta = \frac{M(A_1A_2\xi, A_2A_1A_2\xi)}{\|A_1A_2\xi\|\|A_2A_1A_2\xi\|}$, $cos\alpha = \frac{M(\xi, A_2\xi)}{\|\xi\|\|A_2\xi\|}$, β est l'angle de W_1 et W_2. $\xi \in W_1 \Rightarrow A_1\xi = \xi$,

$\Rightarrow M(\xi, A_2\xi) = M(A_1\xi, A_2\xi) = \mathrm{M}(\xi, A_1A_2\xi)$

Et $M(A_2\xi, A_2\xi) = M(\xi, A_2A_2\xi) = \mathrm{M}(\xi, A_2\xi) = \mathrm{M}(\xi, A_1A_2\xi)$.

Par suite :

$$cos\alpha = \frac{M(\xi, A_2\xi)}{\|\xi\|\|A_2\xi\|} = \frac{\|A_2\xi\|}{\|\xi\|} \text{ et } cos\beta = \frac{M(A_1A_2\xi, A_2A_1A_2\xi)}{\|A_1A_2\xi\|\|A_2A_1A_2\xi\|} = \frac{\|A_2A_1A_2\xi\|}{\|A_1A_2\xi\|}.$$

D'après l'inégalité de Cauchy-Schwarz il vient :

$$\mathrm{M}(\xi, A_1A_2\xi) \leq \|\xi\|\|A_1A_2\xi\| \Rightarrow \frac{\|A_2\xi\|}{\|\xi\|} \leq \frac{\|A_1A_2\xi\|}{\|A_2\xi\|}.$$

De même :

$\|A_1A_2\xi\|^2 = M(A_1A_2\xi, A_1A_2\xi) = \mathrm{M}(A_2\xi, A_1A_2\xi) = \mathrm{M}(A_2\xi, A_2A_1A_2\xi)$..

Encore une fois d'après l'inégalité de Cauchy-Schwarz on obtient :

$\frac{\|A_1A_2\xi\|}{\|A_2\xi\|} \leq \frac{\|A_2A_1A_2\xi\|}{\|A_1A_2\xi\|} \Rightarrow cos\alpha \leq \frac{\|A_1A_2\xi\|}{\|A_2\xi\|} \leq cos\beta.$

Le minimum de α et β obtenu si $co\xi s\alpha = cos\beta$. i.e : pour $M(\xi, A_2A_2\xi) = \|\xi\|\|A_1A_2\xi\|$, ξ et $A_1A_2\xi$ colinéaires, s'il existe un réel λ tel que : $A_1A_2\xi = \lambda$. On a alors : $M(\xi, A_2A_2\xi) = \lambda\mathrm{M}(\xi, \xi)$ d'où :

$\lambda = \frac{M(\xi, A_2A_2\xi)}{\|\xi\|^2} = \frac{\|A_2\xi\|}{\|\xi\|^2} = cos^2\alpha, \alpha$ sera minimum si $cos\alpha$ est maximum i.e : λ es la plus grande valeur propre de A_1A_2.

De même, le vecteur η de W_2 le plus proche de W_1 sera un vecteur propre de A_2A_1 associé à la plus grande valeur propre.■

Proposition 5.3

A_1A_2 est diagonalisable sur $\mathbb{R}$.

Nous démontrons ceci grâce au théorème suivant :

Théorème 5.1

1- Les vecteurs propres de A_1A_2 (resp A_2A_1) associés à des valeurs propres non nulles appartiennent à W_1 $(resp\ W_2)$.
2- La restriction de A_1A_2 à W_1 (resp A_2A_1à W_2) est M-symétrique.

Preuve 5.1

1- Soit u un vecteur propre de A_1A_2 alors : $A_1A_2\ u = \lambda u$, en composant avec A_1, il vient :
$A_1A_1A_2\ u = A_1A_2\ u = \lambda A_1u = \lambda u$ donc si $\lambda \neq 0$, $A_1u = u$ cela veut dire que : $u \in W_1$, donc la restriction de A_1A_2 à W_1 a mêmes valeurs propres, et mêmes vecteurs propres que A_1A_2 . Il suffit de travailler avec cette restriction.
2- Soient u, v deux vecteurs de W_1, alors :

$$\begin{aligned} M(A_1A_2\ u, v) &= M(A_2\ u, A_1v), \text{ car } A_1 \text{ est M-symétrique.} \\ &= M(A_2\ u, v) \text{ car } v \text{ appartient à } W_1 \\ &= M(u, A_2v) \text{ car } A_2\ est\ M\ -\text{symétrique} \\ &= M(A_1u, A_2\ v), \text{ car } u \text{ appartient à } W_1 \\ &= M(\ u, A_1A_2v) \text{ car } A_1 \text{ est M-symétrique.} \end{aligned}$$

Par suite il en découle tous les résultats suivants :

- La restriction de A_1A_2 à W_1 est diagonalisable.
- La dimension d'un sous-espace propre est égale à l'ordre de multiplicité de la valeur propre correspondante.
- Deux vecteurs propres associés à deux valeurs propres différentes sont M-orthogonaux.
- $\mathbb{R}^p$ admet une base orthonormée constituée de vecteurs propres de A_1A_2. (on suppose que $dim\, W_1 = p$).
- On a bien évidemment les mêmes résultats pour A_2A_1. (on suppose que $dim\, W_2 = p$) . ■

Proposition 5.2
A_1A_2 a mêmes valeurs propres que : A_2A_1.

Preuve 5.2
Il suffit de voir que :
$A_1A_2u = \lambda u$, on peut écrire : $A_2A_1(A_2u) = \lambda(A_2u)$.■

<u>Recherche de (ξ_1, η_1) :</u>
Il s'agit de maximiser le cosinus de ξ et de η, où $\xi \in W_1, \eta \in W_2$. ξ et η sont M-normés.
Donc rendre maximale la quantité : $f(\xi, \eta)$=$M(\xi, \eta)$ sous les contraintes : $\begin{cases} g_1(\xi, \eta) = M(\xi, \xi) - 1 \\ g_2(\xi, \eta) = M(\eta, \eta) - 1 \end{cases}$
Ainsi on a :

$$\begin{gathered} f(\xi, \eta)=M(\xi, \eta)=M(A_1\xi, A_2\ \eta) = M(\xi, A_1A_2\ \eta) = \xi^t MA_1A_2\ \eta \\ = M(\eta, \xi\ ) = M(\eta, A_1A_2\xi\ ) = \eta^t MA_1A_2\xi \\ g_1(\xi, \eta) = M(\xi, \xi) - 1 = \xi^t MA_1\xi - 1 \\ g_2(\xi, \eta) = M(\eta, \eta) - 1 = \eta^t MA_2\eta - 1 \end{gathered}$$

Si ce problème admet une solution en (ξ, η), alors il existe deux réels λ,μ tels que les multiplicateurs de Lagrange fonctionnent, ainsi :

$$\begin{cases}\left[\frac{df}{d\xi}-\lambda\frac{dg_1}{d\xi}-\mu\frac{dg_2}{d\xi}\right](\xi,\eta)=0\\ \left[\frac{df}{d\eta}-\lambda\frac{dg_1}{d\eta}-\mu\frac{dg_2}{d\eta}\right](\xi,\eta)=0\end{cases}$$

$\Rightarrow$

$$\begin{cases}MA_1A_2\xi-2\mu MA_1\eta=0\\ MA_1A_2\,\eta-2\lambda MA_1\xi=0\end{cases}$$

M est un isomorphisme alors :

$$\begin{cases}A_1A_2\xi-2\mu A_1\eta=0\\ A_1A_2\,\eta-2\lambda A_1\xi=0\end{cases}$$

D'où :

$$A_2A_1(A_2\eta)=2\lambda A_2A_1\xi=4\lambda\mu(A_2\eta)$$

$\eta\in W_1\Rightarrow A_2\eta=\eta,$ donc celui-ci doit être un vecteur propre deA_2A_1 par rapport à la valeur propre $4\lambda\mu$.

De même, $A_1A_2(A_1\xi)=4\lambda\mu(A_1\xi)$ et ξ doit être vecteur propre de A_1A_2 par rapport à la valeur propre $4\lambda\mu$.

De plus on a :

$f(\xi,\eta)=\xi^tMA_1A_2\eta=2\lambda\xi^tM\xi=2\lambda$

$=\eta^tMA_2A_1\xi=2\mu\eta^tM\eta=2\mu.$

On a donc : $\lambda=\mu$ et $cos^2\beta=f^2(\xi,\eta)=4\lambda^2$. Cette dernière valeur est valeur propre de $A_1A_2(ou\ A_2A_1)$.

f sera maximale si ξ (resp η) est vecteur propre de $A_1A_2(ou\ A_2A_1)$, pour la plus grande valeur propre.

Remarque 5.1

Une valeur propre de $A_1A_2(ou\ A_2A_1)$ devant être égale au carré d'un cosinus, est nécessairement comprise entre 0 et .

<u>Cas où la plus grande valeur propre est égale à 1 :</u>

Proposition 5.3

$W_1\cap W_2$ est le sous-espace propre associé à la valeur propre 1.

Preuve 5.3

Soit (ξ_1,η_1), le couple de variables canoniques associées. Alors, $cos^2(\xi_1,\eta_1)=1$, donc l'angle de (ξ_1,η_1) est égal à 0 ou π et $\xi_1=\eta_1$, ou $\xi_1=-\eta_1$, donc $\xi_1\in W_1\cap W_2$.

<u>Réciproquement :</u>

Soit $u\in W_1\cap W_2$. alors :

$\begin{cases}u\in W_1\Rightarrow A_1u=u.\\ u\in W_2\Rightarrow A_2u=u\\ A_1A_2u=A_1=u\end{cases}$ donc u est vecteur propre de A_1A_2 par rapport à la valeur propre 1. En particulier $dimW_1\cap W_2$ est égale à l'ordre de multiplicité de la valeur propre 1.■

<u>Cas où la plus grande valeur propre λ_1 est simple :</u>

Proposition 5.4

$\eta_1=\frac{A_2\xi_1}{\sqrt{\lambda_1}}$, de même : $\xi_1=\frac{A_1\eta_1}{\sqrt{\lambda_1}}$.

Preuve 5.4

$A_1A_2\xi_1=\lambda_1\xi_1$ composons avec A_2 il vient : $A_2A_1A_2\xi_1=\lambda_1A_2\xi_1$ donc $A_2\xi_1$ est vecteur propre de A_2A_1 par rapport à λ_1, donc si λ_1 est simple alors η_1 est colinéaire à $A_2\xi_1$ il existe un nombre k tel que : $\eta_1=k\ A_2\xi_1$, or η_1 est normé donc :

$\eta_1^tM\eta_1=\eta_1^t\,kM\ A_2\xi_1=\eta_1^t\,kM\ A_2A_1\xi_1=k\sqrt{\lambda_1}$=1. En effet $\eta_1^t\ M\ A_2A_1\xi_1=cos(\xi_1,\eta_1)=\sqrt{\lambda_1}$ d'où les identités de la proposition.■

<u>Recherche des autres variables canoniques :</u>

On cherche (ξ,η), dans (W_1,W_2), M-normés, avec $\xi\perp\xi_1$, $\eta\perp\eta_1$ qui minimise l'angle entre ξ et η. Il s'agit donc de maximiser f sous les contraintes :

$$\begin{cases} M(\xi,\xi)=1 \\ M(\eta,\eta)=1 \\ M(\xi,\xi_1)=0 \\ M(\eta,\eta_1)=0 \end{cases}$$

On montre que la solution ξ_2 (resp η_2) doit être vecteur propre de A_1A_2 (resp A_2A_1) par rapport à la deuxième plus grande valeur propre.

On met ainsi un ensemble de variables canoniques (ξ_k ,η_k), qui sont les vecteurs propres normés de A_1A_2 (resp A_2A_1), associés aux valeurs propres rangées par ordre décroissant.

Facteurs canoniques

Aux variables canoniques (ξ_k ,η_k), on associe les facteurs canoniques (a_k ,b_k), définis par :

$\mathbb{R}^{p*} \xrightarrow{x_1} \mathbb{R}^n \xleftarrow{x_2} \mathbb{R}^{q*}$, $\quad a_k \in \mathbb{R}^{p*}, b_k \in \mathbb{R}^{q*} \quad \xi_k = X_1 a_k, \ \eta_k = X_2 b_k.$

Ecrivons que ξ_k est vecteur propre de A_1A_2 par rapport à la valeur propre λ_k :

$A_1A_2\xi_k = \lambda_k\xi_k$, a_k doit vérifier $A_1A_2X_1a_k = \lambda_k X_1 a_k$, par inversion exprimons A_1A_2 *on a* :

$$X_1(X_1^tMX_1)(X_1^tMX_2)(X_2^tMX_2)(X_2^tMX_1)\ a_k = \lambda_k X_1 a_k \qquad (5.2)$$

On a supposé que : $dimW_1 = p$, ceci implique que x_1 est injective. Alors :

$$(X_1^tMX_1)(X_1^tMX_2)(X_2^tMX_2)(X_2^tM)(X_2^tMX_1)\ \ a_k = \lambda_k a_k \qquad (5.3)$$

Donc a_k est vecteur propre de $(X_1^tMX_1)(X_1^tMX_2)(X_2^tMX_2)(X_2^tM)(X_2^tMX_1)$ par rapport à la valeur propre λ_k.

Proposition 5.5

Le facteur canonique a_k (resp b_k)

Est vecteur propre de $(X_1^tMX_1)(X_1^tMX_2)(X_2^tMX_2)(X_2^tM)(X_2^tMX_1)$ (resp $(X_1^tMX_1)(X_1^tMX_2)(X_2^tMX_2)(X_2^tM)(X_1^tMX_2)$) par rapport à la valeur propre λ_k.

<u>Cas de variables centrées :</u>

On suppose que les p+q variables sont centrées. Soit : $\delta_n = \sum_{i=1..n} f_i$, alors :

$X_1^tM\ \delta_n = 0,\ X_2^tM\ \delta_n = 0,$ alors $X_1^tM\ \ X_j$ s'interprète comme une matrice de covariance que l'on note : V_{ij} soit donc :

$V_{11} = X_1^tM\ X_1,\ V_{12} = X_1^tM\ X_2,\quad V_{22} = X_2^tM\ X_2,\quad V_{21} = (V_{12})^t.$

Les équations des facteurs canoniques s'écrivent :

$$(V_{11})^{-1}V_{12}(V_{22})^{-1}\ V_{21}a_k = \lambda_k a_k$$

$$(V_{22})^{-1}V_{21}(V_{11})^{-1}\ V_{12}b_k = \lambda_k b_k\ .$$

La valeur propre λ_k est égale au carré du coefficient de corrélation entre les variables canoniques ξ_k, η_k.

$$\|\xi_k\| = 1 \text{ donc :}$$

$$a_k^tX_1MX_1a_k = a_k^tV_{11}a_k=1$$

$$b_k^tX_2MX_2b_k = b_k^tV_{22}b_k = 1.$$

Description de l'ensemble des individus

On peut effectuer cette description sur le plan engendré par ξ_1, ξ_2et sur le plan engendré par η_1, η_2. Ces deux ensembles de description seront d'autant plus voisins que λ_1, λ_2 seront proches de 1, car on pourra alors préciser ce que contient $W_1 \cap W_2$.

En général, les résultats sont difficiles à interpréter. L'intérêt de cette méthode est essentiellement théorique. En particulier, on verra que l'analyse des correspondances est un cas particulier d'analyse canonique.

Chapitre 6 Analyse factorielle des correspondances

Cette méthode a été développée en France par l'équipe du professeur J.P. Benzécri. Elle est utilisée pour étudier la correspondance entre deux variables qualitatives.

Sur le plan mathématique deux approches sont possibles :

- L'analyse factorielle des correspondances est un cas particulier d'analyse en composantes principales, la métrique utilisée étant celle du χ^2.
- L'analyse factorielle des correspondances est une variante de l'analyse canonique.

Exemple 6.1

Soit une étude portant sur la colonisation du fleuve Normand la Varenne par des carpes pendant une année.

Des prélèvements ont été effectués sur 12 secteurs par pêche électrique. Les carpes sont dénombrées par classes de taille d'amplitude 10mm, de 200mm à 770mm. Chaque classe est représentée par son centre.

On croise les secteurs avec les classes de taille. A la rencontre de la ligne i et de la colonne j figure le nombre d'carpes pêchées dans le secteur i et dont la taille est représentée par j. c'est un exemple type d'une table de contingence**.**

On utilise également cette méthode pour des tableaux « apparentés » à une table de contingence. En phytosociologie, on établit des tableaux croisant des espèces et des relevés $\left(k(i,j)_{i=1..n,j=1..p}\right)$ $o\grave{u}$ $k(i,j)=1$, si l'espèce i est présente dans le relevé j et 0 sinon. Ces tableaux dits de présence-absence sont soumis à des A.F.C.

A.F.C comme cas particulier d'A.C.P

On suppose que l'on a observé deux caractères qualitatifs A et B sur une population P. soit M de $Card\ n$, (M' de $Card\ p$) l'ensemble des modalités de A (resp de A).

On pose :

$$(i,j) \in M \times M' \qquad x_{ij} = Card\{a \in P, (A \times B)(a) = (i,j)\} \quad x_{ij} \in \mathbb{N}.$$

On obtient ainsi un tableau X à n lignes et p colonnes.

$\mathrm{X} = \{x_{ij}\}_{\mathrm{i}=1..n,j=1..p}$ qui porte le nom de table de contingence.

Distance du khi-deux entre deux lois de probabilité

Soient deux lois μ,ν de probabilité définies sur $(I, P(I))$ où I est fini on pose $Card\ I = n$.

Définition 6.1

La distance du χ^2 de centre p (où p représente la moyenne d'une série statistique $(x_i, p_i)_{I=1..n}$) entre deux lois de probabilité μ,ν définies sur $(I, P(I))$ où I est fini, $Card\ I = n$, $\mu = \sum_{\mathrm{i}=1..\mathrm{n}} \mu_{\mathrm{i}}\, x_i$, $\nu = \sum_{\mathrm{i}=1..\mathrm{n}} \nu_{\mathrm{i}}\, x_i$, $\sum_{\mathrm{i}=1..\mathrm{n}} \mu_{\mathrm{i}} = 1, \sum_{\mathrm{i}=1..\mathrm{n}} \nu_{\mathrm{i}} = 1$, la quantité :

$$d^2_{\chi^2_p}(\mu,\nu) = \|\mu - \nu\|^2_{\chi^2_p},$$

$$d^2_{\chi^2_p}(\mu,\nu) = D_{\frac{1}{p}}(\mu - \nu) = \sum_{i=1..n} \frac{(\mu_{\mathrm{i}} - \nu_{\mathrm{i}})^2}{p_i} \qquad (6.1).$$

Notion de profil

Soit X une table de contingence $X = \left(x_{ij}\right)_{i=1..n,j=1..p}$, $x_{ij} \in \mathbb{N}$. On pose comme au chapitre 3 :

$$x_{i\cdot} = \sum_{j=1..n} x_{ij}\,, \qquad x_{\cdot j} = \sum_{i=1..n} x_{ij}\,, \quad x_{\cdot\cdot} = \sum_{i=1..n} x_{i\cdot} = \sum_{j=1..p} x_{\cdot j} = \sum_{i=1..n, j=1..p} x_{ij},$$

$$D_1 = \left(\delta_{ij} x_{i\cdot}\right)_{i=1..n, j=1..p}, \qquad D_2 = \left(\delta_{ij} x_{\cdot j}\right)_{i=1..n, j=1..p} \quad . \qquad (6.2)$$

Parmi les $x_{i\cdot}$ individus présentant la modalité i de A, x_{ij} présentent la modalité j de B, soit une proportion de $\frac{x_{ij}}{x_{i\cdot}}$. On note cette proportion : f_j^i.

Par définition de la notion de profil on a :

$$f_j^i = \frac{Card\{a \in P, (A \times B)(a) = (i,j)\}}{Card\{a \in P, A(a) = i\}} \qquad (6.3)$$

f_j^i présente la fréquence conditionnelle de la modalité j de B liée par la modalité i de A.

<u>Profil-ligne profil-colonne :</u>

Définition 6.2

On appelle profil-ligne associé à l'individu i (resp profil-colonne associé au caractère j) le vecteur :

$$f_J^i = \left(f_j^i\right)_{j=1..p} \quad (\text{resp } f_i^I = \left(f_j^i\right)_{i=1..n}).$$

La somme des composantes de f_J^i (resp f_i^I) est égale à 1. Par suite f_J^i (resp f_i^I) apparait comme une probabilité sur $(J, P(J)), (resp\ \left(I, P(I)\right)$.

On définit enfin :

$$f_J = \left(f_j\right)_{j=1..p}, \text{ avec : } f_j = \frac{x_{\cdot j}}{x_{\cdot\cdot}}, (\text{resp } f_I = (f_i)_{j=1..n}, f_j = \frac{x_{i\cdot}}{x_{\cdot\cdot}}. \qquad (6.4)$$

Soit le nuage des profil-lignes :

$$N(I) = \left\{\left(f_J^i, f_i\right), i = 1..n\right\},$$

Et le nuage des profils-colonnes :

$$N(J) = \left\{\left(f_I^i, f_j\right), i = 1..n\right\},$$

Proposition 5.6
Le centre de gravité du nuage $N(I)$, est f_J, celui du nuage $N(J)$, est f_I.

Preuve 5.6
En effet le centre de gravité de $N(I)$ a pour composantes :

$$g_j = \sum_{i=1..n} f_J^i f_i = f_j.$$

L' A.F.C sur la table de contingence consiste à effectuer deux A.C.P l'une sur N(I), l'autre sur N(J) muni respectivement des métriques $\chi^2_{f_J}$, $\chi^2_{f_I}$. Nous verrons que les résultats de ces deux A.C.P, sont liés, et que l'on peut effectuer des projections simultanées des deux nuages de profils. Cette particularité permet d'interpréter des proximités de points, d'un nuage en fonction de points de l'autre nuage.

Inertie au centre de gravité des deux nuages

L'inertie au centre de gravité du nuage N(I), s'écrit :

$$I_{f_J}(N(I)) = \sum_{i=1..n} f_i d^2_{\chi^2_{f_J}}\left(f_J^i, f_J\right).$$

En remplaçant par les expressions (6.3) on obtient :

$$I_{f_J}(N(I)) = \frac{1}{x_{\cdot\cdot}} \sum_{i=1..n, j=1..p} \frac{\left(x_{ij} - \frac{x_{i\cdot} x_{\cdot j}}{x_{\cdot\cdot}}\right)^2}{\frac{x_{i\cdot} x_{\cdot j}}{x_{\cdot\cdot}}} \qquad (6.5)$$

Cette formule est symétrique en i, j, elle représente donc aussi l'inertie du nuage $N(J)$ par rapport à son centre de gravité.

Si les deux caractères A et B sont indépendants, alors on a :

$$\forall(i,j) \quad f_{ij} = f_i.f_j \text{ ou encore } f_j = f_j^i.$$

Soit encore : $x_{ij} = \frac{x_i x_{.j}}{x_{..}}$.

D'une certaine manière $I_{f_J}(N(I))$, (ou $I_{f_I}(N(J))$) mesure l'écart à l'indépendance. L'A.C.P sur $N(I)$, l'espace des profils lignes étant muni de la métrique du $\chi^2_{f_J}$, permet de décomposer cet écart le long des axes principaux $\left(\Delta_{u_k^{\perp}}\right)_{k=1..p}$ puisque celui-ci est égal à la somme des $\Delta_{u_k^{\perp}}$.

Enfin, on peut montrer que cette inertie se met sous la forme suivante :

$$I_{f_J}(N(I)) = I_{f_I}(N(J)) = -1 + \sum_{i=1..n, j=1..p} \frac{(x_{ij})^2}{x_i x_{.j}}, \qquad (6.6)$$

Principe d'équivalence distributionnelle

Supposons que, dans le tableau X, les deux colonnes j_1, j_2 soient proportionnelles. Alors les profils correspondants $f_I^{j_1}$, $f_I^{j_2}$ sont égaux. Si on remplace le tableau X par le tableau X_1 obtenu en conservant toutes les colonnes de X autre que j_1 et j_2 et en remplaçant ces deux colonnes par une seule j_0, obtenue en les sommant, le nuage $N(J)$ est inchangé. Les profils-lignes correspondent aux mêmes individus, mais leurs représentations n'est plus la même, il y'a une coordonnée en moins. Une propriété de la distance choisie entre deux profils-lignes quelconques est la conservation de la valeur de cette distance : elle porte le nom d'équivalence distributionnelle. En effet :

$$d^2_{\chi^2_{f_J}}(f_J^i, f_J^{i'}) = \sum_{j \neq j_1, j_2} \frac{1}{f_j}\left(f_j^i - f_j^{i'}\right)^2 + \frac{1}{f_{j_1}}\left(f_{j_1}^i - f_{j_1}^{i'}\right)^2 + \frac{1}{f_{j_2}}\left(f_{j_2}^i - f_{j_2}^{i'}\right)^2 \qquad (6.7)$$

Dans la représentation sur J_0,

$$d^2_{\chi^2_{f_{J_0}}}(f_{j_0}^i, f_{j_0}^{i'}) = \sum_{j \neq j_0} \frac{1}{f_j}\left(f_j^i - f_j^{i'}\right)^2 + \frac{1}{f_{j_0}}\left(f_{j_0}^i - f_{j_0}^{i'}\right)^2 \qquad (6.8)$$

$$\text{Car : } \frac{1}{f_{j_1}}\left(f_{j_1}^i - f_{j_1}^{i'}\right)^2 + \frac{1}{f_{j_2}}\left(f_{j_2}^i - f_{j_2}^{i'}\right)^2 = \frac{1}{f_{j_0}}\left(f_{j_0}^i - f_{j_0}^{i'}\right)^2.$$

Analyse en composantes principales des deux nuages de points

A.C.P non centrée sur N(I) et facteur trivial

Notation

Des identités (6.2) on peut écrire :

$$Y = D_1^{-1}X, \quad Z = D_2^{-1}X^t$$

L'A.C.P sur N(I), l'espace des profils-lignes étant muni de la métrique du $\chi^2_{f_J}$, revient à faire l'A.C.P, sur le tableau Y, la matrice des poids étant $\frac{1}{x_{..}}D_1$, la métrique étant $M_1 = x_{..}D_2^{-1}$.

Un axe principal est engendré par un vecteur propre M_1 −normé de $V_y M_1$, où V_y est la matrice d'inertie de N(I).

Proposition 5.7

f_J est vecteur propre de $V_y M_1$

Preuve 5.7

Soit W le support du nuage, $W = \{M_i, f_J^i = \overrightarrow{OM_i}\}$, W est l'hyperplan affine $D_{\frac{1}{f_J}}$-orthogonal au point G$\{\overrightarrow{OG} = f_J\}$ à l'axe Δ_{f_J}. En effet, soit M le point de W tel que $\overrightarrow{OM} = w$, montrons que : $\overrightarrow{OG} \perp \overrightarrow{GM}$;

$$\forall M \in W, \quad D_{\frac{1}{f_J}}(f_J, w - f_J) = \sum_{j=1..n}(w_j - f_j) = 0.$$

On effectue l'A.C.P sur le nuage N(I), la métrique utilisée étant la métrique du $\chi^2_{f_J}$. Pour déterminer les axes principaux, on doit diagonaliser la matrice $V_y M_1$. Le centre de gravité de N(I) étant f_J

Il vient :

$$V_y = Y^t \frac{D_1}{x_{..}} Y - f_J f_J^t$$

$$\Rightarrow V_y = \left(f_J^1 - f_J, \dots, f_J^n - f_J\right) \frac{D_1}{x_{..}} \begin{pmatrix} \left(f_J^1 - f_J\right)^t \\ \dots \\ \left(f_J^n - f_J\right)^t \end{pmatrix},$$

$f_J$est M_1 −orthogonal à W_1, support du nuage N(I), donc f_J engendre un axe principal, il est donc vecteur propre de $V_y M_1$. ■

Proposition 5.8
f_J est vecteur propre de $V_y M_1$ par rapport à la valeur propre 0.

Preuve 5.8

$V_y M_1 f_J = \left(f_J^1 - f_J, \dots, f_J^n - f_J\right) \frac{D_1}{x_{..}} \begin{pmatrix} \left(f_J^1 - f_J\right)^t \\ \dots \\ \left(f_J^n - f_J\right)^t \end{pmatrix} M_1 f_J = 0,$

En effet : f_J est M_1 −orthogonal à $\left(f_J^i - f_J\right)$ pour tout i.

$M_1 f_J = \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix} \Rightarrow \begin{pmatrix} \left(f_J^1 - f_J\right)^t \\ \dots \\ \left(f_J^n - f_J\right)^t \end{pmatrix} M_1 f_J = 0$, par suite, f_J est vecteur propre de $V_y M_1$ pour la valeur propre 0.

Proposition 5.9
f_J est vecteur propre de $Y^t \left(\frac{D_1}{x_{..}}\right) Y M_1$ pour la valeur propre 1.

Preuve 5.9

$$V_y M_1 f_J = Y^t \left(\frac{D_1}{x_{..}}\right) Y M_1 f_J - f_J {f_J}^t M_1 f_J = 0$$

$$Y^t \left(\frac{D_1}{x_{..}}\right) Y M_1 f_J = f_J {f_J}^t M_1 f_J = f_J. \blacksquare$$

A.C.P sur N(I)

$V_y M_1$ est M_1 −symétrique, soit u un vecteur propre de $V_y M_1$ associé à une valeur propre différente de 0, u est M_1 −orthogonal à f_J. Par suite :

$$V_y M_1 u = Y^t \left(\frac{D_1}{x_{..}}\right) Y M_1 u = f_J {f_J}^t M_1 u = Y^t \left(\frac{D_1}{x_{..}}\right) Y M_1 u.$$

On a le résultat suivant :

Proposition 5.10
$V_y M_1$ et $Y^t \left(\frac{D_1}{x_{..}}\right) Y M_1$ ont même vecteurs propres autres que f_J associés aux mêmes valeurs propres.

Remarque 5.1
f_J est vecteur propre de $V_y M_1$ pour la valeur propre 0 et vecteur propre de $Y^t \left(\frac{D_1}{x_{..}}\right) Y M_1$ pour la valeur propre 1

Dans une A.F.C, il ne sera donc pas nécessaire de centrer avant de déterminer les axes principaux et on aura toujours la valeur propre triviale égale à 1.

Les axes principaux sont engendrés par les vecteurs propres M_1 −orthonormés de $Y^t \left(\frac{D_1}{x_{..}}\right) Y M_1$ qui s'écrit :

$Y^t \left(\frac{D_1}{x_{..}}\right) Y M_1 = X^t \left(\frac{D_1}{x_{..}}\right) X D_2^{-1}$,

Les facteurs principaux sont les vecteurs propres M_1^{-1}-orthonormés :

$$Y\left(x {D_2}^{-1}\right) Y^t \left(\frac{D_1}{x_{..}}\right) = D_1^{-1} X D_2^{-1} X^t = YZ$$

De $\frac{D_1}{x_{..}}$ −norme égal à la racine carrée positive de la valeur propre correspondante.

A.C.P sur N(I)

On peut de même effectuer l'A.C.P sur N(J) la métrique sur l'espace des profils-colonnes étant la métrique $M_2 = x_{..}D_1^{-1}$ie effectuer l'A.C.P sur le tableau Z, la matrice des poids étant $\frac{1}{x_{..}}D_{2.}$ un axe principal est engendré par un vecteur propre M_2-normé de :

$XD_2^{-1}X^tD_1^{-1}$,

Un facteur principal est vecteur propre M_2^{-1} normé de :

$$D_1^{-1}XD_2^{-1}X^t$$

Une composante principale est vecteur propre de :

$D_2^{-1}X^tD_1^{-1}X$,

De $\frac{D_2}{x_{..}}$-norme égale à la racine carrée de la valeur propre correspondante

On constate qu'un facteur principal dans l'analyse de $N(I)$ est colinéaire à une composante principale dans l'analyse de N(J) et réciproquement.

En particulier, les deux analyses conduisent aux mêmes valeurs propres non nulles.

Proposition 5.11
Un facteur principal α dans l'analyse de N(I) et une composante principale b dans l'analyse de $N(J)$ sont tous deux vecteurs propres de ZY. Pour la norme $M_1^{-1} = \frac{D_2}{x_{..}}$, α est de norme 1, et b de norme $\sqrt{\lambda}$, où λ est la valeur propre correspondante.

Proposition 5.12
Un facteur principal β dans l'analyse N(J), et une composante principal α dans l'analyse $N(I)$ sont tous deux vecteurs propres de ZY. Pour la norme $M_1^{-1} = \frac{D_2}{..}$, β est de norme 1, et α de norme $\sqrt{\lambda}$, où λ est la valeur propre correspondante..

L'A.F.C revient donc à diagonaliser YZ, d'ordre n ou ZY d'odre p. En pratique, on diagonalise la plus petite de ces deux matrices.

<u>Tableau récapitulatif :</u>

Nuage de points	$N(I) = \{(f_j^{\,i}, f_i), i = 1..n\}$	$N(J) = \{(f_I^{\,j}, f_i), j = 1..n\}$
Tableau de données	$Y = D_1^{-\&}X$	$Z = D_2^{-\&}X^t$
Matrice des poids	$D_{\frac{1}{F_j}} = \frac{D_1}{x_{..}}$	$D_{\frac{1}{{}^1f_I}} = x_{..}D_2^{-1}$
métrique	$X^tD_1^{-1}XD_6^{6\&}$	$XD_2^{-1}X^tD_1^{-1}$
Vecteurs unitaires des axes factoriels et vecteurs propres de :	$ZY = D_2^{-1}X^tD_1^{-1}X$	$YZ = D_1^{-1}XD_2^{-1}X^t$
Composantes principales et vecteurs propres de :	$YZ = D_1^{-1}XD_2^{-1}X^t$	$ZY = D_2^{-1}X^tD_1^{-1}X$

Propriété 5.13

Dans une A.F.C toutes les valeurs propres sont inférieurs ou égales à 1.

Considérons les séries statistiques : $(a_i, f_{ij})_{i,j\in I\times J}$, $b_j, f_{ij})_{i,j\in I\times J}$ où $f_{ij} = \frac{x_{ij}}{x}$ et $a\ et\ b$ sont les composantes principales respectives des deux nuages par rapport à la même valeur propre λ.

Elles sont de moyenne nulle. En effet :

$$\sum_{i,j} a_i^2 f_{ij} = \frac{1}{x_{..}}\sum_i a_i^2 x_{i.} = var\ a = \lambda.$$

Le coefficient de corrélation entre les deux séries statistiques est égal à :

$r(a,b) = \frac{\sum_{i,j} f_{ij}}{\lambda}$.

$\sum_{i,j} f_{ij}a_ib_j = \sum_i f_i\, a_i \sum_j \frac{f_{ij}}{f_i} b_j = \sum_i f_i a_j \sum_j f_j^i b_j = \sum_i f_i a_i(\sqrt{\lambda}a_i) = \sqrt{\lambda}\sum_i f_i a_i^2 = \lambda\sqrt{\lambda}$.sur un individus du type

Donc : $r(a,b) = \sqrt{\lambda}$., et comme $0 \le r(a,b) \le 1$, alors $0 \le \lambda \le 1$.

Sur le nombres d'axes

Le nuage $N(I)$ comporte n individus ; ceux-ci se situent donc dans un espace de dimension n-1 au plus. Un élément de $N(I)$ appartient à l'hyperplan affine de $\mathbb{R}^p$, défini comme I_p-orthogonal à δ_p, vecteur dont toutes les coordonnées sont égales à 1. Le nuage est donc dans un espace de dimension au plus égale à $\min(p-1, n-1)$. la matrice à diagonaliser a donc au maximum un nombre de valeurs propres non nulles égal à $\min\,(p-1, n-1)$,.

Formules de transition

Soit αune composante principale dans l'analyse de $N(I)$ relativement à la valeur propre λ. α vérifie :

$YZ\alpha = \lambda\alpha.$

Soit b une composante principale dans l'analyse de N(J) relativement à la valeur propre λ. Si λ est valeur propre simple, b est nécessairement colinéaire à $Z\alpha$. Si λ est multiple, on choisit b colinéaire à $b = k\alpha Z$. Où le carré de la $\frac{D_1}{x_{..}}$-norme de a et le carré de la $\frac{D_2}{x_{..}}$ −norme de b sont égaux à λ.

Il vient :

$$\frac{D_2}{x_{..}}(b,b) = k^2\alpha Z^t \frac{D_2}{x_{..}} Z\alpha = \frac{k^2}{x_{..}}\alpha^t X {D_2}^{-1} X D_2^{-1} X^t \alpha.$$

$$\frac{D_b}{x_{..}}(b,b) = \frac{k^2}{x_{..}}\alpha^t (D_1 D_1^{-1}) X D_2^{-1} X\alpha = \frac{k^2}{x_{..}}\alpha^t D_1 YZ\alpha$$

$$\frac{D_2}{x_{..}}(b,b) = k^2\lambda\alpha\left(\frac{D_1}{x_{..}}\right)\alpha = k^2\lambda^2.$$

Puisque $\frac{D_2}{x_{..}}(b,b) = \lambda$. On a :

D'où les formules de transition :

$$\begin{cases} b = \dfrac{1}{\sqrt{\lambda}} Z\alpha \\ a = \dfrac{1}{\sqrt{\lambda}} Yb \end{cases}$$

Ces formules sont des formules pseudo barycentriques (au coefficients $\frac{1}{\sqrt{\lambda}}$, près)

En effet, pour tout j, la jème coordonnée de b vérifie :

$$b_j = \frac{1}{\sqrt{\lambda}} \sum_{i=1..n} f_i^j a_i.$$

b_j apparaît come la moyenne des $(a_i)\quad i = 1..n$, pondérés par les f_i^j, au coefficient $\frac{1}{\sqrt{\lambda}}$, prés.

Ceci, avant de voir une justification théorique, en considérant l'A.F.C. comme un cas particulier de l'analyse canonique, nous autorise à se présenter simultanément les deux nuages N(I) et N(J) en projection sur les axes factoriels, ce qui est d'un très grand intérêt pratique pour l'intérprétation des résultats.

Formule de reconstitution

En appliquant la formule générale vue en A.C.P, il vient :

$$Y = \sum_{k=1} \alpha^k \alpha_k^t \left(D_{\frac{1}{\frac{1}{f_J}}} \right)^{-1}$$

$$= \sum_{k=1..p} \alpha^k \frac{(b^k)^t}{\sqrt{\lambda_k}} D_{f_J}$$

Soit $\forall(i,j)\quad f_j^i = \sum_{k=1..r} \alpha^k \frac{b^k}{\sqrt{\lambda_k}} f_j$., r étant le nombre d'axex factoriels, les valeurs propres étant rangées par ordre décroissant,

$\lambda \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$.

On somme sur tous les facteurs, y compris le facteur trivial associé à la valeur propre 1.

Il vient :

$\frac{x_{ij}}{x_{i.}} = \sum_{k=1..r} \frac{1}{\sqrt{\lambda_k}} \frac{\alpha_i^k b_j^k x_j}{x}$.

On a : $\lambda_i = 1$ et $\forall i\ \alpha_i^1 = 1$ car $b^1 = D_{\frac{1}{f_J}} f_J = \delta_p$ et $\forall j\ b_j^j = 1$ car $\alpha^1 = D_{\frac{1}{f_{f_I}}} f_I = \delta_n$..

Soit :

$x_{ij} = \frac{x_{i.} x.j}{x_{..}} \left(1 + \sum \frac{\alpha_i^k b_j^k}{\sqrt{\lambda_k}}\right)$.

Sur cette formule, apparaît la reconstruction des écarts à l'indépendance.

Deuxième présentation de l'.A.F.C

La table de contingence X a été obtenue par le tri-croisé de données relevées $(x_i^1, x_i^2)_{i=1..n}$, où $x_i^1 et\ x_i^2$ sont des modalités de variables qualitatives A et B.

Mise sous forme disjonctive complète d'une variable qualitative

On introduit la représentation suivante :

Pour une variable qualitative A à n modalités $m_1, m_2, \ldots . m_n$ constituant un ensemble M, on associe les n fonctions indicatrices de ces modalités : $1_{\{m_1\}} \ldots 1_{\{m_k\}}$.

Soit P la population constituée de ces m individus. On note : $P = \{a_1^{\cdots}, \ldots, a_m\}$.

La variable A peut-être considérée comme une application de P dans M.

$A: P \to M, \quad a_i \to A(a_i) = m_k$.

Pour tout k, on a :

$1_{\{m_k\}}$:P→ {0,1} $\quad a_i \to 1_{\{m_k\}}(a_i)$ avec $1_{\{m_k\}}(a_i) = \begin{cases} 1 & si\ A(a_i) = m_k \\ 0 & sinon \end{cases}$

La variable A peut être représentée suivante :

Pour une variable qualitative A à n modalités $m_1, m_2, \ldots, m_n$, constituant un ensemble M, on associe les a fonctions indicatrices de ces modalités : $1_{\{m_1\}}, \ldots, 1_{\{m_n\}}$.

Soit P la population constituée de ces m individus. On note P$\{a_1, \ldots, a_n\}$,

La variable A peut-être représentée par un tableau de données logiques U. Chaque ligne de U correspond à un individu de P, et chaque colonne , à une des n fonctions indications des modalités de A.

Chaque individu répondant à une modalité et une seule de A, dans chaque ligne de U, figure une et une seule valeur 1, les autres étant 0, on dit que A a été mis sous forme disjonctive complète.

De plus, on remarque que :

$$\sum_{i=1..n} 1_{\{m_k\}(a_k)} = 1,$$

Pour tout i variant de 1 à m.

Aux deux variables A et B correspondant respectivement deux matrices U_1, à n lignes et n colonnes et U_2, à n lignes et p colonnes.

On vérifie que :

$X = U_1^t U_2$, $D_1 = U_1^t U_1$, $D_2 = U_2^t U_2$

En effet, l'élément de la i-ième, j-ième colonne de $U_1^t U_2$ est le produit scalaire des fonctions indicatrices $1_{\{m_1\}}, \dots, 1_{\{m'_j\}}$, soit :

$1_{\{m_1\}}, \dots, \rightarrow 1_{\{m'_j\}}$, soit :

$$1_{m_1}.1_{m'_j} = card\{a \in P, A \times B(i) = (m_i, m'_j\}.$$

Ce produit scalaire revient à computer le nombre de co-occurrences.

A.F.C comme cas particulier d'analyse canonique

On peut définir un codage de modalités de A, soit une application φ de M dans F où F est inclus dans $\mathbb{R}$, : $p \rightarrow^A M \rightarrow^{\varphi}$F

$$a_i \rightarrow A(u_i) = m_k \rightarrow \varphi(A_i) = a_k = \sum_{j=1..n} a_j 1_{m_j}(a_i).$$

Soit $\alpha = (a_i)_{i=1..n}$.

On a :

$$U\alpha = \begin{pmatrix} \varphi A(a_i) \\ \dots \\ \varphi A(a_n) \end{pmatrix}.$$

Le vecteur $(\varphi A(a_i), \dots, \varphi A(a_i))$ est combinaison linéaire de n variables indicatrices. IL appartient à $W = \{x \in \mathbb{R}^m, x = U\alpha\}$.

Le problème de la dépendance entre les deux variables qualitatives A et B est celui des relations entre les deux groupes de fonctions indicatrices qui leur sont associées. C'est donc un exemple d'analyse canonique.

Les deux tableaux à analyser sont les tableaux disjonctifs U_1 et U_2, on leur associe les potentiels de prévision.

$W_1 = \{x \in \mathbb{R}^m, x = U_1\alpha, \alpha \in \mathbb{R}^{n*}\}$: sous espace vectoriel de $\mathbb{R}^m$, espace des combinaison linéaire des p fonctions indicatrices de modalités A.

$W_2 = \{x \in \mathbb{R}^m, x = U_2 b, b \in \mathbb{R}^{n*}\}$: sous espace vectoriel de $\mathbb{R}^m$, espace des combinaisons linéaires des p fonctions indicatrices des modalités de B.

On constate que :

$\sum_{i=1..n} 1_{(m_i)} = \delta_m$, $\sum_{i=1..n} 1_{(m'_i)} = \delta_m$, où δ_m est le vecteur de $\mathbb{R}^m$ dont toutes les composantes sont égales à 1.

Par suite $dimW_1 \cap W_2$ est supérieure ou égale à 1.

Les variables canoniques autres que δ_m formant des systèmes orthonormés (pour la métrique choisie dans $\mathbb{R}^m$) de W_1, W_2 sont donc centrées car orthogonaux à δ_m. En choisissant comme métrique sur $\mathbb{R}^m$, la métrique diagonale $\frac{1}{n} I_n$, on a, avec les notations usuelles :

$$V_{11} = \frac{1}{m} U_1^t U_1 = \frac{1}{m} D_1, V_{22} = \frac{1}{m} U_2^t U_2 = \frac{1}{m} D_2$$

$$V_{12} = \frac{1}{m} U_1^t U_2 = \frac{1}{m} X, V_{21} = \frac{1}{m} U_2^t U_1 = \frac{1}{m} X$$

Les facteurs canonique du groupe 1 sont vecteurs propres de

$$V_{11}\, V_{22} V_{12} V_{21}$$

Soit encore : ${D_1}^{-1} X {D_2}^{-1} X^t$, ils sont colinéaires aux composantes principales dans l'analyse des profils-lignes de X.

Les facteurs canoniques du groupe 2 sont vecteurs propres de ${D_2}^{-1} X {D_2}^{-1} X^t$, ils sont colinéaires aux composantes principales dans l'analyse des profils-colonnes de X.

Les valeurs propres sont les carrés des coefficients de corrélation canonique, et donc sont inférieures ou égales à1.

On retrouve la valeur propre triviale $\lambda = 1$, associée au vecteur δ_m élément de $W_1 \cap W_2$.

Les modalités des deux variables qualitatives A et B jouent des rôles identiques, on peut donc faire une représentation simultanée des projections des individus et des variabless sur les axes factoriels.

Mise en élément supplémentaire

A l'issue d'une A.F.C il se peut qu'un individu (ou un caractère) perturbe la détermination des axes factoriels. La parade consiste alors à éliminer du tableau initial la ligne correspondante et à refaire une A.F.C. l'individu que l'on a ainsi mis en élément supplémentaire sera ensuite projeté sur les axes factoriels déterminés sans lui.

Soit π_j , le profil de cet individu. Soit α (resp b) une composante principale dans l'analyse de N(I). la composante de l'individu i de profil f_j^i, sur l'axe factoriel correspondant est donnée par la relation :

$$\alpha_i = \frac{1}{\sqrt{\lambda}} \sum_{j=1..n} f_j^i b_j,$$

La composante de π_j sur cet axe sera $\frac{1}{\sqrt{\lambda}} \sum_{j=1..n} f_j^i b_j$:

$$\frac{1}{\sqrt{\lambda}} \sum_{j=1..n} \pi_j \, b_j.$$

<u>Part d'inertie expliquée, contribution, corrélations, qualité des représentations :</u>

De même qu'un A.C.P, on mesure l'importance relative de l'axe Δu_k par le pourcentage d'inertie expliqué par cet axe, donné par l'expression suivante :

$$k \neq 1, \tau_k = 100 \times \frac{\lambda_k}{\sum_{i=1..n} \lambda_i}$$

Où r est le nombre d'axes factoriels, λ_j la j-iéme plus grande valeur propre 1 n'intervient pas dans le calcul du dénominateur.

Comme en A.C.P on calcule pour chaque individu et chaque caractère, sa contribution à la détermination de l'axe Δu_k par les formules suivantes :

Soient a, b une composante principale dans l'analyse de N(I) (resp N(J)) associée à la valeur propre λ, on a :

$$\lambda = \frac{1}{x_{..}} \sum_{i=1..n} x_i \, (a_i)^2 = \frac{1}{x_{..}} \sum_{i=1..n} x_i \, (b_i)^2.$$

Définition 6.3

On appelle contribution (CTR) du profil-ligne f_j^i (resp profil-colonne) par rapport à l'axe Δu_k la quantité :

$$CTR_k(i) = \frac{\frac{x_i}{x_{..}} \left(a_i{}^2\right)^2}{\lambda_k} \; \left(CTR_k(i) = \frac{\frac{x_i}{x_{..}} \left(b_i{}^2\right)^2}{\lambda_k}\right.$$

Plus la valeur d'une contribution est proche de 1 plus le produit correspondant « tire » l'axe à lui, plus ce point est déterminant pour l'interprétation de l'axe.

Définition 6.4

On appelle corrélation (COR) du profil-ligne f_j^i, par rapport à l'axe Δu_k la quantité :

$$CTR_k(i) = \frac{\left(a_i{}^{(k)}\right)^2}{d^2 \chi^2_{f_J}(f_j^i, f_J)} \; \left(CTR_k(i) = \frac{\left(b_j{}^2\right)^2}{\chi^2_{f_J}(f_j^i, f_J)}\right).$$

Rappelons que l'on a deux écritures possibles pour $d^2 \chi^2_{f_J}(f_j^i, f_J)$:

$d^2 \chi^2_{f_J}\left(f_j^i, f_J\right) = \frac{\sum_{j=1..p} \left(f_j^i - f_j\right)^2}{f_j} = \sum_{k=1..p} \left(a_i{}^{(k)}\right)^2$, où r est le nombre d'axes factoriels et $a_i{}^{(k)}$ désigne la coordonne du profil f_j^i sur l'axe factoriel Δu_k soit la i-ème coordonnée de la k-ième composante principale.

Rappelons enfin que la corrélation ainsi définie est le carré du cosinus de l'angle formé par l'axe Δu_k, est le vecteur $f_j^i - f_j$ plus cette valeur est voisine de 1, plus le profil correspondant est proche de l'axe Δu_k meilleure est sa représentation par sa projection sur l'axe.

Enfin on mesure la quantité de la représentation d'un profil f_j^i sur un sous-espace engendré par $(u_k)_{k\in K}$, avec K contenu dans $\{1, \dots, r\}$ par la quantité : $\sum_{k\in K} COR_k(i)$.

Etude de la répartition des tailles d'carpes sur la Sommam.

La vallée de la Soummam.

Ce travail a été effectué par deux étudiants en environnement :

Les larves d'carpes) ou civelles montent à la surface de la mer au printemps et se dirigent vers l'embouchure des pseudo-fleuves qu'elles remontent pour accomplir leur développement, pendant 4 ou 5 ans.

L'étude concerne la Soummam divisé en 10 secteurs numérotés de secteur 1(sec1) à secteur 10(sec10), d'aval en amont ; le secteur 10 (sec10) regroupe 3 affluents.

Sur l'ensemble des secteurs on a capturé des carpes par la méthode de la pêche électrique. Les poissons pêchés sont soumis à une analyse biométrique. L'étude porte sur la taille de poissons ; celle-ci varie de 205 mm à 765 mm. Le comptage s'est fait par classe de taille de longueur 10 mm, chaque classe étant codée par sa valeur centrale.

Les données sont disposées dans un tableau à double entrée, les secteurs étant en ligne et les classes de tailles en colonnes : à la rencontre de la ligne i et de la colonne j figure le nombre d'carpes pêchées dans le secteur i dans la classe de taille j.

Une première analyse a été faite, dans laquelle, les classes de tailles 705 « tirent » l'axe à eux, rendant difficile l'interprétation des nuages. Elles ont été mises en éléments supplémentaires. Ce sont les résultats de cette deuxième analyse qui sont exposés ici.

Pourcentage d'inertie expliquée :

Remarquons tout d'abord que la première valeur propre est élevée. A partir de la deuxième, la décroissance est lente.

Le premier axe explique à lui seul 46,25% de l'inertie totale, et le plan des deux premiers axes 58,05%.

Etude des poids

En ce qui concerne les secteurs, le secteur 1 a une masse nettement plus élevée. Puis les secteurs 2, 4, 8 et 3 ont des masses du même ordre de grandeur.

Pour les classes de taille, les petites ainsi que les grandes sont faiblement représentées.

Règle d'interprétation des axes factoriels

Pour chaque axe factoriel, on relève les points à la fois des ensembles I et J dont les contributions à l'axe sont les plus élevées.

Il s'agit de relever dans la colonne CTR les plus fortes valeurs. Ceci ne s'entend pas dans l'absolu, les contributions restent souvent d'un ordre de grandeur pas très élevé ; ce sont les plus grandes valeurs relatives que l'on doit retenir.

On regroupe les points des deux ensembles selon leur positionnement sur l'axe. On peut les disposer dans un tableau (cf ci-joint)

Notre attention est ainsi attirée vers des points particuliers. Il s'agit alors de vérifier, sur le tableau initial, si les associations suggérées ainsi que les oppositions sont réellement satisfaites.

On pourra ensuite proposer une interprétation de l'axe, qu'il conviendra d'étayer.

Nous allons expliquer cette démarche.

Sur l'axe 1 de la présente analyse, le secteur sec9 est les points de I à la plus forte, contribution. Cinq classes de taille contribuent à la détermination de cet axe, et ont des projections du côté négatif de l'axe.

NomJ(J)	205	215	235	245	255	265	275	285	295	305	315	325	345	355	365
1	0	0	0	0	0	0	1	4	5	12	9	10	13	10	13
2	1	1	3	6	5	4	7	8	4	5	7	6	8	6	14
3	1	0	1	0	1	0	2	2	4	4	4	5	7	8	8
4	0	0	1	5	0	3	3	5	5	6	6	5	7	9	10
5	0	0	0	1	0	0	0	0	1	0	1	0	1	2	3
6	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0
7	0	0	0	0	0	0	0	0	1	0	1	2	2	1	1
8	0	0	0	0	0	0	0	2	0	1	0	3	2	3	2
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	1	1	0	0	1	1	3

NomJ(J)	375	385	395	405	415	425	435	445	455	465	475	485	495	505
1	25	25	17	28	29	24	22	29	20	27	20	19	26	12
2	8	12	16	18	19	11	10	11	12	5	4	6	3	7
3	5	10	9	9	13	8	9	4	9	7	9	4	11	3
4	7	8	11	6	20	12	8	3	9	6	7	4	16	8
5	0	1	1	3	6	2	4	4	6	1	1	5	3	5
6	1	3	0	1	1	0	4	2	2	2	1	3	1	1
7	3	0	3	3	4	3	0	2	2	4	1	2	3	5
8	2	0	3	2	4	7	3	5	8	4	6	5	3	6
9	0	1	0	0	1	0	0	0	0	1	1	1	0	0
10	0	0	0	0	0	0	0	1	0	0	1	0	0	1
11	0	0	0	0	0	0	0	0	0	0	1	0	1	1
12	5	4	3	4	2	0	0	2	3	2	3	4	3	2

NomJ(J)	515	545	555	565	575	485	595	605	615	625	635	645	655	665
1	7	3	2	3	1	0	2	1	0	2	0	1	0	0
2	3	2	2	1	0	0	3	3	1	0	0	0	0	0
3	4	0	1	1	2	0	0	0	0	0	0	0	0	1
4	6	2	2	2	0	2	1	1	3	1	0	1	0	2
5	9	3	4	8	3	4	7	4	4	1	3	0	1	0
6	1	3	1	2	3	4	1	1	5	6	1	2	2	4
7	5	4	1	3	2	3	2	4	1	2	3	2	2	2
8	6	11	14	9	5	5	3	8	3	8	5	5	7	8
9	3	1	3	1	2	3	4	2	3	1	2	5	5	4
10	0	0	2	3	0	0	1	0	0	0	0	1	1	1
11	1	0	0	0	0	1	1	3	0	0	0	2	0	3

NomJ(J)	675	585	595	705	715	725	735	745	755	765

1	0	0	1	0	1	0	0	0	0	0
2	2	1	0	0	1	1	0	0	0	0
3	1	0	0	0	0	1	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0	0
6	1	1	0	0	2	0	0	1	0	1
7	0	1	0	0	0	0	0	0	0	0
8	3	2	0	0	1	1	0	0	0	0
9	5	1	2	0	1	0	3	0	0	0
10	0	0	1	2	0	0	0	1	0	0
11	0	1	0	1	0	0	0	0	0	0

..

Interprétation du premier axe factoriel

$$\lambda_1 = 0.37, \qquad \lambda_2 = 48.25$$

Poids i explicatifs à coordonnées négatives	Points i explicatifs à coordonnées positives
Sec9 (258) Sec8 (185) Sec 6 (82)	Sec1 (116)) Sec2 (89)
Points j explicatifs à coordonnées négatives	Points j explicatifs à coordonnées positives
655(78) 665(77) 645 (72) 555 (48) 585 (47)	Néant

$\lambda_1 = 0.37 \quad \lambda_2 = 46.25$

Interprétation du deuxième axe factoriel

Points i explicatifs à coordonnées négatives	Points i explicatifs à coordonnées positives
Sec9 (414) Sec2 (214)	Sec 5 (85) Sec1 (70)
Points j explicatifs à coordonnées négatives	
735 (139) 675 (113) 645 (55) 695 (52) 245 (39) 365 (38) 225 (37) 265 (36) 275 (35)	545 (42) 625 (41)

$\lambda_1 = 0.94 \quad \lambda_2 = 11.80$

Parmi celles-ci les classes 655 et 645 ont des projections voisines de celle de sec9. Sur le tableau initial on constate que ces deux classes de taille ont des effectifs non négligeables dans le secteur 9.

Le secteur 9 contribue également à l'axe 2, cette fois en liaison avec la classe 735 (uniquement présente dans ce secteur (effectif 3) et dans le secteur 12 (effectif 1) et la classe 675 dont le poids est faible, mais dont le plus fort effectif est dans le secteur 9.

En conclusion on peut associer le secteur 9 aux classes de taille 655, 645, 735 et 675

Dans les secteur 1 et 2 les tailles 545 et 555 ont des effectifs faibles et les classes 585,655 et 665 des effectifs nuls. Les projections respectives de ces deux secteurs et des classes citées s'opposent sur l'axe 1.

En ce qui concerne le secteur 8 on remarque des effectifs importants des classes 545, 555 ,655, 665 et 585 d'où leur association sur l'axe 1.

Pour le secteur 6 on peut lui associer les classes 665 et 585.

L'axe 1 oppose les secteurs 1 et 2 situés en aval de la rivière aux secteurs 8 et 9 situés beaucoup plus en amont. Les classes de petite taille se projettent au voisinage des secteurs 1 et 2 alors que les lasses de taille 545…665 sont voisines des secteurs 8 et 9. L'axe 1 est-il un axe de taille opposant les secteurs situés en aval à ceux situés en amont ? on constate que le long de cet axe se répartissent les classes de taille 205 à

695 : on peut interpréter le premier axe comme axe de taille . en parallèle, le long de l'axe 1 se répartissent les 10 secteurs : (sec1, sec2, sec3, sec4) très voisins, puis sec5, sec 6 retour vers sec7, puis sec8, sec9, à nouveau un retour vers sec 10. Cependant, il est intéressant de noter que le les classes de grande taille 715,725,735, 755, 765 n'obéissent pas à la règle générale.

Il n'est pas surprennant de trouver ces résultats : pendant la phase de sédentarisation de la vie des carpes, les petites carpes entament une phase de colonisation de la rivière parallèlement à leur croissance et leur passege à l'âge adulte.

Par contre, les points qui sont en rupture avec ce schème global s'avèrent intéressants Essayons d'expliquer les positions dans le secteur 10: dans ces deux secteurs, on a pêché des exemplaires d'carpes assez petites(taille 445,475,55,…) mais aussi de grandes carpes. Dans ces deux secteurs, il semble que ces poissons soient encore en phase de croissance. Il faut cependant émettre des réserves quant à cette interprétation. Dans les relevés étudiés, on observe une prépondérance des classes moyennes 450-550. On peut se poser les questions suivantes : cette prépondérance correspond-t-elle à une réalité ?

Ou bien, peut-être a-t-on un mauvais échantillonnage ? certains difficultés se présentent, inhérentes à la méthode de pêche. En effet, souvent le couvert végétal rivulaire est important, l'épaisseur et la nature du sédiment sont variables. On peu aussi penser que les individus de grande taille échappent plus facilement à la capture.

En ce qui concerne les positions des aguilles de grande taille, on a vu sur le tableau de données initial que 3 individus de taille 735 caractérisent le sec9. Les secteurs 1,2 et 3 contiennent des individus taille 715, 725. Le sec6 des individus de taille 765,715. On retrouve ces rapprochements sur le plan de projection peut expliquer la présence de ces carpes en aval par le fait qu'elles sont arrivées à (maturité et se préparent à partir en mer pour se reproduire.

Sur le plan 1-2 le nuage a une forme parabolique caractéristique de ce que l'on appelle l'effet Guttman. Sur l'axe 1, s'opposent les classes extrêmes ; sur l'axe 2, les moyens s'opposent aux extrêmes.

Pour interpréter les positions des points, il faut voir sur quels sous espaces ils sont le mieux représenter. i.e : sur quels axes les valeurs des corrélations (COR) sont les plus fortes.

Sur le plan (1-2), les secteurs les mieux approchés par leurs projections sont les secteurs 9 (QLT=0.885), sec8(QLT=0.771), sec2 (QLT=0.699) et sec1 (QLT=0.623). Le secteur 6 est bien représenté sur le plan 1-4 (QLT=0.420+0.369=0.789), le secteur 10aussi (QLT=0.264+0.353=0.617).

Pour la classe de taille, on peut relever sur l'axe 1 les classes 305(0.815) 315(0.829) 395(0.768) 415(0.823) 575(0.706) 585(0.721) 655(0.783) : sur le plan 1-2 735(0.272+0.349) 675(0.430+0.345) 645(0.674+0.130) 365(.349+.463) 285 (0.361+0.261) 235(0.252+0.345).

Par rapport à un axe donné, un point peut avoir une contribution assez élevée sans que la corrélation soit très forte. Par exemple, par rapport à l'axe 1 :

	CTR	COR
Sec9	0.258	0.628
Sec8	0.185	0.718

Les valeurs de CTR et de COR dépondent de l'abscisse du point, de sa masse, mais également de la valeur propre correspondante pour CTR, et de la distance du point à l'origine pour COR.

On peut aussi avoir des points bien approchés par leurs projections (« bon » COR) mais qui ne jouent pas un rôle important dans la détermination de l'axe (CTR faible).

<u>Mise en garde dans les interprétations</u>

Interpréter des points ayant des positions extrêmes sur les graphiques, en particulier opérer des regroupements entre points j ne pose aucun problème.

Par contre, Attention, aux points occupant des positions intermédiaires. Nous allons illustrer par un exemple :

Soit la classe de taille 535. Sa projection sur le plan 1-2 est voisine de celle des secteurs 5 et 7, mais il n'y a pas lieu de les lui associer. En effet, la répartition de cette classe est la suivante :

Sec1 8	Sec5 7	Sec4 6	Sec7 5	Sec8 5	Sec 4
Sec9 2	Sec12 2	Secc2 1	Sec3 0	Sec10 0	

Le profil correspondant à cette classe est :

$$f_J^{535} = \left(\frac{8}{40}, \frac{1}{40}, \quad \frac{0.6}{40}, \quad \frac{7}{40}, \quad \frac{4}{40}, \frac{5}{40}, \frac{2}{40}, \quad 0, \quad 0.1/40\right)$$

Sur l'axe 1, on note $G_1(j)$ l'abscisse de j et $F_1(i)$ l'abscisse de i. la formule de transition est la suivante :

$G_1(j) = (1/\sqrt{\lambda_1})\sum_{i=1..12} f_i^j \, F_1(i)$.

Au coefficient $(1/\sqrt{\lambda_1})$ près, la projection de j est le barycentre des projections des 10 secteurs.

Chapitre 7 Régressions linéaire multiple

On a mesuré sur une population P de n individus, p+1 variables numériques, représentées par des vecteurs de $\mathbb{R}^n$, notés : $y, x_1, \dots, x_n$, comme d'habitude, on suppose que n est grand devant p.

On considère ici le cas où y dépend d'une certaine manière des p variables $x_1, \dots, x_p, y$, porte le nom de variable à expliquer.

Exemple 7.1

L'étude porte sur une population de n parcelles agricoles ; y est le rendement en quintaux de blé d'une parcelle cultivée.

x_1 est la quantité d'engrais épandée dans la parcelle avant les semences.

x_2 est la quantité de désherbant répandu dans la parcelle avant les semences.

x_3 est la quantité d'insecticide répandu dans la parcelle avant les semences.

x_4 est la quantité de fongicide répandu dans la parcelle avant les semences.

x_5 est la quantité de raccourcisseur répandu dans la parcelle avant les semences.

On cherche à exprimer y en fonction des p variables x_i , soit une relation du type :

$y = f(x_1 \dots, x_p) + e$.

Où : $f(x_1 \dots, x_p)$ approche au mieux y. Le vecteur e porte le nom de résidu ou encore d'écart entre y et sa valeur approchée $f(x_1 \dots, x_p)$.

Dans le cas de la régression linéaire, on impose à f d'être affine.

Recherche d'un ajustement linéaire

On cherche les coefficients : $a_1 \ldots, a_p$ tels que :

$\forall i \in \{1, \ldots, n\}, y^i = a_0 + a_1 x_1^i + \cdots + a_p x_p^i + e^i$.

Les inconnues $a_0, a_1, \ldots, a_p, e^i$, on affecte une population p_i à x^i, pour tout i, avec $\sum_{i=1\ldots n} p_i = 1$.

<u>Méthode des moindres carrés</u>

On peut poser le problème de la manière suivante :

Trouver $a_0, a_1, \ldots, a_p$ tels que $\sum_{i=1..n} p_i (e^i)^2$ soit minimale, ou en posant :

$\varphi(a_0, a_1, \ldots, a_p) = \sum_{i=1..n} p_i \left(y^i - a_0 + a_1 x_1^i + \cdots + a_p x_p^i\right)^2$ trouver s'il existe, le minimum de φ.

Les solutions sont à rechercher parmi les solutions de :

$\forall k \in \{1, \ldots, p\}, \quad \frac{\partial \varphi}{\partial a_k}(a_0, a_1, \ldots, a_p) = 0.$

Il s'agit de résoudre p+1 équations linéaires.

<u>Solution algébrique</u>

On pose :

$$X = \begin{bmatrix} 1 & x_1^1 \cdots & x_p^1 \\ \vdots & \ddots & \vdots \\ 1 & x_n^1 \cdots & x_p^n \end{bmatrix}, y = \begin{pmatrix} y^1 \\ \ldots \\ y^n \end{pmatrix}, E = \begin{pmatrix} e^1 \\ \ldots \\ e^n \end{pmatrix}, A = \begin{pmatrix} a_0 \\ \cdots \\ a_p \end{pmatrix}.$$

On cherche le vecteur y^*, de la forme $y^* = a_0 \delta_n + a_1 x_1 + \cdots . + a_p x_p$ tel que $y = y^* + e, y^*$ représentant une valeur approchée de y.

On munit $\mathbb{R}^n$, de la métrique définie par D_p on suppose $(\delta_n, x_1, \ldots, x_p)$ linéairement indépendants. On note W le sous espace de $\mathbb{R}^n$, engendré par $(\delta_n, x_1, \ldots, x_p)$ $(dimW = p + 1)$.

Matriciellement, on peut écrire :

$$y = XA + e$$

X et y sot donnés. On cherche A tel que $\|e\|_{D_p}{}^2$, soit minimal. Il est équivalent de chercher y^*, tel que $\|y - y^*\|_{D_p}{}^2$, soit minimal.

On sait que la solution à ce problème est unique et que y^* est la projection D_p-orthogonal de y sur W.

L'opérateur de projection D_p-orthogonale sur W s'écrit : $X(X^t D_p X)^{-1} X^t D_p y$.

Soit : y^*=XA, avec A= $X(X^t D_p X)^{-1} X^t D_p$.

Proposition 7.1

Le vecteur qui minimise $\|y - y^*\|_{D_p}{}^2$, est donné par : y^*=XA.

Avec A= $X(X^t D_p X)^{-1} X^t D_p$.

Proposition 7.2

La moyenne de la série $(e^i, p_i)_{i=1..n}$est nulle.

Preuve 7.1

On sait que $y - y^*$ est orthogonal à W, c'est-à-dire en particulier à δ_n.

$y - y^*$est orthogonal à $\delta_n \Leftrightarrow \sum_{i=1\ldots n} p_i e^i = 0$.

Cas où les individus sont équipondérés :

On a : $D_p = \frac{1}{n} I_n$, les formules se simplifient :

$A = (X^t X)^{-1} X^t y \qquad y^* = X(X^t X)^{-1} X^t y.$

Cas où les variables sont centrées

On peut écrire :

$$y^* = a_0 \delta_n + a_1 x_1 + \cdots . + a_p x_p$$

$$\overline{y^*} = a_0 + a_1 \overline{x_1} + \cdots . + a_p \overline{x_p}$$

$y = y^* + e$ et $\bar{y} = \overline{y^*}$.

$y - \bar{y}\delta_n = a_1(x_1 - \delta_n \overline{x_1}) + \cdots + a_p\left(x_p - \overline{x_p}\delta_n\right) + e.$

Soit en posant X' et y' les tableaux centrés associés à X et y.

$y^* - \bar{y}\delta_n = X'(X'^t D_p X')^{-1} X'^t D_p y'.$

$X'^t D_p X'$ est la matrice de covariance de $\left(x_1, \dots, x_p\right)$. on la note souvent Σ_{XX}.

$X'^t D_p y'$ représente la matrice de covariance de y avec les x_j. On la note Σ_{Xy}.

On peut alors écrire : $y'^* = y^* - \bar{y}\delta_n = X(\Sigma_{XX})^{-1}\Sigma_{Xy}$.

On calcule ensuite a_0 par la formule :

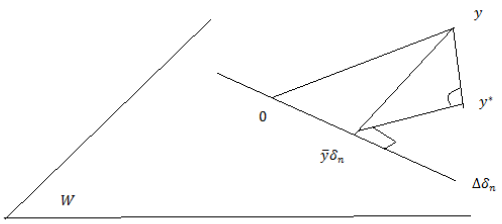
$a_0 = \bar{y} - a_1 \overline{x_1} - \cdots - a_p \overline{x_p}.$

Qualité de l'ajustement

La qualité de l'ajustement se mesure par le coefficient de corrélation multiple $R = r(y, y^*)$.

$R = r(y, y^*) = \frac{cov(y,y^*)}{\sqrt{varyvary^*}}.$

On rappelle que : $vary = \|y - \bar{y}\delta_n\|^2_{D_p,} \qquad vary^* = \|y^* - \bar{y}\delta_n\|^2_{D_p,} \quad cov(y, y^*) = D_p(y - \bar{y}\delta_n, y^* - \bar{y}\delta_n).$



$\Delta\delta_n$ est la droite engendré par δ_n, elle porte le nom de droite des constantes.

On a : $proj_{\Delta\delta_n}(y) = \bar{y}\delta_n proj_{\Delta\delta_n}(y^*)$ d'après le théorème des trois perpendiculaires.

R apparaît comme le cosinus de l'angle des vecteurs $y - \bar{y}\delta_n$ et $y^* - \bar{y}\delta_n$. Si R=1, l'ajustement est parfait.

Régression pas à pas

On peut chercher la régression à expliquer y par seulement q variables explicatives choisies parmi les p.

Pas à pas ascendant

On cherche la variable x_j la plus corrélée avec y. on incorpore successivement les variables qui provoquent un plus fort accroissement de R^2.

Les deux méthodes ne donnent pas nécessairement le même résultat, et ne donnent pas nécessairement les meilleurs régressions à q variables. Leur principal intérêt est d'ordre pratique.

Application à l'ajustement polynomial

Dans le cas d'une seule variable explicative x, une régression simple de y en x peut ne pas être satisfaisante. On peut alors chercher un ajustement du type : $y^* = a_0\delta_n + a_1x + \cdots + a_px^p$.

Exemples

Exemple 7.1 Sous le logiciel R une régression linéaire

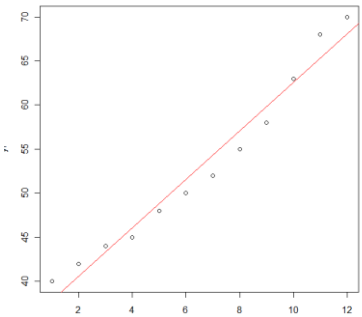
```
> yi <- c(40, 42, 44, 45, 48, 50, 52, 55, 58, 63, 68, 70)
```

```
> xi <- c(1:12)
> xi
 [1]  1  2  3  4  5  6  7  8  9 10 11 12
> plot(xi, yi)
```

```
> plot(xi, yi)
> lm(yi ~ xi)

Call:
lm(formula = yi ~ xi)
```

```
> plot(xi, yi)
> abline(35.076, 2.745, col = 'red')
> plot(xi, yi)
```



Exemple 7.2

Utilisation de la méthode des moindres carrés

<u>Numériquement :</u> sous R

On génère une matrice 10× 10

> matrice<-matrix(1:100,nrow=10,ncol=10;

> matrice<-matrix(1:100,nrow=10,ncol=10);

> matrice;

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	1	11	21	31	41	51	61	71	81	91
[2,]	2	12	22	32	42	52	62	72	82	92
[3,]	3	13	23	33	43	53	63	73	83	93
[4,]	4	14	24	34	44	54	64	74	84	94
[5,]	5	15	25	35	45	55	65	75	85	95
[6,]	6	16	26	36	46	56	66	76	86	96
[7,]	7	17	27	37	47	57	67	77	87	97
[8,]	8	18	28	38	48	58	68	78	88	98
[9,]	9	19	29	39	49	59	69	79	89	99
[10,]	10	20	30	40	50	60	70	80	90	100

```
> cbind(matrice([1,1:10]));
> cbind(matrice[1:10,1]);
     [,1]
 [1,]    1
 [2,]    2
 [3,]    3
 [4,]    4
 [5,]    5
 [6,]    6
 [7,]    7
 [8,]    8
 [9,]    9
[10,]   10

> y<-cbind(matrice[1:8,1]);
> y
     [,1]
[1,]    1
[2,]    2
[3,]    3
[4,]    4
[5,]    5
[6,]    6
[7,]    7
[8,]    8
```

```
> y<-cbind(matrice[1:8,2])
```

Le régresseur X1

```
> y;
     [,1]
[1,]   11
[2,]   12
[3,]   13
[4,]   14
[5,]   15
[6,]   16
```

```
> cbind(matrice[1:8,3],matrice[1:8,4],matrice[1:8,5],matrice[1:8,6],matrice[1:8,7],matrice[1:8,8],matrice[1:8,9],matrice[1:8,10])
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]   21   31   41   51   61   71   81   91
[2,]   22   32   42   52   62   72   82   92
[3,]   23   33   43   53   63   73   83   93
[4,]   24   34   44   54   64   74   84   94
[5,]   25   35   45   55   65   75   85   95
[6,]   26   36   46   56   66   76   86   96
[7,]   27   37   47   57   67   77   87   97
[8,]   28   38   48   58   68   78   88   98
```

X1 est expliquée à partir des régresseurs X2,X3,X4,X5,X6,X7 et X8.

```
> x<-cbind(matrice[1:8,3],matrice[1:8,4],matrice[1:8,5],matrice[1:8,6],matrice[1:8,7],matrice[1:8,8],matrice[1:8,9],matrice[1:8,10])
> c<-t(x)%*%x
```

```
> c;
     [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
[1,]  4844  6804  8764 10724 12684 14644 16604 18564
[2,]  6804  9564 12324 15084 17844 20604 23364 26124
[3,]  8764 12324 15884 19444 23004 26564 30124 33684
[4,] 10724 15084 19444 23804 28164 32524 36884 41244
[5,] 12684 17844 23004 28164 33324 38484 43644 48804
[6,] 14644 20604 26564 32524 38484 44444 50404 56364
[7,] 16604 23364 30124 36884 43644 50404 57164 63924
[8,] 18564 26124 33684 41244 48804 56364 63924 71484
> d<-c^(-1);
> d;
          [,1]         [,2]         [,3]         [,4]         [,5]         [,6]         [,7]         [,8]
[1,] 2.064410e-04 1.469724e-04 1.141031e-04 9.324879e-05 7.883948e-05 6.828735e-05 6.022645e-05 5.386770e-05
[2,] 1.469724e-04 1.045588e-04 8.114249e-05 6.629541e-05 5.604125e-05 4.853427e-05 4.280089e-05 3.827898e-05
[3,] 1.141031e-04 8.114249e-05 6.295643e-05 5.142975e-05 4.347070e-05 3.764493e-05 3.319612e-05 2.968769e-05
[4,] 9.324879e-05 6.629541e-05 5.142975e-05 4.200975e-05 3.550632e-05 3.074653e-05 2.711203e-05 2.424595e-05
[5,] 7.883948e-05 5.604125e-05 4.347070e-05 3.550632e-05 3.000840e-05 2.598482e-05 2.291266e-05 2.049012e-05
[6,] 6.828735e-05 4.853427e-05 3.764493e-05 3.074653e-05 2.598482e-05 2.250023e-05 1.983970e-05 1.774182e-05
[7,] 6.022645e-05 4.280089e-05 3.319612e-05 2.711203e-05 2.291266e-05 1.983970e-05 1.749353e-05 1.564358e-05
[8,] 5.386770e-05 3.827898e-05 2.968769e-05 2.424595e-05 2.049012e-05 1.774182e-05 1.564358e-05 1.398914e-05
> e<-d%*%x
> e
           [,1]        [,2]        [,3]       [,4]       [,5]       [,6]       [,7]       [,8]
[1,] 0.019311864 0.027531726 0.035751589 0.04397145 0.05219131 0.06041118 0.06863104 0.07685090
[2,] 0.013734437 0.019580681 0.025426925 0.03127317 0.03711941 0.04296566 0.04881190 0.05465815
[3,] 0.010656699 0.015193012 0.019729324 0.02426564 0.02880195 0.03333826 0.03787457 0.04241089
[4,] 0.008705821 0.012411766 0.016117711 0.01982366 0.02352960 0.02723555 0.03094149 0.03464744
[5,] 0.007358695 0.010491233 0.013623770 0.01675631 0.01988885 0.02302138 0.02615392 0.02928646
[6,] 0.006372608 0.009085404 0.011798200 0.01451100 0.01722379 0.01993659 0.02264939 0.02536218
[7,] 0.005619568 0.008011818 0.010404067 0.01279632 0.01518857 0.01758082 0.01997307 0.02236531
[8,] 0.005025692 0.007165142 0.009304591 0.01144404 0.01358349 0.01572294 0.01786239 0.02000184
> a<-e%*%y
> a;
```

[,1]

[1,] 5.922675

[2,] 4.212312

[3,] 3.268445

[4,] 2.670139

[5,] 2.256985

[6,] 1.954555

[7,] 1.723598

[8,] 1.541454

> a a été calculé selon la formule $a = (x'x)xy$..

Exemple 7.2

On saisit sous excel les données, colonne 2 et 3, x1 est expliquée par x2 :

1	100	40
2	110	45
3	112	50
4	113	56
5	114	60
6	105	54
7	101	55
8	102	57
9	103	37
10	106	48

Statistiques descriptives (Données quantitatives) :

Variable	Observations	Obs. avec données manquantes	Obs. sans données manquantes	Minimum	Maximum	Moyenne	Ecart-type
40	9	0	9	37,000	60,000	51,333	7,141
100	9	0	9	101,000	114,000	107,333	5,000

Matrice de corrélation :

	100	40
100	1	0,238
40	0,238	1

Régression de la variable 40 :

Coefficients d'ajustement (40) :

Observations	9,000
Somme des poids	9,000
DDL	7,000
R^2	0,057
R^2 ajusté	-0,078
MCE	54,983
RMCE	7,415
MAPE	12,040
DW	1,544
Cp	2,000
AIC	37,801
SBC	38,196
PC	1,482

Analyse de la variance (40) :

Source	DDL	Somme des carrés	Moyenne des carrés	F	Pr > F
Modèle	1	23,120	23,120	0,420	0,537
Erreur	7	384,880	54,983		
Total corrigé	8	408,000			

Calculé contre le modèle Y=Moyenne(Y)

Paramètres du modèle (40) :

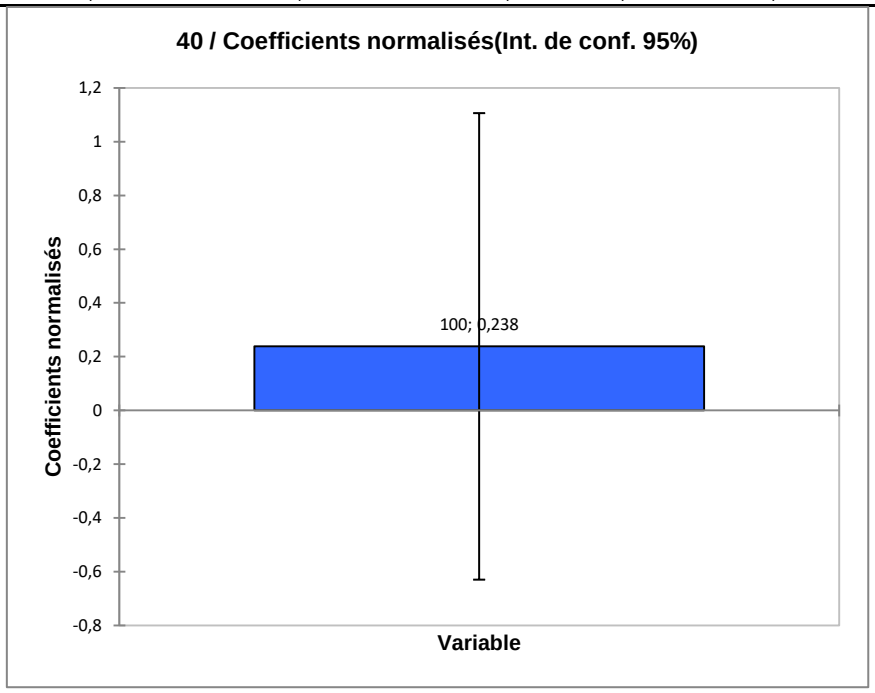
Source	Valeur	Erreur standard	t	Pr > |t|	Borne inférieure (95%)	Borne supérieure (95%)
Constante	14,840	56,332	0,263	0,800	-118,363	148,043
100	0,340	0,524	0,648	0,537	-0,900	1,580

Equation du modèle (40) :

40 = 14,84+0,34*100

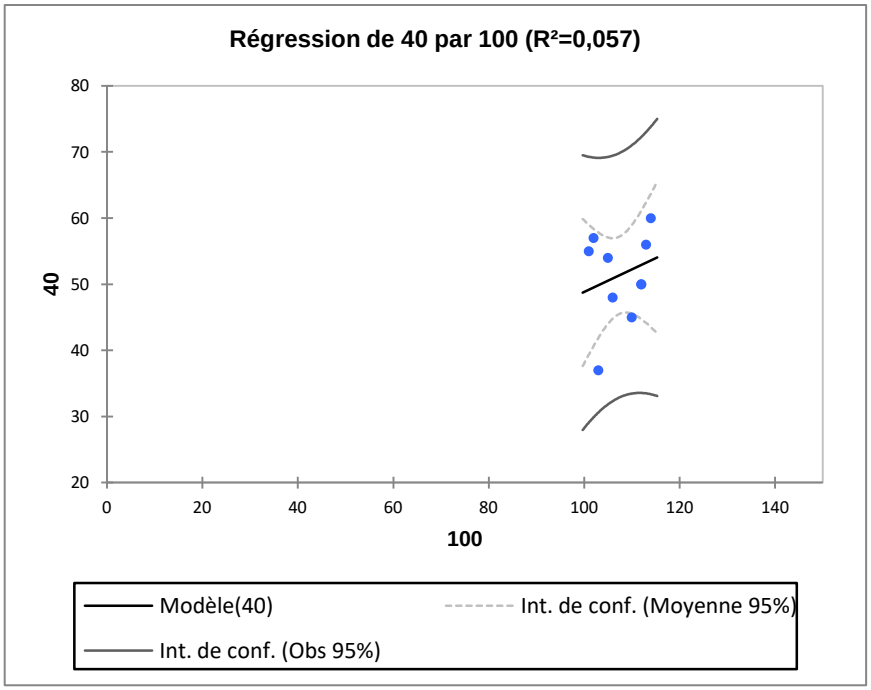
Coefficients normalisés (40) :

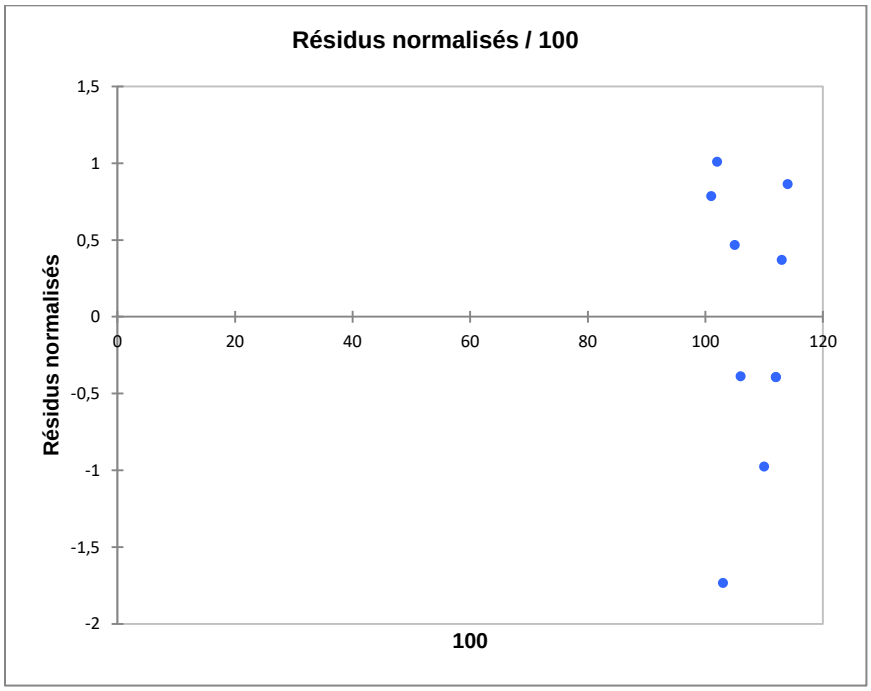
Source	Valeur	Erreur standard	t	Pr > |t|	Borne inférieure (95%)	Borne supérieure (95%)
100	0,238	0,367	0,648	0,537	-0,630	1,106

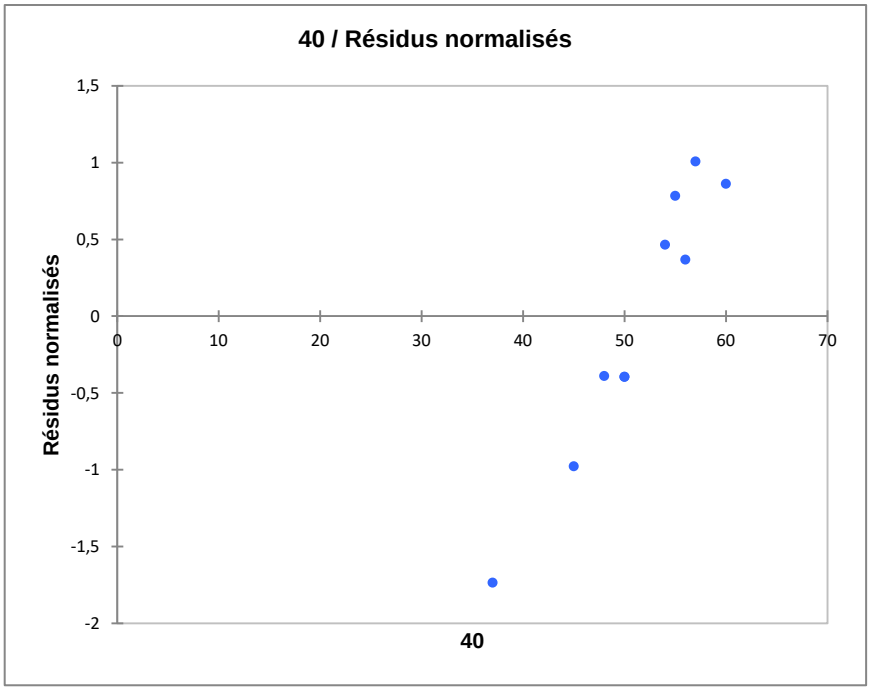


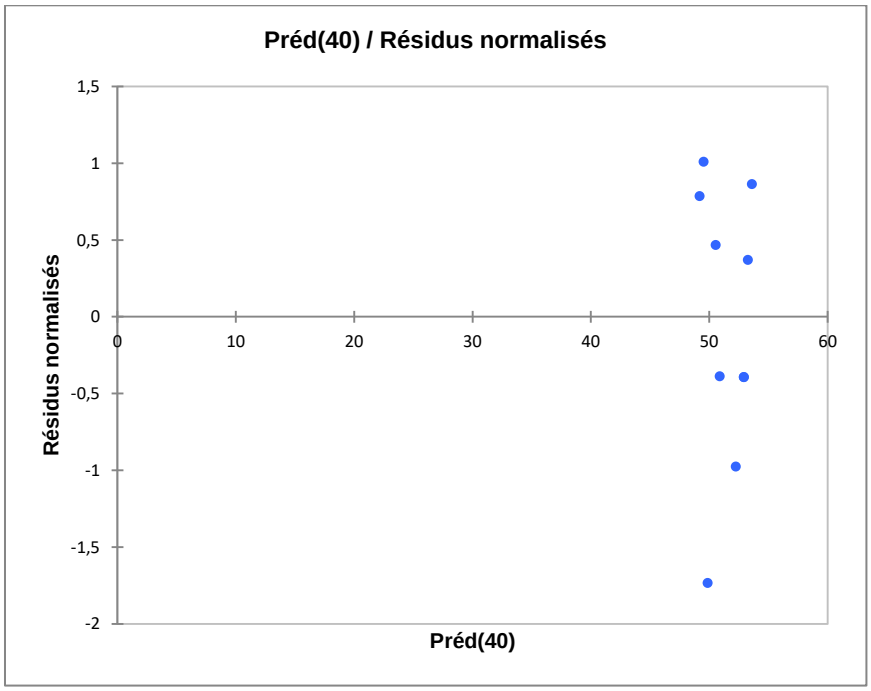
Prédictions et résidus (40) :

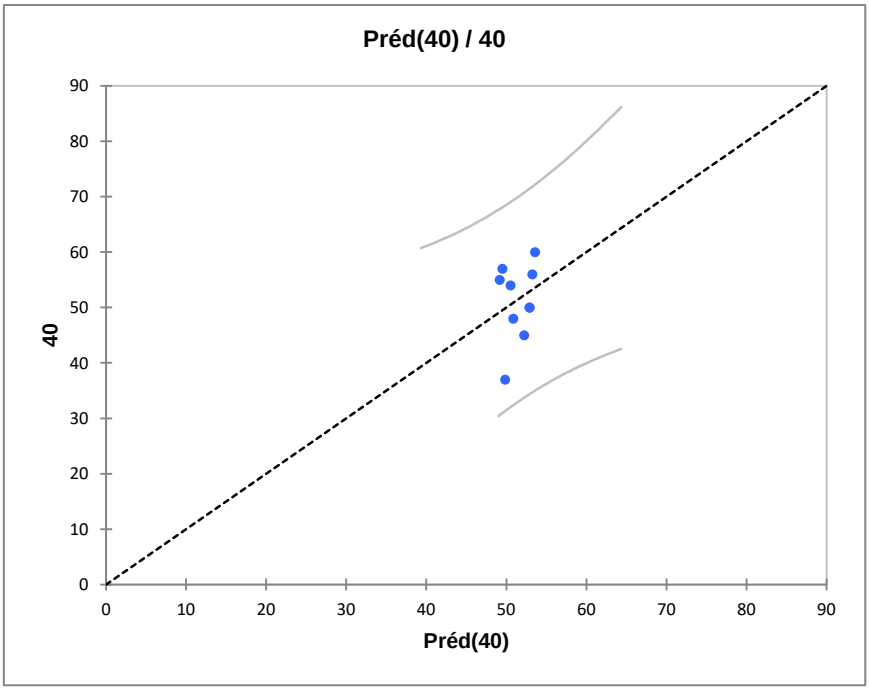
Observation	Poids	100	40	Préd(40)	Résidu	Résidu std.	Ecart-type. (Moyenne)
Obs1	1	110,000	45,000	52,240	-7,240	-0,976	2,840
Obs2	1	112,000	50,000	52,920	-2,920	-0,394	3,478
Obs3	1	113,000	56,000	53,260	2,740	0,370	3,865
Obs4	1	114,000	60,000	53,600	6,400	0,863	4,281
Obs5	1	105,000	54,000	50,540	3,460	0,467	2,758
Obs6	1	101,000	55,000	49,180	5,820	0,785	4,140
Obs7	1	102,000	57,000	49,520	7,480	1,009	3,732
Obs8	1	103,000	37,000	49,860	-12,860	-1,734	3,357
Obs9	1	106,000	48,000	50,880	-2,880	-0,388	2,569

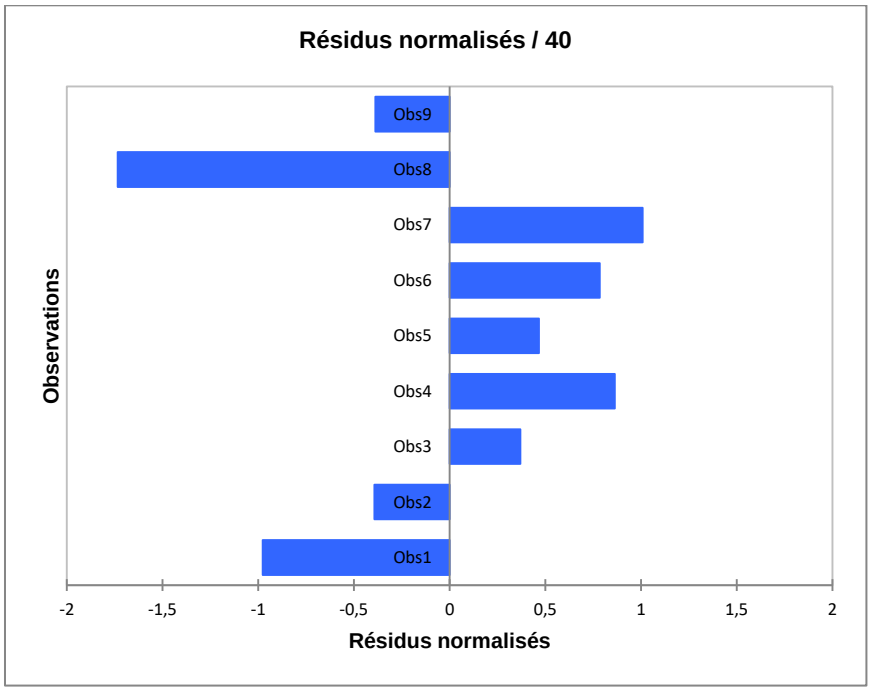

Régression de 40 par 100 (R²=0,057)
80
70
60
50
40
30
20
40
0
20
40
60
80
100
120
140
100
Modèle(40)
Int. de conf. (Moyenne 95%)
Int. de conf. (Obs 95%)


Résidus normalisés / 100
1,5
1
0,5
0
-0,5
-1
-1,5
-2
Résidus normalisés
0
20
40
60
80
100
120
100


40 / Résidus normalisés
Résidus normalisés
1,5
1
0,5
0
-0,5
-1
-1,5
-2
0
10
20
30
40
50
60
70
40

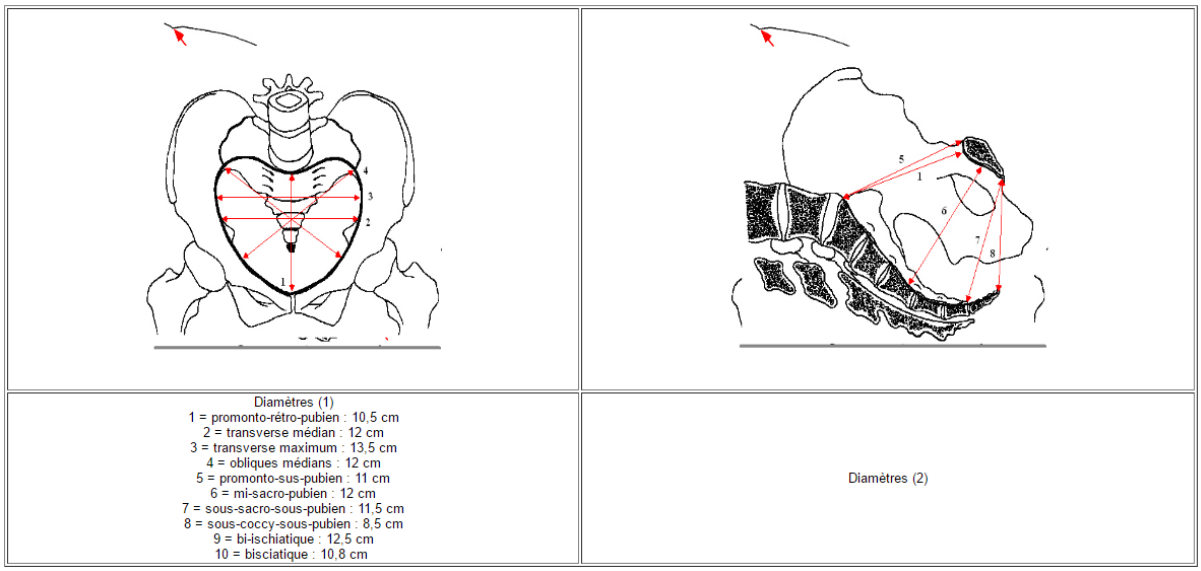

Préd(40) / Résidus normalisés
Résidus normalisés
1,5
1
0,5
0
-0,5
-1
-1,5
-2
0
10
20
30
40
50
60
Préd(40)

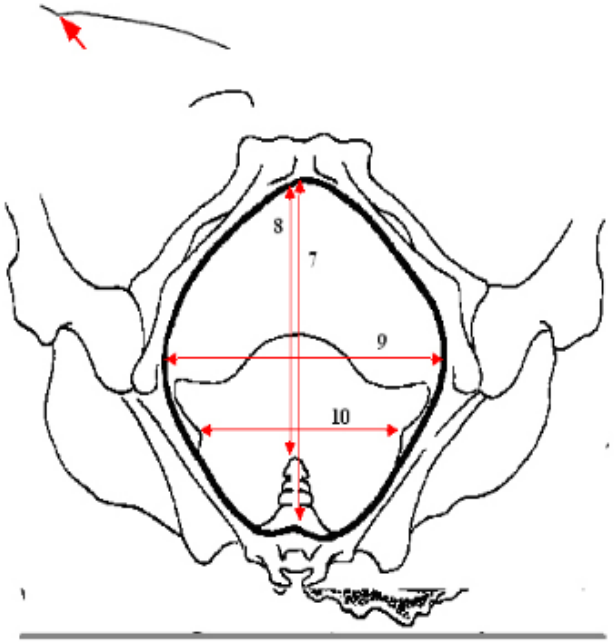

Préd(40) / 40
40
Préd(40)
0
10
20
30
40
50
60
70
80
90


Résidus normalisés / 40
Observations
Résidus normalisés
Obs9
Obs8
Obs7
Obs6
Obs5
Obs4
Obs3
Obs2
Obs1
-2
-1,5
-1
-0,5
0
0,5
1
1,5
2

Exemple 7.3

On utilise ici un exemple issu de la médecine pour traiter une ACP, les diamètres des diaètre pubien dans le squelette, au niveau du bassin voici une imagerie :





Diamètres (3)

Sous R on a les calculs :

Chargement des données et affichage d'un résumé :

library(FactoMineR)

```
on charge les données

summary(DataACP)

# Centrage en lignes du tableau de données :

Tab = log(DataACP[ ,-1]) # Tab = seulement les données métriques

 Tab = t(scale(t(Tab), center=TRUE, scale=FALSE)) # centrage en lignes

 Tab = data.frame(DataACP[,1], Tab)

colnames(Tab)[1] = "Sexe"

# Procéder à l'ACP :

 res.PCA = PCA(Tab, quali.sup=1, graph=FALSE)

 plot.PCA(res.PCA, habillage=1)

 plot.PCA(res.PCA, choix="var")
```

La matrice tab sur laquelle on travaille :

1	2	3	4	5
MH_F_21	F	0.032924380	0.034249870	0.08353337
MH_M_1	M	0.044170777	0.055509451	0.17295784
MH_M_2	M	0.081241557	-0.020627454	0.12471720
MH_M_3	M	0.106451745	0.095336083	0.23970307
MH_M_4	M	-0.057447382	0.043885297	0.17461489
MH_M_5	M	0.005490778	0.026196102	0.17264860
MH_M_6	M	0.003340791	0.003671455	0.17178321
MH_M_7	M	0.104566817	0.019770934	0.23872056
MH_M_8	M	-0.008601073	0.110762155	0.22104388
MH_M_9	M	0.013427275	0.085478049	0.23814470
MH_M_10	M	0.026220608	0.101227265	0.16082217
MH_M_11	M	-0.031235602	0.087978336	0.18517822
MH_M_12	M	-0.060948960	0.134390785	0.15784614
MH_M_13	M	0.095590454	-0.020779701	0.11005805
MH_M_14	M	-0.033914832	0.109492636	0.18358317
MH_M_15	M	0.098173061	0.053233120	0.19691052
MH_M_16	M	0.020031097	0.086173847	0.17996666
MH_M_17	M	-0.072227423	0.085872569	0.19494951

MH_M_18 M -0.034242392 0.036991854 0.17029875

MH_M_19 M 0.009148088 -0.065000253 0.14985919

MH_M_20 M -0.026042530 0.014424359 0.18834594

MH_M_21 M -0.045859100 0.082999264 0.24913919

Kebara 2 Ne 0.024278265 0.100772470 0.19831907

Tabun C1 Ne -0.134585543 0.142411392 0.05947584

Légende

1- Patient
2- Sexe
3- Tropubien
4- Traverse maximum
5- Oblique droit

Les premiers résultats :

Mean :116.16 Mean :127.6 Mean :141.2

3rd Qu.:122.21 3rd Qu.:134.7 3rd Qu.:147.4

Max. :137.42 Max. :142.2 Max. :156.3

Oblique.gauche Promonto.sous.pubien Mi.sacro.pubien Bisciatique

Min. :119.9 Min. :111.4 Min. : 97.6 Min. : 77.34

1st Qu.:133.6 1st Qu.:120.1 1st Qu.:117.5 1st Qu.: 93.47

Median :138.5 Median :127.4 Median :120.9 Median :100.63

Mean :139.0 Mean :127.4 Mean :124.0 Mean :100.88

3rd Qu.:145.7 3rd Qu.:133.5 3rd Qu.:132.0 3rd Qu.:109.29

Résultat de l'ACP :

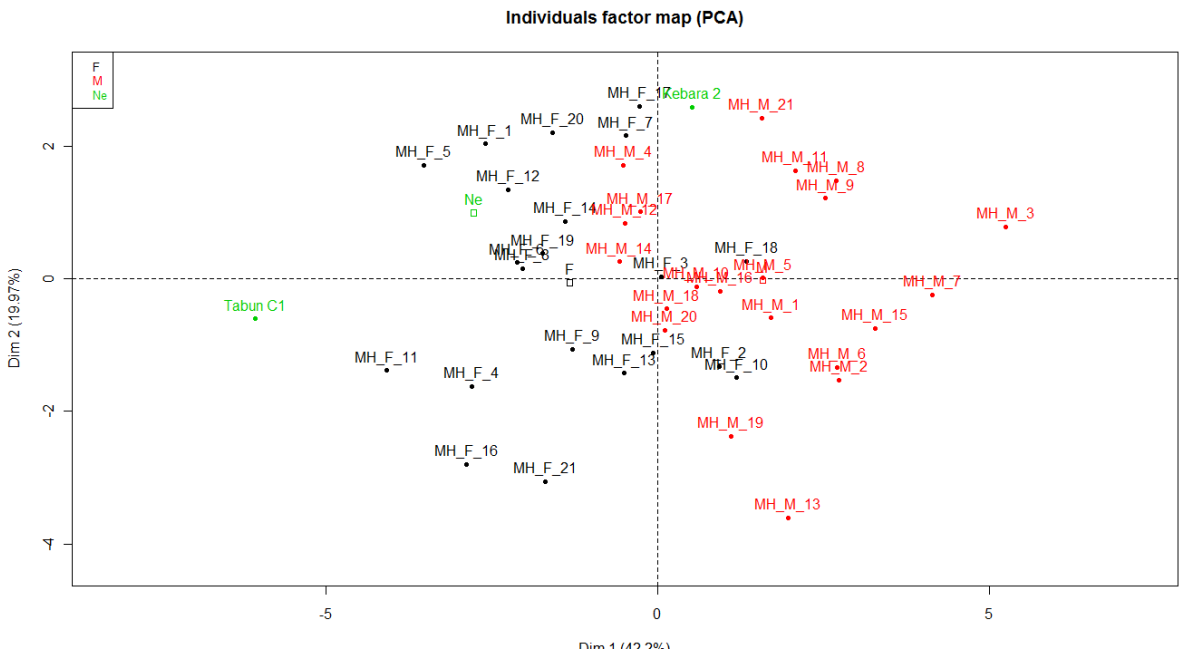
Une fois le changement d'échelle établi:

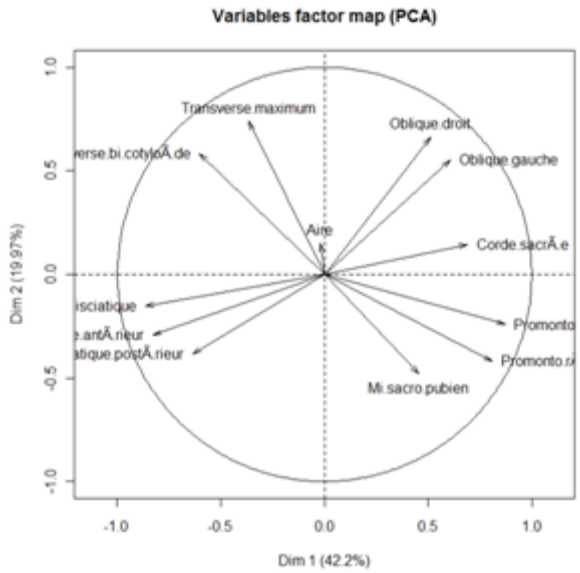
```
> q<-cbind(Tab[1:44,2],Tab[1:44,3])

> q;

            [,1]        [,2]
 [1,] -0.156395355  0.144000543
 [2,]  0.062524028  0.046378977
 [3,] -0.024698663  0.098608071
 [4,] -0.075978309  0.013907743
 [5,] -0.147110423  0.115823067
 [6,] -0.106220755  0.084074731
 [7,] -0.048586685  0.163371461
 [8,] -0.136544055  0.065498826
 [9,] -0.039704520  0.041160641
[10,]  0.034163816  0.018117745
[11,] -0.089591853  0.089362619
[12,] -0.104852267  0.117789754
[13,] -0.062254701  0.021885551
[14,] -0.046692032  0.123490857
[15,] -0.024078235  0.048042279
```

```
[16,] -0.010088548  0.052806586
[17,] -0.094876515  0.149608437
[18,]  0.020368326  0.063037324
[19,] -0.075006158  0.101451124
[20,] -0.103670949  0.118700404
[21,]  0.032924380  0.034249870
[22,]  0.044170777  0.055509451
[23,]  0.081241557 -0.020627454
[24,]  0.106451745  0.095336083
[25,] -0.057447382  0.043885297
[26,]  0.005490778  0.026196102
[27,]  0.003340791  0.003671455
[28,]  0.104566817  0.019770934
[29,] -0.008601073  0.110762155
[30,]  0.013427275  0.085478049
[31,]  0.026220608  0.101227265
[32,] -0.031235602  0.087978336
[33,] -0.060948960  0.134390785
[34,]  0.095590454 -0.020779701
[35,] -0.033914832  0.109492636
[36,]  0.098173061  0.053233120
[37,]  0.020031097  0.086173847
[38,] -0.072227423  0.085872569
[39,] -0.034242392  0.036991854
[40,]  0.009148088 -0.065000253
[41,] -0.026042530  0.014424359
[42,] -0.045859100  0.082999264
[43,]  0.024278265  0.100772470
[44,] -0.134585543  0.142411392
>
```


Individuals factor map (PCA)
F
M
Ne
Kebara 2
Ne
Tabun C1
Dim 2 (19.97%)
Dim 1 (42.2%)


Variables factor map (PCA)
Transverse.maximum
Oblique.droit
Oblique.gauche
Corde.sacrÃ.e
Aire
Mi.sacro.pubien
Dim 2 (19.97%)
Dim 1 (42.2%)

Lexique

A

Analyse factorielle des correspondances, 50

C

Caractere, 4, 5
Cercle des corrélations, 29
Coefficient de correlation, 9
Contribution de l'individu, 31
Covariance, 9

D

Deuxieme axe principal d'inertie, 24

E

Écart type, 6

F

Facteurs principaux, 22, 23
Frequence conditionnelle, 8
Frequence relative, 8

I

Inegalite de cauchy-schwarz, 11
Inertie, 5, 6

M

Matrice d'inertie, 18
Matrice de correlation, 10
Matrice de covariance, 10
Mediane, 6
Modalite, 4
Multiplicateur de lagrange, 22

N

Notion de m-angle, 11
Nuage de points, 5, 6

P

Part d'inertie expliquee, 23
Part d'inertie expliquee par le deuxieme axe, 24
Plan principal, 25
Potentiels de prevision, 45
Premier axe principale d'inertie, 23
Premier facteur principale d'inertie, 23
Premiere composante principale d'inertie, 23
Profil-colonne, 51
Profil-ligne, 51

S

Série statistique, 5, 6
Symbole de kronëcker, 18

T

Table de contingence, 8, 15, 50
Table de contingence, 13
Tableau, 4
Tableaux de marge, 8
Taux d'inertie expliquee, 31
Theoreme de huyghens, 5
Theoreme de huyghens-koenig, 17

V

Variance, 6

,

- ,

Table des matières

Préface 4
Chapitre 1 Statistique descriptive 5
Population 5
Caractère 5
Modalité 5
Caractère quantitatif, qualitatif 5
Caractère statistique 5
Chapitre 2 Statistique descriptive Caractéristiques de position et de dispersion 6
Série statistique 6
Moyenne arithmétique 6
Nuage de points alignés, inertie d'un nuage de points 6
Variance, écart type 7
Médiane 7
Une approche matricielle de la notion de moyenne arithmétique 7
Chapitre 3 Statistique descriptive Description bidimensionnelle 9
Table de contingence 9
Fréquences 9
Fréquence conditionnelle-indépendance 9
Centre de gravité-covariance- coefficient de corrélation 10
Approche géométrique du coefficient de corrélation 11
Chapitre 4 Statistique descriptive multidimensionnelle ; analyse en composantes principales 16
Notations en A.C.P 16
Analyse en composantes principales 16
Notations 17
Schéma de dualité 18
Inertie et moments d'inertie 18
Moment d'inertie par rapport à un sous-espace 20
Moment d'inertie par rapport à un sous-espace affine $\boldsymbol{W1}$ de $\boldsymbol{W}$ 21
Conséquence du théorème de Pythagore 21
But de l' A.C.P 22
Moment d'inertie par rapport à $\Delta\boldsymbol{u}$ et $\Delta\boldsymbol{u}\perp$ 22
Propriété des vecteurs propres et des valeurs propres de VM 23

Résolution du problème d'optimisation, axes et facteurs principaux- Composantes principales 23
Axes et facteurs principaux- composantes principales 24
Dualité 26
Axes principaux d'inertie 26
Facteurs principaux d'inertie 26
Composantes principales 27
Autre présentation de la notion de composante principale 27
Propriété fondamentale d'une composante principale 27
Choix de la métrique M 28
Description du nuage des individus 29
Qualité de la représentation d'un individu sur un axe 30
Corrélation entre composantes et variables initiales 30
Importance d'un individu dans la détermination d'un axe 32
Formule de reconstitution du tableau X 33
Qualité de service du réseau téléphonique 42
Tableau de données de base(rappel) 43
Valeurs propres de la matrice des corrélations 44
Matrice des corrélations des variables 44
Chapitre 5 Analyse canonique 45
Recherche des variables canoniques 46
Facteurs canoniques 50
Description de l'ensemble des individus 50
Chapitre 6 Analyse factorielle des correspondances 51
A.F.C comme cas particulier d'A.C.P 51
Distance du khi-deux entre deux lois de probabilité 51
Notion de profil 51
Inertie au centre de gravité des deux nuages 52
Principe d'équivalence distributionnelle 53
Analyse en composantes principales des deux nuages de points 53
A.C.P non centrée sur N(I) et facteur trivial 53
A.C.P sur N(I) 54
A.C.P sur N(I) 54
Deuxième présentation de l'.A.F.C 57
Mise sous forme disjonctive complète d'une variable qualitative 57

A.F.C comme cas particulier d'analyse canonique 58
Mise en élément supplémentaire 59
Chapitre 7 Régressions linéaire multiple 64
Recherche d'un ajustement linéaire 65
Exemples 68
La matrice tab sur laquelle on travaille : 78
Lexique 84
Bibliographies 88

Bibliographies

Coll, B. J. (1973). *L'analyse des données. Tome 2: l'analyse des correspondances.* France: Dunod.

Cornillon P.A., G. A. (2008). *Statistique avec R.* France: Presses universitaires de Rennes.

Diday E., L. J. (1982). *Eléments d'analyse des données.* France: Dunod.

Escofier B., P. J. (1990). *Analyses factorielles simples et multiples. Objectifs, méthodes et interpretation.* France: Dunod.

F, B. J. (1980). *Pratique de l'analyse des données: Tome 1 : Analyse des correspondances. Exposé élémentaire.* France: Dunod.

F., C. (1984). *Analyse des données.* Canada: Les presses de l'université de Montréal.

G., C. (1965). *Cours de statistique descriptive.* France: Dunod.

G., G. (2003). *Analyse des données.* France: Lavoisier.

G., S. (1990). *Probabilités. Analyse des données de statistique.* France: Technip.

J.M., R. (1973). *Méthodes et programmes d'analyse discrimante.* France: Dunod.

Lebart L., M. A. (1979). *Traitement des données statistiques.* France: Dund.

M., J. (1989). *Exploration informatique et statistiques des données.* France: Dunod informatique.

M., V. (1985). *Analyse des données.* USA: Economica.

Pontier J., D. A. (1990). *Le modèle euclidien en analyse des données.* Belgique: Ellipses.

Sokal R.R., S. P. (1973, 1980). *Principles of numerical taxomy.* USA: Freeman.

Team, R. d. (2008). *A language and environment for statistical computing.* Autriche: Vienne R fondation for statistical computing.