UNIVERSITÉ PIERRE MENDÈS FRANCE

Département STID

L'Analyse en Composantes Principales (ACP)

Enseignant:
Date de rédaction : 25 avril 2005

Analyse factorielle en Composantes Principales : L'Analyse en Composantes Principales est une méthode graphique de statistique descriptive permettant la représentation de **données quantitatives multidimensionnelles**. Cette méthode géométrique a pour objet la description des données contenues dans un tableau individus-caractères numériques : p caractères sont mesurés sur n individus.

1 Rappels

Tableau des données : Les données consistent en p mesures, correspondant à des variables quantitatives $\{X^1, X^2, \ldots, X^p\}$, prises sur n unités $\{E_1, E_2, \ldots, E_n\}$. Le tableau de données, noté $\mathcal{X}$, est de la forme :

$$\mathcal{X} = \begin{array}{c|ccccccc} & X^1 & X^2 & & X^j & & X^p \\ \hline E_1 & x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ E_2 & x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ & & & & & & \\ E_i & x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ & & & & & & \\ E_n & x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{array}$$

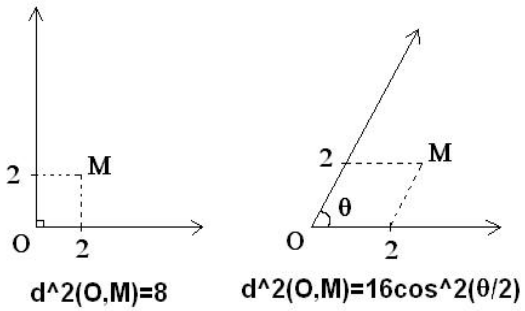
On peut représenter chaque unité E_i par le vecteur dans $\mathbb{R}^p$ de ses mesures sur les p variables : $\boldsymbol{e}_i^T = (x_{i1}, x_{i2}, ..., x_{ip})$. De façon analogue, on peut représenter chaque variable X^j par un vecteur de $\mathbb{R}^n$ dont les composantes sont les valeurs de la variable pour les n unités : $\boldsymbol{x}^{j^T} = (x_{1j}, x_{2j}, ..., x_{nj})$.

Notations

La variable X^j peut se noter de différentes façon équivalentes, suivant la manière dont on la regarde (géométrique, statistique, vectorielle, affine, ...) :

$$X^j = \overrightarrow{x^j} = \boldsymbol{x}^j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}.$$

Nuage de points : Ensemble de points représentés au moyen de leurs coordonnées dans un espace muni d'un système d'axes choisi à l'avance. Choisir un système d'axes pour un espace est équivalent à définir une distance dans cet espace.



Nuage des individus, des variables : L'ensemble des points dans $\mathbb{R}^p$ qui représentent les unités (individus) est appelé traditionnellement *nuage des individus*. En faisant de même dans $\mathbb{R}^n$, chaque variable pourra être représentée par un point de l'espace affine correspondant. L'ensemble de ces points qui représentent les variables est appelé *nuage des variables*.

Remarque importante : L'individu E_i est **défini** (identifié) par le p-uplet $(x_{i1}, x_{i2}, \ldots, x_{ip})$. Cela signifie que deux individus ayant les mêmes valeurs pour toutes les variables X^1 à X^p sont considérés comme étant le même individu (même s'ils sont différents en réalité !). Cela signifie aussi que la ressemblance entre deux individus sera mesurée par la distance entre les points les représentant dans le nuage des individus.
Cela nécessitera de définir ce qu'est une distance entre individus en Statistique.

A préciser par un exemple.

Notions sur les matrices

Une matrice est un tableau rectangulaire de chiffres qui possède deux dimensions : le nombre de ses lignes et le nombre de ses colonnes. On notera une matrice par une lettre majuscule.
Ainsi, la matrice $A = (a_{i,j})_{1 \leq i \leq n; 1 \leq j \leq p}$ est une matrice à n lignes et p colonnes. On écrit souvent $A : n \times p$ ou encore $\underset{n \times p}{A}$. Une matrice est dite diagonale si elle ne contient que des 0 hors de la diagonale.

Voici les opérations possibles sur les matrices :

Somme :

$$\underset{n \times p}{A} + \underset{n \times p}{B} = (a_{i,j}) + (b_{i,j}) = (a_{i,j} + b_{i,j}) = (b_{i,j} + a_{i,j}) = (b_{i,j}) + (a_{i,j}) = B + A \tag{1}$$

Différence :

$$A - B = (a_{i,j} - b_{i,j}) = -B + A \tag{2}$$

Multiplication par un scalaire :

$$kA = k(a_{i,j}) = (ka_{i,j}) = (a_{i,j}k) = Ak \tag{3}$$

Matrice identité :

$$I_n \underset{n \times p}{A} = A = \underset{n \times p}{A} I_p \tag{4}$$

où I_n (resp. I_p) est la matrice identité d'ordre n (resp. p), c'est-à-dire la matrice diagonale qui ne contient que des 1 sur la diagonale et des 0 partout ailleurs.

Produit :

$$\underset{n\times pp\times q}{A}\ \underset{}{B} = \underset{n\times q}{C} = (c_{i,j}) \quad \text{avec } c_{i,j} = \sum_{l=1}^{p} a_{il}b_{lj} \tag{5}$$

En général, on n'a pas $\underset{n\times pp\times n}{A}\ \underset{}{B} = \underset{p\times nn\times p}{B}\ \underset{}{A}$; c'est une grosse différence d'avec les nombres réels.

Transposée :
On note A^T la transposée de A.

$$A : n \times p \Rightarrow A^T : p \times n \tag{6}$$

$$A = (a_{i,j})_{1\leq i\leq n;1\leq j\leq p} \Rightarrow A^T = (a_{j,i})_{1\leq j\leq p;1\leq i\leq n} \tag{7}$$

Les lignes deviennent les colonnes et les colonnes deviennent les lignes.

$$(A^T)^T = A \tag{8}$$

$$(AB)^T = B^T A^T \tag{9}$$

Trace :

$$trace(\underset{n\times n}{A}) = \sum_{i=1}^{n} a_{ii} : 1 \times 1 \tag{10}$$

$$trace(\underset{n\times pp\times n}{A}\ \underset{}{B}) = trace(\underset{p\times nn\times p}{B}\ \underset{}{A}) \tag{11}$$

$$trace(k) = k \tag{12}$$

Inverse :
Une matrice n'est inversible que si elle est carrée, c'est-à-dire : $A : n \times n$.

$$A^{-1} = \frac{1}{\det(A)}(com(A))^T = \frac{1}{\det(A)}com(A^T) \tag{13}$$

$$(A^{-1})^{-1} = A \tag{14}$$

$$(AB)^{-1} = B^{-1}A^{-1} \tag{15}$$

$$(A^T)^{-1} = (A^{-1})^T \tag{16}$$

Valeurs propres et vecteurs propres :

$$A\boldsymbol{x} = \lambda\boldsymbol{x} \Rightarrow \lambda \text{ valeur propre de } A \text{ associée au vecteur propre } \boldsymbol{x}. \tag{17}$$

Les valeurs propres de A sont les racines du polynôme $P(\lambda) = det(A - \lambda I)$.

Exercice 1 *Soient les matrices suivantes :*

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 4 & 1 \end{bmatrix} B = \begin{bmatrix} 3 & 4 \\ 1 & 3 \\ 4 & 1 \end{bmatrix} C = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 3 & 2 \end{bmatrix} \text{ et } D = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{bmatrix} \tag{18}$$

Donner les dimensions des matrices A, B, C et D. Calculer $A + B$, $A - B$, $3A$, AC, CA, A^T, $trace(D)$, $trace(BC)$, $trace(CB)$, D^{-1}. Vérifier que $DD^{-1} = I_3 = D^{-1}D$. Calculer les valeurs propres et les vecteurs propres de D.

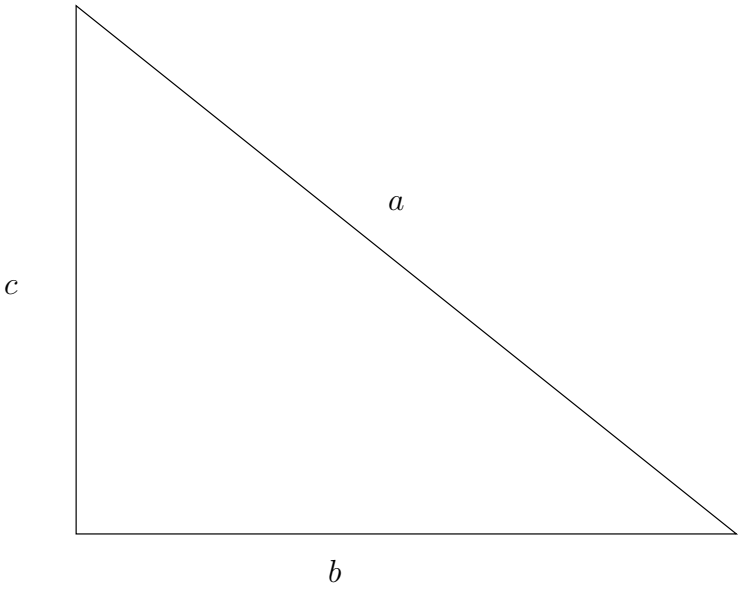
Soient les matrices suivantes :

$$E = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} \text{ et } F = \begin{bmatrix} 3 & 4 & 1 \end{bmatrix} \tag{19}$$

Calucler EF et FE.

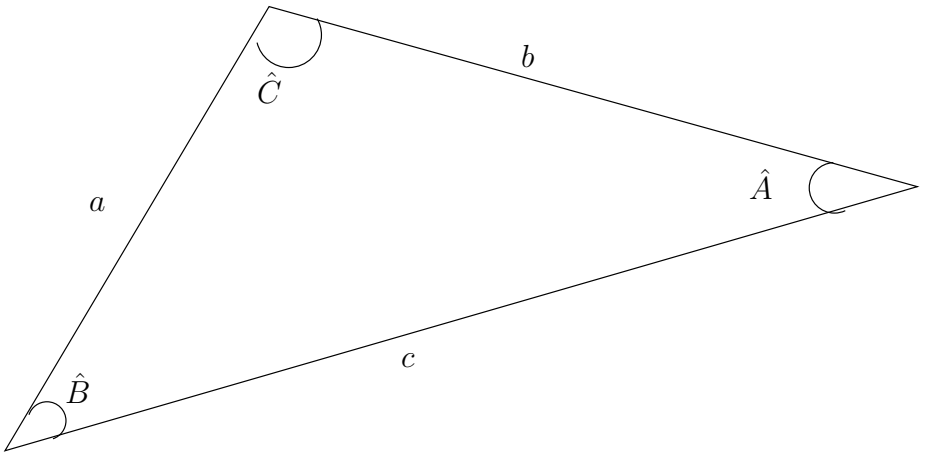
Relations dans les triangles

Théorème de Pythagore :



$$a^2 = b^2 + c^2 \tag{20}$$

Formule d'Al Kashi :



$$a^2 = b^2 + c^2 - 2bc\cos\hat{A} \tag{21}$$

Produit scalaire :

$$\langle \boldsymbol{x}^j, \boldsymbol{x}^k \rangle = \|\boldsymbol{x}^j\|.\|\boldsymbol{x}^k\|.\cos(\boldsymbol{x}^j, \boldsymbol{x}^k). \tag{22}$$

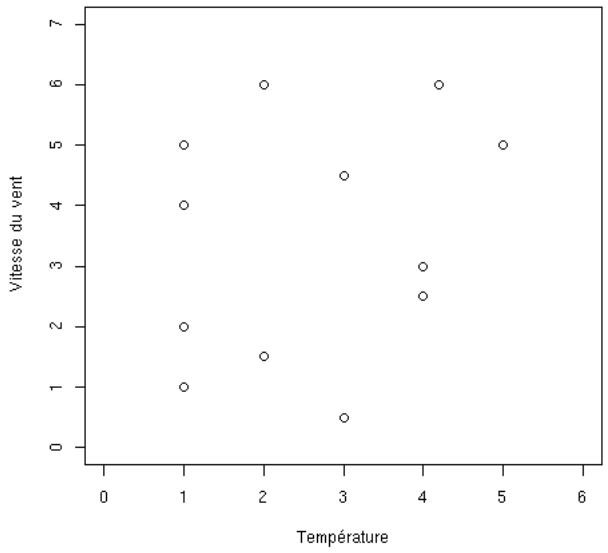
2 Présentation de la méthode

Lorsqu'il n'y a que deux caractères X^1 et X^2, il est facile de représenter, sur un graphique plan, l'ensemble des données : chaque individu E_i est alors un point de coordonnées (x_{i1}, x_{i2}) et le simple examen visuel de l'allure du nuage permet d'étudier l'intensité de la liaison entre X^1 et X^2 et de repérer les individus présentant des caractéristiques voisines.

X^1=température	X^2=Vitesse du vent
1	1
1	2
1	4
1	5
2	1.5
2	6
3	0.5
3	4.5
4	2.5
4	3
4.2	6
5	5

□

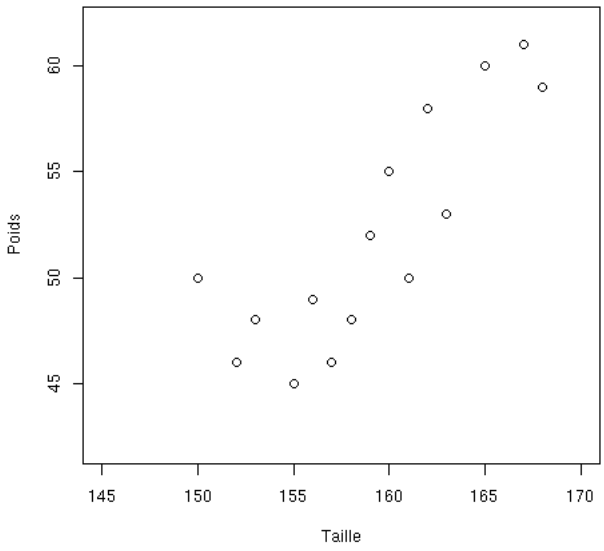
```
> x<-c(1,1,1,1,2,2,3,3,4,4,4.2,5)
> y<-c(1,2,4,5,1.5,6,0.5,4.5,2.5,3,6,5)
> plot(x,y,xlim=c(0,6),ylim=c(0,7),xlab="Température",ylab="Vitesse du vent")
```



X^1=taille	X^2=poids
150	50
152	46
153	48
155	45
156	49
157	46
158	48
159	52
160	55
161	50
162	58
163	53
165	60
167	61
168	59

□

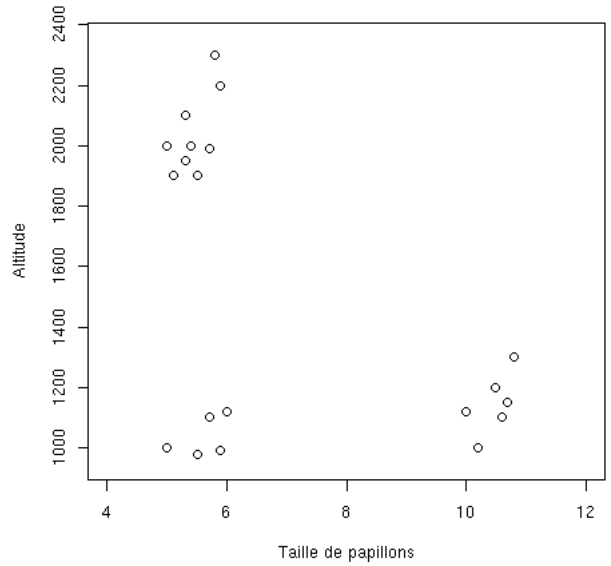
```
> x<-c(150,152,153,155,156,157,158,159,160,161,162,163,165,167,168)
> y<-c(50,46,48,45,49,46,48,52,55,50,58,53,60,61,59)
> plot(x,y,xlim=c(145,170),ylim=c(42,62),xlab="Taille",ylab="Poids")
```



X^1=taille de papillons	X^2=altitude
5	1000
5.5	980
5.7	1100
5.9	990
6	1120
10	1120
10.2	1000
10.5	1200
10.6	1100
10.7	1150
10.8	1300
5	2000
5.1	1900
5.3	1950
5.3	2100
5.4	2000
5.5.	1900
5.7	1990
5.8	2300
5.9	2200

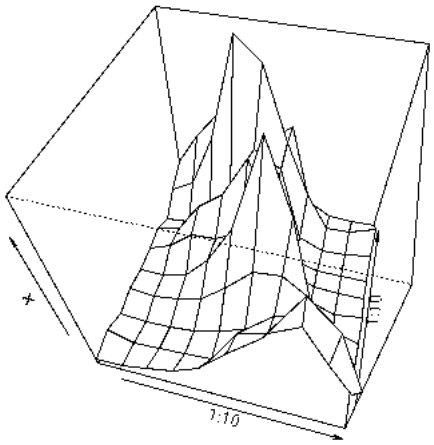
□

```
> x<-c(5,5.5,5.7,5.9,6,10,10.2,10.5,10.6,10.7,10.8,5,5.1,5.3,5.3,5.4,5.5,5.7,5.8,5.9)
> y<-c(1000,980,1100,990,1120,1120,1000,1200,1100,1150,1300,2000,1900,
> 1950,2100,2000,1900,1990,2300,2200)
> plot(x,y,xlim=c(4,12),ylim=c(950,2350),xlab="Taille de papillons",ylab="Altitude")
```



S'il y a trois caractères, l'étude visuelle est encore possible en "dessinant" le nuage de points à l'aide de vues en perspectives (plusieurs logiciels font ça très bien).

```
> x<-matrix(lynx[1:100],nrow=10,ncol=10)
> persp(1:10,1:10,x,theta=15,phi=50)
```



Par contre, dès que p=4, cette façon de procéder n'est plus possible.

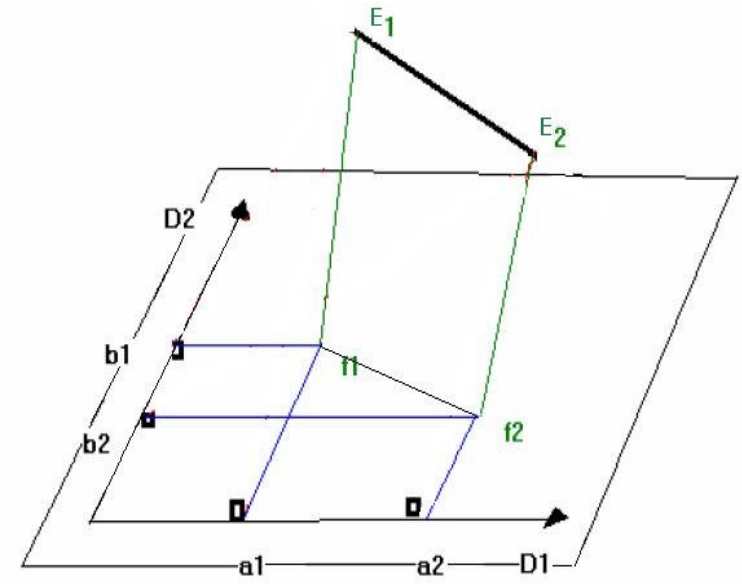
Ainsi, dans le tableau des dépenses de l'Etat français en page 36, chaque année représente un individu décrit par 11 caractères (variables). Les 24 individus forment un nuage (peu visible !) dans un espace à 11 dimensions, puisqu'il y a 11 coordonnées.
L'idée est alors de projeter le nuage de points au complet sur un sous-espace (en général de dimension 2, c'est-à-dire un plan) et de regarder le graphique plan ainsi obtenu. L'idée est que deux points proches dans le nuage d'origine devraient le rester sur le plan où ils sont projetés.

Que ce processus ne vous effraie pas ! C'est celui que vous utilisez lorsque vous réalisez des photographies ; vous passez bien d'un espace à 3 dimensions (celui où nous vivons) à un espace à 2 dimensions : votre photo.

Cependant, selon l'angle sous lequel vous « prenez » votre sujet, toutes vos photos n'apporteront pas la même information sur celui-ci ; il suffit de regarder la figure ci-dessous pour s'en convaincre !

FENELON J.P., 1981, Qu'est-ce que l'analyse des données ?, Lefonen, PARIS.

En effet, toute projection déforme (en général) le nuage de points. L'opération de projection raccourci toujours les distances.



On voit que la distance de f_1 à f_2 est plus courte que celle de E_1 à E_2. Il apparaît donc clair qu'il faut bien choisir ce plan pour que le nuage soit le moins déformé possible par la projection effectuée, c'est-à-dire pour que les distances soient le mieux conservées. Ainsi deux individus proches dans le nuage d'origine auront des chances de le rester sur le plan de projection et les conclusions que l'on pourra tirer sur les projetés auront des chances d'être valables pour les points d'origine.

L'ACP consiste donc à rechercher le meilleur sous-espace sur lequel projeter le nuage de points. Un sous-espace de dimension k est déterminé par la donnée de k axes orthogonaux : nous appelerons **axes principaux** les axes constituant le meilleur sous-espace.

Des critères sont aussi introduits qui permettent de juger de la qualité de la représentation obtenue après projection et dans quelle mesure on peut s'y fier.

Récapitulatif :
On place les points $E_1, \ldots, E_n$ dans $\mathbb{R}^p$. Plus les points sont proches dans $\mathbb{R}^p$, plus les individus se ressemblent. Comme on ne peut pas voir les points dans $\mathbb{R}^p$ (l'oeil humain ne voit que dans $\mathbb{R}^3$), on va projeter le nuage de points sur un plan et observer la répartition des projections des n points sur ce plan.
Si le plan est bien choisi et s'il a un bon pouvoir expliquant, alors plus les projections seront proches sur ce plan, plus les individus seront susceptibles de se ressembler.
L'ACP c'est ça : trouver le bon plan (appelé **plan principal**) sur lequel projeter les points et observer les projections pour en tirer des conclusions sur les individus.

Nous allons maintenant définir les outils mathématiques nécessaires pour mener à bien ce projet.

3 Les données et leurs caractéristiques

Dans cette section, nous utilisons le formalisme matriciel qui permet de simplifier beaucoup de calculs.

3.1 Exemple support

L'exemple qui suit sera utilisé par la suite pour illustrer plusieurs notions théoriques introduites. Les données sont stockées dans la matrice $\mathcal{X}$ ci-dessous. Trois variables ($\boldsymbol{x}^1$=la charge de travail en heures, $\boldsymbol{x}^2$=la distance au travail en km et $\boldsymbol{x}^3$=le salaire en milliers d'euros) sont mesurées sur $n = 5$ individus (Hansen, Jensen, Petersen, Pedersen et Nielsen).

$$\mathcal{X} = \begin{pmatrix} 13.1 & 8.5 & 4.9 \\ 12.1 & 10.5 & 7.9 \\ 14.1 & 11.5 & 10.9 \\ 10.1 & 9.5 & 5.9 \\ 12.1 & 8.5 & 11.9 \\ 13.1 & 11.5 & 9.9 \\ 13.1 & 12.5 & 3.9 \\ 11.1 & 9.5 & 7.9 \\ 9.1 & 7.5 & 10.9 \\ 12.1 & 10.5 & 5.9 \end{pmatrix}. \tag{23}$$

3.2 Le tableau des données

Les observations de p variables quantitatives sur n individus sont rassemblées en un tableau rectangulaire $\mathcal{X}$ à n lignes et p colonnes :

$$\mathcal{X}= \begin{array}{c|ccccccc} & 1 & 2 & & j & & p \\ \hline 1 & x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ 2 & x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ & & & & & & \\ i & x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ & & & & & & \\ n & x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{array}$$

x_{ij} est la valeur prise par la j-ème variable sur le i-ème individu.

Dans une optique purement descriptive on identifiera la variable $\boldsymbol{x}^j$ à la j-ème colonne de $\mathcal{X}$: une variable n'est rien d'autre que la liste des n valeurs qu'elle prend sur les n individus :

$$\boldsymbol{x}^j = \boldsymbol{x}_{.j} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \tag{24}$$

On identifiera de même l'individu i au vecteur $\boldsymbol{e}_i$ à p composantes :

$$\boldsymbol{e}_i = \boldsymbol{x}_{i.} = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}. \tag{25}$$

Application :
Dans l'exemple support, $n = 5$ et $p = 3$.

3.3 La matrice des poids des individus

En général, on accorde le même poids $\frac{1}{n}$ à tous les n individus. Il n'en est pas toujours ainsi et il est utile pour certaines applications de travailler avec des poids p_i éventuellement différents d'un individu à l'autre (échantillons redressés, données regroupées, villes, ...).

Ces poids, qui sont des nombres positifs de somme 1 comparables à des fréquences, sont regroupés dans une matrice diagonale notée D_{p_n} de taille n :

$$D_{p_n} = \begin{pmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_n \end{pmatrix} \tag{26}$$

Dans le cas le plus usuel des poids, on a bien sûr $D_{p_n} = \frac{1}{n} I_n$.

3.4 Point moyen ou centre de gravité

On définit le centre de gravité $\boldsymbol{g}$ du nuage des individus par

$$\boldsymbol{g} = (\overline{x}_{.1}, \ldots, \overline{x}_{.p})^T \tag{27}$$

où $\overline{x}_{.j} = \sum_{i=1}^{n} p_j x_{ij}$ est la moyenne (pondérée) du j-ème caractère (variable) numérique.

On a

$$\boldsymbol{g} = \mathcal{X}^T D_{p_n} \mathbf{1}_n \tag{28}$$

où $\mathbf{1}_n$ désigne le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1.

Exercice 2 *Calculer le centre de gravité $\boldsymbol{g}$ du nuage des 5 points individus de l'exemple support, pour le cas usuel des poids.*

3.5 Données centrées

En ACP, on travaillera toujours sur des données centrées (voire centrées-réduites). On note $\dot{\mathcal{X}}$ la matrice des données centrées. On a

$$\dot{\mathcal{X}} = \mathcal{X} - \mathbf{1}_n \boldsymbol{g}^T = (I_n - \mathbf{1}_n \mathbf{1}_n^T D_{p_n})\mathcal{X}. \tag{29}$$

Centrer les données revient à placer l'origine des axes du nuage des individus au centre de gravité $\boldsymbol{g}$.
Cette transformation n'a aucune incidence sur les définitions de la ressemblance entre individus et de la liaison entre variables. A ce niveau, elle peut être considérée comme un intermédiaire technique qui présente par la suite d'intéressantes propriétés mais qui ne change fondamentalement rien à la problématique.

Exercice 3 *Vérifiez la validité de cette formule.*
Calculer la matrice $\dot{\mathcal{X}}$ sur les données de l'exemple support, en utilisant les deux formules ci-dessus.
Constatez que cela revient à retrancher à chaque valeur la moyenne des valeurs de la colonne dans laquelle se trouve cette valeur.
Tracer le nuage des individus associé aux deux premières colonnes de la matrice $\mathcal{X}$, placer le centre de gravité G sur ce graphique. Utilisez pour cela un repère orthonormé.
Tracer sur un autre graphique, mais avec la même échelle que le graphique précédent, le nuage des individus associé aux deux premières colonnes de la matrice $\dot{\mathcal{X}}$.
Constatez qu'il s'agit bien d'une translation de repère vers le centre de gravité G.
Vous auriez constaté la même chose en procédant sur l'ensemble des variables.
Refaire la même chose avec les 3 variables.

3.6 Matrice de variance-covariance et matrice des corrélations

La matrice de variance-covariance S des données se calcule ainsi :

$$S = \dot{\mathcal{X}}^T D_{p_n} \dot{\mathcal{X}} = \mathcal{X}^T D_{p_n} \mathcal{X} - \boldsymbol{g}\boldsymbol{g}^T \tag{30}$$

On a également

$$\mathcal{X}^T D_{p_n} \mathcal{X} = \sum_{i=1}^{n} p_i \boldsymbol{e}_i \boldsymbol{e}_i^T \tag{31}$$

et donc

$$S = \sum_{i=1}^{n} p_i \boldsymbol{e}_i \boldsymbol{e}_i^T - \boldsymbol{g}\boldsymbol{g}^T \tag{32}$$

Cette dernière formule est utile pour les calculs numériques car elle ne suppose pas la mise en mémoire du tableau $\mathcal{X}$ mais seulement la lecture successive des données.
Remarquez que si l'on note $\boldsymbol{g}^{(n)}$ le point moyen basé sur n individus, alors l'ajout d'un individu supplémentaire $\boldsymbol{e}_{n+1}$ donne $\boldsymbol{g}^{(n+1)} = \frac{n}{n+1}\boldsymbol{g}^{(n)} + \frac{1}{n+1}\boldsymbol{e}_{n+1}$. Le point moyen peut donc aussi se calculer par lecture successive des données.

Exercice 4 *Calculer la matrice S sur les données de l'exemple support, en utilisant les formules ci-dessus.*

Notons $D_{1/s}$ la matrice diagonale $p \times p$ des inverses des écarts-types (ATTENTION!! : ne pas la confondre avec D_{p_n} qui est de taille $n \times n$) :

$$D_{1/s} = \begin{pmatrix} 1/s_1 & & & 0 \\ & 1/s_2 & & \\ & & \ddots & \\ 0 & & & 1/s_p \end{pmatrix} \tag{33}$$

et D_{1/s^2} la matrice diagonale des inverses des variances. Alors, le tableau des données centrées et réduites Z tel que

$$z_{ij} = \frac{x_{ij} - \overline{x}_{.j}}{s_j} \tag{34}$$

est donné par la formule

$$Z = \dot{\mathcal{X}} D_{1/s} \tag{35}$$

Exercice 5 *Calculer la matrice Z des données centrées-réduites de l'exemple support.*

La matrice regroupant tous les coefficients de corrélation linéaire entre les p variables prises deux à deux est notée R :

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ . & 1 & & . \\ . & & & . \\ . & & & . \\ r_{p1} & & & 1 \end{pmatrix} \tag{36}$$

On a

$$R = D_{1/s} S D_{1/s} = Z^T D_{p_n} Z. \tag{37}$$

R est la matrice de variance-covariance des données centrées et réduites et résume la structure des dépendances linéaires entre les p variables.

Exercice 6 *Calculer la matrice R des corrélations sur les données de l'exemple support en utilisant la formule ci-dessus faisant intervenir Z.*

Rappels :

$$var(\boldsymbol{x}^j) = \sum_{i=1}^{n} p_i (x_i^j - \overline{x}_{.j})^2 \tag{38}$$

$$cov(\boldsymbol{x}^j, \boldsymbol{x}^k) = \sum_{i=1}^{n} p_i (x_i^j - \overline{x}_{.j})(x_i^k - \overline{x}_{.k}) \tag{39}$$

$$r(\boldsymbol{x}^j, \boldsymbol{x}^k) = \frac{cov(\boldsymbol{x}^j, \boldsymbol{x}^k)}{\sqrt{var(\boldsymbol{x}^j)}\sqrt{var(\boldsymbol{x}^k)}} \tag{40}$$

4 Définition d'une distance en Statistique

Une métrique est une matrice M symétrique (c'est-à-dire $M^T = M$) définie positive (c'est-à-dire $\boldsymbol{x}^T M \boldsymbol{x} \geq 0, \quad \forall \boldsymbol{x}$) qui permet de définir une distance.

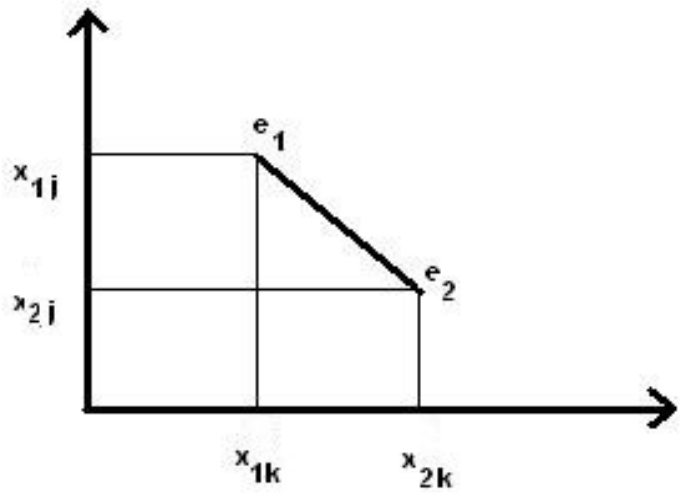
4.1 Métriques sur l'espace des individus $\mathbb{R}^p$

Deux individus $\boldsymbol{e}_1$ et $\boldsymbol{e}_2$ proches dans le nuage de points des individus auront des coordonnées voisines, c'est-à-dire des valeurs proches pour les différentes variables. On pourra alors conclure que ces individus se "ressemblent" (en tout cas en ce qui concerne les variables étudiées).

Mais comment mesurer la distance entre deux individus $\boldsymbol{e}_1$ et $\boldsymbol{e}_2$? Cette question primordiale doit être résolue avant toute étude statistique car les résultats obtenus en dépendent dans une large mesure.

En physique, la distance entre deux points de l'espace se calcule facilement par la formule de Pythagore : la carré de la distance est la somme des carrés des différences des coordonnées, car les dimensions sont de même nature : ce sont des longueurs que l'on mesure avec la même unité.

$$d^2(\boldsymbol{e}_1, \boldsymbol{e}_2) = (x_{1k} - x_{2k})^2 + (x_{1j} - x_{2j})^2. \tag{41}$$



Il n'en est pas de même en statistique où chaque dimension correspond à un caractère qui s'exprime avec son unité particulière : comment calculer la distance entre deux individus décrits par les trois caractères : âge, salaire, nombre d'enfants ?

La formule de Pythagore est alors aussi arbitraire qu'une autre. Si on veut donner des importances différentes à chaque caractère, pourquoi ne pas prendre une formule du type :

$$d^2(\boldsymbol{e}_1, \boldsymbol{e}_2) = a_1(x_{11} - x_{21})^2 + a_2(x_{12} - x_{22})^2 + \ldots + a_p(x_{1p} - x_{2p})^2. \tag{42}$$

ce qui revient à multiplier par $\sqrt{a_i}$ chaque caractère (on prendra bien sûr des a_i positifs).

On donne alors un poids plus important à certains caractères.

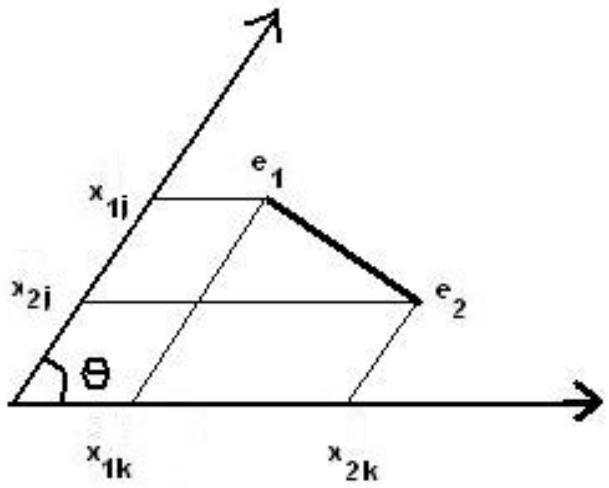
Exercice 7 *Par exemple, si on mesure le salaire et l'âge sur deux individus, on peut vouloir donner 2 fois plus d'importance à la variable âge qu'à la variable salaire.*

	age	*salaire*
$\boldsymbol{e}_1$	20	10
$\boldsymbol{e}_2$	24	15

(43)

Faire le graphique plan des individus avec les axes âge et salaire.
Calculer le carré de la distance entre $\boldsymbol{e}_1$ *et* $\boldsymbol{e}_2$ *:* $d^2(\boldsymbol{e}_1, \boldsymbol{e}_2) = ...$
Maintenant, on veut donner deux fois plus d'importance à la variable âge qu'à la variable salaire.
Faire le graphique de $\boldsymbol{e}_1$ *et* $\boldsymbol{e}_2$ *avec les axes* $\sqrt{4}\times$*age et salaire.*
Calculer alors le carré de la nouvelle distance entre $\boldsymbol{e}_1$ *et* $\boldsymbol{e}_2$ *:* $d^2(\boldsymbol{e}_1, \boldsymbol{e}_2) = ...$

De plus, la formule de Pythagore n'est valable que si les axes sont perpendiculaires, ce que l'on conçoit aisément dans l'espace physique. Mais en statistique ce n'est que par pure convention que l'on représente les caractères par des axes perpendiculaires : on aurait pu tout aussi bien prendre des axes obliques d'angle θ :



La formule donnant la distance fait alors intervenir en plus des carrés des différences des coordonnées les produits des differences (voir les rappels sur les triangles) :

$$d^2(\boldsymbol{e}_1, \boldsymbol{e}_2) = (x_{1k} - x_{2k})^2 + (x_{1j} - x_{2j})^2 - 2(x_{1k} - x_{2k})(x_{1j} - x_{2j})\cos\theta. \tag{44}$$

Sous sa forme la plus générale la distance d entre deux individus peut s'écrire :

$$d^2(\boldsymbol{e}_1, \boldsymbol{e}_2) = \sum_{k=1}^{p} \sum_{j=1}^{p} m_{kj}(x_{1k} - x_{2k})(x_{1j} - x_{2j}) \tag{45}$$

soit en notant M la matrice d'éléments m_{kj} :

$$d^2(\boldsymbol{e}_1, \boldsymbol{e}_2) = (\boldsymbol{e}_1 - \boldsymbol{e}_2)^T M(\boldsymbol{e}_1 - \boldsymbol{e}_2) \tag{46}$$

M peut être n'importe quelle matrice symétrique définie positive de taille $p \times p$. La formule de Pythagore revient à choisir pour M la matrice identité I.

Exercice 8 *Vérifier que les formules (45) et (46) coïncident.*
Calculer le carré de la distance entre les individus 3 et 5 lorsque $M = I$.

Produit scalaire
Ceci revient à définir le produit scalaire de deux vecteurs $\boldsymbol{e}_1$ et $\boldsymbol{e}_2$ de l'espace des individus par :

$$< \boldsymbol{e}_1; \boldsymbol{e}_2 >_M = \boldsymbol{e}_1^T M \boldsymbol{e}_2. \tag{47}$$

On dit que l'on a muni l'espace des individus d'une structure euclidienne, la matrice M s'appelle alors la métrique de l'espace. Le produit scalaire de $\boldsymbol{e}_1$ par lui-même est noté $\|\boldsymbol{e}_1\|_M^2$ et $\|\boldsymbol{e}_1\|_M$, qui est l'analogue de la longueur du vecteur $\boldsymbol{e}_1$, s'appelle la M-norme de $\boldsymbol{e}_1$.

Métriques courantes
Les métriques les plus utilisées en ACP sont les métriques diagonales qui reviennent à pondérer les caractères. Les deux métriques les plus utilisées sont $M = I_p$ et aussi la métrique :

$$M = D_{1/s^2} = \begin{pmatrix} 1/s_1^2 & & & 0 \\ & 1/s_2^2 & & \\ & & \ddots & \\ 0 & & & 1/s_p^2 \end{pmatrix} : p \times p \tag{48}$$

ce qui revient à diviser chaque caractère par son écart-type : entre autres avantages, la distance entre deux individus ne dépend plus des unités de mesure puisque les nombres $\boldsymbol{x}^j/s_j$ sont sans dimension.

Ainsi, si $\boldsymbol{x}^j$ représente l'âge d'un individu, on peut utiliser aussi bien comme unité le mois ou l'année car si $\boldsymbol{x}^j$ est multiplié par 12 (passage de l'âge en années à l'âge en mois), s_j est aussi multiplié par 12 et le rapport reste constant.

Exercice 9 *Calculer le carré de la distance entre les individus 3 et 5 lorsque $M = D_{1/s^2}$.*
Multiplier les éléments de la première colonne par 10 et calculer à nouveau le carré de la distance entre les individus 3 et 5, toujours lorsque $M = D_{1/s^2}$.

Si $M = I_p$ on parle d'ACP non réduite, si $M = D_{1/s^2}$ on parle d'**ACP réduite** (ou **normée**).

4.2 Métrique sur l'espace des variables $\mathbb{R}^n$

On a vu dans la section précédente comment définir une distance sur l'espace des individus au moyen d'une matrice symétrique définie positive.
On peut faire de même dans l'espace des variables, mais dans ce cas le choix de cette matrice (qui sera ici $n \times n$) est immédiat : D_{p_n}. Ce choix est justifié par les considérations pratiques suivantes.

Sur $\mathbb{R}^n$, on définit le produit scalaire entre deux variables par :

$$\langle \boldsymbol{x}^j, \boldsymbol{x}^k \rangle = {\boldsymbol{x}^j}^T D_{p_n} \boldsymbol{x}^k = \sum_{i=1}^{n} p_i x_{ij} x_{ik}. \tag{49}$$

On a alors les relations très utiles suivantes :

Produit scalaire	**Statistique Descriptive**	**Calcul matriciel**
$\left\langle \dot{\boldsymbol{x}}^j, \dot{\boldsymbol{x}}^k \right\rangle$	$cov(\boldsymbol{x}^j, \boldsymbol{x}^k)$	${\dot{\boldsymbol{x}}^j}^T D_{p_n} \dot{\boldsymbol{x}}^k$
$\left\langle \dot{\boldsymbol{x}}^j, \dot{\boldsymbol{x}}^j \right\rangle = \left\| \dot{\boldsymbol{x}}^j \right\|^2$	$var(\boldsymbol{x}^j)$	${\dot{\boldsymbol{x}}^j}^T D_{p_n} \dot{\boldsymbol{x}}^j$
$\cos(\dot{\boldsymbol{x}}^j, \dot{\boldsymbol{x}}^k)$	$r(\boldsymbol{x}^j, \boldsymbol{x}^k)$, la corrélation	

Notez encore ici l'intérêt de centrer les données.

On déduit de la dernière ligne du tableau précédent que l'orthogonalité entre deux variables est équivalente à la non corrélation entre ces variables.

Exercice 10 *Vérifier ces formules.*

5 Inertie

L'inertie d'un nuage de points est une quantité qui caractérise la forme de ce nuage. Cet outil va nous aider à determiner le plan principal.

5.1 Définition et interprétation

On appelle inertie totale du nuage des n points $\{E_1, \ldots, E_n\}$ la moyenne des carrés des distances de ces n points au centre de gravité :

$$\mathcal{I}_{\boldsymbol{g}} = \sum_{i=1}^{n} p_i d^2(E_i, G) = \sum_{i=1}^{n} p_i \|\boldsymbol{e}_i - \boldsymbol{g}\|_M^2 = \sum_{i=1}^{n} p_i (\boldsymbol{e}_i - \boldsymbol{g})^T M (\boldsymbol{e}_i - \boldsymbol{g}). \tag{50}$$

Cette quantité caractéristique du nuage mesure d'une certaine manière l'éloignement des points par rapport à leur centre de gravité, c'est-à-dire la dispersion (l'étalement) globale du nuage. Une inertie nulle ou voisine de zéro signifie que tous les individus sont identiques ou presque et sont confondus avec leur centre de gravité $\boldsymbol{g}$.

Exercice 11 *Calculer l'inertie pour les données (individus) de l'exemple support et pour le cas usuel des poids. Faites-le d'abord pour $M = I$ puis pour $M = D_{1/s^2}$.*

5.2 Relation entre l'inertie et les carrés des distances au plan principal

On définit aussi l'inertie par rapport à un point H (ou $\boldsymbol{h}$ en notation vectorielle) différent du centre de gravité :

$$\mathcal{I}_{\boldsymbol{h}} = \sum_{i=1}^{n} p_i d^2(\boldsymbol{e}_i, \boldsymbol{h}) \tag{51}$$

$\mathcal{I}_{\boldsymbol{h}}$ est reliée à $\mathcal{I}_{\boldsymbol{g}}$ par la formule de Huygens :

$$\mathcal{I}_{\boldsymbol{h}} = \mathcal{I}_{\boldsymbol{g}} + d^2(\boldsymbol{g}, \boldsymbol{h}). \tag{52}$$

$\mathcal{I}_{\boldsymbol{h}}$ est donc toujours supérieure à $\mathcal{I}_{\boldsymbol{g}}$, la valeur minimum étant atteinte lorsque $\boldsymbol{h} = \boldsymbol{g}$.

On en déduit alors que la recherche d'un plan rendant maximum l'inertie des projections des n points est équivalente à la recherche du plan passant "au plus près" de l'ensemble des points du nuage au sens où la moyenne des carrés des distances des points du nuage au plan est minimale.

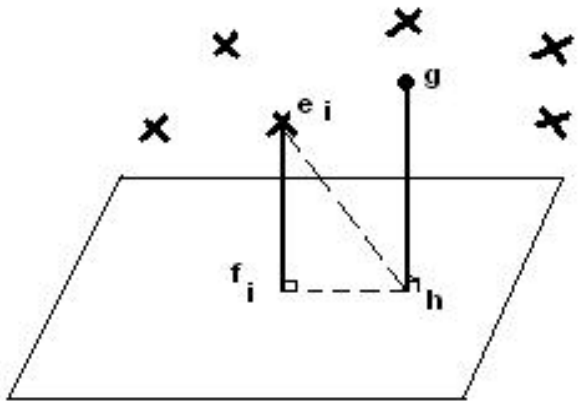
En effet :

Soit $\boldsymbol{h}$ la projection de $\boldsymbol{g}$ sur le plan qui est alors le centre de gravité de projection des points du nuage. Le triangle $\boldsymbol{e}_i; \boldsymbol{f}_i; \boldsymbol{h}$ est rectangle en $\boldsymbol{f}_i$, d'où :

$$d^2(\boldsymbol{e}_i, \boldsymbol{f}_i) = d^2(\boldsymbol{e}_i, \boldsymbol{h}) - d^2(\boldsymbol{f}_i, \boldsymbol{h}) \tag{53}$$

et

$$\sum_{i=1}^{n} p_i d^2(\boldsymbol{e}_i, \boldsymbol{f}_i) = \mathcal{I}_{\boldsymbol{h}} - \sum_{i=1}^{n} p_i d^2(\boldsymbol{f}_i, \boldsymbol{h}) \tag{54}$$



Comme $\mathcal{I}_{\boldsymbol{h}} = \mathcal{I}_{\boldsymbol{g}} + d^2(\boldsymbol{g}, \boldsymbol{h})$, on a

$$\sum_{i=1}^{n} p_i d^2(\boldsymbol{e}_i, \boldsymbol{f}_i) = \mathcal{I}_{\boldsymbol{g}} + d^2(\boldsymbol{g}, \boldsymbol{h}) - \sum_{i=1}^{n} p_i d^2(\boldsymbol{f}_i, \boldsymbol{h}) \tag{55}$$

On voit donc que rendre minimale la moyenne des carrés des distances entre les $\boldsymbol{e}_i$ et les $\boldsymbol{f}_i$ revient à minimiser le terme de droite dans l'équation ci-dessus. Puisque $\mathcal{I}_{\boldsymbol{g}}$ est constant, cela revient à minimiser $d^2(\boldsymbol{g}, \boldsymbol{h}) - \sum_{i=1}^{n} p_i d^2(\boldsymbol{f}_i, \boldsymbol{h})$, qui revient à son tour à minimiser $d^2(\boldsymbol{g}, \boldsymbol{h})$ et maximiser $\sum_{i=1}^{n} p_i d^2(\boldsymbol{f}_i, \boldsymbol{h})$ (lorsqu'on translate le plan contenant $\boldsymbol{h}$ parallèlement à $(\boldsymbol{h}, \boldsymbol{g})$, les distances des $\boldsymbol{f}_i$ à $\boldsymbol{h}$ ne sont pas perturbées).

Cela revient donc à prendre $\boldsymbol{g} = \boldsymbol{h}$ et à maximiser $\sum_{i=1}^{n} p_i d^2(\boldsymbol{f}_i, \boldsymbol{h})$ (qui est l'inertie du nuage projeté).

Réciproquement, on a aussi

$$\sum_{i=1}^{n} p_i d^2(\boldsymbol{f}_i, \boldsymbol{h}) = \mathcal{I}_{\boldsymbol{g}} + d^2(\boldsymbol{g}, \boldsymbol{h}) - \sum_{i=1}^{n} p_i d^2(\boldsymbol{e}_i, \boldsymbol{f}_i). \tag{56}$$

Donc maximiser l'inertie du nuage projeté revient à maximiser $d^2(\boldsymbol{g}, \boldsymbol{h}) - \sum_{i=1}^{n} p_i d^2(\boldsymbol{e}_i, \boldsymbol{f}_i)$. Dans ce cas, en imposant de façon arbitraire le choix $\boldsymbol{h} = \boldsymbol{g}$, on est alors rammené à minimiser $\sum_{i=1}^{n} p_i d^2(\boldsymbol{e}_i, \boldsymbol{f}_i)$.

Remarque :
L'approche par la méthode des moindres carrés conduit à définir de façon unique les axes principaux qui passent alors par G.

Désormais on supposera toujours que le plan principal, et plus généralement les axes principaux, passent par $\boldsymbol{g}$.

5.3 Intérêt de l'inertie

On se rappelera à ce stade ce que l'on cherche à faire. On cherche le plan (passant par $\boldsymbol{g}$) tel que le nuage projeté sur ce plan soit le moins déformé possible par rapport au nuage d'origine. On vient de voir que l'inertie est un nombre qui caractérise la forme d'un nuage. On peut calculer l'inertie du nuage d'origine et l'inertie du nuage projeté sur un plan que l'on se donne (on peut calculer l'inertie du nuage projeté sur différents plan ; on obtient des inertie différentes à chaque fois mais toujours plus petites que celle du nuage d'origine). Plus l'inertie du nuage projeté sera proche de l'inertie du nuage d'origine, moins la déformation engendrée par la projection sera importante. Bref, on cherche le plan pour lequel l'inertie du nuage projeté soit la plus proche de l'inertie du nuage d'origine.
Or on a vu que l'inertie ne pouvait que diminuer par projection, donc il suffit de chercher le plan pour lequel l'inertie du nuage projeté sera maximale (et donc au plus égale à celle de l'inertie du nuage d'origine).

Application
Visionner la vidéo sur l'ACP en 3D.

5.4 Inertie lorsque les variables sont centrées

Il est alors intéressant d'effectuer un changement de repère afin de placer l'origine du nouveau repère au centre de gravité $\boldsymbol{g}$ du nuage. On peut montrer que cela revient à centrer les données.

On travaillera donc toujours avec la matrice $\dot{\mathcal{X}}$ des données centrées.

Notez que centrer les variables ne modifie pas l'inertie du nuage puisqu'il s'agit simplement d'une translation du nuage (ou du repère).

Par la suite, on notera i (ou $\dot{\boldsymbol{x}}_{i.}$) le i-ème individu des données centrées. Il sera identifié à la i-ème ligne de la matrice $\dot{\mathcal{X}}$: $\dot{\boldsymbol{x}}_{i.} = (\dot{x}_{i1}, \ldots, \dot{x}_{ip})^T$.

On notera alors $\mathcal{I}$ l'inertie du nuage des données centrées :

$$\mathcal{I} = \sum_{i=1}^{n} p_i d^2(i, O) = \sum_{i=1}^{n} p_i d^2(\dot{\boldsymbol{x}}_{i.}, \mathbf{0}) \tag{57}$$

où O est l'origine du nouveau repère (qui correspond à G dans l'ancien repère) et $\mathbf{0} = (0, \ldots, 0)^T$ est sa notation vectorielle.

On peut montrer que, lorsque $\boldsymbol{g} = \mathbf{0}$ (données centrées), $\mathcal{I}_{\boldsymbol{g}}$ $(= \mathcal{I})$ est égale à la moyenne des carrés de toutes les distances entre les n points du nuage. On peut alors interpréter le plan principal du nuage de points comme étant le plan qui rend maximum l'inertie de l'ensemble des n points projetés sur lui.

Démonstration :

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \|\boldsymbol{e}_i - \boldsymbol{e}_j\|_M^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (\boldsymbol{e}_i - \boldsymbol{e}_j)^T M (\boldsymbol{e}_i - \boldsymbol{e}_j) && (58)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \left[\boldsymbol{e}_i^T M \boldsymbol{e}_i + \boldsymbol{e}_j^T M \boldsymbol{e}_j - 2\boldsymbol{e}_i^T M \boldsymbol{e}_j\right] && (59)\\
&= 2\underbrace{\sum_{i=1}^{n} p_i \boldsymbol{e}_i^T M \boldsymbol{e}_i}_{I_{\boldsymbol{g}}} - 2\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \underbrace{\boldsymbol{e}_i^T M \boldsymbol{e}_j}_{\langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle_M} && (60)\\
&= 2I_{\boldsymbol{g}} - 2\langle \underbrace{\sum_{i=1}^{n} p_i \boldsymbol{e}_i}_{0}, \underbrace{\sum_{j=1}^{n} p_j \boldsymbol{e}_j}_{0} \rangle_M && (61)\\
&= 2I_{\boldsymbol{g}}. && (62)
\end{aligned}$$

5.5 Quelques formules pour calculer l'inertie

Cas des données ni centrées, ni réduites ($M = I$) :

$$\mathcal{I}_{\boldsymbol{g}} = \sum_{i=1}^{n} p_i d^2(i, G) = \sum_{i=1}^{n} p_i \|\boldsymbol{e}_i - \boldsymbol{g}\|^2 = \sum_{i=1}^{n} p_i (\boldsymbol{e}_i - \boldsymbol{g})^T (\boldsymbol{e}_i - \boldsymbol{g}) = \sum_{i=1}^{n} p_i (\boldsymbol{x}_{i.} - \boldsymbol{g})^T (\boldsymbol{x}_{i.} - \boldsymbol{g}) \tag{63}$$

Cas des données centrées et non réduites ($M = I$) :

$$\mathcal{I}_{\mathbf{0}} = \mathcal{I} = \sum_{i=1}^{n} p_i d^2(i, O) = \sum_{i=1}^{n} p_i d^2(\dot{\boldsymbol{x}}_{i.}, \mathbf{0}) = \sum_{i=1}^{n} p_i \dot{\boldsymbol{x}}_{i.}^T \dot{\boldsymbol{x}}_{i.} \tag{64}$$

et

$$\mathcal{I} = trace(S) = \sum_j Var(\boldsymbol{x}^j) \tag{65}$$

puisque

$$\sum_{i=1}^{n} p_i \dot{\boldsymbol{x}}_{i.}^T \dot{\boldsymbol{x}}_{i.} = trace(\sum_i p_i \dot{\boldsymbol{x}}_{i.}^T \dot{\boldsymbol{x}}_{i.}) = trace(\sum_{i=1}^{n} p_i \dot{\boldsymbol{x}}_{i.} \dot{\boldsymbol{x}}_{i.}^T) = trace(\dot{\mathcal{X}}^T D_{p_n} \dot{\mathcal{X}}) = trace(S) \tag{66}$$

Cas des données centrées et réduites ($M = D_{1/s^2}$) :

$$\mathcal{I} = p \tag{67}$$

puisque

$$\begin{aligned}
\mathcal{I} &= \sum_{i=1}^{n} p_i d^2(i, O) = \sum_{i=1}^{n} p_i d^2(\dot{\boldsymbol{x}}_{i.}, \mathbf{0}) = \sum_{i=1}^{n} p_i \dot{\boldsymbol{x}}_{i.}^T M \dot{\boldsymbol{x}}_{i.} = \sum_{i=1}^{n} p_i \dot{\boldsymbol{x}}_{i.}^T D_{1/s^2} \dot{\boldsymbol{x}}_{i.} && (68)\\
&= trace(\sum_{i=1}^{n} p_i \dot{\boldsymbol{x}}_{i.}^T D_{1/s^2} \dot{\boldsymbol{x}}_{i.}) = trace(\sum_{i=1}^{n} p_i \dot{\boldsymbol{x}}_{i.} \dot{\boldsymbol{x}}_{i.}^T D_{1/s^2}) = trace\left([\sum_{i=1}^{n} p_i \dot{\boldsymbol{x}}_{i.} \dot{\boldsymbol{x}}_{i.}^T] D_{1/s^2} \right) && (69)\\
&= trace\left(\dot{\mathcal{X}}^T D_{p_n} \dot{\mathcal{X}} D_{1/s^2} \right) = trace\left(\dot{\mathcal{X}}^T D_{p_n} \dot{\mathcal{X}} D_{1/s} D_{1/s} \right) && (70)\\
&= trace\left(D_{1/s} \dot{\mathcal{X}}^T D_{p_n} \dot{\mathcal{X}} D_{1/s} \right) = trace\left(Z^T D_{p_n} Z \right) = trace(R) = p. && (71)
\end{aligned}$$

Lorsque les données sont centrées (réduites ou non), notez que l'inertie lorsque $p = 1$ est égale à la variance des variables considérées. Lorsque $p > 1$, l'inertie est donc une sorte de mesure de dispersion globale, une variance totale en quelque sorte.

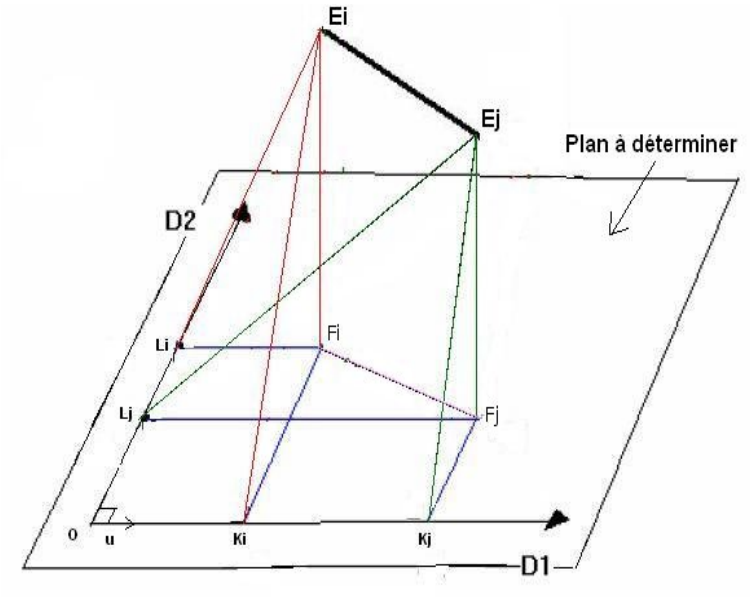
Exercice 12 *Calculer l'inertie des variables centrées pour les données de l'exemple support et pour le cas usuel des poids. Vérifier que vous trouvez la même valeur qu'à l'exercice 11.*

6 Recherche des axes principaux et des composantes principales.

On travaille avec les variables centrées (donc le centre de gravité G, ou en notation vectorielle $\boldsymbol{g}$, du nuage de points est confondu avec l'origine O du repère) et avec les poids p_i.

6.1 Recherche des axes principaux dans $\mathbb{R}^p$

On rappelle que l'on projette les points individus $\boldsymbol{e_i}$ (notés encore E_i) sur un plan comme le montre la figure ci-dessous.



Il faudra évidemment choisir le plan de projection sur lequel les distances seront en moyenne le mieux conservées. Comme l'opération de projection raccourcit toujours les distances : $d(F_i, F_j) \leq d(E_i, E_j)$, on se fixera pour critère de rendre maximale la moyenne des carrés des distances entre les projections $F_1, F_2, \ldots, F_n$.

Pour déterminer ce plan que l'on appelle le **plan principal**, il suffit de trouver deux droites D_1 et D_2. Si D_1 et D_2 sont perpendiculaires, on a :

$$d^2(F_i, F_j) = d^2(K_i, K_j) + d^2(L_i, L_j) \tag{72}$$

où les K_i et les L_i sont les projections des E_i (et aussi des F_i : projeter orthogonalement E_i sur F_i puis F_i sur K_i et L_i est équivalent à projeter orthogonalement directement E_i sur K_i et L_i) sur D_1 et D_2 respectivement.

La moyenne des carrés des distances entre les F_i est donc égale à la moyenne des carrés des distances entre les K_i plus la moyenne des carrés des distances entre les L_i :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j d^2(F_i, F_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j d^2(K_i, K_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j d^2(L_i, L_j).$$

Soit d'après (58),

$$\mathcal{I}(F) = \mathcal{I}(K) + \mathcal{I}(L). \tag{73}$$

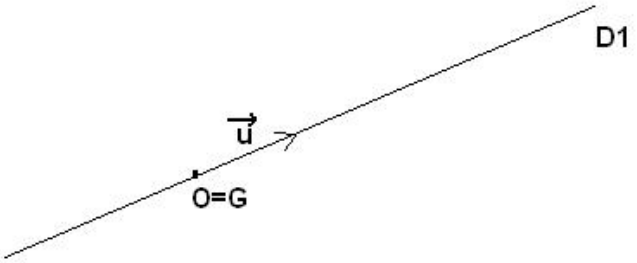
La méthode consiste alors à chercher tout d'abord D_1, rendant maximale $\mathcal{I}(K)$ puis D_2 perpendiculaire à D_1, rendant maximale $\mathcal{I}(L)$.

On peut continuer en dehors du plan et on trouvera alors $D_3, D_4, \ldots, D_p$ perpendiculaires entre elles : les D_i sont les **axes principaux** du nuage.

On recherche d'abord la droite D_1 de l'espace des individus $\mathbb{R}^p$, telle que le nuage projeté sur cette 1ère droite déforme le moins possible le nuage initial des points. On a vu que cela veut dire qu'il faut maximiser l'inertie des points projetés. La droite D_1 doit donc permettre de :

$$\text{maximiser } \sum_{i=1}^{n} p_i \left\| \overrightarrow{O\,K_i} \right\|^2,$$

inertie des points projetés sur D_1, où chaque point K_i est la projection orthogonale de E_i sur la droite. On **choisit** de faire passer la droite D_1 par le point moyen (centre du nuage appelé aussi centre de gravité ou barycentre). On définit D_1 par le point $O = G$ et par un vecteur directeur $\vec{u} = \boldsymbol{u} = (u_1, \ldots, u_p)^T$ avec $\|\boldsymbol{u}\| = 1$.



On a

$$\overline{OK}_i = \langle \vec{u}\,, \overrightarrow{OE_i} \rangle = \boldsymbol{u}^T \dot{\boldsymbol{x}}_{i.}$$

et donc

$$\sum_{i=1}^{n} p_i \|\vec{OK}_i\|^2 = (\dot{\mathcal{X}}\,\boldsymbol{u})^T\, D_{p_n} \dot{\mathcal{X}}\,\boldsymbol{u} = \boldsymbol{u}^T\, \dot{\mathcal{X}}^T D_{p_n} \dot{\mathcal{X}}\,\boldsymbol{u} = \boldsymbol{u}^T S \boldsymbol{u} \tag{74}$$

On est donc rammené à maximiser en $\boldsymbol{u}$ la quantité $\boldsymbol{u}^T S \boldsymbol{u}$.

Soient $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ les valeurs propres de S et soient $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ les vecteurs propres associés. Puisque $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p\}$ forme une base de $\mathbb{R}^p$, on peut écrire

$$\boldsymbol{u} = \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \cdots + \alpha_p \boldsymbol{v}_p \text{ avec } \|\boldsymbol{u}\|^2 = \sum_{j=1}^{p} \alpha_j^2 = 1$$

On sait d'autre part que, pour tout j, $S\,\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j$ donc

$$S\,\boldsymbol{u} = \alpha_1\,\lambda_1\,\boldsymbol{v}_1 + \alpha_2\,\lambda_2\,\boldsymbol{v}_2 + \cdots + \alpha_p\,\lambda_p\,\boldsymbol{v}_p\ .$$

Puisque $\boldsymbol{v}_i^T \boldsymbol{v}_j = \mathbb{1}(i = j)$, on a

$$\boldsymbol{u}^T\, S\, \boldsymbol{u} = \lambda_1\, \alpha_1^2 + \lambda_2\, \alpha_2^2 + \cdots + \lambda_p\, \alpha_p^2$$

ce dont on déduit, du fait que les λ_j sont classés par ordre décroissant :

$$\boldsymbol{u}^T\, S\, \boldsymbol{u} \leq \lambda_1\, (\alpha_1^2 + \alpha_2^2 + \cdots + \alpha_p^2) = \lambda_1$$

λ_1 est la borne supérieure de $\boldsymbol{u}^T\, S\, \boldsymbol{u}$. Cette borne est effectivement atteinte quand $\alpha_1 = 1$ et $\alpha_2 = \cdots = \alpha_p = 0$ auquel cas, $\boldsymbol{u} = \boldsymbol{v}_1$.

λ_1 est l'inertie du nuage sur le premier axe D_1.

La droite D_1 recherchée est définie par le vecteur propre $\vec{v}_1$ associé à la plus grande valeur propre de la matrice S. Cette droite passe par G=O.
Les composantes de $\boldsymbol{v}_1$ sont les coordonnées du premier axe dans la base canonique.
D_1 est appelée le **premier axe principal**.

Les droites $D_2, \ldots, D_p$ définies par les vecteurs $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$ sont les axes principaux suivants.
Les coordonnées de l'individu i dans la nouvelle base des $\vec{v}_j$ sont dans la ligne i de la matrice

$$C_V = \dot{\mathcal{X}}\, V \ .$$

Remarques

- Si les données sont centrées réduites, on remplace S par R au dessus.

- La matrice V (resp. W) des vecteurs propres de S (resp. de R dans le cas centré-réduit) est la matrice de la nouvelle base (exprimée dans la base canonique) dans laquelle on va "regarder" le nuage. En particulier, les deux premières colonnes de cette matrice (la première colonne est le vecteur propre associé à la plus grande valeur propre, la deuxième colonne est le vecteur propre associé à la deuxième plus grande valeur propre, etc ...) sont les vecteurs qui déterminent (dans la base canonique) les deux axes du plan principal.
  Ces vecteurs propres sont appelés les axes (ou facteurs) principaux de l'ACP.

- Les composantes principales sont les coordonnées des points individus dans cette nouvelle base. Ces coordonnés sont regroupées dans la matrice C_V des composantes principales et on a $C_V = \dot{\mathcal{X}}V$ (resp. $C_W = ZW$ lorsque les données sont centrées réduites).

- Il convient de noter que les vecteurs propres ainsi définis sont uniques seulement au signe près :
  $\boldsymbol{x}$ vecteur propre de la matrice A associé à la plus grande valeur propre λ_1
  $\Leftrightarrow A\boldsymbol{x} = \lambda_1 \boldsymbol{x}$
  $\Leftrightarrow A(-\boldsymbol{x}) = \lambda_1(-\boldsymbol{x})$
  $-\boldsymbol{x}$ vecteur propre de la matrice A associé à la plus grande valeur propre λ_1

  $\boldsymbol{y}$ orthogonal à $\boldsymbol{x}$ et vecteur propre de A associé à la deuxième plus grande valeur propre λ_2
  $\Leftrightarrow A\boldsymbol{y} = \lambda_2 \boldsymbol{y}$ et $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$
  $\Leftrightarrow A(-\boldsymbol{y}) = \lambda_2(-\boldsymbol{y})$ et $\langle \boldsymbol{x}, -\boldsymbol{y} \rangle = 0$
  $-\boldsymbol{y}$ orthogonal à $\boldsymbol{x}$ et vecteur propre de A associé à la deuxième plus grande valeur propre λ_2

  Cela aura des conséquences sur les graphiques plans obtenus (quatres configurations différentes seront possibles) mais ce qui importe c'est la lecture simultanée du graphique plan des individus et du graphique plan des variables. Il importe peu que l'individu i soit du côté positif de l'axe 1 dans une configuration et du côté négatif de l'axe 1 dans une autre. Ce qui compte c'est avec quelles variables d'origine est corrélé l'axe 1.

  Par exemple :
  Configuration 1 :
  L'individu i est du côté positif de l'axe 1.

L'axe 1 est corrélé positivement avec la variable X^1.

Configuration 2 :
Le même individu i est du côté négatif de l'axe 1.
Alors l'axe 1 sera forcément corrélé négativement avec la variable X^1 (et on pourra le vérifier sur le graphique).

Préciser ce qui précede avec des graphiques ...

Exercice 13 *Pour les données de l'exemple support, les valeurs propres et les vecteurs propres de la matrice de variance-covariance (resp. de la matrice des corrélations) sont :*

```
> n<-10
> X<-matrix(c(13.1,8.5,4.9,12.1,10.5,7.9,14.1,11.5,10.9,10.1,9.5,5.9,12.1,8.5,11.9,13.1,11.5,9.9,13.
> unn<-as.matrix(rep(1,n))
> g<-(1/n)*t(X)%*%unn
> Xpoint<-X-unn%*%t(g)
> S<-t(Xpoint)%*%Xpoint/n
> eigen(S,symmetric=T)
$values
[1] 7.4465483 3.3085163 0.6749353

$vectors
           [,1]       [,2]        [,3]
[1,] -0.1375708 -0.6990371  0.70172743
[2,] -0.2504597 -0.6608892 -0.70745703
[3,]  0.9583028 -0.2730799 -0.08416157

> Dunsurs<-diag(1/(sd(X)*sqrt((n-1)/n)))
> Z<-Xpoint%*%Dunsurs
> R<-t(Z)%*%Z/n
> eigen(R,symmetric=T)
$values
[1] 1.7687741 0.9270759 0.3041499

$vectors
           [,1]       [,2]       [,3]
[1,]  0.6420046 0.38467229  0.6632174
[2,]  0.6863616 0.09713033 -0.7207450
[3,] -0.3416692 0.91792861 -0.2016662

```

Calculer les matrices C_V et C_W des composantes principales, respectivement pour les données centrées et les données centrées réduites.

6.2 Les composantes principales

A l'axe D_1 on associe une nouvelle variable notée $\boldsymbol{c}^1$ et appelée **première composante principale**. On a $\boldsymbol{c}^1 = (c_{11}, \ldots, c_{n1})^T$ avec

$$c_{i1} = \overline{OK}_i = \boldsymbol{v}_1^T \dot{\boldsymbol{x}}_{i.}$$

La variable $\boldsymbol{c}^1$ est une variable centrée :

$$\sum_{i=1}^{n} p_i c_{i1} = \sum_{i=1}^{n} p_i \sum_{j=1}^{p} v_{j1} \dot{x}_{ij} = \sum_{j=1}^{p} v_{j1} \underbrace{\sum_{i=1}^{n} p_i \dot{x}_{ij}}_{=0} = 0,$$

et on a :

$$\mathrm{Var}(\boldsymbol{c}^1) = \sum_{i=1}^{n} p_i c_{i1}^2 = \sum_{i=1}^{n} p_i (\overline{OK}_i)^2 = \lambda_1$$

La première composante principale est le vecteur dont les composantes sont les coordonnées des n individus sur l'axe principal. C'est une nouvelle variable, meilleure combinaison linéaire des variables d'origine.

On définit les autres composantes principales $\boldsymbol{c}^2, \ldots, \boldsymbol{c}^p$ de la même manière.

$$C_V = [\boldsymbol{c}^1 | \ldots | \boldsymbol{c}^p] : n \times p$$

$$\mathrm{Var}(\boldsymbol{c}^j) = \sum_{i=1}^{n} p_i c_{ij}^2 = \lambda_j \ .$$

$cov(\boldsymbol{c}^j, \boldsymbol{c}^{j'}) = 0$, lorsque j est différent de j'

Exercice 14 *Vérifier les deux formules précédentes sur les données de l'exemple support.*

6.3 Autres formules pour calculer l'inertie

Les données sont centrées (ou centrées réduites).

$$\mathcal{I} = \sum_{j=1}^{p} Var(\boldsymbol{c}^j) = \sum_{j=1}^{p} \lambda_j \tag{75}$$

Exercice 15 *Calculer l'inertie pour les données de l'exemple support dans le cas centré et dans le cas centré-réduit. Comparez avec les résultats trouvés précédemment.*

6.4 Récapitulatif

La matrice $\dot{\mathcal{X}}$ (resp. Z) contient les coordonnées des individus centrés (resp. centrés-réduits) dans la base canonique. La matrice V des vecteurs propres de S (resp. W des vecteurs propres de R) contient les "coordonnées" du nouveau repère dans la base canonique. La matrice $C_V = \dot{\mathcal{X}} V$ dans le cas centré uniquement (resp. $C_W = ZW$ dans le cas centré réduit) contient les coordonnées des individus dans le nouveau repère.

Exercice 16 *Tracer le nuage des individus dans le plan principal (constitué des deux premiers axes du nouveau repère).*

7 Critères de qualité dans l'ACP

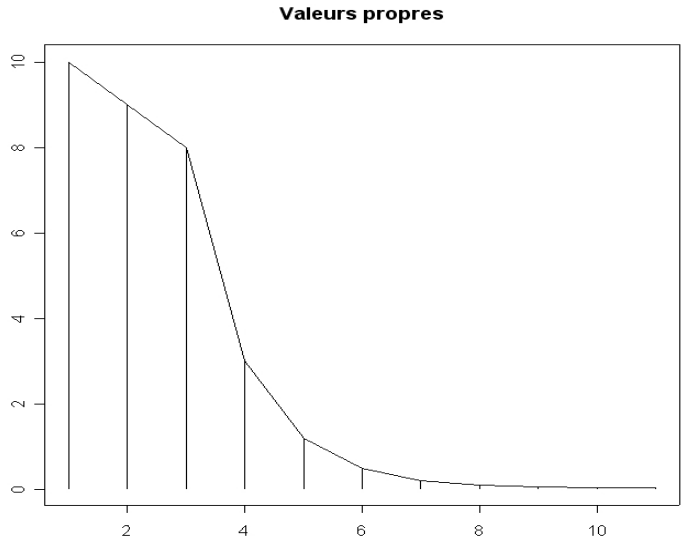
7.1 Nombre d'axes à retenir

Le principal intérêt de l'ACP consistant à réduire la dimension de l'espace des individus le choix du nombre d'axes à retenir est un point essentiel qui n'a pas hélas de solution rigoureuse. Remarquons tout d'abord que la réduction de dimension n'est possible que s'il y a redondance entre les variables $\boldsymbol{x}^1$, $\boldsymbol{x}^2$, ..., $\boldsymbol{x}^p$: si celles-ci sont indépendantes, ce qui est un résultat fort intéressant en soi, l'ACP sera inefficace à réduire la dimension.

Le critère suivant est connu sous le nom de critère de Kaiser.

Il est bon de ne retenir que les valeurs propres supérieure à 1 (ou 0.8 par exemple) dans le cas d'une ACP sur données centrées réduites : cela revient à ne regarder un axe que si la part de variation qu'il explique est supérieure ou, au moins, du même ordre de grandeur que celle d'une seule variable initiale (qui a pour variance 1).

On préconise également de détecter sur le diagramme des valeurs propres l'existence d'un coude, ce qui n'est pas toujours aisé en pratique.



Exercice 17 *Tracer le diagramme des valeurs propres pour les données de l'exemple support.*

Un autre critère habituellement utilisé est celui du pourcentage d'inertie totale expliquée. On mesure la qualité du sous-espace défini par les axes D_1 à D_k par :

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\mathcal{I}} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \ldots + \lambda_p} \tag{76}$$

où λ_i, valeur propre de la matrice de variance (ou de corrélation si les données sont centrées réduites), représente l'inertie expliquée par l'axe D_i.

Ainsi :
* $\frac{\lambda_1}{\sum_h \lambda_h}$ représente la part d'inertie expliquée par la première composante principale.
* $\frac{\lambda_1+\lambda_2}{\sum_h \lambda_h}$ représente la part d'inertie expliquée par le premier plan principal.
* etc. ...

Ces coefficients permettent de juger du nombre d'axes nécessaires pour obtenir un bon résumé.

Rappelons que si les données sont centrées-réduites, on a $\mathcal{I} = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_h \lambda_h = p$.

Si par exemple $\frac{\lambda_1+\lambda_2}{\mathcal{I}} = 0.9$, on conçoit clairement que le nuage de points est presque applati sur un sous-espace à deux dimensions et qu'une représentation du nuage dans le plan des deux premiers axes principaux (c'est-à-dire dans le plan principal) sera très satisfaisante.

L'application du pourcentage d'inertie doit toutefois faire intervenir le nombre de variables initiales : un % de 10% n'a pas le même intérêt sur un tableau de 20 variables et sur un tableau de 100 variables.
Il reste néanmoins l'obligation de ne retenir que les composantes interprétables et l'usage des corrélations avec les variables actives et supplémentaires joue ici un grand rôle.

Exercice 18

Complétez le tableau qui suit, obtenu après avoir effectué une ACP normée des données de l'exemple support. C'est-à-dire qu'il vous faut donner les valeurs numériques de total, λ_1, Prop1, Prop2, PC1, PC2.

	Valeur propre	Proportion	Proportion cumulée (% d'inertie)
1	λ_1	*Prop1*	*PC1*
2	0.9270759	*Prop2*	*PC2*
3	0.3041499	0.1013833	1.000
TOTAL	*total*		

7.2 Qualité de représentation et contribution d'un individu

Le pourcentage d'inertie expliquée est un critère global qui doit être complété par d'autres considérations plus locales.

7.2.1 Qualité de représentation des individus sur les sous-espaces principaux

Supposons que le plan (D_1, D_2) des deux premiers axes porte une inertie totale importante (valeur de $\lambda_1 + \lambda_2$ élevée) et que, en projection sur ce plan, deux individus soient très proches : la figure ci-dessous montre que cette proximité est illusoire si les deux individus se trouvent éloignés dans l'orthogonal du plan principal.
Il faut en fait envisager pour chaque individu $\boldsymbol{e}_i$ la qualité de sa représention. Celle-ci est souvent définie par le cosinus de l'angle entre le plan principal et le vecteur $\boldsymbol{e}_j$. Si ce cosinus est grand, $\boldsymbol{e}_j$ est voisin du plan, on pourra alors examiner la position de sa projection sur le plan par rapport à d'autres points ; si ce cosinus est faible on se gardera de toute conclusion.

Un petit angle α correspond à une grande valeur de $cos^2(\alpha)$.

Notons aussi que cette mesure du cosinus est d'autant meilleure que $\boldsymbol{e}_i$ est éloigné de $\boldsymbol{g}$; si $\boldsymbol{e}_i$ est proche de $\boldsymbol{g}$, la valeur du cosinus peut ne pas être significative. Donc si le cosinus au carré est proche de zéro, il faudra regarder la distance au carré de $\boldsymbol{e}_i$ à $\boldsymbol{g}$ (c'est-à-dire $p_i \dot{\boldsymbol{x}}_{i.}^T \dot{\boldsymbol{x}}_{i.}$) et s'assurer qu'elle n'est pas trop petite par rapport à l'inertie afin de ne pas conclure à tort que le point est mal représenté sur l'axe.

On note QLT_{i1} la qualité de représentation de l'individu i sur l'axe D_1 que l'on définit par :

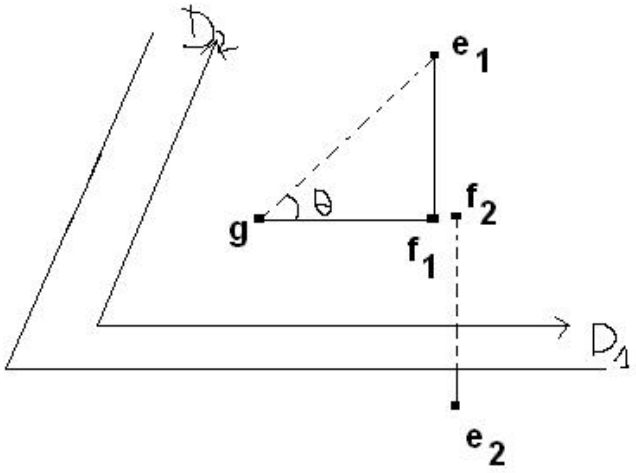


FIG. 1 – Erreur de perspective

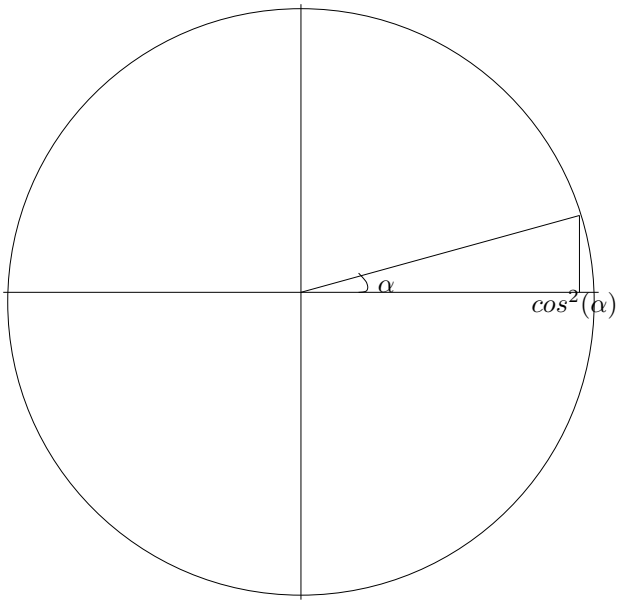


FIG. 2 – Cercle trigonométrique

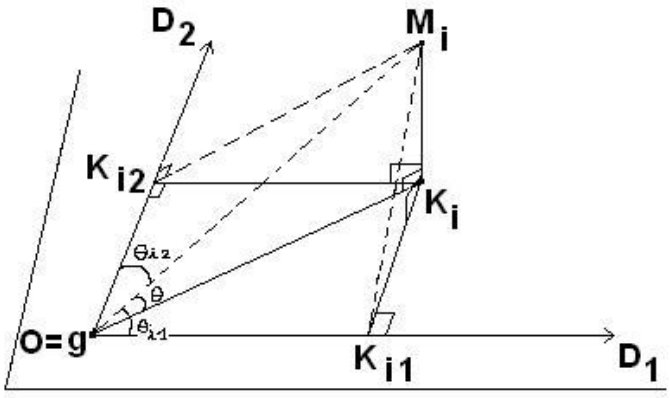
$$QLT_{i1} = \cos^2 \theta_{i1} = \frac{c_{i1}^2}{\sum_j c_{ij}^2} \tag{77}$$

où $\sum_j c_{ij}^2 = d^2(i, O)$. On appelle parfois *contribution relative* la quantité QLT_i.

On note $QLT_{i,1-2}$ la qualité de représentation de l'individu i sur le plan principal (D_1, D_2). On

a :

$$QLT_{i,1-2} = QLT_{i1} + QLT_{i2} = \cos^2\theta_{i1} + \cos^2\theta_{i2} = \frac{c_{i1}^2 + c_{i2}^2}{\sum_j c_{ij}^2} = \cos^2\theta. \qquad (78)$$



puisque $\cos^2\theta = \frac{OK_i^2}{OM_i^2} = \frac{OK_{i1}^2+OK_{i2}^2}{OM_i^2} = \frac{OK_{i1}^2}{OM_i^2} + \frac{OK_{i2}^2}{OM_i^2} = \cos^2\theta_{i1} + \cos^2\theta_{i2}$.
Ces relations sont vraies car le triangle $OK_{i2}M_i$ est rectangle en K_{i2}, le triangle $OK_{i1}M_i$ est rectangle en K_{i1}, le triangle OK_iM_i est rectangle en K_i, et car D_1 et D_2 sont perpendiculaires.

Exercice 19 *Calculer les qualités de représentation des individus de l'exemple support, sur le premier axe et sur le premier plan principal.*

Bien que moins utilisée, une mesure liée à la distance entre $\boldsymbol{e}_i$ et l'espace F_k semble préférable. En particulier, la quantité :

$$\frac{d(\boldsymbol{e}_i, \boldsymbol{f}_i)}{\sqrt{\mathcal{I} - \lambda_1 - \lambda_2 - \ldots - \lambda_k}}(signe(c_i^{k+1})) \qquad (79)$$

qui compare la distance entre $\boldsymbol{e}_i$ et F_k à la moyenne des carrés des distances de tous les individus à F_k présente un intérêt statistique certain (on peut la comparer à une variable de Laplace-Gauss réduite).

7.2.2 Contribution des individus aux axes

On a

$$Var(\boldsymbol{c}^1) = \sum_{i=1}^{n} p_i c_{i1}^2 = \lambda_1. \qquad (80)$$

La contribution de l'individu i à l'axe 1 est la proportion de la variance de $\boldsymbol{c}^1$ expliquée par l'individu i :

$$\mathrm{CTR}_{i1} = \frac{p_i c_{i1}^2}{\lambda_1} \ .$$

On appelle parfois *contribution absolue* la quantité CRT_i.
La contribution de l'individu i au plan principal est

$$\mathrm{CTR}_{i,1-2} = \frac{p_i c_{i1}^2 + p_i c_{i2}^2}{\lambda_1 + \lambda_2} = \frac{p_i c_{i1}^2 + p_i c_{i2}^2}{Var(\boldsymbol{c}^1) + Var(\boldsymbol{c}^2)}$$

C'est la proportion de l'inertie du nuage de points individus projeté sur le plan engendré par $\boldsymbol{c}^1$ et $\boldsymbol{c}^2$ expliquée par l'individu i.

D'une façon générale, on considèrera comme importante une contribution qui excède le poids p_i de l'individu concerné.

Exercice 20 *Calculer les contributions des individus de l'exemple support au premier axe et au premier plan principal.*

7.3 Qualité de représentation et contribution d'une variable

7.3.1 Qualité de représentation des variables

$$C = \dot{\mathcal{X}} V \Rightarrow CV^T = \dot{\mathcal{X}} V V^T = \dot{\mathcal{X}} I = \dot{\mathcal{X}} \Rightarrow \dot{\boldsymbol{x}}^j = \sum_h v_{jh} \boldsymbol{c}^h.$$

On note $P\,\dot{\boldsymbol{x}}^j$ la variable projection orthogonale de $\dot{\boldsymbol{x}}^j$ sur $\boldsymbol{c}^1$:

$$P\,\dot{\boldsymbol{x}}^j = v_{j1}\,\boldsymbol{c}^1 \ .$$

La qualité de représentation de $\boldsymbol{x}^j$ sur le premier facteur $\boldsymbol{c}^1$ est :

$$\mathrm{QLT}_{j1} = \cos^2 \alpha_{j1} = \frac{\|P\,\dot{\boldsymbol{x}}^j\|^2}{\|\dot{\boldsymbol{x}}^j\|^2} = \frac{v_{j1}^2\,\|\boldsymbol{c}^1\|^2}{\|\dot{\boldsymbol{x}}^j\|^2} = \frac{v_{j1}^2\,\mathrm{Var}(\boldsymbol{c}^1)}{\mathrm{Var}(\boldsymbol{x}^j)} = \frac{v_{j1}^2\,\lambda_1}{\mathrm{Var}(\boldsymbol{x}^j)} \ .$$

(on se rappellera que $\boldsymbol{c}^1$ est centrée : voir page 26).

Exercice 21 *Calculer les qualités de représentation des variables de l'exemple support sur le premier axe et sur le premier plan principal.*

Remarque 1 *La qualité de représentation de $\boldsymbol{x}^j$ sur le premier facteur $\boldsymbol{c}^1$ est égale au carré du coefficient de corrélation linéaire entre $\boldsymbol{x}^j$ et $\boldsymbol{c}^1$: $r^2(\boldsymbol{x}^j, \boldsymbol{c}^1)$.*

7.3.2 Contribution des variables

On définit la contribution de la variable $\dot{\boldsymbol{x}}^j$ au premier facteur $\boldsymbol{c}^1$ par

$$\mathrm{CTR}_{j1} = \frac{\mathrm{Var}(P\dot{\boldsymbol{x}}^j)}{\mathrm{Var}(\boldsymbol{c}^1)} = \frac{v_{j1}^2\,\lambda_1}{\lambda_1} = v_{j1}^2 \ .$$

Exercice 22 *Calculer les contributions des variables de l'exemple support au premier axe et au premier plan principal.*

8 Interprétation des axes principaux

L'ACP construit de nouvelles variables (les $\boldsymbol{c}^j$), artificielles, et des représentations graphiques permettant de visualiser les relations entre variables (cercle des corrélations), ainsi que l'existence éventuelle de groupes d'individus et de groupes de variables. Il faut constamment garder à l'esprit que le facteur d'ordre s ($s > 1$) traduit les tendances résiduelles non prises en compte par les facteurs précédents.

L'interprétation des résultats est une phase délicate qui doit se faire en respectant une démarche dont les éléments sont les suivants.

8.1 Corrélations entre composantes et variables initiales

La méthode la plus naturelle pour donner une signification à une composante principale $\boldsymbol{c}$ est de la relier aux variables initiales $\boldsymbol{x}^j$ en calculant les coefficients de corrélation linéaire $r(\boldsymbol{c}, \boldsymbol{x}^j)$ et en s'intéressant aux plus forts coefficients en valeur absolue.

Dire que $\boldsymbol{c}^1$ est très corrélée (positivement) avec une variable $\boldsymbol{x}^j$ signifie que les individus ayant une forte coordonnée positive sur l'axe 1 (déterminé par $\boldsymbol{c}^1$) sont caractérisés par une valeur de $\boldsymbol{x}^j$ nettement supérieure à la moyenne (rappelons que l'origine des axes principaux représente le centre de gravité du nuage).

Lorsqu'on choisit la métrique D_{1/s^2}, ce qui revient à travailler sur données centrées réduites et donc à chercher les valeurs propres et vecteurs propres de R, le calcul de $r(\boldsymbol{c}, \boldsymbol{x}^j)$ est particulièrement simple. En effet :

$$r(\boldsymbol{c}, \boldsymbol{x}^j) = r(\boldsymbol{c}, \boldsymbol{z}^j) = \frac{\boldsymbol{c}^T D_{p_n} \boldsymbol{z}^j}{\sqrt{Var(\boldsymbol{c})}} \tag{81}$$

comme $Var(\boldsymbol{c}) = \lambda$:

$$r(\boldsymbol{c}, \boldsymbol{x}^j) = \frac{\boldsymbol{c}^T D_{p_n} \boldsymbol{z}^j}{\sqrt{\lambda}} \tag{82}$$

or $\boldsymbol{c} = Z\boldsymbol{w}$ où $\boldsymbol{w}$, appelé facteur principal associé à $\boldsymbol{c}$, est vecteur propre de R associé à la valeur propre λ :

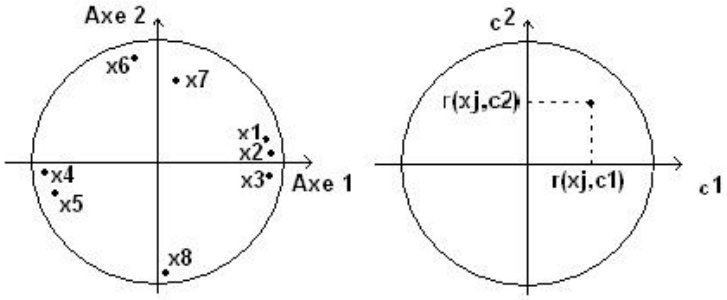
$$r(\boldsymbol{c}, \boldsymbol{x}^j) = \frac{\boldsymbol{w}^T Z^T D_{p_n} \boldsymbol{z}^j}{\sqrt{\lambda}} = \frac{(\boldsymbol{z}^j)^T D_{p_n} Z\boldsymbol{w}}{\sqrt{\lambda}}. \tag{83}$$

$(\boldsymbol{z}^j)^T D_{p_n} Z$ est la j^{eme} ligne de $Z^T D_{p_n} Z = R$ donc $(\boldsymbol{z}^j)^T D_{p_n} Z\boldsymbol{w}$ est la j^{eme} composante de $R\boldsymbol{w}$. Comme $R\boldsymbol{w} = \lambda\boldsymbol{w}$, il vient :

$$r(\boldsymbol{c}, \boldsymbol{x}^j) = \sqrt{\lambda} w_j. \tag{84}$$

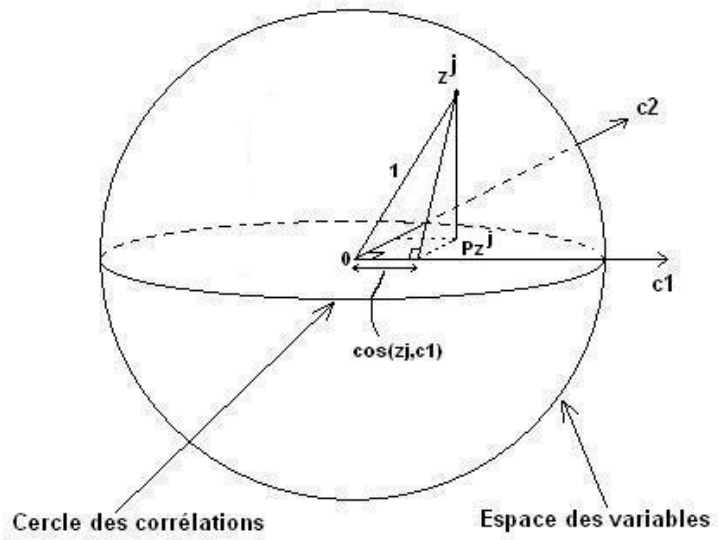
Ces calculs s'effectuent pour chaque composante principale. Pour un couple de composantes principales $\boldsymbol{c}^1$ et $\boldsymbol{c}^2$ par exemple on synthétise usuellement les corrélations sur une figure appelée "cercle des corrélations" où chaque variable $\boldsymbol{x}^j$ est repérée par un point d'abscisse $r(\boldsymbol{c}^1, \boldsymbol{x}^j)$ et d'ordonnée $r(\boldsymbol{c}^2, \boldsymbol{x}^j)$.

Ainsi la figure suivante montre une première composante principale très corrélée positivement avec les variables 1, 2 et 3, anticorrélée avec les variables 4 et 5 et non corrélée avec 6, 7 et 8.

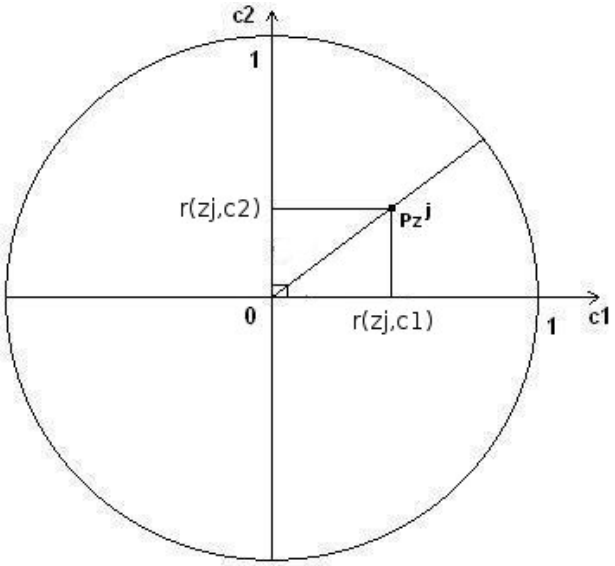


Par contre, la deuxième composante principale oppose la variable 8 aux variables 6 et 7.

Dans le cas de la métrique D_{1/s^2}, c'est-à-dire rappelons-le, de l'ACP sur données centrées réduites, le cercle des corrélations n'est pas seulement une représentation symbolique commode : c'est la projection de l'ensemble des variables centrées réduites sur le sous-espace engendré par $\boldsymbol{c}^1$ et $\boldsymbol{c}^2$. En effet, les $\boldsymbol{z}^j$ étant de variance un, sont situées sur la surface de la sphère unité de l'espace $\mathbb{R}^n$ des variables. Projetons l'extrémité du vecteur $\boldsymbol{z}^j$ sur le sous-espace de dimension 2 engendré par $\boldsymbol{c}^1$ et $\boldsymbol{c}^2$ (qui sont orthogonales). La projections $P\boldsymbol{z}^j$ tombe à l'intérieur du grand cercle intersection de la sphère avec le plan $\boldsymbol{c}^1$; $\boldsymbol{c}^2$. La projection se faisant avec la métrique D_{1/s^2} de l'espace des variables, $\boldsymbol{z}^j$ se projette sur l'axe engendré par $\boldsymbol{c}^1$ en un point d'abscisse $\cos(\boldsymbol{z}^j, \boldsymbol{c}^1)$ ce qui n'est autre que le coefficient de corrélation linéaire $r(\boldsymbol{x}^j, \boldsymbol{c}^1)$.



Le cercle des corrélations est donc, dans l'espace des variables, le pendant exact de la projection des individus sur le premier plan principal : on montre en effet qu'on l'obtient en projetant dans l'espace des caractères, les caractères centrés réduits sur le plan engendré par $\boldsymbol{c^1}$ et $\boldsymbol{c^2}$.



Donc Pz^j est le point de coordonnées $(r(\boldsymbol{z}^j, \boldsymbol{c}^1), r(\boldsymbol{z}^j, \boldsymbol{c}^2))$ dans le repère $(\boldsymbol{c}^1, \boldsymbol{c}^2)$.

Remarque :

- On se gardera d'interpréter des proximités entre points variables, si ceux-ci ne sont pas proches de la circonférence.
  Il est facile de s'en persuader en regardant le graphique précédent de la sphère.

- Comme $\lambda_k = \sum_{j=1}^{p} r^2(\boldsymbol{c}^k, \boldsymbol{x}^j)$ on appelle parfois contribution de la variable j à l'axe k le rapport :
$$\frac{r^2(\boldsymbol{c}^k, \boldsymbol{x}^j)}{\lambda_k} = (w_{jk})^2 \tag{85}$$
  mais cette quantité ne présente que peu d'intérêt en ACP et n'apporte rien de plus que le coefficient de corrélation.

Exercice 23 *Tracer le cercle des corrélations des variables de l'exemple support.*

8.2 Effet "taille"

Lorsque toutes les variables $\boldsymbol{x}^j$ sont corrélées positivement entre elles, la première composante principale définit un "facteur de taille". On sait qu'une matrice symétrique ayant tous ses termes positifs admet un premier vecteur propre dont toutes les composantes sont de même signe (théorème de Frobenius) : si on les choisit positives la première composante principale est alors corrélée positivement avec toutes les variables et les individus sont rangés sur l'axe 1 par valeurs croissantes de l'ensemble des variables (en moyenne). Si de plus les corrélations entre variables sont toutes de même ordre la première composante principale est proportionnelle à la moyenne des variables initiales :
$$\frac{1}{p}\sum_{j=1}^{p} \boldsymbol{x}^j. \tag{86}$$
La deuxième composante principale différencie alors des individus de "taille" semblable : on l'appelle facteur de "forme".

9 Individus et variables supplémentaires

Les interpétations fondées sur les remarques précédentes présentent le défaut d'être tautologiques : on explique les résultats à l'aide des données qui ont servi à les obtenir. On risque alors de prendre pour une propriété des données ce qui n'est qu'un artefact dû à la méthode : il n'est pas étonnant par exemple de trouver de fortes corrélations entre la première composante principale $\boldsymbol{c}^1$ et certaines variables puisque $\boldsymbol{c}^1$ maximise

$$\sum_{j=1}^{p} r^2(\boldsymbol{c}, \boldsymbol{x}^j). \tag{87}$$

On n'est donc pas sûr d'avoir découvert un phénomène significatif.
Par contre si on trouve une forte corrélation entre une composante principale et une variable qui n'a pas servi à l'analyse, le caractère probant de ce phénomène sera bien plus élevé. D'où la pratique courante de partager en deux groupes l'ensemble des variables : d'une part les variables "actives" qui servent à déterminer les axes principaux, d'autres part les variables "passives" ou supplémentaires que l'on relie *a posteriori* aux composantes principales.
On distinguera le cas des variables numériques supplémentaires de celui des variables qualitatives supplémentaires.
Les variables numériques supplémentaires peuvent être placées dans les cercles de corrélation : il suffit de calculer le coefficient de corrélation entre chaque variable supplémentaire $\boldsymbol{y}$ et les composantes principales $\boldsymbol{c}^1, \boldsymbol{c}^2, \ldots$. On peut alors utiliser les résultats du chapitre précédent pour détecter une corrélation significative.
Une variable qualitative supplémentaire correspond à la donnée d'une partition des n individus en k catégories : on peut faire apparaître par des symboles différents les individus de chaque catégorie sur les plans principaux. En général on se contente de représenter chaque catégorie par son centre de gravité : on peut alors mesurer au moyen du rapport de corrélation la liaison entre une variable qualitative supplémentaire et une composante principale et vérifier son caractère significatif au moyen du F de Fisher-Snedecor.
On peut également ne pas faire participer à l'analyse une partie des individus (on calcule les corrélations sans eux) ce qui permettra de vérifier sur cet échantillon-test des hypothèses formulées après une ACP sur les individus actifs. Il est d'ailleurs immédiat de positionner de nouveaux individus sur les axes principaux puisqu'il suffit de calculer des combinaisons linéaires de leurs caractéristiques.

10 Exemples traités et commentés

10.1 Les dépenses de l'état

TAB. 1 – Dépenses de l'Etat francais entre 1872 et 1971.

	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ACO	DEF	DET	DIV
1872	18	0.5	0.1	6.7	0.5	2.1	2	0	26.4	41.5	2.1
1880	14.1	0.8	0.1	15.3	1.9	3.7	0.5	0	29.8	31.3	2.5
1890	13.6	0.7	0.7	6.8	0.6	7.1	0.7	0	33.8	34.4	1.7
1900	14.3	1.7	1.7	6.9	1.2	7.4	0.8	0	37.7	26.2	2.2
1903	10.3	1.5	0.4	9.3	0.6	8.5	0.9	0	38.4	27.2	3
1906	13.4	1.4	0.5	8.1	0.7	8.6	1.8	0	38.5	25.3	1.9
1909	13.5	1.1	0.5	9	0.6	9	3.4	0	36.8	23.5	2.6
1912	12.9	1.4	0.3	9.4	0.6	9.3	4.3	0	41.1	19.4	1.3
1920	12.3	0.3	0.1	11.9	2.4	3.7	1.7	1.9	42.4	23.1	0.2
1923	7.6	1.2	3.2	5.1	0.6	5.6	1.8	10	29	35	0.9
1926	10.5	0.3	0.4	4.5	1.8	6.6	2.1	10.1	19.9	41.6	2.3
1929	10	0.6	0.6	9	1	8.1	3.2	11.8	28	25.8	2
1932	10.6	0.8	0.3	8.9	3	10	6.4	13.4	27.4	19.2	0
1935	8.8	2.6	1.4	7.8	1.4	12.4	6.2	11.3	29.3	18.5	0.4
1938	10.1	1.1	1.2	5.9	1.4	9.5	6	5.9	40.7	18.2	0
1947	15.6	1.6	10	11.4	7.6	8.8	4.8	3.4	32.2	4.6	0
1950	11.2	1.3	16.5	12.4	15.8	8.1	4.9	3.4	20.7	4.2	1.5
1953	12.9	1.5	7	7.9	12.1	8.1	5.3	3.9	36.1	5.2	0
1956	10.9	5.3	9.7	7.6	9.6	9.4	8.5	4.6	28.2	6.2	0
1959	13.1	4.4	7.3	5.7	9.8	12.5	8	5	26.7	7.5	0
1962	12.8	4.7	7.5	6.6	6.8	15.7	9.7	5.3	24.5	6.4	0.1
1965	12.4	4.3	8.4	9.1	6	19.5	10.6	4.7	19.8	3.5	1.8
1968	11.4	6	9.5	5.9	5	21.1	10.7	4.2	20	4.4	1.9
1971	12.8	2.8	7.1	8.5	4	23.8	11.3	3.7	18.8	7.2	0

PVP = PouVoirs Publics, AGR = AGRiculture, CMI = CoMmerce et Industrie,
TRA = TRAnsports, LOG = LOGement et amenagement du territoire,
EDU = EDUcation et culture, ACS = ACtion Sociale, ACO = Anciens Combattants,
DEF = DEFense nationale, DET = DETte, DIV = DIVers.

Si les phases de calcul sont effectuées automatiquement par des programmes informatiques, la lecture des documents obtenus nécessite une certaine méthode afin d'éviter des interprétations erronées. Nous avons choisi pour analyser le tableau des dépenses de l'Etat la métrique D_{1/s^2} ce qui revient à centrer et réduire les 11 caractères. Voici les données centrées réduites.

```
> round(Z,2)
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11]
 [1,]  2.64 -0.91 -0.86 -0.66 -0.83 -1.50 -0.83 -1.03 -0.53  1.83  0.89
 [2,]  0.86 -0.73 -0.86  2.83 -0.49 -1.19 -1.27 -1.03 -0.06  1.00  1.28
 [3,]  0.63 -0.79 -0.72 -0.62 -0.80 -0.54 -1.21 -1.03  0.48  1.25  0.50
 [4,]  0.95 -0.18 -0.50 -0.58 -0.66 -0.49 -1.18 -1.03  1.02  0.58  0.99
 [5,] -0.87 -0.30 -0.79  0.40 -0.80 -0.28 -1.15 -1.03  1.11  0.66  1.77
 [6,]  0.54 -0.36 -0.77 -0.09 -0.78 -0.26 -0.88 -1.03  1.13  0.51  0.70
 [7,]  0.59 -0.54 -0.77  0.28 -0.80 -0.18 -0.42 -1.03  0.89  0.36  1.38
 [8,]  0.31 -0.36 -0.81  0.44 -0.80 -0.12 -0.15 -1.03  1.48  0.02  0.11
 [9,]  0.04 -1.03 -0.86  1.45 -0.37 -1.19 -0.91 -0.57  1.66  0.32 -0.96
[10,] -2.11 -0.48 -0.16 -1.31 -0.80 -0.83 -0.88  1.38 -0.17  1.30 -0.28
[11,] -0.78 -1.03 -0.79 -1.55 -0.52 -0.64 -0.80  1.40 -1.42  1.84  1.09
[12,] -1.01 -0.85 -0.74  0.28 -0.71 -0.35 -0.47  1.81 -0.31  0.55  0.80
[13,] -0.74 -0.73 -0.81  0.23 -0.23  0.01  0.46  2.20 -0.39  0.00 -1.15
[14,] -1.56  0.37 -0.57 -0.21 -0.61  0.47  0.41  1.69 -0.13 -0.05 -0.76
[15,] -0.96 -0.54 -0.61 -0.98 -0.61 -0.08  0.35  0.39  1.43 -0.08 -1.15
[16,]  1.55 -0.24  1.35  1.25  0.87 -0.22  0.00 -0.21  0.27 -1.19 -1.15
[17,] -0.46 -0.42  2.80  1.65  2.83 -0.35  0.02 -0.21 -1.31 -1.23  0.31
[18,]  0.31 -0.30  0.68 -0.17  1.95 -0.35  0.14 -0.09  0.80 -1.14 -1.15
[19,] -0.60  2.01  1.29 -0.29  1.35 -0.10  1.08  0.08 -0.28 -1.06 -1.15
[20,]  0.41  1.46  0.75 -1.06  1.40  0.49  0.93  0.17 -0.49 -0.95 -1.15
[21,]  0.27  1.64  0.79 -0.70  0.68  1.10  1.43  0.25 -0.79 -1.04 -1.06
[22,]  0.09  1.40  1.00  0.32  0.49  1.83  1.70  0.10 -1.43 -1.28  0.60
[23,] -0.37  2.43  1.24 -0.98  0.25  2.14  1.73 -0.02 -1.40 -1.21  0.70
[24,]  0.27  0.49  0.71  0.07  0.01  2.65  1.90 -0.14 -1.57 -0.98 -1.15
```

Les facteurs principaux s'obtiennent donc en diagonalisant la matrice de corrélation R.

10.1.1 Valeurs propres, facteurs et composantes principales

On trouve au moyen d'un programme standard d'ACP :

```
> round(valp,2) # Valeurs propres
 [1] 4.97 2.05 1.29 0.99 0.71 0.56 0.20 0.13 0.06 0.04 0.00
> round(valp/11*100,2)  # % d'inertie
 [1] 45.21 18.64 11.73  9.03  6.44  5.08  1.86  1.14  0.56  0.32  0.00
> round(cumsum(valp)/11*100,2)  # % d'inertie cumulée
 [1]  45.21  63.85  75.57  84.61  91.04  96.12  97.98  99.12  99.68 100.00
[11] 100.00
```

La somme des valeurs propres est égale au nombre de caractères puisque $M = D_{1/s^2}$, soit ici 11. On vérifie que la dernière valeur propre est nulle, ce qui était attendu puisque les caractères sont sont liés par une relation linéaire (leur somme vaut 100).

Les deux premières valeurs propres représentant environ 64% de l'inertie, nous résumerons les données par les deux premières composantes principales.

Il est difficile de donner une réponse générale à la question : à partir de quel pourcentage peut-on négliger les composantes principales restantes ? Cela dépend tout d'abord du nombre de caractères : un premier axe expliquant 45% de l'inertie avec 11 caractères est plus intéressant que si p avait été égal à 5. Si R ne contient que des termes peu différents de zéro, il ne faut pas s'attendre à trouver des valeurs propres très élevées : on ne peut réduire efficacement le nombre de caractères que si ceux-ci étaient très corrélés. En fait, seul l'examen de la signification des composantes principales, et surtout l'expérience, permettent de savoir quelles sont les composantes à conserver.
Les deux premiers vecteurs propres $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$ de R sont ici les suivants :

```
> vecp[,1:2]
             [,1]         [,2]
 [1,] -0.07783386 -0.516681046
 [2,]  0.36703067 -0.004133249
 [3,]  0.37379156 -0.237624806
 [4,] -0.06150644 -0.440408207
 [5,]  0.32357810 -0.277781415
 [6,]  0.35282174  0.095339841
 [7,]  0.41848234  0.070182389
 [8,]  0.12957308  0.564035365
 [9,] -0.27458025 -0.150965222
[10,] -0.39852348  0.210559788
[11,] -0.24578306 -0.078479897
```

La somme des carrés de leur composantes vaut 1 et on peut vérifier que $R\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$.

```
> R
              [,1]        [,2]         [,3]        [,4]        [,5]       [,6]
 [1,]  1.000000000 -0.08456147 -0.001829121  0.23274025  0.03561605 -0.1500397
 [2,] -0.084561467  1.00000000  0.601120182 -0.27583846  0.43573158  0.7313206
 [3,] -0.001829121  0.60112018  1.000000000  0.09193172  0.89135190  0.4678457
 [4,]  0.232740254 -0.27583846  0.091931724  1.00000000  0.16610414 -0.2131541
 [5,]  0.035616047  0.43573158  0.891351897  0.16610414  1.00000000  0.2323607
 [6,] -0.150039669  0.73132063  0.467845734 -0.21315407  0.23236075  1.0000000
 [7,] -0.131402582  0.80568293  0.621981491 -0.20307154  0.48776783  0.8749779
 [8,] -0.686901167  0.04428292  0.022731574 -0.31322317  0.04471192  0.1569665
 [9,]  0.101050140 -0.44836614 -0.537273880  0.15797549 -0.37856133 -0.5240575
[10,]  0.033556338 -0.69491720 -0.804117505 -0.14834035 -0.75803796 -0.6702498
[11,]  0.149324296 -0.27720187 -0.347406886  0.11436882 -0.43793515 -0.2486460
            [,7]        [,8]        [,9]       [,10]       [,11]
 [1,] -0.1314026 -0.68690117  0.10105014  0.03355634  0.14932430
 [2,]  0.8056829  0.04428292 -0.44836614 -0.69491720 -0.27720187
 [3,]  0.6219815  0.02273157 -0.53727388 -0.80411750 -0.34740689
 [4,] -0.2030715 -0.31322317  0.15797549 -0.14834035  0.11436882
 [5,]  0.4877678  0.04471192 -0.37856133 -0.75803796 -0.43793515
 [6,]  0.8749779  0.15696654 -0.52405752 -0.67024982 -0.24864596
 [7,]  1.0000000  0.28819417 -0.56715016 -0.80819753 -0.52959488
 [8,]  0.2881942  1.00000000 -0.41685323 -0.04936630 -0.37746819
 [9,] -0.5671502 -0.41685323  1.00000000  0.26163585  0.02041298
[10,] -0.8081975 -0.04936630  0.26163585  1.00000000  0.55393211
[11,] -0.5295949 -0.37746819  0.02041298  0.55393211  1.00000000
```

Pour obtenir les composantes principales $\boldsymbol{c}_1$ et $\boldsymbol{c}_2$, on applique la formule $\boldsymbol{c} = Z\boldsymbol{v}$. Ainsi pour l'année 1872, dont on avait calculé plus haut les valeurs des coordonnées centrées réduites, il suffit de multiplier chaque coordonnée par la composante du premier vecteur propre et en faire la somme, pour obtenir la valeur de $\boldsymbol{c}_1$, soit ici 2.9.
On peut vérifier que $\boldsymbol{c}_1$ et $\boldsymbol{c}_2$ sont de moyenne nulle et ont pour variances respectives 4.97 et 2.05 (aux arrondis près), c'est-à-dire les deux premières valeurs propres.

10.1.2 Représentation des individus dans le plan principal

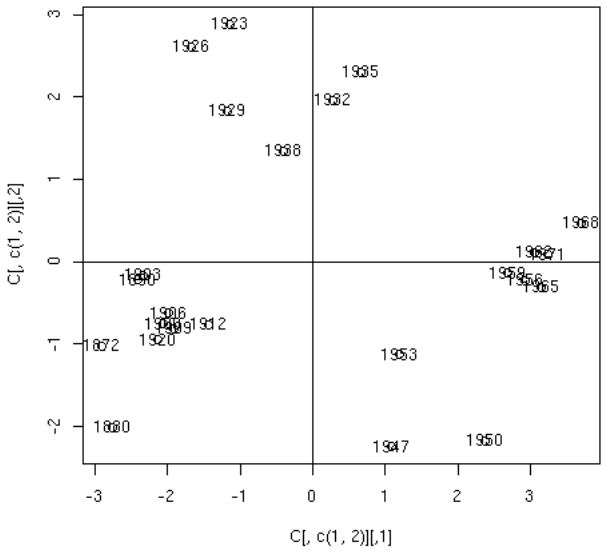
Les composantes $\boldsymbol{c}_1$ et $\boldsymbol{c}_2$ donnent les coordonnées des individus sur le plan principal.

```
> Z%*%vecp[,1:2]
             [,1]        [,2]
 [1,] -2.9005388 -1.02442948
 [2,] -2.7673894 -2.01195321
 [3,] -2.4163158 -0.22401426
 [4,] -2.0566342 -0.75515473
 [5,] -2.3378578 -0.16724592
 [6,] -1.9851416 -0.62613750
 [7,] -1.9073550 -0.81222235
 [8,] -1.4310705 -0.76841935
 [9,] -2.1391748 -0.95590969
[10,] -1.1429101  2.88394967
[11,] -1.6740880  2.61095464
[12,] -1.1734318  1.83119786
[13,]  0.2706376  1.95931965
[14,]  0.6590494  2.29620943
[15,] -0.4023984  1.34296396
[16,]  1.0812810 -2.25116584
[17,]  2.3728087 -2.17540030
```

```
[18,]  1.2037765 -1.13432281
[19,]  2.9279665 -0.23069051
[20,]  2.6861971 -0.14018403
[21,]  3.0547171  0.11078962
[22,]  3.1430131 -0.31123287
[23,]  3.6956105  0.46665277
[24,]  3.2392487  0.08644524
```

On obtient la configuration suivante.

```
> C<-Z%*%vecp
> plot(C[,c(1,2)])
> text(C[,c(1,2)],depenses.ind)
> abline(h=0)
> abline(v=0)
```



On voit immédiatement apparaître quatre groupes d'individus bien séparés :
groupe 1 : avant la première guerre mondiale ; groupe 2 : entre les deux guerres ; groupe 3 : l'après-guerre 1947-1950-1953 ; groupe 4 : la période 1956 à 1971.
La figure obtenue étant en projection, il ne faut pas confondre proximités sur le plan principal et proximités dans l'espace, une erreur de perspective est toujours possible comme le montre la figure 1.
Il faut donc examiner la qualité de la représentation de chaque point : ceci se fait en considérant l'angle θ entre le vecteur $\boldsymbol{e}_i$ et sa projection $\boldsymbol{f}_i$. Le critère de qualité communément utilisé est le carré du cosinus de l'angle avec le plan : un cosinus égal à 1 indique que $\boldsymbol{e}_i$ et $\boldsymbol{f}_i$ sont confondus ; un cosinus voisin de zéro doit mettre en garde l'utilisateur contre toute conclusion hâtive, sauf si $\boldsymbol{e}_i$ est à une distance faible du centre de gravité.

Dans notre exemple on trouve les valeurs suivantes :

```
> cos2<-round(C^2/apply(C^2,MARGIN=1,FUN=sum),2)
> rownames(cos2)<-depenses.ind
> apply(cos2[,1:2],MARGIN=1,FUN=sum)
1872 1880 1890 1900 1903 1906 1909 1912 1920 1923 1926 1929 1932 1935 1938 1947
0.53 0.69 0.80 0.68 0.57 0.78 0.71 0.53 0.53 0.79 0.66 0.63 0.47 0.80 0.29 0.65
1950 1953 1956 1959 1962 1965 1968 1971
0.46 0.34 0.73 0.75 0.89 0.74 0.69 0.65
```

Dans l'ensemble presque tous les points sont bien représentés sauf peut-être les années 1938 et 1953 (un cosinus carré de 0.3 correspond à un angle de 57°).
Lorsque de nombreux points sont mal représentés c'est en général parce que l'inertie du plan principal est trop faible : il faut alors considérer les composantes principales suivantes et regarder les plans principaux définis par les axes 1, 3 ; 2, 3 ; etc ...

10.1.3 L'interprétation des composantes principales et des axes principaux

Quelle signification concrète donner à des caractères qui sont des combinaisons des caractères de départ ? C'est sans doute un des points les plus délicats des analyses de données. Deux approches doivent généralement être utilisées : on considère, d'une part, les corrélations avec les caractères initiaux et, d'autre part, des individus typiques.

A) Le cercle des corrélations.

Le calcul des corrélations entre les composantes principales et les caractères initiaux est très simple à effectuer, dans le cas de la métrique D_{1/s^2} : on montre que le coefficient de corrélation linéaire entre $\boldsymbol{x}^j$ et $\boldsymbol{c}^k$ est égal à la j-ème composante du k-ème vecteur propre $\boldsymbol{v}_k$ multiplié par $\sqrt{\lambda_k}$. On en déduit que la somme des carrés des corrélations de $\boldsymbol{c}_k$ avec les $\boldsymbol{x}^j$ vaut λ_k. On trouve ici :

```
> correl<-round(sqrt(valp[1:2])*vecp[,1:2],2)
> row.names(correl)<-depenses.var
> correl
     [,1]  [,2]
PVP -0.17 -0.74
AGR  0.53 -0.01
CMI  0.83 -0.34
TRA -0.09 -0.98
LOG  0.72 -0.40
EDU  0.51  0.21
ACS  0.93  0.10
ACO  0.19  1.26
DEF -0.61 -0.22
DET -0.57  0.47
DIV -0.55 -0.11
```

La première composante principale est très corrélée positivement avec les pourcentages du budget consacré à l'action sociale, au commerce et industrie, à l'agriculture et très négativement avec les pourcentages consacrés à la défense, au remboursement de la dette.

L'opposition de ces deux groupes de caractères, que l'on retrouve sur le tableau R, est donc le trait dominant. Ceci permet d'interpréter la position des individus sur le plan principal : plus un point se situe à droite sur le graphique plus il s'écarte de la moyenne par de fortes valeurs des caractères ACS, CMI, AGR, ce qui est concomitant avec des valeurs inférieures à la moyenne des caractères DET et DEF. Aux points situés à gauche du graphique correspondent évidemmment des phénomènes inverses.

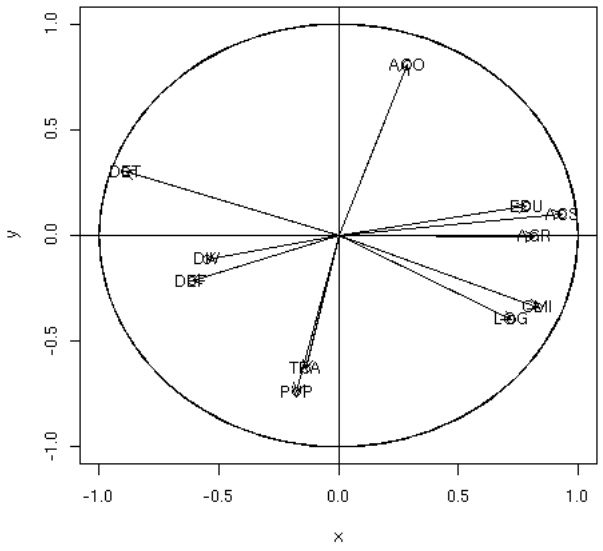
La deuxième composante principale dont l'importance est près de 2.5 fois moindre traduit essentiellement l'opposition entre le budget des anciens combattants et celui des pouvoirs publics. Si on représente chaque caractère par un point dont les coordonnées sont ses corrélations avec $\boldsymbol{c}_1$ et $\boldsymbol{c}_2$, les caractères initiaux s'inscrivent alors à l'intérieur d'un cercle de rayon 1 appelé cercle des corrélations car $\boldsymbol{c}_1$ et $\boldsymbol{c}_2$ étant non corrélées on montre que :

$$r^2(\boldsymbol{c}_1; \boldsymbol{x}^j) + r^2(\boldsymbol{c}_2; \boldsymbol{x}^j) \leq 1. \tag{88}$$

L'examen de cette figure permet d'interpréter les composantes principales et de repérer rapidement les groupes de caractères liés entre eux ou opposés, à condition toutefois que les points soient proches de la circonférence. Cette représentation joue pour les caractères le même rôle que le plan principal pour les individus : on montre en effet que l'on obtient exactement cette figure en projetant dans l'espace des caractères, les caratères centrés réduits sur le plan engendré par $\boldsymbol{c}_1$ et $\boldsymbol{c}_2$.

□

```
>   theta<-seq(0,20,.05)
> x<-cos(theta)
> y<-sin(theta)
> plot(x,y,type="l")
> abline(h=0,v=0)
> p<-ncol(depenses)
> A<-matrix(NA,nrow=p,ncol=p)
> for(i in 1:p) A[,i]<-(vecp[,i])*sqrt(valp[i])
> points(A[,1:2])
> text(A[,1:2],depenses.var)
> arrows(rep(0,p),rep(0,p),A[,1],A[,2],length=0.1)
```



B) La place et l'importance des individus.

Si on remarque que le long de l'axe 1 les années s'échelonnent à peu près selon l'ordre chronologique on met en évidence un phénomème d'évolution temporelle de la structure des dépenses de l'Etat (vers plus de social, moins de dettes et une moindre part à la défense nationale), ce qui enrichit l'étude des corrélations. De même il n'est peut-être pas inintéressant de noter que l'axe 2 qui oppose les dépenses en faveur des anciens combattants à celles des pouvoirs publics oppose en fait les deux après-guerre.

On peut d'ailleurs chercher quels sont les individus qui caractérisent le plus fortement un axe en calculant la "contribution" d'un point à l'axe numéro k que l'on définit comme $p_i c_{ik}^2/\lambda_k$, c'est la part de variance de $\boldsymbol{c}_k$ due à l'individu i. On trouve ici, mais nous ne reproduisons pas le détail des calculs, que pour l'axe 1 les contributions dominantes sont celles de 1968 et 1872 et pour l'axe 2, 1923, 1926 et 1947.

Ces considérations ne sont valables que parce que les individus présentent dans cet exemple un intérêt en eux-mêmes. Dans d'autres cas, en particulier ceux où les individus ont été obtenus par tirage au hasard pour un sondage, on a affaire à des êtres anonymes n'ayant d'intérêt que par leur ensemble et non par leur individualité ; l'ACP se résumera alors souvent à l'étude des caractères, c'est-à-dire au cercle des corrélations. Le fait que quelques individus puissent avoir des contributions importantes à la formation d'un des premiers axes principaux peut alors être un grave défaut car le fait de retirer ces individus risque de modifier profondément les résultats : il y a alors tout intérêt à effectuer l'ACP en éliminant cet individu quitte à le faire figurer ensuite sur les graphiques en point supplémentaire (car il est facile de calculer ses coordonnées), à condition qu'il ne s'agisse pas d'une donnée aberrante qui a ainsi été mise en évidence.

Notons enfin la possibilité de représenter sur les plans principaux des groupes d'individus possédant un trait particulier, par exemple l'ensemble des années représentant la IVème République. Ceci s'effectue très simplement en plaçant sur le graphique le centre de gravité des individus concernés dont les coordonnées se calculent aisément.
Dans l'état actuel de la technique informatique, on peut traiter des tableaux où le nombre de caractères est de quelques centaines pour un nombre d'individus en principe illimité, puisque la phase essentielle de calcul se réduit à la diagonalisation d'une matrice d'ordre p.

10.2 Les poissons d'Amiard

Tab. 2 – Les poissons d'Amiard.

	ray	rab	rao	ran	raf	rat	rar	rae	ram	poi	lon	las	lat	la	lam	dia
1	10	65	65	107	7	76	16	142	1	132	214	197	54	47	18	11
2	9	43	39	67	29	113	10	99	2	122	220	198	49	44	16	10
3	6	47	71	95	11	192	9	121	2	129	220	198	49	45	17	11
4	7	70	40	66	8	310	10	90	2	133	225	199	52	48	15	11
5	8	59	67	100	14	289	4	244	1	57	168	149	37	37	9	9
6	8	46	55	112	17	115	8	153	1	59	178	160	38	35	11	9
7	7	47	36	87	16	100	4	162	1	59	176	156	40	36	11	9
8	11	79	46	95	20	106	10	141	4	47	176	165	39	31	10	8
9	13	80	64	155	42	192	9	169	3	72	182	164	40	39	12	10
10	21	150	115	146	49	229	9	233	5	79	200	179	45	38	12	9
11	12	91	84	138	22	590	9	220	2	80	185	163	43	41	12	11
12	14	120	76	125	21	309	9	617	5	72	175	158	40	39	13	10
13	14	142	86	135	34	523	9	211	10	75	189	169	42	39	18	10
14	23	92	80	132	49	459	9	197	2	52	164	147	36	35	12	9
15	13	85	64	124	20	318	9	191	4	86	195	175	41	39	16	10
16	14	106	67	110	31	115	9	248	6	87	210	170	46	40	17	10
18	32	224	260	314	36	107	13	461	3	72	181	164	41	36	13	9
19	22	162	218	318	25	884	5	590	2	63	175	160	38	35	12	9
20	31	195	208	350	73	109	5	809	11	49	170	154	39	33	12	8
21	15	127	119	197	23	99	7	157	2	107	204	185	47	45	15	11
22	22	160	256	282	12	102	11	690	3	83	190	176	42	44	14	9
23	24	162	231	308	51	1031	17	558	2	82	194	168	42	39	14	10
24	19	64	163	229	16	109	8	345	1	91	190	172	44	42	13	11

24 poissons ont été répartis dans 3 aquariums radio-contaminés par un même polluant.
À ces trois aquariums correspondent des durées différentes de contact avec le polluant radio-actif.

groupe 1 : poissons 1 à 8 ;
groupe 2 : poissons 9 à 17 ;
groupe 3 : poissons 18 à 24.
Le poisson numéro 17 est mort en cours d'expérience.
Un poisson est repéré par 16 caractères formant deux groupes.
Le premier groupe, mesuré en fin d'expérience, comprend les variables de radio-activité (des yeux, des branchies, des opercules, des nageoires, du foie, du tube digestif, des reins, des écailles et des muscles) ;
le deuxième est constitué des variables de taille (poids, longueur, longueur standard, longueur de la tête, largeur, longueur du museau, diamètre des yeux).

Question : On veut savoir si la contamination du poisson est liée avec la durée de contact avec le polluant, et si la taille du poisson influe sur sa radio-contamination.

L'ACP :
Une analyse en composantes principales normée est effectuée. On utilise l'ACP normée car les variables ne sont pas toutes mesurées dans les mêmes unités. La métrique $M = Id_{16}$ à été utilisée : on donne la même importance à toute les variables. De même, tous les poissons ont le même poids, 1/23.

remarque Si toutes les variables étaient mesurées dans les mêmes unités, c'est à l'utilisateur de choisir d'effectuer une ACP réduite ou non. Dans une ACP non réduite, les variances des variables vont influer sur les résultats. Ainsi si l'on considère que la variation des variables est une information importante dans l'analyse, une ACP non réduite est le bon choix. Cependant, si quelques variables ont une très forte variance par rapport aux autres, elles seules vont influencer l'ACP et de l'information pourra être masquée : il est intéressant alors d'effectuer aussi une ACP réduite, qui ramène la variance des variables à 1 et ainsi l'analyse n'est pas influencée par la variance des variables.

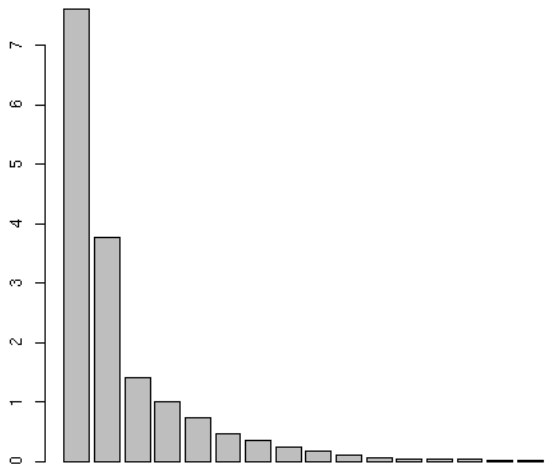
ANALYSE avec R en utilisant le package ADE4 :

```
> # il faut d'abord télécharger puis charger le module ADE4:
> library(ade4)

> #On entre les données:
> data<-matrix(c(10,65,65,107,7,76,16,142,1,132,214,197,54,47,18,11,
+ 9,43,39,67,29,113,10,99,2,122,220,198,49,44,16,10,
+ 6,47,71,95,11,192,9,121,2,129,220,198,49,45,17,11,
+ 7,70,40,66,8,310,10,90,2,133,225,199,52,48,15,11,
+ 8,59,67,100,14,289,4,244,1,57,168,149,37,37,9,9,
+ 8,46,55,112,17,115,8,153,1,59,178,160,38,35,11,9,
+ 7,47,36,87,16,100,4,162,1,59,176,156,40,36,11,9,
+ 11,79,46,95,20,106,10,141,4,47,176,165,39,31,10,8,
+ 13,80,64,155,42,192,9,169,3,72,182,164,40,39,12,10,
+ 21,150,115,146,49,229,9,233,5,79,200,179,45,38,12,9,
+ 12,91,84,138,22,590,9,220,2,80,185,163,43,41,12,11,
+ 14,120,76,125,21,309,9,617,5,72,175,158,40,39,13,10,
+ 14,142,86,135,34,523,9,211,10,75,189,169,42,39,18,10,
+ 23,92,80,132,49,459,9,197,2,52,164,147,36,35,12,9,
+ 13,85,64,124,20,318,9,191,4,86,195,175,41,39,16,10,
+ 14,106,67,110,31,115,9,248,6,87,210,170,46,40,17,10,
+ 32,224,260,314,36,107,13,461,3,72,181,164,41,36,13,9,
+ 22,162,218,318,25,884,5,590,2,63,175,160,38,35,12,9,
+ 31,195,208,350,73,109,5,809,11,49,170,154,39,33,12,8,
+ 15,127,119,197,23,99,7,157,2,107,204,185,47,45,15,11,
```

```
+ 22,160,256,282,12,102,11,690,3,83,190,176,42,44,14,9,
+ 24,162,231,308,51,1031,17,558,2,82,194,168,42,39,14,10,
+ 19,64,163,229,16,109,8,345,1,91,190,172,44,42,13,11),ncol=16,byrow=T)

> dimnames(data)[[2]]<-c('ray','rab','rao','ran','raf','rat','ras','rae','ram','poi',
+ 'lon','las','lat','la','lam','dia')
> data.dudi <- dudi.pca(data,scannf = FALSE)
> symnum(cor(data))
    ry rb rao rn rf rt rs rae rm p ln ls lt la lm d
ray 1
rab +  1
rao +  +  1
ran +  +  B   1
raf ,  .  .   .  1
rat            .     1
ras                     1
rae ,  ,  +   +  .        1
ram .  .          .        .   1
poi .  .      .   .     .   .      1
lon .         .   .     .   .      * 1
las .             .     .   .      * B  1
lat .         .   .     .   .      * *  *  1
la  .             .     .       .  * +  +  +  1
lam                     .          , ,  ,  ,  ,  1
dia .  .          .     .   .   .  + ,  ,  ,  +  ,  1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

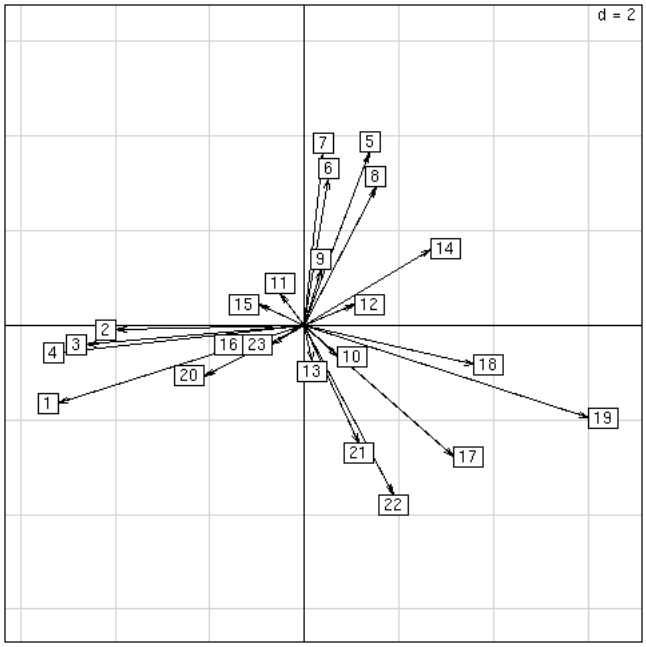
□

```
> barplot(data.dudi$eig)
```



□

```
> s.corcircle(data.dudi$co, lab = dimnames(data)[[2]], clabel=0.8,full = TRUE)
> s.arrow(data.dudi$li)
```




```
> data.dudi$eig # Valeurs propres
 [1] 7.607100492 3.763340863 1.405396247 1.001566207 0.731006724 0.455803745
 [7] 0.344391076 0.229504711 0.166091333 0.095884718 0.065308844 0.044954926
[13] 0.039067627 0.034340601 0.009335404 0.006906482
> contrib <- inertia.dudi(data.dudi,col.inertia=TRUE,row.inertia=TRUE)
> contrib$col.rel # CTR variables
    Comp1 Comp2 con.tra
ray  5178 -3808     625
rab  4108 -4704     625
rao  3239 -5079     625
ran  4383 -4233     625
raf  4035  -926     625
rat   741  -531     625
ras  -822 -3295     625
rae  4759 -2652     625
ram  1517  -732     625
poi -8034 -1649     625
lon -7423 -1575     625
las -7229 -1603     625
lat -7407 -1690     625
la  -7081 -1526     625
lam -3998 -2992     625
dia -6118  -638     625
> contrib$col.abs # CTA variables
    Comp1 Comp2
ray   681  1012
rab   540  1250
rao   426  1350
```

```
ran   576  1125
raf   530   246
rat    97   141
ras   108   875
rae   626   705
ram   199   195
poi  1056   438
lon   976   419
las   950   426
lat   974   449
la    931   405
lam   526   795
dia   804   169
> contrib$row.rel # CTR individus
   Axis1 Axis2 con.tra
1  -8327  -828     885
2  -8559    -5     501
3  -9505   -76     608
4  -9342  -120     753
5   1086  7725     470
6    230  8651     300
7    105  9266     390
8   1414  5377     432
9    337  4016      96
10   848  -741     157
11  -536   865     140
12  2010   343     153
13    23  -490     327
14  5002  1801     390
15 -3333   681      76
16 -2671  -268     168
17  4320 -3357     624
18  5833  -315     596
19  7214  -763    1357
20 -4982 -1274     245
21   867 -4118     413
22  1282 -4639     749
23  -792  -259     169
> contrib$row.abs # CTA individus
   Axis1 Axis2
1   1550   311
2    903     1
3   1215    20
4   1479    38
5    107  1545
6     15  1103
7      9  1535
8    129   988
9      7   164
10    28    49
11    16    51
12    65    22
13     2    68
14   411   299
15    53    22
```

```
16    94    19
17   567   891
18   731    80
19  2058   440
20   256   132
21    75   722
22   202  1478
23    28    19
> data.dudi$li
        Axis1       Axis2
1  -5.2073839 -1.64186159
2  -3.9740127 -0.09326168
3  -4.6108986 -0.41127622
4  -5.0874981 -0.57667154
5   1.3710155  3.65668934
6   0.5042002  3.09029683
7   0.3881823  3.64557319
8   1.4997537  2.92455686
9   0.3454857  1.19215504
10  0.6996999 -0.65414758
11 -0.5249197  0.66681399
12  1.0648313  0.44015157
13  0.1658738 -0.76878784
14  2.6805262  1.60828666
15 -0.9674451  0.43726073
16 -1.2837866 -0.40678333
17  3.1505961 -2.77733061
18  3.5767618 -0.83110219
19  6.0010714 -1.95124172
20 -2.1172012 -1.07051238
21  1.1472460 -2.50026530
22  1.8806353 -3.57688191
23 -0.7027334 -0.40166030
> data.dudi$co
         Comp1      Comp2
ray  0.7195815 -0.6171038
rab  0.6409406 -0.6858242
rao  0.5691158 -0.7127034
ran  0.6620693 -0.6506312
raf  0.6351939 -0.3043795
rat  0.2721933 -0.2305422
ras -0.2867513 -0.5739869
rae  0.6898420 -0.5149525
ram  0.3895018 -0.2706238
poi -0.8963099 -0.4061323
lon -0.8615698 -0.3969191
las -0.8502059 -0.4003350
lat -0.8606221 -0.4110423
la  -0.8414838 -0.3905861
lam -0.6322775 -0.5469913
dia -0.7821925 -0.2525112
```

On donne ci-dessous les moyennes et variances des variables avant réduction ainsi que la matrice des corrélations.

```
Col.:   1 | Mean:  1.5435e+01 | Variance: 5.3898e+01
Col.:   2 | Mean:  1.0504e+02 | Variance: 2.5491e+03
Col.:   3 | Mean:  1.0913e+02 | Variance: 5.2150e+03
Col.:   4 | Mean:  1.6487e+02 | Variance: 7.5767e+03
Col.:   5 | Mean:  2.7217e+01 | Variance: 2.5939e+02
Col.:   6 | Mean:  2.8161e+02 | Variance: 6.4208e+04
Col.:   7 | Mean:  9.0870e+00 | Variance: 9.7316e+00
Col.:   8 | Mean:  2.9774e+02 | Variance: 4.2711e+04
Col.:   9 | Mean:  3.2609e+00 | Variance: 6.8885e+00
Col.:  10 | Mean:  8.2087e+01 | Variance: 6.6425e+02
Col.:  11 | Mean:  1.9048e+02 | Variance: 3.0651e+02
Col.:  12 | Mean:  1.7070e+02 | Variance: 2.3543e+02
Col.:  13 | Mean:  4.2783e+01 | Variance: 2.2083e+01
Col.:  14 | Mean:  3.9435e+01 | Variance: 1.9463e+01
Col.:  15 | Mean:  1.3565e+01 | Variance: 6.1588e+00
Col.:  16 | Mean:  9.7391e+00 | Variance: 8.8847e-01


---------------------- Correlation matrix -------------------

[  1] 1000
[  2]  882 1000
[  3]  857  829 1000
[  4]  877  825  959 1000
[  5]  700  588  370  497 1000
[  6]  219  282  288  310  240 1000
[  7]  164  173  210   94    6  167 1000
[  8]  743  745  810  832  416  264   -1 1000
[  9]  378  522  149  239  590  -24 -136  386 1000
[ 10] -386 -307 -202 -304 -444 -157  422 -381 -285 1000
[ 11] -379 -264 -235 -339 -333 -178  412 -422 -140  938 1000
[ 12] -361 -257 -188 -294 -395 -240  426 -397 -185  943  953 1000
[ 13] -348 -249 -223 -314 -362 -253  437 -385 -158  947  940 931 1000
[ 14] -370 -282 -142 -259 -494 -145  357 -287 -306  933  829 829  862 1000
[ 15] -147  -17  -86 -150 -164  -39  460 -192  204  748  762 712  723  680 1000
[ 16] -411 -360 -254 -297 -440   64  303 -395 -324  803  677 629  714  843  621 1000
```

Choix du nombre d'axes :
La première valeur propre explique 47.5% de l'inertie (près de la moitié de l'inertie totale qui vaut 16 ici). La seconde valeur propre explique 23.5% et la troisième seulement 8%. A elle deux, les deux premières valeurs propres expliquent 71% de l'inertie ce qui est suffisant d'autant plus que la troisième valeur explique 3 fois moins d'inertie que la seconde, c'est-à-dire presque rien (et cela engendre une cassure dans l'histogramme des valeurs propres donné en cours). Cependant, on pourrait regarder ce qui se passe sur le troisième axe, qui doit montrer un phénomène «discret».

Etude des variables :
La première chose à faire est de consulter les contributions absolues qui nous disent quelles sont les variables qui contribuent à la constructions des axes. Un des critères de choix est de comparer les CTA des variables sur chacun des axes avec 1/p (car la somme des CTA sur un axe vaut 1) : les variables pour lesquelles la CTA est plus grande que 1/p contribuent fortement à la construction des axes. Ces variables permettent de donner un nom aux axes. Ensuite, il faut repérer quelles sont les variables qui sont bien représentées. Cela ce fait avec les CTR (sur axe, ou sur le plan) mais cela se repère aussi sur le cercle des corrélations en ACP normée : les variables dont la

flèche atteint presque le cercle sont très bien représentées. Moins elles sont proches du cercles moins elles sont bien représentées. De plus, pour savoir quelles sont les variables qui caractérisent un axe, on regarde l'angle formé par cette variable et l'axe. Ici, dans le premier plan principal (1,2), on s'aperçoit que les variables de taille sauf la variable 15 (lam) caractérisent l'axe 1, et plus particulièrement la variable 10 (poi) qui possède la CTA la plus élevé (0.1056). Toutes ses variables sont bien représentées sur cet axe avec une CTR supérieure à 0.7, et sont donc bien représentées dans le premier plan principal, et dans n'importe quel plan contenant l'axe 1. Ces variables ne seront représentées par aucun autre axe. Les variables de radio activité 1 et 8 (dans les yeux et les ecailles) contribuent aussi à cet axe mais dans une moindre mesure. On constate aussi les variables poi,la, la,lac,los,lon sont fortement corrélées positivement : la mesure d'une seule de ces variable suffit pour en déduire les autres. Le premier axe est caractérisé par la taille et le poids des poissons. D'autre part, les variables 1à 4, 7 et 8 caractérisent l'axe 2 (CTA>0.0625). Ce sont des variables de radio-activité. Seules la radioactivité dans le foie, le tube digestif et le muscle ne caractérisent pas l'axe 2. Le foie, le tube digestif et les muscles ne réagissent pas de la même manière à la radio activité que les autres organes. (Ces trois variables caractérisent l'axe3 d'après le tableau des contributions). Les variables 1 à 4 et 8 sont bien représentées dans le premier plan principal. La variable 7 ne l'est pas très bien. On constate aussi une forte corrélation positive entre les variables rao, rae, ran, ray et rab (ce qui se vérifie dans la matrice des corrélations), alors que la variable rar leur est presque orthogonale (elle n'a aucun lien (linéaire) avec les autres variables). On peut dire que l'axe 2 représente la radio activité «à la surface du poisson» (yeux, branchies, opercules, nageoires, écailles) : les organes au contact direct avec l'eau contaminée.
Enfin, on peut voir sur le cercle que les variables de radio activité caractérisant l'axe 2 sont quasiment orthogonales aux variables de tailles caractérisant l'axe 1.
Etude des individus :
Les poissons 1 à 4 et 18 à 20 (17 à 19 dans le tableau des contributions) ont une forte CTA (>0.0434) sur l'axe 1. Les poissons 5 à 8 et 20 à 23 (19 à 21 dans le tableau des contributions) ont une forte CTA sur l'axe 2. Ils ont fortement contribués à la construction des axes. Les poissons 1à 4 et 20 (et 19) sont très bien représentés sur l'axe 1, les poissons 5 à 7 sont très bien représentés sur l'axe 2, et les poissons 18, 8, 22 et 23 sont bien représentés dans le plan. On distingue plusieurs groupes de poissons :
Les poissons 1 à 4 : ils ont les mêmes caractéristiques.
Les poissons 5 à 8 : ils ont les mêmes caractéristiques.
Si l'on revient à l'origine des poissons, on constate que les trois aquariums sont assez bien différenciés.

Interprétation :
Il faut mettre en relation le graphique des individus et le graphique des variables. On peut lire les caractéristiques des poissons à partir de leur position dans le plan et à partir de la direction des variables dans le cercle des corrélations. Les individus à droite du graphique auront de fortes valeurs pour les variables caractérisant l'axe 1 positivement. Les individus à gauche auront de faibles valeurs par rapport à la moyenne pour ces mêmes variables. Inversement elles auront de fortes valeurs par rapport à la moyenne pour les variables caractérisant l'axe 1 négativement. Un raisonnement similaire est fait avec le deuxième axe. Des individus sur la première diagonale du plan seront caractérisés par les variables de l'axe 1 et 2 et plus particulièrement par celles se situant aussi sur la première diagonale sur cercle des corrélations correspondant. Ainsi, sur cet exemple : Les poissons 1 à 4 (aquarium 1) sont des poissons beaucoup plus gros que les autres (fortes valeurs pour les variables de taille) et moyennement contaminés car leur coordonnée sur l'axe est presque nulle. Les poissons 5 à 8 (aquarium 1) sont des poissons peu contaminés par rapport aux autres et de taille moyenne. Le poisson 20 (aquarium 3) est un très petit poisson plutôt fortement contaminé. Le poisson 18 est poisson fortement contaminé et plutôt petit. Le poisson 19 est un poisson plutôt petit et moyennement contaminé. Les poissons 22 et 23 sont fortement contaminés et de taille moyenne. On se gardera de donner une interprétation pour les autres poissons qui sont mal représentés. Cependant, on constate que l'aquarium 2 n'a pas influencé les résultats des axes 1 et 2 car se situe au centre du graphique. Il ne doit pas y avoir de poissons aux caractéristiques spéciales dans cet aquarium. On remarque que la distinction

des aquariums se fait le long de la première diagonale qui est la direction des variables des radio-activités mesurées sur les organes au contact de l'eau. Le premier plan principal oppose principalement les aquariums 1 et 3 (contact court et contact long avec la radio contamination) où la différence de radio contamination est mesurable sur le poisson à la sa surface : plus un poisson est exposé plus est contaminé en surface. De plus, ce plan principale met en avant une opposition dans la taille des poissons de l'aquarium 1 : il y a de très gros poissons (1 à 4) et des poissons de taille moyenne (5 à 8). Peut être peut on dire que ces deux groupes de poissons n'ont réagit de la même manière à la radio activité : les poissons 1 à 4 ont fortement grossis (pas freinés par la contamination ?), les poissons 5 à 8 ont un peu grandit seulement (freinés par la contamination ?). Tandis que dans l'aquarium 3, les poissons sont plutôt de petite taille par rapport à la moyenne (croissance freinées par la contamination de longue durée ?).

11 Formulaire

Matrice des données brutes :

$$X = (x_{ij})_{1\leq i\leq n; 1\leq j\leq p} = [\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p] = \begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_n^T \end{pmatrix}.$$

On note $\boldsymbol{x}^j = \boldsymbol{x}_{.j} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}$ la j-ème variable et $\boldsymbol{e}_i^T = \boldsymbol{x}_{i.} = (x_{i1}, \ldots, x_{ip})$ le i-ème individu.

Matrices des poids des individus : $D_{p_n} = diag(p_1, \ldots, p_n)$. Cas usuel des poids : $D_{p_n} = \frac{1}{n} I_n$.

Centre de gravité : $\boldsymbol{g} = X^T D_{p_n} \mathbf{1}_n = (\overline{x}_{.1}, \ldots, \overline{x}_{.p})^T$.

Données centrées : $\dot{X} = X - \mathbf{1}_n \boldsymbol{g}^T$.

Données centrées réduites : $Z = \dot{X} D_{1/s}$ avec $D_{1/s} = diag(1/s_1, \ldots, 1/s_p)$.

Matrice de variance-covariance : $S = \dot{X}^T D_{p_n} \dot{X}$.

Matrice des corrélations : $R = D_{1/s} S D_{1/s} = Z^T D_{p_n} Z$.

Distance entre individus : $d^2(\boldsymbol{e}_1, \boldsymbol{e}_2) = (\boldsymbol{e}_1 - \boldsymbol{e}_2)^T M (\boldsymbol{e}_1 - \boldsymbol{e}_2)$ avec $M = I$ ou $M = D_{1/s^2}$.

Produit scalaire entre deux variables :

$$\langle \boldsymbol{x}^j, \boldsymbol{x}^k \rangle = {\boldsymbol{x}^j}^T D_{p_n} \boldsymbol{x}^k = \sum_{i=1}^{n} p_i x_i^j x_i^k. \tag{89}$$

Produit scalaire	**Statistique Descriptive**	**Calcul matriciel**
$\left\langle \dot{\boldsymbol{x}}^j, \dot{\boldsymbol{x}}^k \right\rangle$	$cov(\boldsymbol{x}^j, \boldsymbol{x}^k)$	${\dot{\boldsymbol{x}}^j}^T D_{p_n} \dot{\boldsymbol{x}}^k$
$\left\langle \dot{\boldsymbol{x}}^j, \dot{\boldsymbol{x}}^j \right\rangle = \left\| \dot{\boldsymbol{x}}^j \right\|^2$	$var(\boldsymbol{x}^j)$	${\dot{\boldsymbol{x}}^j}^T D_{p_n} \dot{\boldsymbol{x}}^j$
$\cos(\dot{\boldsymbol{x}}^j, \dot{\boldsymbol{x}}^k)$	$r(\boldsymbol{x}^j, \boldsymbol{x}^k)$, la corrélation	

On note V la matrice des vecteurs propres de S. On note $\lambda_j(S)$ la j-ème valeur propre de S.

On note W la matrice des vecteurs propres de R. On note $\lambda_j(R)$ la j-ème valeur propre de R.

On note $C_V = \dot{X}V = [\boldsymbol{c}_V^1 | \dots | \boldsymbol{c}_V^p]$ la matrice des composantes principales pour les données centrées.

On note $C_W = ZW = [\boldsymbol{c}_W^1 | \dots | \boldsymbol{c}_W^p]$ la matrice des composantes principales pour les données centrées réduites.

Inertie (cas des données ni centrées, ni réduites ($M = I$)) :

$$\mathcal{I}_{\boldsymbol{g}} = \sum_i p_i d^2(i, G) = \sum_i p_i \|\boldsymbol{e}_i - \boldsymbol{g}\|^2 = \sum_i p_i (\boldsymbol{e}_i - \boldsymbol{g})^T (\boldsymbol{e}_i - \boldsymbol{g}) = \sum_i p_i (\boldsymbol{x}_{i.} - \boldsymbol{g}^T)(\boldsymbol{x}_{i.} - \boldsymbol{g}^T)^T$$

Inertie (cas des données centrées et non réduites ($M = I$)) :

$$\mathcal{I}_{\mathbf{0}} = \mathcal{I} = \sum_{i=1}^{n} p_i d^2(i, O) = \sum_{i=1}^{n} p_i d^2(\dot{\boldsymbol{x}}_{i.}^T, \mathbf{0}) = \sum_i p_i \dot{\boldsymbol{x}}_{i.} \dot{\boldsymbol{x}}_{i.}^T$$

et

$$\mathcal{I} = trace(S) = \sum_j Var(\boldsymbol{x}^j) = \sum_{j=1}^{p} Var(\boldsymbol{c}_V^j) = \sum_{j=1}^{p} \lambda_j(S).$$

Inertie (cas des données centrées et réduites ($M = D_{1/s^2}$)) :

$$\mathcal{I} = trace(R) = p = \sum_{j=1}^{p} Var(\boldsymbol{z}^j) = \sum_{j=1}^{p} Var(\boldsymbol{c}_W^j) = \sum_{j=1}^{p} \lambda_j(R).$$

λ_j représente l'inertie des points projetés sur le j-ème axe du nouveau repère, c'est-à-dire sur la j-ème composante principale.

Qualité de représentation de l'individu i sur l'axe D_1 : $QLT_{i1} = \frac{c_{i1}^2}{\sum_j c_{ij}^2}$.
Qualité de représentation de l'individu i sur l'axe D_2 : $QLT_{i2} = \frac{c_{i2}^2}{\sum_j c_{ij}^2}$.
Qualité de représentation de l'individu i sur le plan $(D_1; D_2)$: $QLT_{i,1-2} = QLT_{i1} + QLT_{i2}$.

Contribution de l'individu i à l'axe 1 : $CTR_{i1} = \frac{\frac{1}{n} c_{i1}^2}{\lambda_1}$.
Contribution de l'individu i à l'axe 2 : $CTR_{i2} = \frac{\frac{1}{n} c_{i2}^2}{\lambda_2}$.
Contribution de l'individu i au premier plan : $CTR_{i,1-2} = \frac{\frac{1}{n} c_{i1}^2 + \frac{1}{n} c_{i2}^2}{\lambda_1 + \lambda_2}$.

Mettre des références biblio!!!

Table des matières