

Université Mohamed Khider, Biskra
Faculté des Sc. Exactes et Sc. de la Nature et la Vie
Département de Mathématiques
Master 1: 2023/2024

Analyse en Composantes Principales (ACP)

Abdelhakim Necir

Il est important de souligner que ce polycopié est la composition des documents, ci-dessous, que nous l'avons enrichi par des rappels de certaines notions d'algèbre linéaire et quelques applications.

Charlotte Baey (2019). Analyse de données. <https://baeyc.github.io/teaching/>

Alboukadel Kassambara (2019). <http://www.sthda.com/french/>

François Husson, Sébastien Lê et Jérôme Pagès (2009). Analyse des données avec R, Presses Universitaires de Rennes, 224 p. (ISBN 978-2-7535-0938-2)

1. Introduction

L'**analyse en composantes principales (ACP)**, ou *principal component analysis (PCA)* en anglais, permet d'analyser et de visualiser un jeu de données contenant des individus décrits par plusieurs variables quantitatives, voir **Alboukadel Kassambara (2019)**.

C'est une méthode statistique qui permet d'explorer des données dites multivariées (données avec plusieurs variables). Chaque variable pourrait être considérée comme une dimension différente. Si vous avez plus de 3 variables dans votre jeu de données, il pourrait être très difficile de visualiser les données dans une "hyper-espace" multidimensionnelle.

L'analyse en **composantes principales** est utilisée pour extraire et de visualiser les informations importantes contenues dans une table de données multivariées. L'ACP synthétise cette information en seulement quelques nouvelles variables appelées composantes principales. Ces nouvelles variables correspondent à une combinaison linéaire des variables originels. Le nombre de composantes principales est inférieur ou égal au nombre de variables d'origine.

L'information contenue dans un jeu de données correspond à la variance ou l'*inertie totale* qu'il contient. L'objectif de l'ACP est d'identifier les directions (i.e., axes principaux ou composantes principales) le long desquelles la variation des données est maximale.

En d'autres termes, l'ACP réduit les dimensions d'une donnée multivariée à deux ou trois composantes principales, qui peuvent être visualisées graphiquement, en perdant le moins possible d'information. Lorsque le nombre de variables observées est élevé, il devient difficile de proposer une représentation graphique des données et des relations existant entre les variables.

2. Notations

Les notations suivantes seront utilisées par la suite :

- Une matrice de m lignes et p colonnes: $\mathcal{M}(m \times p)$.
- Un élément (ou vecteur) de l'espace \mathbb{R}^m :

$$\mathbf{e} := \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_m \end{pmatrix} \in \mathcal{M}(m \times 1).$$

- Le vecteur unitaire de \mathbb{R}^m :

$$\mathbf{1}_m := \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathcal{M}(m \times 1).$$

- La matrice-poids est une matrice diagonale de $\mathbb{R}^m \times \mathbb{R}^m$:

$$\mathbf{D}_\ell := \begin{pmatrix} \ell_1 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & \ell_{m-1} & 0 \\ 0 & \cdots & 0 & \ell_m \end{pmatrix} \in \mathcal{M}(m \times m)$$

où $\ell = (\ell_1, \dots, \ell_m)^t$.

- La matrice identité \mathbf{Id}_m est une matrice \mathbf{D}_1 :

$$\mathbf{Id}_m := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} \in \mathcal{M}(m \times m).$$

- La norme d'un vecteurs $\mathbf{e} = (x_1, \dots, x_m)^t$ de \mathbb{R}^m :

$$\|\mathbf{e}\| = \sqrt{x_1^2 + \dots + x_m^2}.$$

- La distance entre deux point $\mathbf{e} = (x_1, \dots, x_m)^t$ et $\mathbf{f} = (y_1, \dots, y_m)^t$ de \mathbb{R}^m :

$$d(\mathbf{e}, \mathbf{f}) = \|\mathbf{e} - \mathbf{f}\| = \sqrt{(x_1 - y_1)^2 + \dots + (x_m - y_m)^2}.$$

- Le produit scalaire de deux vecteurs $\mathbf{e} = (x_1, \dots, x_m)^t$ et $\mathbf{f} = (y_1, \dots, y_m)^t$ de \mathbb{R}^m :

$$\langle \mathbf{e}, \mathbf{f} \rangle = \mathbf{e}^t \mathbf{f} = x_1 y_1 + \dots + x_m y_m = \langle \mathbf{f}, \mathbf{e} \rangle.$$

En particulier $\langle \mathbf{e}, \mathbf{e} \rangle = \|\mathbf{e}\|^2$.

- L'opérateur de projection orthogonal d'un vecteur $\mathbf{e} = (x_1, \dots, x_m)^t$ de \mathbb{R}^m sur une droite vectorielle dirigée par le vecteur \mathbf{u} :

$$\mathbf{Proj}_{\mathbf{u}} \mathbf{e} = \frac{\langle \mathbf{e}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} = \frac{\langle \mathbf{e}, \mathbf{u} \rangle}{\|\mathbf{u}\|^2} \mathbf{u} = \frac{\mathbf{e}^t \mathbf{u}}{\mathbf{u}^t \mathbf{u}} \mathbf{u}.$$

En particulier, si \mathbf{u} est un vecteur unitaire, $\|\mathbf{u}\| = 1$, alors

$$\mathbf{Proj}_{\mathbf{u}} \mathbf{e} = \langle \mathbf{e}, \mathbf{u} \rangle \mathbf{u}.$$

- Une famille de vecteurs $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ est dite orthogonale si: $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$, pour $i \neq j$.
- Une famille de vecteurs $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ est dite orthonormale si: $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$, pour $i \neq j$ et $\langle \mathbf{u}_i, \mathbf{u}_i \rangle = 1$.
- Si $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ est une famille libre de l'espace vectoriel euclidien \mathbb{R}^m (donc une base, mais n'est pas nécessairement orthogonale), alors en utilisant le procédé de Gram-Schmidt, on peut construire de cette dernière, une autre famille $\{\mathbf{u}_1^*, \dots, \mathbf{u}_m^*\}$ formant une base orthonormale du même espace, via ce

processus:

$$\begin{array}{lll}
 \mathbf{u}_1 = \mathbf{v}_1 & \rightarrow & \mathbf{u}_1^* = \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|} \\
 \mathbf{u}_2 = \mathbf{v}_2 - \mathbf{Proj}_{\mathbf{u}_1} \mathbf{v}_2 & \rightarrow & \mathbf{u}_2^* = \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} \\
 \mathbf{u}_3 = \mathbf{v}_3 - \mathbf{Proj}_{\mathbf{u}_1} \mathbf{v}_3 - \mathbf{Proj}_{\mathbf{u}_2} \mathbf{v}_3 & \rightarrow & \mathbf{u}_3^* = \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|} \\
 \vdots & \vdots & \vdots \\
 \mathbf{u}_m = \mathbf{v}_m - \sum_{j=1}^{m-1} \mathbf{Proj}_{\mathbf{u}_j} \mathbf{v}_m & \rightarrow & \mathbf{u}_m^* = \frac{\mathbf{u}_m}{\|\mathbf{u}_m\|}
 \end{array}$$

- Deux matrices carrées \mathbf{A} et \mathbf{B} sont dites semblables s'il existe une matrice inversible \mathbf{P} telle que $\mathbf{A} = \mathbf{PBP}^{-1}$, où \mathbf{P}^{-1} est la matrice inverse de \mathbf{P} .
- Une matrice diagonalisable est une matrice carrée semblable à une matrice diagonale. Cette propriété est équivalente à l'existence **d'une base de vecteurs propres**, ce qui permet de définir de manière analogue un endomorphisme diagonalisable d'un espace vectoriel. Plus précisément, une matrice carrée \mathbf{M} à coefficients dans un corps \mathbf{K} est dite diagonalisable sur \mathbf{K} s'il existe une matrice inversible \mathbf{P} et une matrice diagonale \mathbf{D} à coefficients dans \mathbf{K} satisfaisant la relation :

$$\mathbf{M} = \mathbf{PDP}^{-1},$$

Dans ce cas, chaque vecteur colonne \mathbf{v} de la matrice \mathbf{P} est un vecteur propre pour la matrice \mathbf{M} , c'est-à-dire qu'il existe un scalaire λ sur la diagonale de \mathbf{D} tel que $\mathbf{Mv} = \lambda\mathbf{v}$.

- **Théorème de diagonalisation:** Une matrice $\mathbf{A} \in \mathcal{M}(m \times m)$ est diagonalisable si et seulement si tout ses **vecteurs propres sont linéairement indépendants**.
- Une matrice carrée $\mathbf{M} \in \mathcal{M}(m \times m)$ **orthogonale** est une matrice **unitaire à coefficients réels**, c'est à dire elle vérifie $\mathbf{M}^t\mathbf{M} = \mathbf{M}\mathbf{M}^t = \mathbf{Id}_m$. Une matrice \mathbf{M} est orthogonale si et seulement si elle est inversible et son inverse est égale à sa transposée : $\mathbf{M}^{-1} = \mathbf{M}^t$. Une matrice carrée est orthogonale si et seulement si ses vecteurs colonnes sont orthogonaux deux à deux et de norme 1. Ainsi une matrice orthogonale représente une base orthonormale. Le déterminant d'une matrice orthogonale est de carré 1, c'est-à-dire qu'il est égal à +1 ou -1.
- Une matrice symétrique \mathbf{M} est une matrice carrée qui est égale à sa propre transposée: $\mathbf{M} = \mathbf{M}^t$.

- Pour une matrice \mathbf{M} carrée symétrique réelle, il existe une matrice diagonale \mathbf{D} , formée des valeurs propres, et une matrice \mathbf{P} orthogonale, en colonnes formée des vecteurs propres orthonormés, telle que $\mathbf{M} = \mathbf{P}\mathbf{D}\mathbf{P}^t$, où \mathbf{P}^t est la matrice transposée de \mathbf{P} .
- Les vecteurs propres d'une matrice symétrique réelle qui correspondent à des valeurs propres distinctes **sont orthogonaux deux à deux**.
- Les matrices symétriques réelles sont diagonalisables par une matrice orthogonale.
- Une matrice $\mathcal{M}(m \times m)$ qui à m valeurs propres distinctes est diagonalisable, en particulier les matrices symétriques réelles

3. Procédure de l'ACP

3.1. Matrice des observations. Le point de départ de l'ACP est un nuage de points $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ de \mathbb{R}^p . Chaque point correspondant à un individu pour lequel on observe p variables quantitatives $(\mathbf{X}_1, \dots, \mathbf{X}_p)$. On note x_{ij} la valeur de la j -ème variable \mathbf{X}_j mesurée sur le i -ème individu \mathbf{e}_i , ainsi $\mathbf{X}_j = (x_{1j}, \dots, x_{nj})^t \in \mathbb{R}^n$ avec $1 \leq i \leq n$ et $1 \leq j \leq p$, désigne le vecteur contenant les observations de la variable j pour tous les individus, et $\mathbf{e}_i = (x_{i1}, \dots, x_{ip})^t \in \mathbb{R}^p$ désigne le vecteur des observations de l'individu i . Ces observations sont rassemblées dans une matrice notée \mathbf{X}^* , dont les colonnes correspondent aux vecteurs $(\mathbf{X}_1, \dots, \mathbf{X}_p)$:

$$\mathbf{X}^* = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \in \mathcal{M}(n \times p).$$

On obtient donc en ligne, les observations de chaque individu, et en colonne, les observations de chaque variable.

Application 3.1. *Les notes, sur 10 de 6 élèves, en trois matières; mathématiques, Français et sciences naturelles:*

$$\mathbf{X}^* = \begin{pmatrix} 8 & 1 & 0 \\ 4 & 6 & 5 \\ 6 & 8 & 7 \\ 10 & 4 & 7 \\ 8 & 2 & 5 \\ 0 & 3 & 6 \end{pmatrix} \in \mathcal{M}(6 \times 3).$$

3.2. Choix de l'origine et réduction des variables. Afin de faciliter les représentations graphiques du nuage de points, on choisit en général comme origine le centre de gravité du nuage. Le nuage se retrouve alors centré autour de cette nouvelle origine. Pour cela, on définit le centre de gravité du nuage des individus par :

$$\mathbf{g} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_{i1} \\ \frac{1}{n} \sum_{i=1}^n x_{ij} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n x_{ip} \end{pmatrix} \in \mathcal{M}(p, 1).$$

Autrement dit, le centre de gravité \mathbf{g} est le vecteur dont la j -ème coordonnée g_j qui correspond à la valeur moyenne de la variable j sur les n individus. On se retrouve alors avec une version *centrée* de la matrice \mathbf{X}^* , que l'on note

$$\mathbf{X} = \mathbf{X}^* - \mathbf{1}_n \mathbf{g}^t.$$

Pus précisément, on a

$$\mathbf{1}_n \mathbf{g}^t = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} g_1 & g_2 & \cdots & g_p \end{pmatrix} = \begin{pmatrix} g_1 & \cdots & g_p \\ \vdots & \ddots & \vdots \\ g_1 & \cdots & g_p \end{pmatrix} \in \mathcal{M}(n, p),$$

ansi

$$\mathbf{X} = \begin{pmatrix} x_{11} - g_1 & \cdots & x_{1p} - g_p \\ \vdots & \ddots & \vdots \\ x_{n1} - g_1 & \cdots & x_{np} - g_p \end{pmatrix} \in \mathcal{M}(n, p).$$

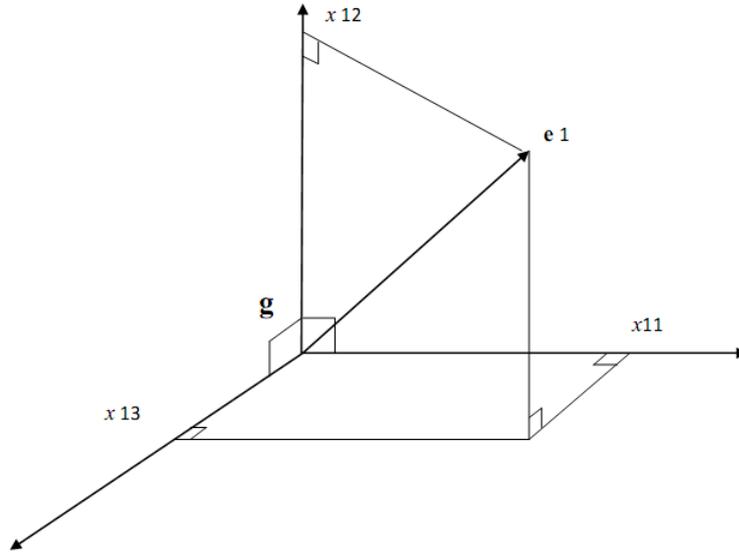


Fig.1

De même, il est souvent plus commode de travailler avec des variables réduites (c'est-à-dire dont l'écart-type vaut 1), afin de les rendre comparables entre elles et de gommer les effets d'échelle. En d'autres termes, on divise les coordonnées de chaque colonne par l'écart-type correspondant. Plus précisément, notons

$$s_j^2 := \frac{1}{n} \sum_{i=1}^n (x_{ij} - g_j)^2, \quad j = 1, \dots, p,$$

la variance empirique de la variable \mathbf{X}_j , et $1/s = (1/s_1, \dots, 1/s_p)^t$. On définit la matrice poids suivante:

$$\mathbf{D}_{1/s} := \begin{pmatrix} 1/s_1 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & 1/s_{p-1} & 0 \\ 0 & \cdots & 0 & 1/s_p \end{pmatrix} \in \mathcal{M}(p \times p).$$

La version centrée et réduite de la matrice des observations est

$$\mathbf{Z} = \mathbf{X}\mathbf{D}_{1/s} = \begin{pmatrix} \frac{x_{11}-g_1}{s_1} & \cdots & \frac{x_{1p}-g_p}{s_p} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1}-g_1}{s_1} & \cdots & \frac{x_{np}-g_p}{s_p} \end{pmatrix} \in \mathcal{M}(n, p).$$

On définit, respectivement, les matrices de variance-covariance et de corrélation associées au vecteur $(\mathbf{X}_1, \dots, \mathbf{X}_p)$, par

$$\mathbf{V} = \frac{1}{n} \mathbf{X}^t \mathbf{X} \in \mathcal{M}(p \times p) \quad \text{et} \quad \mathbf{R} = \frac{1}{n} \mathbf{Z}^t \mathbf{Z} \in \mathcal{M}(p \times p).$$

On note que les deux matrices \mathbf{V} et \mathbf{R} sont symétriques:

$$\mathbf{V}^t = \frac{1}{n} (\mathbf{X}^t \mathbf{X})^t = \frac{1}{n} \mathbf{X}^t \mathbf{X} = \mathbf{V}$$

et

$$\mathbf{R}^t = \frac{1}{n} (\mathbf{Z}^t \mathbf{Z})^t = \frac{1}{n} \mathbf{Z}^t \mathbf{Z} = \mathbf{R}.$$

Application: Pour notre exemple, le centre de gravité du nuage est

$$\mathbf{g} = \frac{1}{6} \begin{pmatrix} 8 + 4 + 6 + 10 + 8 + 0 \\ 1 + 6 + 8 + 4 + 2 + 3 \\ 0 + 5 + 7 + 7 + 5 + 6 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 36 \\ 24 \\ 30 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \\ 5 \end{pmatrix},$$

ainsi la matrice centrée du nuage est

$$\mathbf{X} = \begin{pmatrix} 8 - 6 & 1 - 4 & 0 - 5 \\ 4 - 6 & 6 - 4 & 5 - 5 \\ 6 - 6 & 8 - 4 & 7 - 5 \\ 10 - 6 & 4 - 4 & 7 - 5 \\ 8 - 6 & 2 - 4 & 5 - 5 \\ 0 - 6 & 3 - 4 & 6 - 5 \end{pmatrix} = \begin{pmatrix} 2 & -3 & -5 \\ -2 & 2 & 0 \\ 0 & 4 & 2 \\ 4 & 0 & 2 \\ 2 & -2 & 0 \\ -6 & -1 & 1 \end{pmatrix}.$$

Calculons la matrice de variance-covariance \mathbf{V} . Le transposé de \mathbf{X} est

$$\mathbf{X}^t = \begin{pmatrix} 2 & -2 & 0 & 4 & 2 & -6 \\ -3 & 2 & 4 & 0 & -2 & -1 \\ -5 & 0 & 2 & 2 & 0 & 1 \end{pmatrix},$$

donc

$$\mathbf{V} = \frac{1}{6} \mathbf{X}^t \mathbf{X} = \frac{1}{6} \begin{pmatrix} 2 & -2 & 0 & 4 & 2 & -6 \\ -3 & 2 & 4 & 0 & -2 & -1 \\ -5 & 0 & 2 & 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & -3 & -5 \\ -2 & 2 & 0 \\ 0 & 4 & 2 \\ 4 & 0 & 2 \\ 2 & -2 & 0 \\ -6 & -1 & 1 \end{pmatrix}.$$

Après le calcul on trouve

$$\mathbf{V} = \begin{pmatrix} 10.667 & -1.333 & -1.333 \\ -1.333 & 5.666 & 3.666 \\ -1.333 & 3.666 & 5.666 \end{pmatrix}.$$

Calculons la matrice de corrélation \mathbf{R} . L'écart-type de chaque variable colonne sont:

$$\begin{aligned} s_1 &= \sqrt{\frac{1}{6} ((8-6)^2 + (4-6)^2 + (6-6)^2 + (10-6)^2 + (8-6)^2 + (0-6)^2)} \\ &= 3.2660 \left(\text{vérification } \sqrt{10.667} = 3.266 \right), \end{aligned}$$

$$\begin{aligned} s_2 &= \sqrt{\frac{1}{6} ((1-4)^2 + (6-4)^2 + (8-4)^2 + (4-4)^2 + (2-4)^2 + (3-4)^2)} \\ &= 2.3805 \left(\text{vérification } \sqrt{5.6667} = 2.3805 \right), \end{aligned}$$

$$\begin{aligned} s_3 &= \sqrt{\frac{1}{6} ((0-5)^2 + (5-5)^2 + (7-5)^2 + (7-5)^2 + (5-5)^2 + (6-5)^2)}, \\ &= 2.3805 \left(\text{vérification } \sqrt{5.6667} = 2.3805 \right). \end{aligned}$$

Par conséquent $1/s = (0.306, 0.420, 0.420)^t$, et la matrice poids est

$$\mathbf{D}_{1/s} = \begin{pmatrix} \frac{1}{s_1} & 0 & 0 \\ 0 & \frac{1}{s_2} & 0 \\ 0 & 0 & \frac{1}{s_3} \end{pmatrix} = \begin{pmatrix} 0.306 & 0 & 0 \\ 0 & 0.420 & 0 \\ 0 & 0 & 0.420 \end{pmatrix}.$$

Ainsi, la matrice de nuage centré-réduit est

$$\mathbf{Z} = \mathbf{XM} = \begin{pmatrix} 2 & -3 & -5 \\ -2 & 2 & 0 \\ 0 & 4 & 2 \\ 4 & 0 & 2 \\ 2 & -2 & 0 \\ -6 & -1 & 1 \end{pmatrix} \begin{pmatrix} 0.306 & 0 & 0 \\ 0 & 0.420 & 0 \\ 0 & 0 & 0.420 \end{pmatrix}.$$

Le calcul donne

$$\mathbf{Z} = \begin{pmatrix} 0.612 & -1.260 & -2.100 \\ -0.612 & 0.840 & 0 \\ 0 & 1.680 & 0.840 \\ 1.224 & 0 & 0.840 \\ 0.612 & -0.840 & 0 \\ -1.837 & -0.420 & 0.420 \end{pmatrix}$$

Ainsi la matrice de corrélation est

$$\mathbf{R} = \frac{1}{6} \mathbf{Z}^t \mathbf{Z} = \frac{1}{6} \begin{pmatrix} 0.612 & -0.612 & 0 & 1.224 & 0.612 & -1.837 \\ -1.260 & 0.840 & 1.680 & 0 & -0.840 & -0.420 \\ -2.100 & 0 & 0.840 & 0.840 & 0 & 0.420 \end{pmatrix} \times \begin{pmatrix} 0.612 & -1.260 & -2.100 \\ -0.612 & 0.840 & 0 \\ 0 & 1.680 & 0.840 \\ 1.224 & 0 & 0.840 \\ 0.612 & -0.840 & 0 \\ -1.837 & -0.420 & 0.420 \end{pmatrix}$$

Le calcul donne

$$\mathbf{R} = \begin{pmatrix} 0.999 & -0.171 & -0.171 \\ -0.171 & 0.999 & 0.647 \\ -0.171 & 0.647 & 0.999 \end{pmatrix}.$$

3.3. Espace des individus et espace des variables. Deux points de vue peuvent être adoptés pour l'étude du nuage de points :

- on peut associer à chaque individu i le vecteur \mathbf{e}_i contenant ses observations sur les p variables (i.e. la i -ème ligne de \mathbf{X}). Ce vecteur appartient à un espace vectoriel de dimension p : c'est l'espace des individus.

- on peut associer à chaque variable j le vecteur \mathbf{X}_j contenant les observations de la variable j (i.e. la j -ème colonne de \mathbf{X}). Ce vecteur appartient à un espace vectoriel de dimension n : c'est l'espace des variables.

3.4. Choix d'une distance. L'étape suivante consiste à introduire une distance entre les points de chacun des deux espaces définis ci-dessus. En ACP, on utilise par convention la distance euclidienne classique. Autrement dit, la distance entre deux individus, représentés par les vecteurs \mathbf{e}_i et $\mathbf{e}_{i'}$, est donnée par :

$$d^2(\mathbf{e}_i, \mathbf{e}_{i'}) = \sum_{j=1}^p (x_{ij} - x_{i'j})^2.$$

Le choix de la distance euclidienne permet de donner le même poids à chacune des variables. Cependant, il est possible de considérer une métrique définie par une matrice symétrique définie positive $\mathbf{M} \in \mathcal{M}(p \times p)$ de la façon suivante:

$$d^2(\mathbf{e}_i, \mathbf{e}_{i'}) = (\mathbf{e}_i - \mathbf{e}_{i'})^t \mathbf{M} (\mathbf{e}_i - \mathbf{e}_{i'}).$$

La distance euclidienne classique correspond donc au cas $\mathbf{M} = \mathbf{Id}_p$.

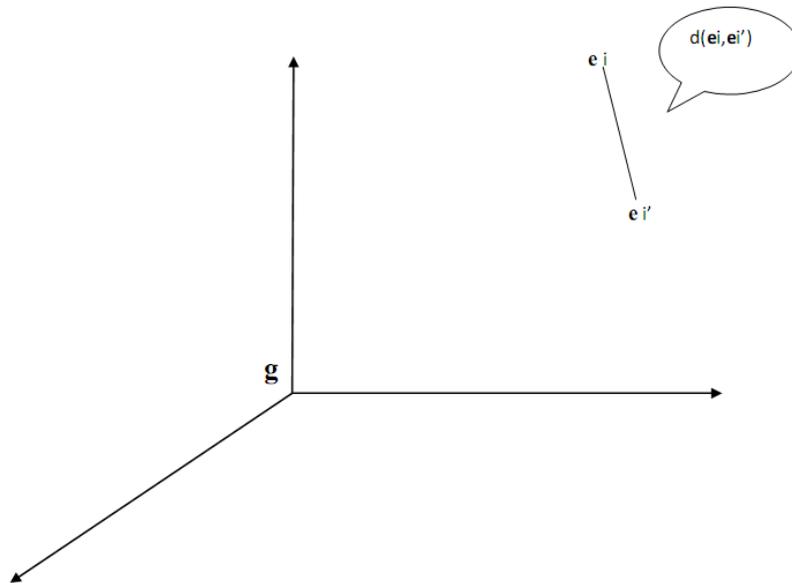


Fig.2

3.5. **Inertie.** On définit ensuite la notion d'inertie totale:

$$I_T = \frac{1}{n} \sum_{i=1}^n d^2(\mathbf{e}_i, \mathbf{g}) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - g_j)^2.$$

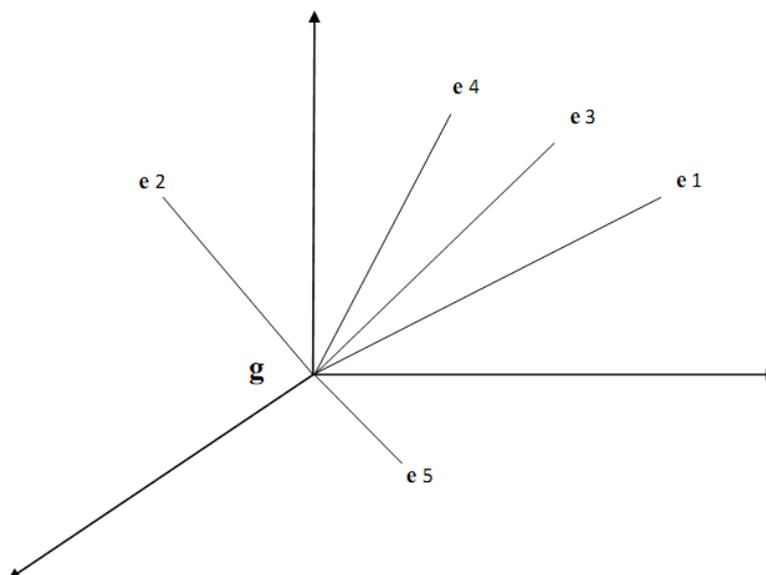


Fig. 3

L'inertie *mesure la dispersion des points du nuage* par rapport à son centre de gravité: plus l'inertie est grande, plus le nuage est dispersé, et à l'inverse plus l'inertie est petite, plus le nuage est concentré autour de son centre de gravité. L'inertie du

nuage de points I_T s'interprète facilement en termes de variance. En effet, on a :

$$\begin{aligned} I_T &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - g_j)^2 = \sum_{j=1}^p \left[\frac{1}{n} \sum_{i=1}^n (x_{ij} - g_j)^2 \right] \\ &= \sum_{j=1}^p \left[\frac{1}{n} \sum_{i=1}^n \left(x_{ij} - \frac{1}{n} \sum_{i=1}^n x_{ij} \right)^2 \right], \end{aligned}$$

ainsi on a

$$I_T = \sum_{j=1}^p s_j^2.$$

L'inertie totale du nuage de points est donc égale à la somme des variances des variables. En d'autres termes,

$$I_T = \text{Trace } \mathbf{V}.$$

Notons que si les variables sont réduites, i.e. si leur écart-type vaut 1, alors la trace de la matrice \mathbf{V} vaut exactement p . Autrement dit, ce n'est plus la matrice de variance-covariance \mathbf{V} qui joue un rôle, mais la matrice de corrélation \mathbf{R} . Il est aussi important de noter que les valeurs propres de \mathbf{V} et de \mathbf{R} pas nécessairement des mêmes. Donc l'inertie totale basée sur la matrice des données \mathbf{Z} est en général différente de celle basée sur \mathbf{X} . Mais le résultat final de l'analyse, que ce soit via \mathbf{X} ou \mathbf{Z} , est le même.

Application:

- Si le nuage est centré, on a:

$$I_T = (3.2660)^2 + (2.3805)^2 + (2.3805)^2 = 22.$$

- Si le nuage est centré-réduit, on a:

$$I_T = 3 \text{ (nombre de colonnes).}$$

Inertie par rapport à un sous-espace vectoriel

Notons \mathbf{E} un sous-espace vectoriel passant par le centre de gravité \mathbf{g} . L'inertie du nuage de points par rapport à ce sous-espace vectoriel \mathbf{E} est définie par:

$$I_{\mathbf{E}} = \frac{1}{n} \sum_{i=1}^n d^2(\mathbf{e}_i, \text{Proj}_{\mathbf{E},i}),$$

où $\text{Proj}_{\mathbf{E},i}$ désigne la projection orthogonale de \mathbf{e}_i sur \mathbf{E} .

Décomposition de l'inertie

Considérons une décomposition de \mathbb{R}^p en deux sous-espaces vectoriels supplémentaires, orthogonaux, et passant par et passant par \mathbf{g} , c'est à dire

$$\mathbb{R}^p = \mathbf{E} \oplus \mathbf{E}^\perp,$$

où \mathbf{E}^\perp désigne l'espace orthogonal à \mathbf{E} . On peut montrer facilement que :

$$I_T = I_E + I_{E^\perp}. \quad (3.1)$$

En effet, par le théorème de Pythagore, on a

$$d^2(\mathbf{e}_i, \mathbf{Proj}_{E,i}) + d^2(\mathbf{e}_i, \mathbf{Proj}_{E^\perp,i}) = d^2(\mathbf{e}_i, \mathbf{g}).$$

En projetant le nuage de points sur un sous-espace \mathbf{E} , on perd l'inertie mesurée par I_E , et on ne conserve que celle mesurée par I_{E^\perp} .

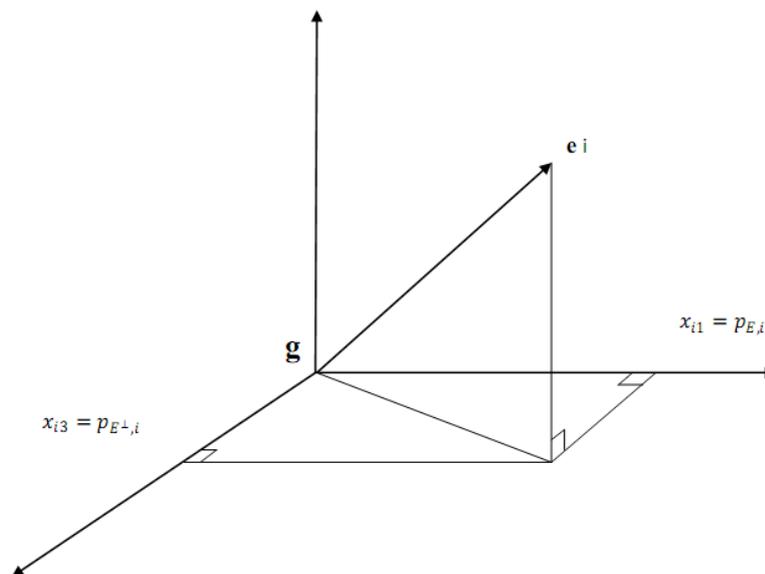


Fig.4

Dans le cas particulier où \mathbf{E} est un axe (i.e. un sous-espace vectoriel de dimension 1), I_{E^\perp} mesure l'étirement du nuage selon cet axe. On parle également de l'inertie portée par l'axe \mathbf{E} . Si on note \mathbf{u} le vecteur directeur unitaire de l'axe \mathbf{E} , i.e. $\|\mathbf{u}\| = 1$, on a l'expression matricielle suivante pour l'inertie expliquée par cet axe :

$$I_{E^\perp} = \mathbf{u}^t \mathbf{V} \mathbf{u} = \langle \mathbf{u}, \mathbf{V} \mathbf{u} \rangle. \quad (3.2)$$

En effet,

$$\begin{aligned}
I_{E^\perp} &= \frac{1}{n} \sum_{i=1}^n d^2(\mathbf{e}_i, \mathbf{Proj}_{E^\perp, i}) \\
&= \frac{1}{n} \sum_{i=1}^n \|\mathbf{e}_i - \mathbf{Proj}_{E^\perp, i}\|^2 = \frac{1}{n} \sum_{i=1}^n \|\mathbf{Proj}_{E, i}\|^2 \text{ (relation de Chasles)} \\
&= \frac{1}{n} \sum_{i=1}^n \left\| \frac{\langle \mathbf{e}_i, \mathbf{u} \rangle}{\|\mathbf{u}\|^2} \mathbf{u} \right\|^2 = \frac{1}{n} \sum_{i=1}^n \|\langle \mathbf{e}_i, \mathbf{u} \rangle \mathbf{u}\|^2 \text{ (car } \|\mathbf{u}\| = 1) \\
&= \frac{1}{n} \sum_{i=1}^n \langle \mathbf{e}_i, \mathbf{u} \rangle^2 \|\mathbf{u}\|^2 = \frac{1}{n} \sum_{i=1}^n \langle \mathbf{e}_i, \mathbf{u} \rangle^2 = \frac{1}{n} \sum_{i=1}^n (\mathbf{e}_i^t \mathbf{u})^2.
\end{aligned}$$

Comme $\mathbf{e}_i^t \mathbf{u}$ est un scalaire, alors $\mathbf{e}_i^t \mathbf{u} = (\mathbf{e}_i^t \mathbf{u})^t = \mathbf{u}^t \mathbf{e}_i$, par conséquent

$$I_{E^\perp} = \frac{1}{n} \sum_{i=1}^n (\mathbf{u}^t \mathbf{e}_i) (\mathbf{e}_i^t \mathbf{u}) = \mathbf{u}^t \left[\frac{1}{n} \sum_{i=1}^n \mathbf{e}_i \mathbf{e}_i^t \right] \mathbf{u}.$$

D'après la notion du produit matricielle:

$$\mathbf{X}^t \mathbf{X} = \sum_{i=1}^n \mathbf{e}_i \mathbf{e}_i^t,$$

ainsi

$$I_{E^\perp} = \mathbf{u}^t \left[\frac{1}{n} \mathbf{X}^t \mathbf{X} \right] \mathbf{u} = \mathbf{u}^t \mathbf{V} \mathbf{u}.$$

La décomposition de l'inertie obtenue en (3.1) est connue sous le nom de **Théorème de Huygens**. On peut le généraliser au cas d'une décomposition de \mathbb{R}^p comme somme de p sous-espaces de dimension 1 et orthogonaux entre eux :

$$\mathbb{R}^p = \mathbf{E}_1 \oplus \mathbf{E}_2 \oplus \dots \oplus \mathbf{E}_p.$$

Ainsi

$$I_T = I_{E_1^\perp} + I_{E_2^\perp} + \dots + I_{E_p^\perp}.$$

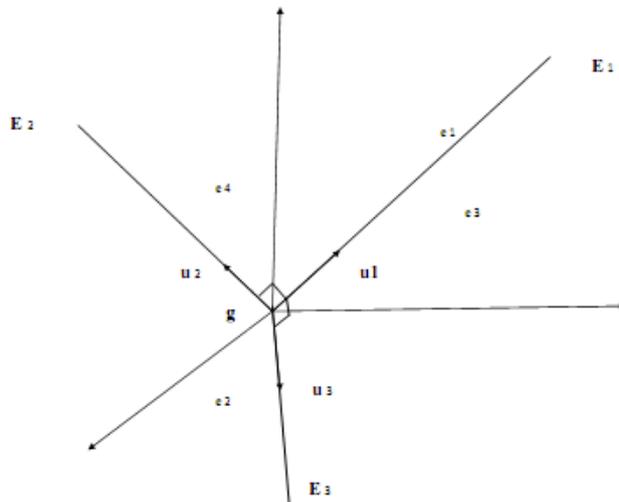


Fig. 5

3.6. Principes généraux. L'objectif d'une analyse en composantes principales est de projeter le nuage de points sur des sous-espaces affines de dimension plus petite, tels que la projection selon chacun de ces sous-espaces soit la plus fidèle possible. Autrement dit, on souhaite minimiser la distance entre les points du nuage initial et leurs projections selon le sous-espace considéré. En reprenant les notations de la section précédente, on cherche alors un sous-espace \mathbf{E} tel que l'inertie I_E soit minimale, ou de façon équivalente, tel que l'inertie expliquée par \mathbf{E} , I_E soit maximale.

Plus précisément, on cherche en **ACP** une décomposition de \mathbb{R}^p en sous-espaces orthogonaux de dimension \mathbb{R}^p notés $\mathbf{E}_1, \dots, \mathbf{E}_p$ passant par le centre de gravité du nuage, et tels que l'inertie expliquée par chacun de ces axes soit maximale. Pour cela, on procède de façon itérative en cherchant d'abord le premier axe \mathbf{E}_1 tel que $I_{E_1^\perp}$ soit maximale, puis l'axe \mathbf{E}_2 orthogonal à \mathbf{E}_1 et tel que $I_{E_2^\perp}$ soit maximale, et ainsi de suite.

Autrement dit, on effectue un changement de repère dans \mathbb{R}^p tel que, dans ce nouveau système de représentation, le premier axe explique le plus possible l'inertie totale du nuage, le deuxième axe le plus possible l'inertie non prise en compte par le premier axe, et ainsi de suite.

3.6.1. Recherche du premier axe principal. On cherche donc un axe \mathbf{E}_1 , passant par \mathbf{g} , et tel que I_{E_1} soit minimum, ou de façon équivalente tel que $I_{E_1^\perp}$ soit maximum. Notons \mathbf{u}_1 le vecteur directeur unitaire de l'axe \mathbf{E}_1 . On cherche donc \mathbf{u}_1 tel que

$I_{E_1^\perp}$ soit maximum sous la contrainte $\|\mathbf{u}_1\|^2 = 1$. On obtient alors le problème d'optimisation sous contrainte suivant :

$$\left\{ \begin{array}{l} \max_{\mathbf{u}_1} I_{E_1^\perp} \\ \|\mathbf{u}_1\|^2 = 1 \end{array} \right. \iff \left\{ \begin{array}{l} \max_{u_1} \mathbf{u}_1^t \mathbf{V} \mathbf{u}_1 \\ \mathbf{u}_1^t \mathbf{u}_1 = 1 \end{array} \right. .$$

Pour résoudre ce problème de maximisation sous contraintes, on utilise la méthode des multiplicateurs de Lagrange. On doit alors trouver \mathbf{u}_1 vérifiant

$$\frac{\partial}{\partial \mathbf{u}_1} (\mathbf{u}_1^t \mathbf{V} \mathbf{u}_1 - \lambda_1 (\mathbf{u}_1^t \mathbf{u}_1 - 1)) = 0.$$

En utilisant la dérivée matricielle

$$2\mathbf{V}\mathbf{u}_1 - 2\lambda_1 \mathbf{u}_1 = 0 \iff \mathbf{V}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1.$$

Autrement dit, \mathbf{u}_1 est un vecteur propre de la matrice de variance-covariance des données \mathbf{V} . Or, d'après l'équation (3.2) on a:

$$I_{E_1^\perp} = \mathbf{u}_1^t \mathbf{V} \mathbf{u}_1 = \mathbf{u}_1^t (\lambda_1 \mathbf{u}_1) = \lambda_1 \mathbf{u}_1^t \mathbf{u}_1 = \lambda_1 \|\mathbf{u}_1\|^2 = \lambda_1.$$

Comme on cherche à maximiser $I_{E_1^\perp}$ on va également chercher à maximiser λ_1 et on choisira donc la plus grande valeur propre de la matrice \mathbf{V} .

Finalement, le vecteur directeur unitaire de l'axe \mathbf{E}_1 est le vecteur propre associé à la plus grande valeur propre de la matrice \mathbf{V} .

3.6.2. *Recherche des axes suivants.* Une fois que le premier axe a été identifié, on cherche l'axe \mathbf{E}_2 , orthogonal à \mathbf{E}_1 et tel que l'inertie I_{E_2} soit minimale, ou de façon équivalent tel que $I_{E_2^\perp}$ soit maximale. En notant \mathbf{u}_2 le vecteur directeur unitaire de \mathbf{E}_2 , on doit alors résoudre le système suivant :

$$\left\{ \begin{array}{l} \max_{\mathbf{u}_2} \mathbf{u}_2^t \mathbf{V} \mathbf{u}_2 \\ \|\mathbf{u}_2\|^2 = \mathbf{u}_2^t \mathbf{u}_2 = 1 \\ \mathbf{u}_2 \perp \mathbf{u}_1 \iff \langle \mathbf{u}_1, \mathbf{u}_2 \rangle = \mathbf{u}_2^t \mathbf{u}_1 = 0. \end{array} \right. .$$

Par rapport au cas précédent, on a donc une contrainte de plus car le nouvel axe doit être orthogonal au premier. On utilise la encore la méthode des multiplicateurs de Lagrange. On cherche donc \mathbf{u}_2 qui vérifiant

$$\frac{\partial}{\partial \mathbf{u}_2} (\mathbf{u}_2^t \mathbf{V} \mathbf{u}_2 - \lambda_2 (\mathbf{u}_2^t \mathbf{u}_2 - 1) - \mu \mathbf{u}_2^t \mathbf{u}_1) = 0,$$

c'est-à-dire :

$$2\mathbf{V}\mathbf{u}_2 - 2\lambda_2 \mathbf{u}_2 - \mu \mathbf{u}_1 = 0. \tag{3.3}$$

En multipliant à gauche par \mathbf{u}_1^t et en utilisant le fait que \mathbf{u}_1 et \mathbf{u}_2 sont orthogonaux, et que \mathbf{u}_1 est de norme 1, on obtient :

$$\begin{aligned} 2\mathbf{u}_1^t \mathbf{V} \mathbf{u}_2 - 2\lambda_2 \mathbf{u}_1^t \mathbf{u}_2 - \mu \mathbf{u}_1^t \mathbf{u}_1 &= 0 \implies 2\mathbf{u}_1^t \mathbf{V} \mathbf{u}_2 - \mu \mathbf{u}_1^t \mathbf{u}_1 = 0 \\ \implies 2(\mathbf{V}^t \mathbf{u}_1)^t \mathbf{u}_2 - \mu \mathbf{u}_1^t \mathbf{u}_1 &= 0 \\ \implies 2(\mathbf{V} \mathbf{u}_1)^t \mathbf{u}_2 - \mu \mathbf{u}_1^t \mathbf{u}_1 &= 0 \\ \implies 2\lambda_1 \mathbf{u}_1^t \mathbf{u}_2 - \mu \mathbf{u}_1^t \mathbf{u}_1 &= 0 \\ \implies \mu &= 0. \end{aligned}$$

Finalement, en reprenant l'équation (3.3), on obtient :

$$\mathbf{V} \mathbf{u}_2 = \lambda_2 \mathbf{u}_2.$$

Le vecteur directeur unitaire de \mathbf{E}_2 est donc également un vecteur propre de \mathbf{V} . Comme on cherche à maximiser l'inertie $I_{\mathbf{E}_2^\perp}$ on va raisonner de même que pour le premier axe et maximiser λ_2 . On choisira donc pour \mathbf{u}_2 , le vecteur propre associé à la deuxième plus grande valeur propre de \mathbf{V} .

On raisonne de même pour trouver les axes suivants, dont les vecteurs directeurs unitaires sont tous des vecteurs propres de la matrice \mathbf{V} , associés aux valeurs propres ordonnées par ordre décroissant.

La matrice \mathbf{V} étant symétrique, elle possède bien p vecteurs propres qui forment une base orthogonale de \mathbb{R}^p .

Les axes $\mathbf{E}_1, \mathbf{E}_2, \dots, \mathbf{E}_p$ sont appelées axes factoriels, ou axes principaux d'inertie.

Application: calculons les valeurs propres de la matrice variance-covariance \mathbf{V} . Pour cela on cherche les racines du polynôme caractéristique de \mathbf{V} , défini par le déterminant de $\mathbf{V} - \lambda \mathbf{Id}_3$ que l'on note $\det(\mathbf{V} - \lambda \mathbf{Id}_3)$. On a

$$\mathbf{V} - \lambda \mathbf{Id}_3 = \begin{pmatrix} \frac{32}{3} & -\frac{4}{3} & -\frac{4}{3} \\ -\frac{4}{3} & \frac{17}{3} & \frac{11}{3} \\ -\frac{4}{3} & \frac{11}{3} & \frac{17}{3} \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Alors

$$\det(\mathbf{V} - \lambda \mathbf{Id}_3) = \begin{vmatrix} \frac{32}{3} - \lambda & -\frac{4}{3} & -\frac{4}{3} \\ -\frac{4}{3} & \frac{17}{3} - \lambda & \frac{11}{3} \\ -\frac{4}{3} & \frac{11}{3} & \frac{17}{3} - \lambda \end{vmatrix} = -\lambda^3 + 22\lambda^2 - 136\lambda + 192.$$

Les solutions de l'équation $-\lambda^3 + 22\lambda^2 - 136\lambda + 192 = 0$, sont

$$\lambda_1 = 12 > \lambda_2 = 8 > \lambda_3 = 2.$$

Le vecteur propre associé à la valeur propre $\lambda_1 = 12$ est une solution du système

$$\begin{pmatrix} \frac{32}{3} & -\frac{4}{3} & -\frac{4}{3} \\ -\frac{4}{3} & \frac{17}{3} & \frac{11}{3} \\ -\frac{4}{3} & \frac{11}{3} & \frac{17}{3} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = 12 \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

Ce système se réduit à $a = -2b$ et $b = c$. Observons que

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} -2b \\ b \\ b \end{pmatrix} = b \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}.$$

Donc $\mathbf{u}_1^* = \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}$ est le premier vecteur propre. Pour qu'il soit un vecteur

directeur du premier axe factoriel (principal), il doit être de norme 1. Nous avons

$$\|\mathbf{u}_1^*\| = \sqrt{(-2)^2 + 1^2 + 1^2} = \sqrt{6}, \text{ ainsi}$$

$$\mathbf{u}_1 = \mathbf{u}_1^* / \|\mathbf{u}_1^*\| = \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix} / \sqrt{6},$$

est le vecteur directeur (unitaire) du premier axe principal \mathbf{E}_1 . De la même manière, on trouve les vecteurs propres associés à $\lambda_2 = 8$ et $\lambda_3 = 2$ sont

$$\mathbf{u}_2^* = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ et } \mathbf{u}_3^* = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}.$$

En normalisant, on trouve

$$\mathbf{u}_2 = \mathbf{u}_2^* / \|\mathbf{u}_2^*\| = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} / \sqrt{3} \text{ et } \mathbf{u}_3 = \mathbf{u}_3^* / \|\mathbf{u}_3^*\| = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} / \sqrt{2}.$$

Les deux derniers vecteurs directeurs (unitaires) des deux axes principaux \mathbf{E}_2 et \mathbf{E}_3 sont orthogonaux deux à deux à \mathbf{E}_1 . On peut vérifier facilement que les produits scalaires:

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0, i \neq j \text{ et } \langle \mathbf{u}_i, \mathbf{u}_i \rangle = 1.$$

En d'autres termes $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ forme une base pour \mathbb{R}^3 , ainsi $\mathbb{R}^3 = \mathbf{E}_1 \oplus \mathbf{E}_2 \oplus \mathbf{E}_3$.

Remarque: Si les valeurs propres ne sont pas distinctes, alors les axes principaux peuvent être déterminés en utilisant le procédé de Gram-Schmidt.

3.6.3. *Contribution des axes à l'inertie totale.* Par le théorème de Huygens, on a :

$$I_T = I_{\mathbf{E}_1^\perp} + I_{\mathbf{E}_2^\perp} + \dots + I_{\mathbf{E}_p^\perp} = \lambda_1 + \lambda_2 + \dots + \lambda_p.$$

L'inertie expliquée par l'axe \mathbf{E}_j est :

$$I_{\mathbf{E}_j^\perp} = \lambda_j,$$

et le pourcentage d'inertie expliquée par cet axe, aussi appelé contribution relative, est égale à :

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \dots + \lambda_p}.$$

De la même façon, on peut définir le pourcentage d'inertie expliquée par le sous-espace

$$\mathbf{F}_h = \mathbf{E}_1 \oplus \mathbf{E}_2 \oplus \dots \oplus \mathbf{E}_h$$

par:

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_h}{\lambda_1 + \lambda_2 + \dots + \lambda_p}.$$

Ces pourcentages d'inertie expliquée permettent de rendre compte de la part de variabilité du nuage expliquée par le sous-espace considéré. En particulier, si les dernières valeurs propres sont faibles, les axes associés expliqueront une faible part de l'inertie totale et pourront être négligés. Dans la pratique, on représentera alors le nuage de points dans un sous-espace de dimension $d < p$, où d est choisi tel que l'inertie expliquée par les d premiers axes soit proche de 1, c'est à dire proche de 100/100.

Application:

- L'inertie totale du nuage $(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_6)$ par rapport au centre de gravité \mathbf{g} est:

$$I_T = \lambda_1 + \lambda_2 + \lambda_3 = 12 + 8 + 2 = 22.$$

- Les inerties expliquées par les axes \mathbf{E}_1 , \mathbf{E}_2 et \mathbf{E}_3 sont respectivement

$$I_{\mathbf{E}_1^\perp} = \lambda_1 = 12, \quad I_{\mathbf{E}_2^\perp} = \lambda_2 = 8, \quad \text{et} \quad I_{\mathbf{E}_3^\perp} = \lambda_3 = 2.$$

- Les pourcentages d'inerties expliquées (contribution relative) par \mathbf{E}_1 , \mathbf{E}_2 et \mathbf{E}_3 sont respectivement

$$I_{\mathbf{E}_1^\perp}(r) = \frac{12}{12 + 8 + 2} = 54\%, \quad I_{\mathbf{E}_2^\perp}(r) = \frac{8}{12 + 8 + 2} = 36\%$$

et

$$I_{\mathbf{E}_3^\perp}(r) = \frac{2}{12 + 8 + 2} = 9\%.$$

- Le pourcentage d'inertie expliquée par le sous-espace $\mathbf{F}_2 = \mathbf{E}_1 \oplus \mathbf{E}_2$ est

$$I_{\mathbf{F}_2^\perp} = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_p} = \frac{12 + 8}{12 + 8 + 2} = 90\%.$$

3.7. Représentation des individus dans les nouveaux axes. Notons y_{ij} la coordonnée de l'individu i selon l'axe \mathbf{E}_j . Par définition, on a :

$$y_{ij} = \langle \mathbf{e}_i, \mathbf{u}_j \rangle = \mathbf{e}_i^t \mathbf{u}_j.$$

En notant \mathbf{y}_i le vecteur des coordonnées de l'individu i dans la nouvelle base et U la matrice des vecteurs propres, on obtient alors :

$$\mathbf{y}_i^t = \mathbf{e}_i^t \mathbf{U}.$$

En notant \mathbf{Y} la matrice contenant en ligne les coordonnées des individus dans la nouvelle base, on a :

$$\mathbf{Y} = \mathbf{XU}, \quad (3.4)$$

Application: rappelons que la matrice des vecteurs propres est

$$\mathbf{U} = \begin{pmatrix} -2/\sqrt{6} & 1/\sqrt{3} & 0 \\ 1/\sqrt{6} & 1/\sqrt{3} & -1/\sqrt{2} \\ 1/\sqrt{6} & 1/\sqrt{3} & 1/\sqrt{2} \end{pmatrix}.$$

Les coordonnées des individus dans le nouveau repère $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ sont définis par les colonnes de cette matrice

$$\mathbf{Y} = \begin{pmatrix} 2 & -3 & -5 \\ -2 & 2 & 0 \\ 0 & 4 & 2 \\ 4 & 0 & 2 \\ 2 & -2 & 0 \\ -6 & -1 & 1 \end{pmatrix} \begin{pmatrix} -2/\sqrt{6} & 1/\sqrt{3} & 0 \\ 1/\sqrt{6} & 1/\sqrt{3} & -1/\sqrt{2} \\ 1/\sqrt{6} & 1/\sqrt{3} & 1/\sqrt{2} \end{pmatrix}.$$

Le calcul donne

$$\mathbf{Y} = \begin{pmatrix} -4.899 & -3.464 & -1.414 \\ 2.449 & 0 & -1.414 \\ 2.449 & 3.464 & -1.414 \\ -2.449 & 3.464 & 1.414 \\ -2.449 & 0 & 1.414 \\ 4.899 & -3.464 & 1.414 \end{pmatrix}.$$

3.7.1. *Qualité de représentation des individus.* Comme on travaille avec des projections du nuage initial dans un sous-espace, il faut faire attention à l'interprétation géométrique du nouveau nuage de points obtenu. En effet, si le fait que deux points projetés soient éloignés l'un de l'autre assure que les deux points initiaux (i.e. non projetés) sont également éloignés l'un de l'autre. Le fait que deux points projetés soient proches l'un de l'autre n'implique pas forcément que les points initiaux le soient aussi. Pour interpréter correctement la proximité entre deux points projetés, il faut s'assurer au préalable que ces points sont bien représentés dans le sous-espace sur lequel on les a projetés. Pour cela, on s'intéresse à l'angle entre le point e_i et le sous-espace sur lequel on projette, et plus précisément au cosinus carré de cet angle. Ainsi, si ce cosinus carré est proche de 1, cela veut dire que l'individu est bien représenté par sa projection sur le sous-espace. Par suite, si deux individus bien représentés par leurs projections sur un sous-espace et ont des projections qui sont proches, alors on peut en déduire qu'ils sont également proches dans l'espace entier. Le cosinus au carré de l'angle α_{ij} entre \mathbf{e}_i et \mathbf{u}_j est donné par :

$$\cos^2 \alpha_{ij} = \frac{\langle \mathbf{e}_i, \mathbf{u}_j \rangle^2}{\|\mathbf{e}_i\|^2} = \frac{(\mathbf{e}_i^t \mathbf{u}_j)^2}{\|\mathbf{e}_i\|^2} = \frac{y_{ij}^2}{\|\mathbf{e}_i\|^2}. \quad (3.5)$$

De même, on peut calculer le cosinus au carré de l'angle $\alpha_{ijj'}$ entre e_i et le sous-espace engendré par les axes \mathbf{E}_j et $\mathbf{E}_{j'}$ par:

$$\cos^2 \alpha_{ijj'} = \cos^2 \alpha_{ij} + \cos^2 \alpha_{ij'}.$$

Comme $\|\mathbf{e}_i\| = \|\mathbf{y}_i\|$ (il s'agit simplement d'un changement de base, la norme du vecteur étant inchangée), on a

$$\sum_{j=1}^p \cos^2 \alpha_{ij} = 1.$$

Application: calculons tout d'abord la norme au carré $\|\mathbf{e}_i\|^2$, $i = 1, 2, \dots, 6$. Rappelons que

$$\mathbf{X} = \begin{pmatrix} 2 & -3 & -5 \\ -2 & 2 & 0 \\ 0 & 4 & 2 \\ 4 & 0 & 2 \\ 2 & -2 & 0 \\ -6 & -1 & 1 \end{pmatrix}.$$

Alors

$$\|\mathbf{e}_1\|^2 = 2^2 + (-3)^2 + (-5)^2 = 38, \quad \|\mathbf{e}_2\|^2 = (-2)^2 + (2)^2 + (0)^2 = 8,$$

$$\|\mathbf{e}_3\|^2 = 0^2 + (4)^2 + (2)^2 = 20, \quad \|\mathbf{e}_4\|^2 = (4)^2 + (0)^2 + (2)^2 = 20,$$

$$\|\mathbf{e}_5\|^2 = 2^2 + (-2)^2 + (0)^2 = 8 \text{ et } \|\mathbf{e}_6\|^2 = (-6)^2 + (-1)^2 + (1)^2 = 38.$$

Le cosinus au carré des angles, entre chaque individu et chaque composante, est récapitulé dans le tableau suivant:

$$\cos^2 \alpha = \begin{pmatrix} \frac{y_{11}^2}{\|\mathbf{e}_1\|^2} & \frac{y_{12}^2}{\|\mathbf{e}_1\|^2} & \frac{y_{13}^2}{\|\mathbf{e}_1\|^2} \\ \frac{y_{21}^2}{\|\mathbf{e}_2\|^2} & \frac{y_{22}^2}{\|\mathbf{e}_2\|^2} & \frac{y_{23}^2}{\|\mathbf{e}_2\|^2} \\ \frac{y_{31}^2}{\|\mathbf{e}_3\|^2} & \frac{y_{32}^2}{\|\mathbf{e}_3\|^2} & \frac{y_{33}^2}{\|\mathbf{e}_3\|^2} \\ \frac{y_{41}^2}{\|\mathbf{e}_4\|^2} & \frac{y_{42}^2}{\|\mathbf{e}_4\|^2} & \frac{y_{43}^2}{\|\mathbf{e}_4\|^2} \\ \frac{y_{51}^2}{\|\mathbf{e}_5\|^2} & \frac{y_{52}^2}{\|\mathbf{e}_5\|^2} & \frac{y_{53}^2}{\|\mathbf{e}_5\|^2} \\ \frac{y_{61}^2}{\|\mathbf{e}_6\|^2} & \frac{y_{62}^2}{\|\mathbf{e}_6\|^2} & \frac{y_{63}^2}{\|\mathbf{e}_6\|^2} \end{pmatrix} = \begin{pmatrix} \frac{(-2\sqrt{6})^2}{38} & \frac{(-2\sqrt{3})^2}{38} & \frac{(-\sqrt{2})^2}{38} \\ \frac{(\sqrt{6})^2}{8} & \frac{(0)^2}{8} & \frac{(-\sqrt{2})^2}{8} \\ \frac{(\sqrt{6})^2}{20} & \frac{(2\sqrt{3})^2}{20} & \frac{(-\sqrt{2})^2}{20} \\ \frac{(-\sqrt{6})^2}{20} & \frac{(2\sqrt{3})^2}{20} & \frac{(\sqrt{2})^2}{20} \\ \frac{(-\sqrt{6})^2}{8} & \frac{(0)^2}{8} & \frac{(\sqrt{2})^2}{8} \\ \frac{(2\sqrt{6})^2}{38} & \frac{(-2\sqrt{3})^2}{38} & \frac{(\sqrt{2})^2}{38} \end{pmatrix}.$$

Après le calcul, on obtient

$$\cos^2 \alpha = \begin{pmatrix} 0.631 & 0.315 & 0.0526 \\ 0.75 & 0 & 0.25 \\ 0.3 & 0.6 & 0.1 \\ 0.3 & 0.6 & 0.1 \\ 0.75 & 0 & 0.25 \\ 0.631 & 0.31579 & 0.0526 \end{pmatrix}.$$

3.7.2. *Interprétation des nouveaux axes en fonction des individus.* On rappelle que l'inertie expliquée par un axe \mathbf{E}_j est:

$$I_{\mathbf{E}_j^\perp} = \frac{1}{n} \sum_{i=1}^n d^2(\mathbf{e}_i, \text{Pr } \mathbf{o}_{\mathbf{E}_j^\perp, i}) = \frac{1}{n} \sum_{i=1}^n \langle \mathbf{e}_i, \mathbf{u}_j \rangle^2 = \lambda_j, \quad j = 1, \dots, p.$$

Comme tous les individus ont le même poids, ils contribuent chacun à cette inertie à hauteur de

$$\text{ca}(i, j) = \frac{1}{n} y_{ij}^2.$$

On parle de *contribution absolue*. Autrement dit, plus la projection d'un individu sur un axe donné est grande, plus il contribuera à expliquer l'axe en question. Inversement, un individu dont la projection sur un axe est faible, et donc proche du centre de gravité, contribuera peu à l'inertie portée par cet axe. On peut aussi

s'intéresser à la *contribution relative* d'un individu à un axe \mathbf{E}_j . L'inertie portée par l'axe étant λ_j , la contribution relative est donnée par:

$$\text{cr}(i, j) = \frac{\text{ca}(i, j)}{\lambda_j} = \frac{y_{ij}^2}{n\lambda_j}.$$

Application:

- La matrice de contributions absolues des individus ($\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_6$) à l'inertie

$$\mathbf{ca} = \frac{1}{6} \begin{pmatrix} (-2\sqrt{6})^2 & (-2\sqrt{3})^2 & (-\sqrt{2})^2 \\ (\sqrt{6})^2 & 0^2 & (-\sqrt{2})^2 \\ (\sqrt{6})^2 & (2\sqrt{3})^2 & (-\sqrt{2})^2 \\ (-\sqrt{6})^2 & (2\sqrt{3})^2 & (\sqrt{2})^2 \\ (-\sqrt{6})^2 & 0^2 & (\sqrt{2})^2 \\ (2\sqrt{6})^2 & (-2\sqrt{3})^2 & (\sqrt{2})^2 \end{pmatrix} = \begin{pmatrix} 4 & 2 & 0.33 \\ 1 & 0 & 0.33 \\ 1 & 2 & 0.33 \\ 1 & 2 & 0.33 \\ 1 & 0 & 0.33 \\ 4 & 2 & 0.33 \end{pmatrix}.$$

- Rappelons que les valeurs propres sont $\lambda_1 = 12 > \lambda_2 = 8 > \lambda_3 = 2$. Alors la matrice de contributions relatives des individus ($\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_6$) à l'inertie portée par chaque axe $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3$ est récapitulée au tableau suivant:

$$\mathbf{cr} = \begin{pmatrix} 4/12 & 2/8 & 0.33/2 \\ 1/12 & 0/8 & 0.33/2 \\ 1/12 & 2/8 & 0.33/2 \\ 1/12 & 2/8 & 0.33/2 \\ 1/12 & 0/8 & 0.33/2 \\ 4/12 & 2/8 & 0.33/2 \end{pmatrix} = \begin{pmatrix} 0.33 & 0.25 & 0.16 \\ 0.08 & 0 & 0.16 \\ 0.08 & 0.25 & 0.16 \\ 0.08 & 0.25 & 0.16 \\ 0.08 & 0 & 0.16 \\ 0.33 & 0.25 & 0.16 \end{pmatrix}.$$

3.7.3. Reconstruction et approximation de la matrice de données (théorème d'Eckart-Young). Rappelons que $Xu_j = \mathbf{y}_j$. En multipliant les deux membres de cette équation par u_j^t , on obtient $Xu_ju_j^t = \mathbf{y}_ju_j^t$. Comme u_j sont des vecteurs orthonormés, alors que $\sum_{j=1}^p u_ju_j^t = \mathbf{Id}_p$, ce qui implique que

$$\mathbf{X} = \sum_{j=1}^p \mathbf{y}_ju_j^t.$$

En forme matricielle, cette dernière est équivalente à $X = \mathbf{Y}U^t$. De même, en utilisant le fait que $\mathbf{V}u_j = \lambda_ju_j$, on aura

$$\mathbf{V} = \sum_{j=1}^p \lambda_j \mathbf{u}_ju_j^t = \mathbf{U}\mathbf{D}_\lambda \mathbf{U}^t, \quad (3.6)$$

où

$$\mathbf{D}_\lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_p \end{pmatrix}.$$

Dans ce contexte, le fameux *Théorème la décomposition en valeurs singulière (SVD)*, dit que

$$\mathbf{X} = \sum_{j=1}^r \sqrt{\lambda_j} \mathbf{p}_j \mathbf{q}_j^t,$$

où $\mathbf{p}_j \in \mathcal{M}(n \times 1)$ et $\mathbf{q}_j \in \mathcal{M}(p \times 1)$ sont les vecteurs propres (orthonormés) de $\mathbf{X}\mathbf{X}^t \in \mathcal{M}(n \times n)$ et $\mathbf{X}^t\mathbf{X} \in \mathcal{M}(p \times p)$. Sous forme matricielle, l'équation précédente peut être réécrite comme suit:

$$\mathbf{X} = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}^t,$$

où $\mathbf{P} \in \mathcal{M}(n \times n)$ et $\mathbf{Q} \in \mathcal{M}(p \times p)$ sont les deux matrices orthogonales associées aux vecteurs propres \mathbf{p}_j et \mathbf{q}_j , respectivement, et $r := \min(n, p)$ est le rang de \mathbf{X} , et

$$\mathbf{\Lambda} \in \mathcal{M}(n \times p) := \begin{cases} \mathbf{D}_{\sqrt{\lambda}} & \text{si } n = p; \\ (\mathbf{D}_{\sqrt{\lambda}}, \mathbf{0}) & \text{si } n < p; \\ \begin{pmatrix} \mathbf{D}_{\sqrt{\lambda}} \\ 0 \end{pmatrix} & \text{si } n > p. \end{cases}$$

L'ACP permet de reconstruire la matrice de données X , et de construire des approximations de X de rang inférieur, c'est-à-dire de rang k avec $k < r$. Ce problème est connu sous le nom d'approximation de rang faible. La meilleure approximation de rang k de la matrice X est la matrice \mathbf{A}_k vérifiant

$$\mathbf{A}_k = \arg \min_{A_k} \|\mathbf{X} - \mathbf{A}_k\|^2, \text{ avec } \text{rang}(A_k) = k. \quad (3.7)$$

Elle s'obtient à l'aide des k premiers vecteurs propres de la matrice \mathbf{V} , i.e. les vecteurs propres associés aux k plus grandes valeurs propres de \mathbf{V} . Plus précisément

$$\mathbf{A}_k = \mathbf{X}\mathbf{U}_k\mathbf{U}_k^t,$$

où $\mathbf{U}_k \in \mathcal{M}(p \times k)$ contenant en colonnes les k premiers vecteurs propres i.e $\mathbf{U}_k = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k]$. Ce résultat est connu sous le nom de théorème de *Eckart-Young* (1936). L'erreur d'approximation est donnée par la somme des valeurs propres restantes:

$$\|\mathbf{X} - \mathbf{A}_k\|^2 = \sum_{j=k}^p \lambda_j.$$

Application 3.2. Rappelons que la matrice des vecteurs propres est

$$\mathbf{U} = \begin{pmatrix} -2/\sqrt{6} & 1/\sqrt{3} & 0 \\ 1/\sqrt{6} & 1/\sqrt{3} & -1/\sqrt{2} \\ 1/\sqrt{6} & 1/\sqrt{3} & 1/\sqrt{2} \end{pmatrix}.$$

La meilleure approximation de rang 2 de la matrice \mathbf{X} est $\mathbf{A}_2 = \mathbf{X}\mathbf{U}_2\mathbf{U}_2^t$, où

$$\mathbf{U}_2 = \begin{pmatrix} -2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{3} \end{pmatrix}.$$

Alors

$$\mathbf{A}_2 = \begin{pmatrix} 2 & -3 & -5 \\ -2 & 2 & 0 \\ 0 & 4 & 2 \\ 4 & 0 & 2 \\ 2 & -2 & 0 \\ -6 & -1 & 1 \end{pmatrix} \begin{pmatrix} -2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{3} \end{pmatrix} \begin{pmatrix} -\frac{1}{3}\sqrt{6} & \frac{1}{6}\sqrt{6} & \frac{1}{6}\sqrt{6} \\ \frac{1}{3}\sqrt{3} & \frac{1}{3}\sqrt{3} & \frac{1}{3}\sqrt{3} \end{pmatrix}.$$

Après un calcul on trouve:

$$\mathbf{A}_2 = \begin{pmatrix} 2 & -4 & -4 \\ -2 & 1 & 1 \\ 0 & 3 & 3 \\ 4 & 1 & 1 \\ 2 & -1 & -1 \\ -6 & 0 & 0 \end{pmatrix}.$$

L'erreur de l'approximation est

$$\|\mathbf{X} - \mathbf{A}_2\|^2 = \sum_{j=2}^3 \lambda_j = \lambda_2 + \lambda_3 = 8 + 2 = 10.$$

3.8. Représentation des variables. La recherche des axes factoriels nous a permis d'identifier une nouvelle base orthonormée de l'espace \mathbb{R}^p . Les nouveaux axes obtenus sont des combinaisons linéaires des axes initiaux, et on peut donc associer aux variables initiales les combinaisons linéaires correspondantes afin de former de "nouvelles variables", appelées *composantes principales*. On peut alors s'intéresser à l'interprétation de ces "nouvelles variables", en cherchant à les relier aux "anciennes variables".

Définition 3.1. On note $\mathbf{c}_1, \dots, \mathbf{c}_p$ les composantes principales, avec \mathbf{c}_k la composante principale correspond à l'axe \mathbf{E}_k . On a

$$\mathbf{c}_k = \sum_{j=1}^p u_{kj} \mathbf{X}_j = \mathbf{X} \mathbf{u}_k.$$

Proposition 3.1. Les composantes principales vérifient :

- $E[\mathbf{c}_j] = 0$
- $\text{Var}[\mathbf{c}_j] = \lambda_j$
- $\text{Cov}[\mathbf{c}_j, \mathbf{c}_k] = 0, j \neq k.$

Démonstration 3.1. Les composantes principales sont centrées comme combinaisons linéaires de variables centrées (rappelons que l'on a supposé que le nuage de points était centré). On se place maintenant dans \mathbb{R}^n pour calculer la covariance entre deux composantes principales:

$$\text{Cov}[\mathbf{c}_j, \mathbf{c}_k] = \frac{1}{n} (\mathbf{X} \mathbf{u}_k)^t (\mathbf{X} \mathbf{u}_j) = \frac{1}{n} \mathbf{u}_k^t \mathbf{X}^t \mathbf{X} \mathbf{u}_j = \mathbf{u}_k^t \mathbf{V} \mathbf{u}_j = \lambda_j \mathbf{u}_k^t \mathbf{u}_j = \begin{cases} 0 & \text{si } k \neq j \\ \lambda_j & \text{si } k = j \end{cases}.$$

Application 3.3. Les composantes principales de notre exemple sont définies par les colonnes de cette matrice

$$\mathbf{C} = \mathbf{X} \mathbf{U} = \begin{pmatrix} 2 & -3 & -5 \\ -2 & 2 & 0 \\ 0 & 4 & 2 \\ 4 & 0 & 2 \\ 2 & -2 & 0 \\ -6 & -1 & 1 \end{pmatrix} \begin{pmatrix} -2/\sqrt{6} & 1/\sqrt{3} & 0 \\ 1/\sqrt{6} & 1/\sqrt{3} & -1/\sqrt{2} \\ 1/\sqrt{6} & 1/\sqrt{3} & 1/\sqrt{2} \end{pmatrix}.$$

Après le calcul, on obtient

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 \\ -4.89 & -3.46 & -1.41 \\ 2.44 & 0 & -1.41 \\ 2.44 & 3.46 & -1.41 \\ -2.44 & 3.46 & 1.41 \\ -2.44 & 0 & 1.41 \\ 4.89 & -3.46 & 1.41 \end{pmatrix} = \mathbf{Y}$$

3.8.1. *Qualité de la représentation et cercle des corrélations.* Afin d'étudier le lien entre les composantes principales et les variables initiales, on s'intéresse aux corrélations entre les nouvelles et les anciennes variables. On définit d'abord la covariance entre une variable initiale X_j et une composante principale \mathbf{c}_k :

$$\begin{aligned} \mathbf{Cov}(\mathbf{X}_j, \mathbf{c}_k) &= \frac{1}{n} \mathbf{c}_k^t \mathbf{X}_j = \frac{1}{n} (\mathbf{X} \mathbf{u}_k)^t \mathbf{X}_j = \frac{1}{n} \mathbf{u}_k^t \mathbf{X}^t \mathbf{X}_j \\ &= \frac{1}{n} \mathbf{u}_k^t \mathbf{X}^t \mathbf{X} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{u}_k^t \mathbf{V} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \lambda_k \mathbf{u}_k^t \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \lambda_k u_{kj}. \end{aligned}$$

On en déduit la corrélation entre X_j et \mathbf{c}_k :

$$\mathbf{COR}[\mathbf{X}_j, \mathbf{c}_k] = \frac{\mathbf{Cov}(\mathbf{X}_j, \mathbf{c}_k)}{\sqrt{\mathbf{Var}(\mathbf{X}_j) \mathbf{Var}(\mathbf{c}_k)}} = \frac{\lambda_k u_{kj}}{\sqrt{s_j^2 \lambda_k}} = \sqrt{\lambda_k} \frac{u_{kj}}{s_j}.$$

Sous forme matricielle, on a

$$\mathbf{COR} = \mathbf{D}_{\sqrt{\lambda}} \mathbf{U}^t \mathbf{D}_{1/\sigma} \in \mathcal{M}(p \times p).$$

Si les variables sont réduites, alors

$$\mathbf{COR}[\mathbf{X}_j, \mathbf{c}_k] = \sqrt{\lambda_k} u_{kj}.$$

Sous forme matricielle, on a

$$\mathbf{COR} = \mathbf{D}_{\sqrt{\lambda}} \mathbf{U} \in \mathcal{M}(p \times p).$$

Théorème. Nous avons

$$\sum_{k=1}^p \mathbf{COR}^2[\mathbf{X}_j, \mathbf{c}_k] = \sum_{k=1}^p \cos^2 \alpha_{jk} = 1, \quad j = 1, \dots, p,$$

où α_{jk} est l'angle entre les deux vecteurs \mathbf{X}_j et \mathbf{c}_k .

Preuve. D'après la décomposition spectrale (3.6), on a $\mathbf{V} = \sum_{k=1}^p \lambda_k \mathbf{u}_k \mathbf{u}_k^t$. Ceci implique, en particulier, l'égalité des diagonaux des deux matrices \mathbf{V} et $\sum_{k=1}^p \lambda_k \mathbf{u}_k \mathbf{u}_k^t$, ainsi on a

$$\sigma_j^2 = \mathbf{Var}(\mathbf{X}_j) = \sum_{k=1}^p \lambda_k u_{kj}^2, \quad j = 1, \dots, p.$$

Par conséquent

$$\sum_{k=1}^p \mathbf{COR}^2[\mathbf{X}_j, \mathbf{c}_k] = \sum_{k=1}^p \lambda_k \frac{u_{kj}^2}{s_j^2} = \frac{\sum_{k=1}^p \lambda_k u_{kj}^2}{s_j^2} = \frac{s_j^2}{s_j^2} = 1. \quad (\text{CQFD})$$

Remarque. Si

$$\text{COR}^2[\mathbf{X}_j, \mathbf{c}_1] + \text{COR}^2[\mathbf{X}_j, \mathbf{c}_2] \simeq 1$$

alors

$$\sum_{k=3}^p \text{COR}^2[\mathbf{X}_j, \mathbf{c}_k] \simeq 0,$$

et par conséquent $\text{COR}[\mathbf{X}_j, \mathbf{c}_k]$, $j = 3, \dots, p$ sont aussi proches de zéro. Ceci explique que la variable \mathbf{X}_j a une forte corrélation avec les premières composantes principales et non corrélées avec le reste des composantes.

La quantité $\text{COR}[\mathbf{X}_j, \mathbf{c}_k]$ donne la qualité de représentation de la variable j sur l'axe \mathbf{E}_k . Plus elle est proche de 1 en valeur absolue, plus la variable est bien représentée par l'axe k . Le signe de la corrélation permet de savoir si la variable contribue positivement ou négativement à la définition de l'axe k . On représente souvent graphiquement ces corrélations en dimension 2, à l'aide de ce que l'on appelle un cercle des corrélations. Plus précisément, on choisit deux axes \mathbf{E}_j et \mathbf{E}_k , et l'on trace les points dont les coordonnées sont les corrélations de chacune des variables avec les j -ème et k -ème composantes principales. Un exemple est fourni dans la figure 1.1.

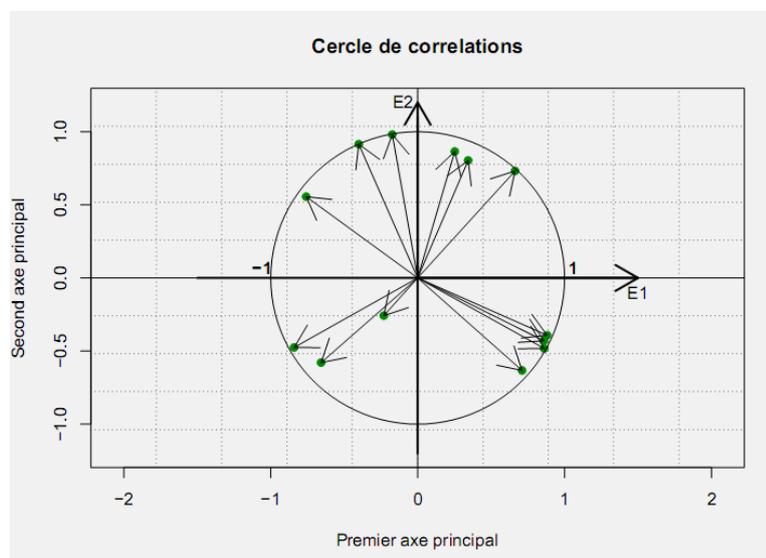


Figure 1.1 – Exemple de cercle des corrélations sur le premier plan factoriel.

L'abscisse correspond à la corrélation entre chaque variable initiale et la composante principale \mathbf{c}_1 , et l'ordonnée à la corrélation entre \mathbf{c}_2 .

Plus une variable soit proche du cercle de corrélation, mieux elle est représentée par le plan considéré. Si deux variables proches du cercle de corrélation (et donc bien représentées dans le plan considéré) sont proches l'une de l'autre, alors elles sont

elles-mêmes fortement corrélées positivement. Inversement, deux variables proches du cercle de corrélation mais symétriquement opposées par rapport à l'origine seront fortement corrélées négativement.

Application: Rappelons que

$$\mathbf{U} = \begin{pmatrix} -2/\sqrt{6} & 1/\sqrt{3} & 0 \\ 1/\sqrt{6} & 1/\sqrt{3} & -1/\sqrt{2} \\ 1/\sqrt{6} & 1/\sqrt{3} & 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} -0.81 & 0.57 & 0 \\ 0.40 & 0.57 & -0.70 \\ 0.40 & 0.57 & 0.70 \end{pmatrix}.$$

Nous avons

$$\mathbf{D}_{\sqrt{\lambda}} = \begin{pmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{8} & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix} = \begin{pmatrix} 3.46 & 0 & 0 \\ 0 & 2.82 & 0 \\ 0 & 0 & 1.41 \end{pmatrix},$$

et

$$\mathbf{D}_{1/\sigma} = \begin{pmatrix} 1/3.26 & 0 & 0 \\ 0 & 1/2.38 & 0 \\ 0 & 0 & 1/2.61 \end{pmatrix} = \begin{pmatrix} 0.30 & 0 & 0 \\ 0 & 0.42 & 0 \\ 0 & 0 & 0.38 \end{pmatrix}.$$

Comme $\mathbf{COR} = \mathbf{D}_{\sqrt{\lambda}} \mathbf{U}^t \mathbf{D}_{1/\sigma}$, alors

$$\mathbf{COR} = \begin{pmatrix} 3.46 & 0 & 0 \\ 0 & 2.82 & 0 \\ 0 & 0 & 1.41 \end{pmatrix} \begin{pmatrix} -0.81 & 0.57 & 0 \\ 0.40 & 0.57 & -0.70 \\ 0.40 & 0.57 & 0.70 \end{pmatrix}^T \begin{pmatrix} 0.30 & 0 & 0 \\ 0 & 0.42 & 0 \\ 0 & 0 & 0.38 \end{pmatrix}.$$

Après le calcul, on trouve

$$\mathbf{COR} = \begin{pmatrix} -0.84078 & 0.58128 & 0.52592 \\ 0.48222 & 0.67511 & 0.61081 \\ 0 & -0.41454 & 0.37506 \end{pmatrix}.$$

3.9. En pratique. Dans la pratique, on retient un nombre $q < p$ d'axes principaux, sur lesquels on va projeter notre nuage de points. On doit alors proposer une interprétation des nouveaux axes obtenus, ou de façon équivalente des composantes principales. On peut s'aider pour cela des outils définis dans les sections précédentes (contribution des individus et des variables dans la définition des axes). Plus précisément, on réalisera les étapes successives suivantes :

- (1) centrage de la matrice de données
- (2) réduction de la matrice de données si nécessaire

- (3) calcul des valeurs propres de \mathbf{V} , et choix du nombre d'axes à retenir en fonction du pourcentage d'inertie que l'on souhaite conserver. On peut également utiliser des règles "empiriques" comme la règle de Kaiser dans le cas d'une **ACP** sur variables réduites, qui consiste à ne garder que les axes pour lesquels la valeur propre λ_j est plus grande que $1/p$, ou la règle du "coude" qui consiste à repérer un "coude" dans le graphe des valeurs propres (voir un exemple sur la figure 1.2).
- (4) interprétation des nouvelles variables, à l'aide des cercles de corrélations : attention, les variables doivent être proches du bord du cercle pour être bien représentées dans le plan factoriel considéré.
- (5) complément pour l'interprétation des nouveaux axes à l'aide des individus et de leurs contributions à la fabrication des axes. Par exemple, on peut retenir les individus dont la contribution est supérieure à la contribution moyenne. Attention, si on souhaite comparer des individus entre eux, il faut d'abord s'assurer qu'ils sont bien représentés dans le plan factoriel considéré.

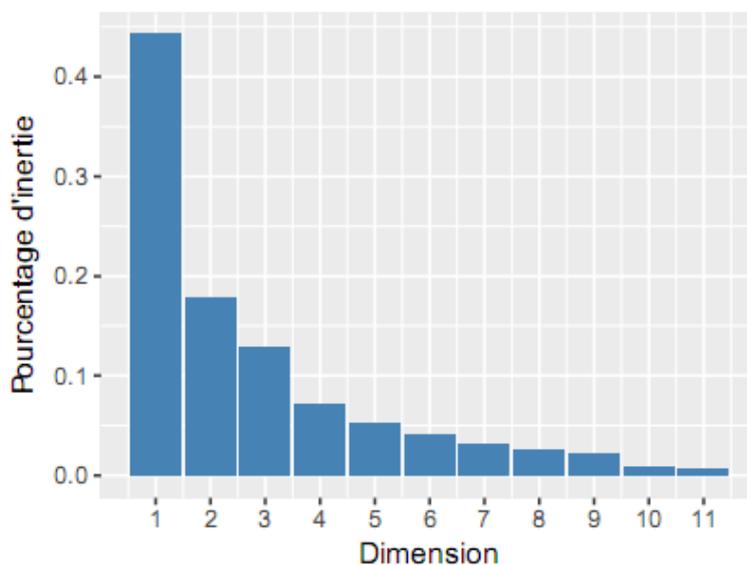


Figure 1.2 – Éboulis des valeurs propres, exprimé en pourcentage de l'inertie totale.

On peut observer un "coude" entre les dimensions 3 et 4, et on gardera ainsi 3 axes.

3.10. Effet taille. Lorsque toutes les variables \mathbf{X}_j sont corrélées positivement entre elles, la première composante principale définit ce que l'on appelle un effet taille. En effet, on sait qu'une matrice symétrique dont tous les termes sont positifs admet un premier vecteur propre dont toutes les composantes sont de même signe (c'est le théorème de Frobenius). La première composante principale est alors corrélée positivement avec toutes les variables. Si un tel effet se produit, on veillera à ne

pas considérer cet axe dans l'interprétation des résultats, que l'on débutera avec le deuxième axe factoriel.

3.11. Notations générales. Dans ce chapitre, on a allégé les notations en considérant la distance euclidienne classique entre les individus vus comme points de \mathbb{R}^p , et en affectant à chacun des individus un poids $1/n$. Cependant il est possible de placer l'ACP dans un cadre un peu plus général que l'on présente ici. De façon générale, on peut munir \mathbb{R}^p de la métrique définie par la matrice \mathbf{M} de dimension $p \times p$. On définit alors le produit scalaire par :

$$\langle \mathbf{X}_i, \mathbf{X}_j \rangle_M = \mathbf{X}_j^t \mathbf{M} \mathbf{X}_i.$$

De même, on peut affecter à chaque individu des poids différents. On choisit alors une matrice de poids \mathbf{D}_p de dimension $n \times n$. La matrice de variance-covariance s'écrit alors :

$$\mathbf{V} = \mathbf{X}^t \mathbf{D}_p \mathbf{X},$$

et on a :

$$I_T = \text{Trace}(\mathbf{V}\mathbf{M})$$

L'ACP consiste alors à chercher les valeurs propres de la matrice $\mathbf{V}\mathbf{M}$.

4. Références

- **Brigitte Escofier, Jérôme Pagès (2008).** Analyses factorielles simples et multiples : Objectifs, méthodes et interprétation.
- **Charlotte Baey (2019).** Analyse de données. <https://baeyc.github.io/teaching/>
- **Mohamed El Merouani (2019):** <https://elmerouani.jimdofree.com/analyse-des-donn%C3%A9es/>
- **Gilbert Saporta (2006).** Probabilités, analyse des données et statistique. Editions Technip.
- **Alboukadel Kassambara (2019).** <http://www.sthda.com/french/>
- **François Husson, Sébastien Lê et Jérôme Pagès (2009).** Analyse des données avec R, Presses Universitaires de Rennes, 224 p. (ISBN 978-2-7535-0938-2)
- **Michel Volle (1997).** Analyse des données, Economica, 4e édition, , (ISBN 2-7178-3212-2)
- **Ludovic Lebart, Morineau Alain, Piron Marie (1995).** Statistique exploratoire multidimensionnelle, Dunod Paris,

- **Jérôme Pagès (2013)**. Analyse factorielle multiple avec R, EDP sciences, Paris, 253 p. (ISBN 978-2-7598-0963-9)