Exercice 1 Considérons un échantillon de $n = 5$ individus où chaque individu $\boldsymbol{x}_i \in \mathbb{R}^d$ est décrit par $d = 3$ variables réelles. Cet échantillon est représenté par la matrice $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5)^t$ suivante :

$$\mathbf{X} = \sqrt{10} \begin{pmatrix} 2 & 2 & 3 \\ 3 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 1 & 4 \\ 2 & 1 & 3 \end{pmatrix}$$

On va faire une ACP centrée réduite de ce jeu de données.

1. Calculer l'individu moyen (le centre de gravité du nuage de données) $\bar{\boldsymbol{x}}$
2. Calculer la matrice $\mathbf{Y}$ des données centrées
3. Calculer les écarts types σ_j de chacune des variables
4. Calculer la matrice $\mathbf{Z}$ des données centrées-réduites
5. Calculer la matrice de variance-covariance $\mathbf{\Sigma}$ de $\mathbf{Z}$ et la matrice de corrélation $\mathbf{R}$ de $\mathbf{X}$. Commenter.
6. Effectuer une décomposition spectrale de la matrice de corrélation $\mathbf{R}$: déterminer les valeurs propres λ_j associées aux vecteurs propres non-nuls $\mathbf{u}_j$ de $\mathbf{R}$.
7. Déterminer les facteurs principaux $\mathbf{f}_j$ et les axes principaux $\mathbf{a}_j$ du nuage des individus. Vérifier leurs propriétés statistiques
8. Calculer pour chacun des axes factoriels, l'inertie du jeu de données projetées sur l'axe considéré, et la part d'inertie qu'il explique.
9. Calculer les composantes principales $\mathbf{c}_j$ pour les individus. Comment s'interprètent les composantes principales en fonction des variables de départ. Vérifier leur propriétés statistiques.
10. Représenter graphiquement le nuage des individus sur le plan factoriel défini par les deux premiers axes factoriels. Commenter.
11. Représenter graphiquement le nuage des variables sur le plan factoriel défini par les deux premiers axes factoriels. Commenter.

Solution 1

1. L'individu moyen est obtenu en faisant la moyenne des lignes du tableau $\mathbf{X}$: $\bar{\boldsymbol{x}} = \sum_{i=1}^{n} \boldsymbol{x}_i / n = \sqrt{10}(2, 1, 3)^T$
2. La matrice $\mathbf{Y}$ des données centrées est obtenue en soustrayant à chaque ligne de $\mathbf{X}$ la moyenne $\bar{\boldsymbol{x}}$:

$$\mathbf{Y} = \mathbf{X} - (\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}, \bar{\boldsymbol{x}})^T = \sqrt{10} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

3. Le calcul des écarts-type σ_j (racines carrées des variances σ_j^2) de chacune des variables peut se faire de deux façons. La première en appliquant la définition de la variance pour chaque variable :

$$\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_{ij}^2}$$

pour $j = 1, \ldots, 3$ et $n = 5$.

La deuxième en calculant directement la matrice de variances-covariances et en exploitant ainsi la formulation vectorielle on trouve directement toutes les variances (et donc les écarts type) car celles-ci sont les éléments diagonaux de la matrice de variances-covariances :

$$\mathbf{\Sigma}_X = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i\boldsymbol{y}_i^T = \frac{1}{n}\mathbf{Y}^T\mathbf{Y} = \frac{10}{5}\begin{pmatrix} 2 & 1 & -1 \\ 1 & 2 & 0 \\ -1 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 4 & 0 \\ -2 & 0 & 4 \end{pmatrix}$$

donc $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)^T = (2, 2, 2)^T$

4. La matrice $\mathbf{Z}$ des données centrées-réduites est de terme général $z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} = \frac{y_{ij}}{\sigma_j}$. Cela revient donc à diviser chaque colonne de $\mathbf{Y}$ par l'écart type de la variable correspondante :

$$\mathbf{Z} = \frac{\sqrt{10}}{2}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \frac{1}{2}\mathbf{Y}$$

5. La matrice de variance-covariance $\mathbf{\Sigma}$ de $\mathbf{Z}$: $\mathbf{Z}$ étant la matrice centrée-réduite de $\mathbf{X}$ donc sa matrice de covariance de terme correspond à la matrice de corrélation de $\mathbf{X}$. En effet :

$$\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{z}_i - \bar{\boldsymbol{z}})(\boldsymbol{z}_i - \bar{\boldsymbol{z}})^T = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{z}^T = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\boldsymbol{y}_i\frac{1}{2}\boldsymbol{y}^T = \frac{1}{n}\sum_{i=1}^{n}(\frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}}{2})(\frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}}{2})^T$$

$$\mathbf{R} = \frac{1}{n}\sum_{i=1}^{n}(\frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}}{\boldsymbol{\sigma}})(\frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}}{\boldsymbol{\sigma}})^T = \frac{1}{n}\sum_{i=1}^{n}(\frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}}{2})(\frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}}{2})^T = \mathbf{\Sigma}$$

car $\mathbf{Z}$ est centrée et donc sa moyenne suivant les lignes $\bar{\boldsymbol{z}}$ est le vecteur nul. Pour le calcul on trouve donc :

$$\mathbf{\Sigma} = \mathbf{R} = \frac{1}{5}\mathbf{Z}^T\mathbf{Z} = \frac{1}{5}\frac{10}{4}\begin{pmatrix} 2 & 1 & -1 \\ 1 & 2 & 0 \\ -1 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 1/2 & -1/2 \\ 1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix}$$

6. L'ACP centrée réduite nécessite le calcul des valeurs propres λ_j associées aux vecteurs propres non-nuls $\mathbf{u}_j$ de la matrice de corrélation $\mathbf{R}$. On résout l'équation $\mathbf{Ru} = \lambda\mathbf{u}$. Pour les valeurs propres, cela revient à résoudre le système $\det(\mathbf{R} - \lambda\mathbf{I}) = 0$:

$$\begin{vmatrix} 1-\lambda & 1/2 & -1/2 \\ 1/2 & 1-\lambda & 0 \\ -1/2 & 0 & 1-\lambda \end{vmatrix} = (1-\lambda)\begin{vmatrix} 1-\lambda & 0 \\ 0 & 1-\lambda \end{vmatrix} - 1/2\begin{vmatrix} 1/2 & -1/2 \\ 0 & 1-\lambda \end{vmatrix} - 1/2\begin{vmatrix} 1/2 & -1/2 \\ 1-\lambda & 0 \end{vmatrix}$$

$= (1-\lambda)(1-\lambda)^2 - \frac{1}{4}(1-\lambda) - \frac{1}{4}(1-\lambda) = (1-\lambda)\left((1-\lambda)^2 - \frac{1}{2}\right) = (1-\lambda)(1-\lambda-\frac{1}{\sqrt{2}})(1-\lambda+\frac{1}{\sqrt{2}})$
et on obtient les trois valeurs propres (selon l'ordre décroissant) : $\lambda_1 = 1+\frac{1}{\sqrt{2}}, \lambda_2 = 1, \lambda_3 = 1-\frac{1}{\sqrt{2}}$.

7. Pour déterminer les vecteurs propres $\mathbf{u}_j$ on résout $\mathbf{Ru} = \lambda\mathbf{u}$. En posant $\mathbf{u} = (x, y, z)^t$, on a, pour $\lambda_2 = 1$:

$$\begin{pmatrix} 1 & 1/2 & -1/2 \\ 1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \Rightarrow \begin{cases} x + \frac{1}{2}y - \frac{1}{2}z = x \\ \frac{1}{2}x + y = y \\ -\frac{1}{2}x + z = z \end{cases}$$

On peut ainsi facilement identifier d'après l'un des deux dernières équations que $x = 0$, et d'après la première en déduire que $y = z$. Donc une solution pour le vecteur propre associé à λ_2 est $\mathbf{u}_2 = (0, 1, 1)^T$. Pour $\lambda_1 = 1 + \frac{1}{\sqrt{2}}$:

$$\begin{cases} x + \frac{1}{2}y - \frac{1}{2}z = x + \frac{1}{\sqrt{2}}x \\ \frac{1}{2}x + y = y + \frac{1}{\sqrt{2}}y \\ -\frac{1}{2}x + z = z + \frac{1}{\sqrt{2}}z \end{cases}$$

On peut ainsi facilement identifier d'après les deux dernières en les sommant que $y = -z$ et d'après la deuxième identifier que $x = \sqrt{2}y$. Donc une solution pour le vecteur propre associé à λ_1 est $\mathbf{u}_1 = (\sqrt{2}, 1, -1)^T$ (pour $z = 1$).

Pour $\lambda_3 = 1 - \frac{1}{\sqrt{2}}$:

$$\begin{cases} x + \frac{1}{2}y - \frac{1}{2}z = x - \frac{1}{\sqrt{2}}x \\ \frac{1}{2}x + y = y - \frac{1}{\sqrt{2}}y \\ -\frac{1}{2}x + z = z - \frac{1}{\sqrt{2}}z \end{cases}$$

On peut aussi facilement identifier d'après les deux dernières en les sommant que $y = -z$ et d'après la deuxième identifier que $x = -\sqrt{2}y$. Donc une solution pour le vecteur propre associé à λ_3 est $\mathbf{u}_3 = (-\sqrt{2}, 1, -1)^T$.

8. Les facteurs principaux $\mathbf{f}_j$ sont déterminés par les vecteurs propres de $\mathbf{R}$. Il y en a donc $d = 3$ qui sont : $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$:
$\mathbf{u}_1 = (\sqrt{2}, 1, -1)^T$; $\mathbf{u}_2 = (0, 1, 1)^T$; $\mathbf{u}_3 = (-\sqrt{2}, 1, -1)^T$.
On appelle axes principaux $\mathbf{a}_j$ les vecteurs propres de $\mathbf{R}$ (facteurs principaux) normés à 1. Donc $\mathbf{a}_j = \frac{1}{||\mathbf{u}_j||_2}\mathbf{u}_j = \frac{1}{\sqrt{\sum_{k=1}^{d=3} u_{jk}^2}}\mathbf{u}_j$. Ainsi on obtient :
$\mathbf{a}_1 = \frac{1}{2}(\sqrt{2}, 1, -1)^T$; $\mathbf{a}_2 = \frac{1}{\sqrt{2}}(0, 1, 1)^T$; $\mathbf{a}_3 = \frac{1}{2}(-\sqrt{2}, 1, -1)^T$.
Les facteurs ainsi que les axes sont orthogonaux : $\mathbf{a}_j^T\mathbf{a}_k = \mathbf{f}_j^T\mathbf{f}_k = 0$ pour tout $j \neq k$ (facile à vérifier)/
Les axes sont normés (condition vérifiée car par construction ici).

9. L'inertie du jeu de données projetées sur chaque axe est données par la valeur propre associé à chaque axe, dans un ordre décroissant. Ainsi l'inertie expliquée par l'axe 1 est $\lambda_1 = 1+\frac{1}{\sqrt{2}}$, celle expliquée par l'axe 2 est $\lambda_2 = 1$ et celle expliquée par l'axe 3 est $\lambda_3 = 1 - \frac{1}{\sqrt{2}}$.
La part d'inertie expliquée par l'axe $\mathbf{a}_j$ est donnée en % par $I_j = \frac{100\lambda_j}{\lambda_1+\lambda_3+\lambda_3}\% = \frac{100\lambda_j}{3}\%$.
Ainsi $I_1 = \frac{100(1+\frac{1}{\sqrt{2}})}{3}\% \approx 56.6\%$; $I_2 = \frac{100}{3}\% \approx 33.3\%$ et $I_3 = \frac{100(1-\frac{1}{\sqrt{2}})}{3}\% \approx 10.1\%$

10. Les composantes principales $\mathbf{c}_j$ sont les variables définies par les axes principaux (représentant donc la projection des données $\mathbf{Z}$ centrées réduites dans ce cas, sur les axes principaux $\mathbf{a}_j$) : $\mathbf{c}_j = \mathbf{Z}\mathbf{a}_j$ pour $j = 1, \ldots, 3$. Ainsi :

$$\mathbf{c}_1 = \frac{\sqrt{10}}{2}\frac{1}{2}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} \sqrt{2} \\ 1 \\ -1 \end{pmatrix} = \frac{\sqrt{10}}{4}\begin{pmatrix} 1 \\ \sqrt{2}+1 \\ -\sqrt{2}-1 \\ -1 \\ 0 \end{pmatrix}$$

$$\mathbf{c}_2 = \frac{\sqrt{10}}{2}\frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \frac{\sqrt{10}}{2\sqrt{2}}\begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 0 \end{pmatrix}$$

$$\mathbf{c}_3 = \frac{\sqrt{10}}{2}\frac{1}{2}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} -\sqrt{2} \\ 1 \\ -1 \end{pmatrix} = \frac{\sqrt{10}}{4}\begin{pmatrix} 1 \\ -\sqrt{2}+1 \\ \sqrt{2}-1 \\ -1 \\ 0 \end{pmatrix}$$

Les composantes principales $\mathbf{Z}\mathbf{u}_j$ sont ainsi des combinaisons linéaires des variables de départs $\boldsymbol{z}_i$.

Les composantes principales sont centrées ($\bar{\mathbf{c}}_j = 0$, facile vérifier), orthogonales deux à deux et donc non corrélées entre elles ($\mathbf{c}_j^T\mathbf{c}_k = 0$ pour tout $j \neq k$, facile à vérifier), et leurs variances sont égales aux valeurs propres qui leur sont associées : $\frac{1}{n}\mathbf{c}_j^T\mathbf{c}_j^T = \lambda_j$

Exercice 2 Considérons un échantillon de n individus où chaque individu $\boldsymbol{x}_i \in \mathbb{R}^d$ est décrit par $d = 3$ variables réelles qui ont pour matrice de corrélation

$$\mathbf{R} = \begin{pmatrix} 1 & \rho & -\rho \\ \rho & 1 & \rho \\ -\rho & \rho & 1 \end{pmatrix}$$

ave $-1 \leq \rho \leq 1$.

On va faire une ACP centrée-réduite de ce jeu de données.

1. Effectuer une décomposition spectrale de la matrice de corrélation $\mathbf{R}$: déterminer les valeurs propres λ_j associées aux vecteurs propres non-nuls $\mathbf{u}_j$ de $\mathbf{R}$.
2. Quelles sont les valeurs possibles pour ρ. Justifier que ρ doit vérifier $-1 \leq \rho \leq 1$.
3. Calculer pour chacun des axes factoriels, l'inertie du jeu de données projetées sur l'axe considéré, et la part d'inertie qu'il explique. Faire une représentation graphique.

4. Calculer les composantes principales $\mathbf{c}_j$ pour les individus. Comment s'interprètent les composantes principales en fonction des variables de départ. Vérifier leur propriétés statistiques.
5. Comment s'interprète en fonction des données d'origines $\boldsymbol{x}_i$ l'unique composante à retenir dans ce cas.