Exercice ACP

On considère le tableau $\mathbf{R}$ de notes sur 20 suivant ($n = 9$ individus, $p = 5$ variables) :

	Mathématiques	Sciences	Français	Latin	Musique
Jean	6	6	5	5,5	8
Aline	8	8	8	8	9
Annie	6	7	11	9,5	11
Monique	14,5	14,5	15,5	15	8
Didier	14	14	12	12	10
André	11	10	5,5	7	13
Pierre	5,5	7	14	11,5	10
Brigitte	13	12,5	8,5	9,5	12
Evelyne	9	9,5	12,5	12	18

Le tableau des moyennes par matière est :

	Mathématiques	Sciences	Français	Latin	Musique
Moyenne	9,67	9,83	10,2	10,1	11,0

On désire soumettre le tableau $\mathbf{R}$ à une ACP. Pour cela on est conduit à rechercher les vecteurs propres de la matrice ${}^T\mathbf{X}\,\mathbf{X}$ des covariances empiriques des cinq variables, qui est

$${}^T\mathbf{X}\,\mathbf{X} = \begin{bmatrix} & \text{Math.} & \text{Sciences} & \text{Français} & \text{Latin} & \text{Musique} \\ \text{Math.} & 11{,}4 & 9{,}92 & 2{,}66 & 4{,}82 & 0{,}111 \\ \text{Sciences} & & 8{,}94 & 4{,}12 & 5{,}48 & 0{,}056 \\ \text{Français} & & & 12{,}1 & 9{,}29 & 0{,}389 \\ \text{Latin} & & & & 7{,}91 & 0{,}667 \\ \text{Musique} & & & & & 8{,}67 \end{bmatrix}$$

i) Indiquer la transformation qui permet de passer de la matrice $\mathbf{R}$ à la matrice $\mathbf{X}$. Calculer la première ligne de $\mathbf{X}$.

ii) Les trois plus grandes valeurs propres de la matrice ${}^T\mathbf{X}\,\mathbf{X}$ des variances-covariances sont $\lambda_1 = 28{,}253$, $\lambda_2 = 12{,}075$ et $\lambda_3 = 8{,}616$. Quels sont les taux d'inertie expliquée par chacun des trois axes factoriels correspondants ? En limitant la représentation à l'espace des 3 premiers facteurs, quel est le taux d'inertie totale expliquée par cette représentation ? Que peut-on en conclure ?

iii) Les trois premiers vecteurs propres normés de ${}^T\mathbf{X}\,\mathbf{X}$ sont donnés dans le tableau ci-dessous :

	1	2	3
Maths	0,515	-0,567	-0,051
Sciences	0,507	-0,372	-0,014
Français	0,492	0,650	0,108
Latin	0,485	0,323	0,023
Musique	0,031	0,113	-0,992

Calculer les coordonnées de « Jean » sur les trois axes factoriels.

iv) Calculer les coefficients de corrélation linéaire entre le premier facteur et les 5 variables.

v) Les corrélations entre les variables et les deux autres facteurs sont données ci-dessous :

	Facteur 2	Facteur 3
Maths	-0,584	-0,045
Sciences	-0,432	-0,014
Français	0,651	0,091
Latin	0,399	0,024
Musique	0,133	-0,990

Donner brièvement une interprétation possible pour les 3 facteurs.

vi) En utilisant les résultats obtenus à la première et à la troisième question, calculer l'indice ponctuel de qualité de la représentation de « Jean » sur le premier axe factoriel, puis sur le plan des deux premiers facteurs, puis sur l'espace des trois premiers facteurs. Conclure.

Solution

i) On passe de la matrice $\mathbf{R}$ à la matrice $\mathbf{X}$ par centrage par rapport aux variables, soit :

$$x_{ij} = \frac{r_{ij} - \bar{r}_j}{\sqrt{n-1}},\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, p$$

Pour la première ligne de $\mathbf{X}$ on trouve donc $x_{11} = (6 - 9{,}67)/2{,}828 = -1{,}298$, $x_{12} = (6 - 9{,}83)/2{,}828 = -1{,}354$, $x_{13} = (5 - 10{,}2)/2{,}828 = -1{,}839$, $x_{14} = (5{,}5 - 10{,}1)/2{,}828 = -1{,}626$, $x_{15} = (8 - 11)/2{,}828 = -1{,}06$

ii) On a $\sum_{\alpha=1}^{5} \lambda_\alpha = \text{trace}\left({}^T\mathbf{X}\,\mathbf{X}\right) = 11{,}4 + 8{,}94 + 12{,}1 + 7{,}91 + 8{,}67 = 49{,}02$. Les taux d'inertie expliquée par les trois premiers axes factoriels sont donc :

$$\tau_1 = \frac{\lambda_1}{\sum_{\alpha=1}^{5}\lambda_\alpha} = \frac{28{,}253}{49{,}02} = 0{,}576,\ \tau_2 = \frac{\lambda_2}{\sum_{\alpha=1}^{5}\lambda_\alpha} = \frac{12{,}075}{49{,}02} = 0{,}246,$$

$$\tau_3 = \frac{\lambda_3}{\sum_{\alpha=1}^{5}\lambda_\alpha} = \frac{8{,}616}{49{,}02} = 0{,}176$$

Le taux d'inertie totale expliquée par cette représentation est la somme des taux calculés à la question précédente.

$$\tau = \frac{\lambda_1 + \lambda_2 + \lambda_3}{\sum_{\alpha=1}^{5}\lambda_\alpha} = 0{,}576 + 0{,}246 + 0{,}176 = 0{,}998$$

Le nuage est pratiquement dans un espace à 3 dimensions.

iii) Les coordonnées des individus sur l'axe factoriel α (valeur propre λ_α) sont données par $\mathbf{X}\mathbf{u}_\alpha$ (ce sont les composantes). La coordonnée du premier individu sur l'axe factoriel α est donc le produit de la première ligne ${}^T\mathbf{L}_1$ de la matrice $\mathbf{X}$, calculée à la question 1, par $\mathbf{u}_\alpha$.

Pour le premier axe de vecteur $\mathbf{u}_1$:

$$\begin{aligned} {}^T\mathbf{L}_1\,\mathbf{u}_1 &= \begin{bmatrix} -1{,}298 & -1{,}354 & -1{,}839 & -1{,}626 & -1{,}060 \end{bmatrix} \cdot \\ &\quad \cdot {}^T\begin{bmatrix} 0{,}515 & 0{,}507 & 0{,}492 & 0{,}485 & 0{,}031 \end{bmatrix} \\ &= -2{,}317 \end{aligned}$$

On trouve, de même, pour le deuxième axe $\mathbf{u}_2$, ${}^T\mathbf{L}_1\,\mathbf{u}_2 = -0{,}566$, et pour le troisième axe $\mathbf{u}_3$, ${}^T\mathbf{L}_1\,\mathbf{u}_3 = 0{,}850$.

iv) Commençons par calculer les coordonnées des variables sur le premier axe factoriel $\mathbf{v}_1$ de l'analyse du nuage des variables, obtenu par ${}^T\mathbf{X}\,\mathbf{v}_1 = \sqrt{\lambda_1}\,\mathbf{u}_1$. Puisque $\sqrt{\lambda_1} = \sqrt{28{,}253} = 5{,}315$, on a

$$\begin{array}{r} \text{Maths} \\ \text{Sciences} \\ \text{Français} \\ \text{Latin} \\ \text{Musique} \end{array} \begin{bmatrix} 0{,}515 \cdot 5{,}315 = 2{,}737 \\ 0{,}507 \cdot 5{,}315 = 2{,}695 \\ 0{,}492 \cdot 5{,}315 = 2{,}615 \\ 0{,}485 \cdot 5{,}315 = 2{,}578 \\ 0{,}031 \cdot 5{,}315 = 0{,}165 \end{bmatrix}$$

Le coefficient de corrélation linéaire $\rho_{j\alpha}$ entre la variable j et l'axe factoriel α est le cosinus de l'angle β_j ainsi formé.

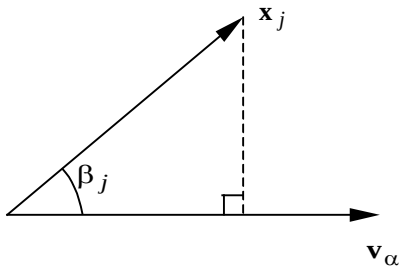

Figure A.1. *Angle formé entre la variable j et l'axe factoriel α*

β_j est l'angle entre la variable et sa projection sur l'axe factoriel :

$$\cos\left(\beta_j\right) = \frac{{}^T\mathbf{x}_j\,\mathbf{v}_\alpha}{\left\|\mathbf{x}_j\right\|}$$

Or, $\mathbf{u}_\alpha = {}^T\mathbf{X}\,\mathbf{v}_\alpha \big/ \sqrt{\lambda_\alpha}$ d'où ${}^T\mathbf{x}_j\,\mathbf{v}_\alpha = \sqrt{\lambda_\alpha}\,\mathbf{u}_{j\alpha}$. De plus, $\mathbf{x}_j = \dfrac{\mathbf{r}_j - \bar{\mathbf{r}}_j}{\sqrt{n-1}}$ d'où $\left\|\mathbf{x}_j\right\|^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(r_{ij} - \bar{r}_j\right)^2 = \sigma_j^2$ (σ_j est l'écart-type de la variable j) et donc $\left\|\mathbf{x}_j\right\| = \sigma_j$. Finalement, $\rho_{j\alpha} = \sqrt{\lambda_\alpha}\,\mathbf{u}_{j\alpha} \big/ \sigma_j$.

Il suffit donc de diviser chaque composante de la matrice précédente par l'écart type de la variable correspondante. Le tableau suivant donne les corrélations entre les variables et le premier facteur :

$$\begin{array}{r} \text{Maths} \\ \text{Sciences} \\ \text{Français} \\ \text{Latin} \\ \text{Musique} \end{array} \begin{bmatrix} 2{,}737 / \sqrt{11{,}4} = 0{,}811 \\ 2{,}695 / \sqrt{8{,}94} = 0{,}901 \\ 2{,}615 / \sqrt{12{,}1} = 0{,}752 \\ 2{,}578 / \sqrt{7{,}91} = 0{,}916 \\ 0{,}165 / \sqrt{8{,}67} = 0{,}056 \end{bmatrix}$$

v) Le premier facteur est fortement et positivement corrélé avec les quatre matières principales : c'est un facteur de taille. Le deuxième facteur oppose les matières scientifiques et littéraires. Le troisième facteur est fortement corrélé avec la musique et caractérise l'« aptitude artistique » (musicale ici).

vi) L'indice ponctuel est défini par $\cos^2\left(\theta_i\right)$ où θ_i est l'angle entre le vecteur individu (centré) $\mathbf{L}_i$ et le sous-espace vectoriel de la représentation $\mathbf{l}_i$ (ici le premier axe factoriel).

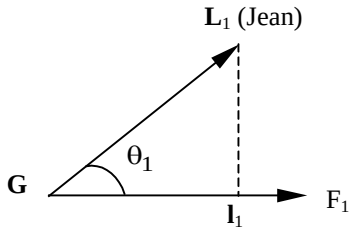

Figure A.2. *Angle formé entre l'individu Jean et le premier axe factoriel*

On a donc $\cos^2(\theta_i) = \| \mathbf{Gl}_i \|^2 / \| \mathbf{GL}_i \|^2$, $\| \mathbf{Gl}_1 \|^2 = (-2{,}904)^2 = 8{,}433$,

$$\| \mathbf{GL}_1 \|^2 = \sum_{j=1}^{5} x_{1j}^2$$

Ainsi, $\cos^2(\theta_1) = \| \mathbf{Gl}_1 \|^2 / \| \mathbf{GL}_1 \|^2 = 0{,}890$ pour le premier axe. Notons maintenant $\mathbf{l}_{1j}$ la projection de l'individu 1 sur le j-ième axe factoriel. Considérons le plan des deux premiers axes factoriels :

$$\cos^2(\theta_1) = \frac{\| \mathbf{Gl}_{11} \|^2 + \| \mathbf{Gl}_{12} \|^2}{\| \mathbf{GL}_1 \|^2} = 0{,}923$$

Considérons le plan des trois premiers axes factoriels :

$$\cos^2(\theta_1) = \frac{\| \mathbf{Gl}_{11} \|^2 + \| \mathbf{Gl}_{12} \|^2 + \| \mathbf{Gl}_{13} \|^2}{\| \mathbf{GL}_1 \|^2} = 1{,}000$$

Jean est déjà bien représenté sur le premier axe factoriel et idéalement dans l'espace des 3 premiers axes puisque $\cos^2(\theta_1)$ vaut 1,000.