Ecole supérieure d'agro-développement international
istom

Faire une ACP à la main

Promotion 115

Table des matières

Aide-mémoire : l'ACP normée **2**

Exercice 1 – guidé : azote, rendement et argile sur cinq parcelles de maïs **4**

Exercice 2 – en autonomie : le climat de cinq villes du Cameroun **19**

Indications **21**

Corrections **22**

Correction de l'exercice 1 22

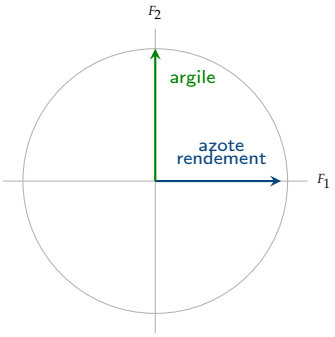
Ce TD comporte **deux exercices**, à faire **entièrement à la main**, calculatrice en main :

- **Exercice 1 (guidé)** : une ACP complète, pas à pas, sur cinq parcelles de maïs et trois variables. Des encadrés de cours expliquent chaque formule, et des lignes ✓ **Contrôle** permettent de vérifier ses résultats avant de continuer ;
- **Exercice 2 (en autonomie)** : une ACP d'entraînement sur le climat de cinq villes du Cameroun. Les questions sont moins détaillées : on s'inspire de la démarche de l'exercice 1.

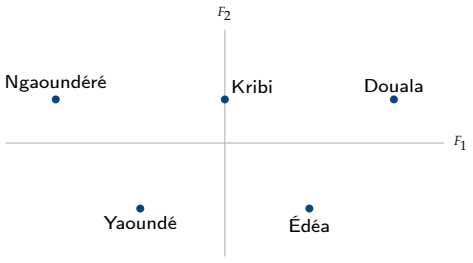
Refaire une ACP à la main, c'est comprendre **d'où viennent les résultats** qu'affiche un logiciel, et donc savoir les lire avec esprit critique :

- **les calculs** : centrage-réduction, matrice des corrélations, valeurs et vecteurs propres, coordonnées des individus et des variables, $\cos^2$, contributions ;
- **l'interprétation** : nommer les axes, lire la carte des individus et le cercle des corrélations, repérer les points mal représentés ou trop influents, rédiger une conclusion agronomique.

Un **aide-mémoire** (page 2) rassemble les formules et les règles de lecture des graphiques. Les indications et les corrections sont en fin de document.



Exercice 1 – guidé
cinq parcelles de maïs : le cercle des corrélations

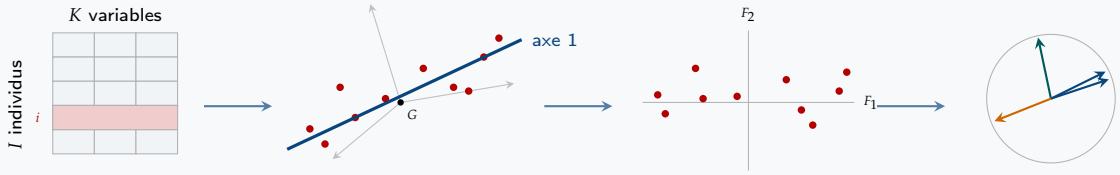


Exercice 2 – en autonomie
climat de cinq villes du Cameroun : la carte des villes

Aide-mémoire : l'ACP normée

Ces deux pages rassemblent tout ce qui sert dans les quatre exercices : le principe de l'ACP, les formules « à la main » et leur équivalent sous R (`FactoMineR`), puis les règles de lecture des graphiques. Les notations sont celles du TD : I individus, K variables, z_{ik} la valeur centrée réduite de l'individu i pour la variable k.

Le principe : photographier un nuage de points sous son meilleur angle



1. le tableau
centré, réduit

2. le nuage des I points
axe 1 : direction d'étalement maximal

3. la carte (plan 1-2)
la « photographie » la plus fidèle

4. le cercle
les variables

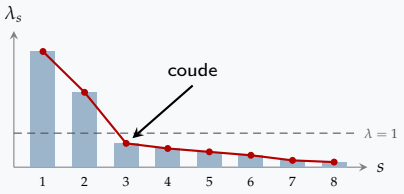
L'ACP cherche, dans l'espace à K dimensions, les directions orthogonales $U_1, U_2, \ldots$ le long desquelles le nuage est le plus étalé : la première porte la plus grande part d'inertie λ_1, la deuxième la plus grande part restante λ_2, etc. On lit ensuite les **individus** sur la carte et les **variables** sur le cercle, en faisant des allers-retours entre les deux.

Les étapes, à la main et sous R

Étape	À la main	Sous R (`res <- PCA(...)`)
Centrer, réduire	$z_{ik} = \dfrac{x_{ik} - \overline{X}_k}{\sigma_k}$ (σ_k : division par I)	`scale.unit = TRUE`
Distance au centre	$d^2(i,G) = \sum_k z_{ik}^2$ $\quad \mathcal{I}_{\text{tot}} = \frac{1}{I}\sum_i d^2(i,G) = K$	`res$ind$dist` (donne d)
Corrélations	$r_{kl} = \frac{1}{I}\sum_i z_{ik} z_{il}$ $\quad R = \frac{1}{I} Z^{\mathrm{T}} Z$	`cor(...)`
Valeurs propres	$\det(R - \lambda \mathrm{I}) = 0$ $\quad \sum_s \lambda_s = K$ $\quad \% = 100\,\lambda_s / K$	`res$eig`
Vecteurs propres	$R\,U_s = \lambda_s\,U_s$ $\quad \lVert U_s \rVert = 1$	`res$svd$V`
Coordonnées des individus	$F_s(i) = \sum_k u_{sk} z_{ik}$ $\quad \mathrm{Var}(F_s) = \lambda_s$	`res$ind$coord`
Qualité (individus)	$\cos^2_s(i) = \dfrac{F_s(i)^2}{d^2(i,G)}$ $\quad \sum_s \cos^2_s(i) = 1$	`res$ind$cos2`
Contribution (individus)	$\mathrm{contrib}_s(i) = 100\,\dfrac{F_s(i)^2}{I\,\lambda_s}$ $\quad$ moyenne $\dfrac{100}{I}$	`res$ind$contrib`
Coordonnées des variables	$r(X_k, F_s) = \sqrt{\lambda_s}\,u_{sk}$	`res$var$coord` (`$cor`)
Qualité (variables)	$\cos^2_s(X_k) = r^2(X_k, F_s)$ $\quad \sum_s r^2(X_k, F_s) = 1$	`res$var$cos2`
Contribution (variables)	$\mathrm{contrib}_s(X_k) = 100\,\dfrac{r^2(X_k, F_s)}{\lambda_s}$ $\quad$ moyenne $\dfrac{100}{K}$	`res$var$contrib`

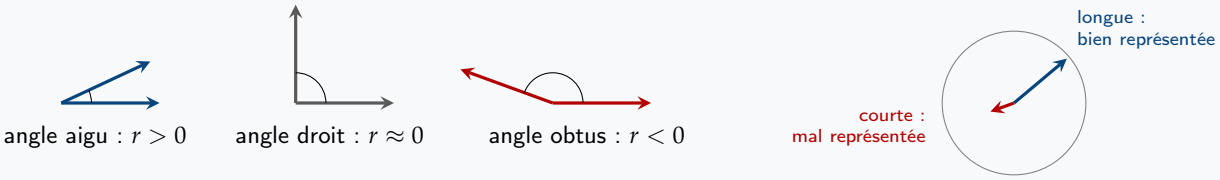
✓ **Contrôle** deux façons de retrouver la même valeur propre : $\lambda_s = \frac{1}{I}\sum_i F_s(i)^2$ (côté individus) $= \sum_k r^2(X_k, F_s)$ (côté variables).

Combien d'axes ? Deux critères à croiser



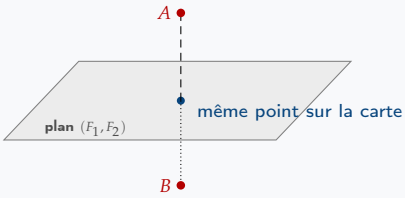
- **Kaiser** : garder les axes tels que $\lambda_s \geq 1$ (ils résument au moins autant qu'une variable de départ).
- **Coude** : garder les axes situés avant la cassure de l'éboulis.
- Regarder aussi le **% d'inertie cumulé**, et ne garder que des axes que l'on sait **interpréter**.

Lire le cercle des corrélations



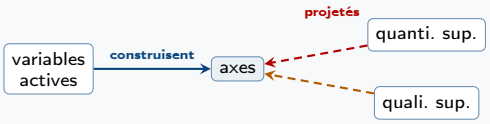
- La coordonnée d'une flèche sur un axe est la **corrélation** de la variable avec cet axe ; on nomme l'axe avec les variables qui y **contribuent** le plus (lues *axe par axe*).
- L'angle entre deux flèches approche leur corrélation ($r \approx \cos\theta$) **seulement si les deux flèches arrivent près du cercle**.

Lire la carte des individus



- Deux points proches sur la carte ne se ressemblent que s'ils sont **tous deux bien représentés** ($\cos^2$ sur le plan proche de 1) ; sinon, comme A et B, ils peuvent être éloignés.
- Un point situé dans la direction d'une flèche a une forte valeur de cette variable.
- Un individu de **contribution** très supérieure à $100/I$ « tire » l'axe : vérifier la donnée, refaire l'ACP en le mettant en supplémentaire.

Éléments supplémentaires



- Ils ne construisent pas les axes : on les place ensuite pour *interpréter* (ce que l'on cherche à expliquer, variables qualitatives).
- Une modalité est placée au barycentre de ses individus ; sa position sur un axe est significative si $|\text{v.test}| > 2$ (`res$quali.sup$v.test`, `dimdesc`).

Les pièges classiques. (1) Faire une ACP non normée sur des variables d'unités différentes : la variable de plus grande variance écrase les autres. (2) Interpréter la proximité de points ou l'angle de flèches mal représentés. (3) Croire qu'un axe classe du « meilleur » au « moins bon » : il oppose souvent deux profils. (4) S'inquiéter d'un axe inversé : son sens est arbitraire (un logiciel peut le retourner). (5) Nommer les axes avec les contributions *au plan* : elles mélangent les deux axes. (6) Conclure à une causalité : l'ACP décrit des associations.

Exercice 1 ACP guidée à la main : azote, rendement et argile sur cinq parcelles de maïs

Dans un essai de fertilisation conduit en zone forestière (région du Centre, Cameroun), un technicien a suivi **cinq parcelles de maïs** $P_1, \dots, P_5$. Pour chacune, il a relevé trois variables ($K = 3$) :

- X_1 **(azote) :** dose d'azote apportée (kg N/ha) ;
- X_2 **(rendement) :** rendement en grain (t/ha) ;
- X_3 **(argile) :** teneur du sol en argile (%).

Parcelle	X_1 : azote (kg N/ha)	X_2 : rendement (t/ha)	X_3 : argile (%)
P_1	30	2,5	17
P_2	50	3,5	17
P_3	60	4	32
P_4	70	5,5	17
P_5	90	4,5	17

À la main

L'objectif est de mener **entièrement à la main** une ACP normée de ce tableau ($I = 5$ individus, $K = 3$ variables), puis d'en interpréter les résultats. Les données ont été choisies pour que tous les calculs « tombent juste » : à chaque étape, des lignes ✓ **Contrôle** vous permettent de vérifier vos résultats avant de continuer. Les encadrés bleus rappellent le cours nécessaire. Toutes les variances sont des variances « population » (division par $I = 5$).

1. Préparer le tableau : centrer et réduire

a) Calculer les moyennes $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_3$ et les écarts-types σ_1, σ_2, σ_3.

✓ **Contrôle** les trois écarts-types sont des nombres entiers.

b) Compléter le tableau des données centrées réduites $z_{ik} = \dfrac{x_{ik} - \overline{X}_k}{\sigma_k}$.

Parcelle	Z_1 (azote)	Z_2 (rendement)	Z_3 (argile)	$d^2(P_i, G)$
P_1				
P_2				
P_3				
P_4				
P_5				

Que représente la dernière colonne ?

Chaque parcelle P_i est un point de l'espace à trois dimensions. Par exemple :

- P_1 a pour coordonnées $(z_{11}\,;z_{12}\,;z_{13})$;
- P_2 a pour coordonnées $(z_{21}\,;z_{22}\,;z_{23})$, etc.

Après centrage-réduction, le **centre de gravité** G du nuage de points est l'origine $(0\,;0\,;0)$: c'est la « parcelle moyenne », celle qui aurait exactement la moyenne pour chacune des trois variables.

Le nombre $d^2(P_i, G)$ est le **carré de la distance** entre la parcelle P_i et cette parcelle moyenne. Il se calcule avec le théorème de Pythagore, étendu à trois dimensions. Par exemple, pour P_1 :

$$d^2(P_1, G) = (z_{11} - 0)^2 + (z_{12} - 0)^2 + (z_{13} - 0)^2 = z_{11}^2 + z_{12}^2 + z_{13}^2.$$

Plus $d^2(P_i, G)$ est grand, plus la parcelle P_i s'écarte de la moyenne sur l'ensemble des variables, c'est-à-dire plus elle est *atypique*. Au contraire, plus $d^2(P_i, G)$ est proche de 0, plus la parcelle P_i est « ordinaire ».

✓ **Contrôle** toutes les valeurs z_{ik} sont des multiples de 0,5, et chaque colonne Z_1, Z_2, Z_3 a une moyenne nulle.

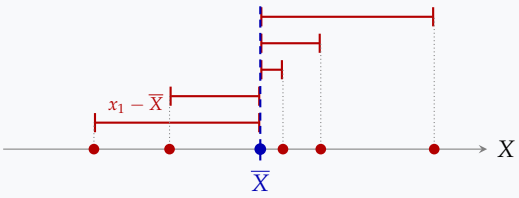
c) **L'inertie totale.**

Pourquoi mesurer la dispersion du nuage par la moyenne des d^2 ?

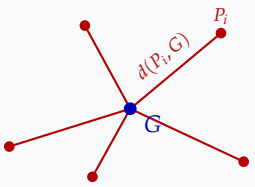
Avec *une seule* variable, on sait déjà mesurer la dispersion : c'est la **variance**, moyenne des carrés des écarts à la moyenne,

$$\mathrm{Var}(X) = \frac{1}{I}\sum_{i=1}^{I}(x_i - \overline{X})^2.$$

L'écart $|x_i - \overline{X}|$ n'est rien d'autre que la *distance* entre l'individu et le point moyen, sur une droite. Avec plusieurs variables, chaque parcelle est un point de l'espace et le point moyen est G : on fait exactement la même chose, en remplaçant l'écart par la distance $d(P_i, G)$.



une variable : variance
= moyenne des (écarts)^2



plusieurs variables : inertie
= moyenne des (distances à G)^2

- *Pourquoi des carrés ?* Pour la même raison que pour la variance : les écarts « à gauche » et « à droite » ne doivent pas se compenser. De plus, Pythagore donne directement d^2 comme une somme de carrés, sans racine : on pourra ainsi découper l'inertie, morceau par morceau (question 5).
- *Pourquoi le mot « inertie » ?* Il vient de la mécanique : si chaque parcelle est une petite masse $\frac{1}{I}$, l'inertie mesure la difficulté à faire tourner le nuage autour de G. Un nuage étalé a une grande inertie, un nuage ramassé autour de G une inertie faible.
- Avec une seule variable, $d^2(P_i, G) = (x_i - \overline{X})^2$ et l'inertie *est* la variance : l'inertie est la « variance du nuage ».

Calculer

$$\mathcal{I}_{\text{tot}} = \frac{1}{I}\sum_{i=1}^{I} d^2(P_i, G)$$

et vérifier qu'elle vaut $K = 3$. Pourquoi ce résultat était-il prévisible ?

d) Sans calcul supplémentaire, quelle parcelle est la plus « atypique » ? Quelle parcelle a la valeur la plus extrême pour *une seule* variable, et pour quelle variable ?

2. La matrice des corrélations

a) **Les coefficients de corrélation.**

D'où vient la formule de calcul ?

On part de la définition du coefficient de corrélation entre deux variables A et B mesurées sur les I parcelles :

$$r(A,B) = \frac{\mathrm{Cov}(A,B)}{\sigma_A\,\sigma_B}, \qquad \text{avec} \qquad \mathrm{Cov}(A,B) = \frac{1}{I}\sum_{i=1}^{I}(a_i - \overline{A})(b_i - \overline{B}).$$

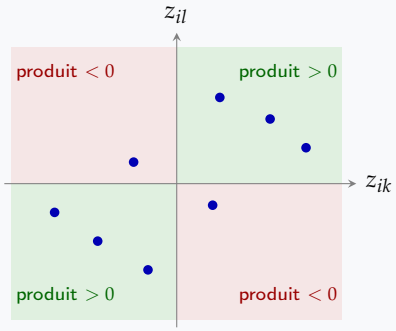
- *Centrer et réduire ne change pas une corrélation* : on soustrait une constante (cela ne change pas les écarts à la moyenne) et on divise par un nombre positif (cela se simplifie entre le numérateur et le dénominateur). Donc $r(X_k, X_l) = r(Z_k, Z_l)$.
- Les variables Z_k et Z_l sont *centrées* : $\overline{Z}_k = \overline{Z}_l = 0$, et la covariance se réduit à la moyenne

des produits, $\mathrm{Cov}(Z_k, Z_l) = \frac{1}{I}\sum_i z_{ik}\, z_{il}$.

- Elles sont *réduites* : $\sigma_{Z_k} = \sigma_{Z_l} = 1$, et le dénominateur vaut 1.

Il reste donc, pour des variables centrées réduites,

$$r_{kl} = r(X_k, X_l) = \frac{1}{I}\sum_{i=1}^{I} z_{ik}\, z_{il}.$$



Le produit $z_{ik}\, z_{il}$ est **positif** quand la parcelle est du même côté de la moyenne pour les deux variables (au-dessus pour les deux, ou au-dessous pour les deux), **négatif** sinon. Si la plupart des points sont dans les zones vertes, la moyenne des produits est positive : $r > 0$. Si les points se répartissent autant dans les zones vertes que rouges, les produits se compensent : $r \approx 0$.

Calculer r_{12}, r_{13} et r_{23}.

b) Écrire la matrice des corrélations R.

c) Interpréter : quelles variables sont liées, et dans quel sens ? Ce résultat vous paraît-il plausible d'un point de vue agronomique ?

3. Les valeurs propres : combien d'information par axe ?

Idée clé. Chaque axe de l'ACP porte une part de l'inertie totale ; cette part est sa **valeur propre** λ_s. Les valeurs propres sont les solutions de l'équation $\det(R - \lambda\, \mathrm{I}) = 0$, où I est la matrice identité.

a) Calculer le polynôme caractéristique $P(\lambda) = \det(R - \lambda\, \mathrm{I})$ de la matrice des corrélations R (on pourra développer selon la dernière ligne), puis le factoriser.

b) En déduire les trois valeurs propres $\lambda_1 \geq \lambda_2 \geq \lambda_3$.

✓ **Contrôle** $\lambda_1 + \lambda_2 + \lambda_3 = \mathrm{Tr}(R) = 3$ (la somme des valeurs propres est égale à l'inertie totale).

Combien d'axes garder ? (1) Le critère de Kaiser

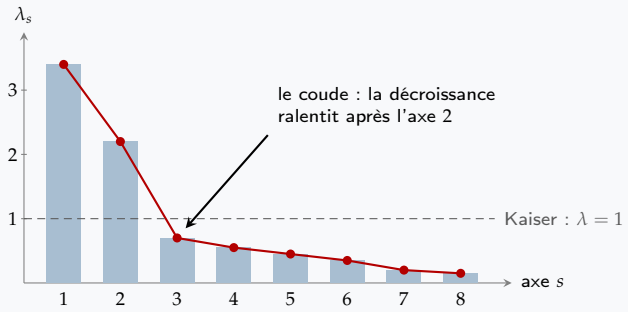
En ACP normée, chacune des K variables a une variance égale à 1 : elle « pèse » donc 1 dans l'inertie totale, qui vaut K. Une valeur propre λ_s mesure l'information portée par l'axe s.

- Un axe tel que $\lambda_s \geq 1$ résume *au moins autant qu'une variable de départ* : il mérite d'être conservé.
- Un axe tel que $\lambda_s < 1$ résume moins qu'une seule variable : on l'écarte.

Le critère de Kaiser consiste donc à retenir les axes dont la valeur propre est supérieure ou égale à 1 (ici, avec $K = 3$ variables, les axes qui portent au moins $\frac{1}{3} \approx 33{,}3\,\%$ de l'inertie).

Combien d'axes garder ? (2) Le critère du coude

On trace l'**éboulis des valeurs propres** : un diagramme en bâtons des valeurs propres $\lambda_1 \geq \lambda_2 \geq \ldots$ en fonction du numéro de l'axe. En général, les valeurs propres diminuent d'abord vite (les premiers axes portent la structure des données), puis lentement (les derniers axes ne portent plus que de petites variations sans grande signification). Le **critère du coude** consiste à garder les axes situés *avant* la cassure de l'éboulis, là où la décroissance ralentit nettement.



Exemple fictif avec $K = 8$ variables : on garde les axes 1 et 2 (les deux critères sont ici d'accord). Ces critères sont des aides, pas des règles absolues : on les confronte, on regarde le pourcentage d'inertie cumulé, et l'on garde surtout les axes que l'on sait interpréter.

c) Calculer le pourcentage d'inertie de chaque axe, $100 \times \lambda_s / K$, et le pourcentage cumulé du plan (F_1, F_2). Combien d'axes retient le critère de Kaiser ?

d) Tracer l'éboulis des valeurs propres de cette ACP. Le critère du coude est-il concluant ici ? Combien d'axes décidez-vous finalement de retenir ?

4. Les vecteurs propres : la direction des axes

Rappel : le produit scalaire (utile dans tout l'exercice)

Soient deux vecteurs $\vec{u} = (u_1\,;u_2\,;\ldots;u_n)$ et $\vec{v} = (v_1\,;v_2\,;\ldots;v_n)$ de $\mathbb{R}^n$ (ici $n = 3$).

- **Définition.** Leur produit scalaire est le *nombre*

$$\vec{u} \cdot \vec{v} = u_1v_1 + u_2v_2 + \cdots + u_nv_n = \sum_{k=1}^{n} u_k\, v_k.$$

  Par exemple, $(1\,;2\,;-1) \cdot (3\,;0\,;4) = 1 \times 3 + 2 \times 0 + (-1) \times 4 = -1$.

- **Norme.** $\vec{u} \cdot \vec{u} = u_1^2 + \cdots + u_n^2 = \|\vec{u}\|^2$: c'est le carré de la longueur de $\vec{u}$ (Pythagore). Un vecteur est *unitaire* si $\|\vec{u}\| = 1$.

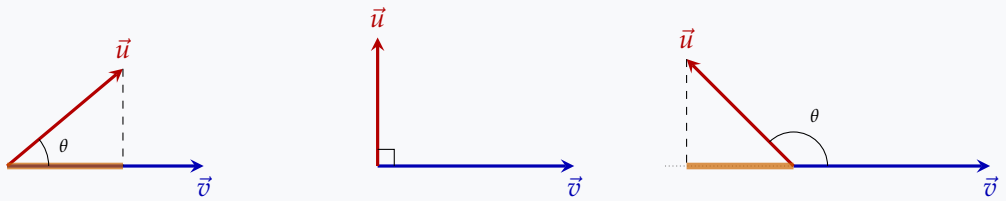
- **Règles de calcul.** $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$, et $(a\,\vec{u} + b\,\vec{w}) \cdot \vec{v} = a\,(\vec{u} \cdot \vec{v}) + b\,(\vec{w} \cdot \vec{v})$ pour tous nombres a et b : on développe comme un produit ordinaire.

- **Formule géométrique.** Si θ est l'angle entre les deux vecteurs,

$$\vec{u} \cdot \vec{v} = \|\vec{u}\| \times \|\vec{v}\| \times \cos\theta.$$

  Le signe du produit scalaire indique donc si l'angle est aigu, droit ou obtus ; en particulier, $\vec{u}$ et $\vec{v}$ sont **orthogonaux** (perpendiculaires) si et seulement si $\vec{u} \cdot \vec{v} = 0$.

- **Projection.** Si $\vec{v}$ est *unitaire*, $\vec{u} \cdot \vec{v} = \|\vec{u}\| \cos\theta$ est la longueur (comptée avec un signe) de la projection orthogonale de $\vec{u}$ sur la droite dirigée par $\vec{v}$. C'est ce qui servira à calculer les coordonnées des parcelles sur les axes (question 5).



θ aigu
$\vec{u} \cdot \vec{v} > 0$

$\theta = 90°$
$\vec{u} \cdot \vec{v} = 0$

θ obtus
$\vec{u} \cdot \vec{v} < 0$

En orange : la projection de $\vec{u}$ sur la droite de $\vec{v}$. Lorsque $\vec{v}$ est unitaire, sa longueur signée est exactement $\vec{u} \cdot \vec{v}$.

a) Pour chaque valeur propre, résoudre le système $RU = \lambda U$, où $U = (u_1\,;u_2\,;u_3)$, et en déduire un vecteur propre *unitaire*, c'est-à-dire tel que $\|U\| = \sqrt{u_1^2+u_2^2+u_3^2} = 1$. On notera U_1, U_2, U_3 les vecteurs obtenus pour λ_1, λ_2, λ_3, et u_{sk} la k-ième coordonnée de U_s : $U_s = (u_{s1}\,;u_{s2}\,;u_{s3})$.

✓ **Contrôle** chaque vecteur propre a une coordonnée nulle.

b) Vérifier que ces trois vecteurs sont deux à deux orthogonaux (produit scalaire nul). Ils définissent les directions des trois nouveaux axes.

5. Les individus : coordonnées et carte des parcelles

Notations

- L'**axe** s ($s = 1,2,3$) est la droite qui passe par G et qui a pour direction le vecteur propre unitaire U_s. On le note aussi F_s :

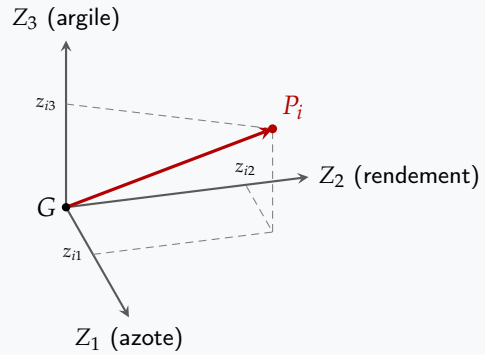
  F_1 : axe 1, de direction U_1, F_2 : axe 2, de direction U_2, F_3 : axe 3, de direction U_3.

- $F_s(i)$ désigne la **coordonnée de la parcelle P_i sur l'axe** s. Par exemple, $F_2(4)$ est la coordonnée de P_4 sur l'axe 2.
- F_s désigne aussi la colonne $\big(F_s(1)\,;\ldots;F_s(5)\big)$ des coordonnées des cinq parcelles. C'est une *nouvelle variable*, appelée **composante principale**, dont on peut calculer la moyenne, la variance ou la corrélation avec les autres variables.

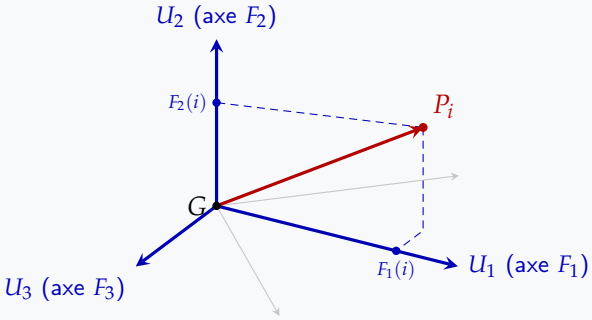
a) **La formule des coordonnées.**

Changer de repère

Au départ, chaque parcelle P_i est repérée dans le repère $(G\,;Z_1,Z_2,Z_3)$ des variables centrées réduites : ses coordonnées sont $(z_{i1}\,;z_{i2}\,;z_{i3})$, lues dans le tableau de la question 1. L'ACP fournit un **nouveau repère** orthonormé $(G\,;U_1,U_2,U_3)$, formé des vecteurs propres de la question 4 : même origine G, mêmes points, mais des axes *tournés*. **On cherche** les coordonnées $\big(F_1(i)\,;F_2(i)\,;F_3(i)\big)$ du *même* vecteur $\overrightarrow{GP_i}$ dans ce nouveau repère.



au départ :
$P_i = (z_{i1}\,;z_{i2}\,;z_{i3})$

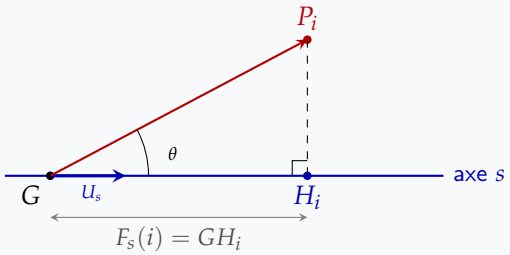


on cherche :
$P_i = \big(F_1(i)\,;F_2(i)\,;F_3(i)\big)$

À gauche, les coordonnées de départ se lisent en projetant P_i sur les axes gris Z_1, Z_2, Z_3. À droite, le même point est projeté sur les axes bleus U_1, U_2 (et U_3, non dessiné pour ne pas surcharger la figure).

D'où vient la formule des coordonnées ?

L'axe s passe par G et a pour direction le vecteur $U_s = (u_{s1}\,;u_{s2}\,;u_{s3})$, qui est *unitaire* ($\|U_s\| = 1$). La coordonnée $F_s(i)$ de P_i sur l'axe est la position de son projeté orthogonal H_i sur cet axe (comptée positivement dans le sens de U_s, négativement dans l'autre sens) :



D'après le rappel de la question 4 (projection sur un vecteur unitaire), cette coordonnée est exactement le **produit scalaire** entre le vecteur de la parcelle et le vecteur unitaire de l'axe : dans le triangle rectangle $G\,H_i\,P_i$, $GH_i = GP_i \times \cos\theta = \overrightarrow{GP_i} \cdot U_s$. Il se calcule composante par composante :

$$F_s(i) = \overrightarrow{GP_i} \cdot U_s = z_{i1}\,u_{s1} + z_{i2}\,u_{s2} + z_{i3}\,u_{s3} = \sum_{k=1}^{K} u_{sk}\,z_{ik},$$

où u_{sk} est la k-ième composante du vecteur propre U_s. (C'est pour cela qu'on impose $\|U_s\| = 1$: avec un vecteur de longueur 2, toutes les coordonnées seraient doublées.)

Exemple. Une parcelle fictive située à un écart-type au-dessus de la moyenne pour l'azote et à la moyenne pour le reste a pour vecteur $(1\,;0\,;0)$; sur un axe de direction $U = \frac{1}{\sqrt{2}}(1\,;1\,;0)$, sa coordonnée vaut $1 \times \frac{1}{\sqrt{2}} + 0 \times \frac{1}{\sqrt{2}} + 0 \times 0 = \frac{1}{\sqrt{2}} \approx 0{,}71$.

À l'aide des vecteurs propres trouvés à la question 4, montrer que

$$F_1(i) = \frac{z_{i1} + z_{i2}}{\sqrt{2}}, \qquad F_2(i) = z_{i3}, \qquad F_3(i) = \frac{z_{i1} - z_{i2}}{\sqrt{2}}.$$

b) Compléter le tableau.

Astuce. Calculer d'abord $\sqrt{2}\,F_1$ et $\sqrt{2}\,F_3$, qui sont des entiers, puis les carrés (par exemple $F_1^2 = (\sqrt{2}\,F_1)^2/2$).

	Coordonnées				Carrés		
Parcelle	$\sqrt{2}\,F_1$	F_1 (approx.)	F_2	$\sqrt{2}\,F_3$	F_1^2	F_2^2	F_3^2
P_1							
P_2							
P_3							
P_4							
P_5							

✓ **Contrôle** pour chaque parcelle, $F_1^2 + F_2^2 + F_3^2 = d^2(P_i, G)$ (dernière colonne de la question 1) : changer d'axes ne change pas les distances.

c) Vérifier que $\mathrm{Var}(F_1) = \lambda_1$ et $\mathrm{Var}(F_2) = \lambda_2$.

d) Tracer la carte des individus dans le plan (F_1, F_2) (même échelle sur les deux axes), en indiquant sur chaque axe son pourcentage d'inertie. Que remarquez-vous pour P_4 et P_5 ?

6. Les variables : le cercle des corrélations

Idée clé. F_1 et F_2 sont de nouvelles variables (une valeur par parcelle). La coordonnée d'une variable sur un axe est simplement sa **corrélation** avec cette nouvelle variable.

a) **Les coordonnées des variables.**

D'où vient la formule de calcul ?

On applique la définition du coefficient de corrélation, rappelée à la question 2 a), à $A = X_k$ (une variable de départ) et $B = F_s$ (la nouvelle variable « coordonnée sur l'axe s ») :

- centrer et réduire ne change pas une corrélation : $r(X_k, F_s) = r(Z_k, F_s)$, et l'on peut travailler avec les valeurs z_{ik} du tableau de la question 1 ; Z_k est centrée ($\overline{Z}_k = 0$) et réduite ($\sigma_{Z_k} = 1$) ;
- F_s est *centrée*, c'est-à-dire de moyenne nulle. En effet, la colonne F_s est une combinaison des colonnes Z_1, Z_2, Z_3 : $F_s(i) = u_{s1}\, z_{i1} + u_{s2}\, z_{i2} + u_{s3}\, z_{i3}$. En faisant la moyenne sur les I parcelles, les coefficients u_{sk} (qui ne dépendent pas de i) sortent de la somme :

$$\overline{F}_s = \frac{1}{I}\sum_{i=1}^{I} F_s(i) = u_{s1}\,\overline{Z}_1 + u_{s2}\,\overline{Z}_2 + u_{s3}\,\overline{Z}_3 = u_{s1}\times 0 + u_{s2}\times 0 + u_{s3}\times 0 = 0;$$

- F_s a pour variance λ_s, comme on l'a vérifié à la question 5 c) ; son écart-type est donc $\sigma_{F_s} = \sqrt{\lambda_s}$.

Les deux moyennes étant nulles, la covariance se réduit à la moyenne des produits, $\mathrm{Cov}(Z_k, F_s) = \frac{1}{I}\sum_i z_{ik}\, F_s(i)$, et il reste

$$r(X_k, F_s) = \frac{\mathrm{Cov}(Z_k, F_s)}{\sigma_{Z_k}\,\sigma_{F_s}} = \frac{\dfrac{1}{I}\displaystyle\sum_{i=1}^{I} z_{ik}\, F_s(i)}{1\times\sqrt{\lambda_s}} = \frac{1}{\sqrt{\lambda_s}}\left(\frac{1}{I}\sum_{i=1}^{I} z_{ik}\, F_s(i)\right).$$

Lecture. Le numérateur est grand et positif quand les parcelles qui ont une forte valeur de X_k sont aussi celles qui ont une forte coordonnée sur l'axe s ; la flèche de X_k pointe alors dans le sens positif de l'axe.

Calculer ainsi les six coordonnées $r(X_k, F_1)$ et $r(X_k, F_2)$, $k = 1, 2, 3$.

Astuce. Une corrélation ne change pas si l'on multiplie une colonne par un nombre positif ; on peut donc utiliser $\sqrt{2}\,F_1$ à la place de F_1, à condition de diviser par *son* écart-type.

Variable	$r(X_k, F_1)$	$r(X_k, F_2)$
X_1 (azote)		
X_2 (rendement)		
X_3 (argile)		

b) **Un raccourci.**

D'où vient le raccourci $r(X_k, F_s) = \sqrt{\lambda_s}\, u_{sk}$?

On part de la formule de a) et l'on remplace $F_s(i)$ par $\sum_{l=1}^{K} u_{sl}\, z_{il}$ (question 5 a)) :

$$\begin{aligned}
\frac{1}{I}\sum_{i=1}^{I} z_{ik}\, F_s(i) &= \frac{1}{I}\sum_{i=1}^{I} z_{ik}\Big(\sum_{l=1}^{K} u_{sl}\, z_{il}\Big) && \text{(1) on remplace } F_s(i)\\
&= \sum_{l=1}^{K} u_{sl}\Big(\frac{1}{I}\sum_{i=1}^{I} z_{ik}\, z_{il}\Big) && \text{(2) } u_{sl} \text{ ne dépend pas de } i \text{ : on échange les sommes}\\
&= \sum_{l=1}^{K} r_{kl}\, u_{sl} && \text{(3) question 2 a)}\\
&= \big(R\, U_s\big)_k && \text{(4) } k\text{-ième ligne du produit de } R \text{ par } U_s\\
&= \lambda_s\, u_{sk} && \text{(5) car } R\, U_s = \lambda_s\, U_s \text{ (question 4).}
\end{aligned}$$

En divisant par $\sqrt{\lambda_s}$ comme dans la formule de a) :

$$r(X_k, F_s) = \frac{\lambda_s\, u_{sk}}{\sqrt{\lambda_s}} = \sqrt{\lambda_s}\, u_{sk}.$$

Toute la « magie » est à l'étape (5) : c'est parce que U_s est un vecteur *propre* de R que le calcul se simplifie autant.

Vérifier l'étape (5) pour $k = 1$ et $s = 1$ avec la matrice R et le vecteur U_1. Retrouver ensuite les six coordonnées de a) avec ce raccourci. Lequel des deux calculs est le plus rapide ?

c) Vérifier que, pour chaque axe,

$$\sum_{k=1}^{K} r^2(X_k, F_s) = \lambda_s.$$

Calculer aussi $r(X_k, F_3)$ et vérifier que, pour chaque variable,

$$r^2(X_k, F_1) + r^2(X_k, F_2) + r^2(X_k, F_3) = 1.$$

Comment comprendre ces deux relations ?

Un carré de corrélation $r^2(X_k, F_s)$ est la **part de la variance de X_k « expliquée » par l'axe s** (comme le coefficient de détermination R^2 d'une régression linéaire simple de Z_k sur F_s). Rangeons ces parts dans un tableau : une ligne par variable, une colonne par axe.

	axe 1	axe 2	axe 3	**total de la ligne**
X_1	$r^2(X_1, F_1)$	$r^2(X_1, F_2)$	$r^2(X_1, F_3)$	$1 = \mathrm{Var}(Z_1)$
X_2	$r^2(X_2, F_1)$	$r^2(X_2, F_2)$	$r^2(X_2, F_3)$	$1 = \mathrm{Var}(Z_2)$
X_3	$r^2(X_3, F_1)$	$r^2(X_3, F_2)$	$r^2(X_3, F_3)$	$1 = \mathrm{Var}(Z_3)$
total de la colonne	λ_1	λ_2	λ_3	$K = 3$

- **En colonne : $\sum_k r^2(X_k, F_s) = \lambda_s$.** Avec le raccourci de b), la démonstration tient en une ligne : $\sum_k r^2(X_k, F_s) = \sum_k \lambda_s\, u_{sk}^2 = \lambda_s\, \|U_s\|^2 = \lambda_s$. Cela donne un *second sens* à la valeur propre : λ_s est la quantité totale de variance des variables que l'axe s capte. Un axe tel que $\lambda_s = 1{,}8$ résume à lui seul l'équivalent de 1,8 variable de départ : c'est exactement le raisonnement du critère de Kaiser. L'axe 1 est, parmi toutes les directions possibles, celle qui rend cette somme la plus grande : c'est la nouvelle variable *la plus corrélée à l'ensemble* des variables de départ.
- **En ligne : $\sum_s r^2(X_k, F_s) = 1$.** La variance de chaque variable (égale à 1) est *entièrement répartie* entre les K axes, sans perte ni double compte : les axes étant non corrélés entre eux, leurs parts s'ajoutent (c'est un théorème de Pythagore, démontré à la question 9). Ce que cela apporte : $r^2(X_k, F_1) + r^2(X_k, F_2)$ est la part de la variable X_k visible sur le plan (F_1, F_2), et ce qui manque pour aller jusqu'à 1 est « caché » sur l'axe 3. On y reviendra à la question 7 (qualité de représentation des variables).

- Le total général vaut $K = 3$ des deux façons : c'est l'inertie totale de la question 1, répartie entre les variables (en ligne) ou entre les axes (en colonne).

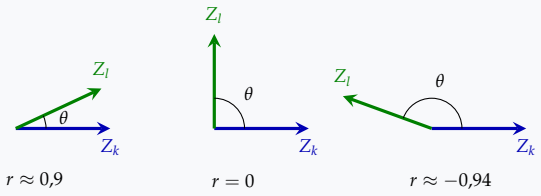
d) **Le cercle des corrélations.**

Pourquoi une variable a-t-elle pour coordonnées $\big(r(X_k, F_1)\,;\, r(X_k, F_2)\big)$?

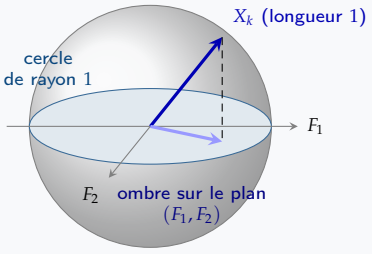
Changeons de point de vue. Au lieu de voir les parcelles comme des points dans l'espace des variables, voyons chaque *variable* comme un vecteur : la colonne $Z_k = (z_{1k}\,;\ldots;z_{5k})$, qui a $I = 5$ composantes. Sur ces colonnes, on utilise le produit scalaire « moyen » $\langle A, B\rangle = \frac{1}{I}\sum_i a_i\, b_i$ (le produit scalaire du rappel de la question 4, divisé par I). Alors :

- $\langle Z_k, Z_k\rangle = \frac{1}{I}\sum_i z_{ik}^2 = \mathrm{Var}(Z_k) = 1$: chaque variable centrée réduite est un vecteur de **longueur** 1 ;
- $\langle Z_k, Z_l\rangle = \frac{1}{I}\sum_i z_{ik}\, z_{il} = r_{kl}$ (question 2 a)). D'après la formule géométrique du produit scalaire, $r_{kl} = 1 \times 1 \times \cos\theta$: **le coefficient de corrélation est le cosinus de l'angle entre les deux variables** ;
- la colonne F_s a pour longueur $\sqrt{\mathrm{Var}(F_s)} = \sqrt{\lambda_s}$; le vecteur unitaire de l'axe s est donc $F_s/\sqrt{\lambda_s}$, et $\big\langle Z_k, F_s/\sqrt{\lambda_s}\big\rangle = \frac{1}{\sqrt{\lambda_s}} \cdot \frac{1}{I}\sum_i z_{ik}\, F_s(i) = r(X_k, F_s)$ (formule de a).

Comme pour les parcelles (question 5 a)), la coordonnée d'un vecteur sur un axe est son produit scalaire avec le vecteur unitaire de l'axe : **la coordonnée de X_k sur l'axe s est donc $r(X_k, F_s)$**. Les variables étant des vecteurs de longueur 1, leurs extrémités sont sur une sphère de rayon 1 ; vue sur le plan (F_1, F_2), chaque flèche est l'ombre de cette flèche de longueur 1, et tombe donc *à l'intérieur* du cercle de rayon 1.



$r_{kl} = \cos\theta$: angle aigu, droit ou obtus



Tracer le cercle des corrélations : un cercle de **rayon** 1 et, pour chaque variable, une flèche partant de l'origine jusqu'au point de coordonnées $\big(r(X_k, F_1)\,;\, r(X_k, F_2)\big)$. Que remarquez-vous pour l'azote et le rendement ?

7. Les aides à l'interprétation

Pourquoi des aides à l'interprétation ?

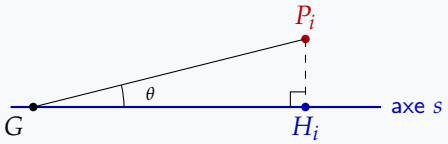
La carte de la question 5 est une *photographie* en deux dimensions d'un nuage qui en compte trois. Comme toute photographie, elle peut tromper : deux parcelles qui paraissent proches sur la carte peuvent être éloignées dans la réalité, si elles diffèrent selon la direction qui n'est pas photographiée (ici l'axe 3). Deux indicateurs permettent de contrôler la lecture : la **qualité de représentation** ($\cos^2$), qui dit si un point est *bien vu* sur la carte, et la **contribution**, qui dit s'il a *construit* l'axe.

La qualité de représentation $\cos^2$

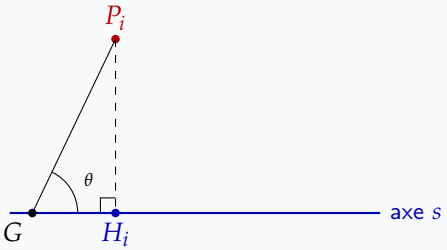
Plaçons-nous dans l'espace des parcelles. Le vecteur $\overrightarrow{GP_i}$ relie le centre de gravité à la parcelle ; sa longueur au carré est $d^2(P_i, G)$ (question 1). Sa projection orthogonale sur l'axe s est le vecteur $\overrightarrow{GH_i}$, de longueur

$|F_s(i)|$. Le triangle $G\,H_i\,P_i$ est rectangle en H_i ; si θ est l'angle entre $\overrightarrow{GP_i}$ et l'axe, alors

$$\cos_s^2(P_i) = \cos^2\theta = \frac{GH_i^2}{GP_i^2} = \frac{F_s(i)^2}{d^2(P_i,G)}.$$



θ petit : $\cos^2\theta$ proche de 1
→ point **bien représenté**



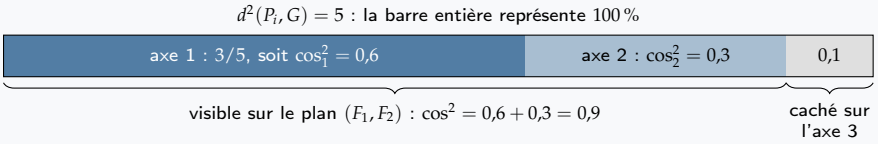
θ grand : $\cos^2\theta$ proche de 0
→ point **mal représenté**

Ce que mesure le $\cos^2$. Le théorème de Pythagore, dans ce triangle rectangle, donne $GP_i^2 = GH_i^2 + H_iP_i^2$, c'est-à-dire

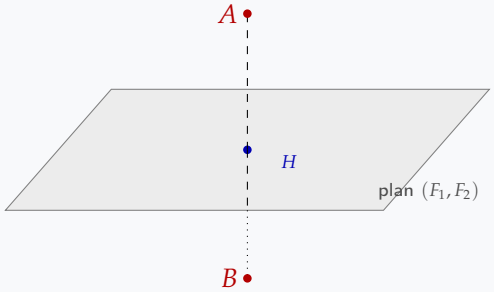
$$\underbrace{d^2(P_i,G)}_{\text{vraie distance}^2\text{ au centre}} = \underbrace{F_s(i)^2}_{\text{conservé par la projection}} + \underbrace{H_iP_i^2}_{\text{perdu (écrasé) par la projection}}.$$

Le $\cos_s^2(P_i)$ est la **fraction de $d^2(P_i,G)$ conservée quand on projette la parcelle sur l'axe**. Si $\cos^2 = 0{,}9$, la projection garde 90 % du carré de la distance au centre : sur l'axe, la parcelle apparaît pratiquement à sa vraie distance de G, on peut se fier à sa position. Si $\cos^2 = 0{,}1$, la projection écrase 90 % de cette distance : la parcelle *paraît* proche du centre alors qu'elle en est loin.

Comme les trois axes sont orthogonaux, Pythagore s'applique encore dans l'espace complet : $d^2(P_i,G) = F_1(i)^2 + F_2(i)^2 + F_3(i)^2$. Le carré de la distance se partage entre les trois axes, et $\cos_1^2(P_i) + \cos_2^2(P_i) + \cos_3^2(P_i) = 1$. Exemple fictif d'une parcelle avec $d^2 = 5$, $F_1^2 = 3$, $F_2^2 = 1{,}5$ et $F_3^2 = 0{,}5$:



La qualité de représentation sur le **plan** (F_1,F_2) est la somme $\cos_1^2 + \cos_2^2$: la part de d^2 visible sur la carte. Ce qui manque pour aller jusqu'à 1 est « caché » sur l'axe 3.



A (au-dessus du plan) et B (au-dessous) se projettent au **même point** H : sur la carte, ils sont confondus. Pourtant ils sont éloignés l'un de l'autre. Seul un $\cos^2$ sur le plan nettement inférieur à 1 permet de s'en douter.

Règle de lecture. On n'interprète la proximité de deux points sur la carte que s'ils sont *tous les deux* bien représentés (somme des $\cos^2$ sur le plan proche de 1).

La contribution

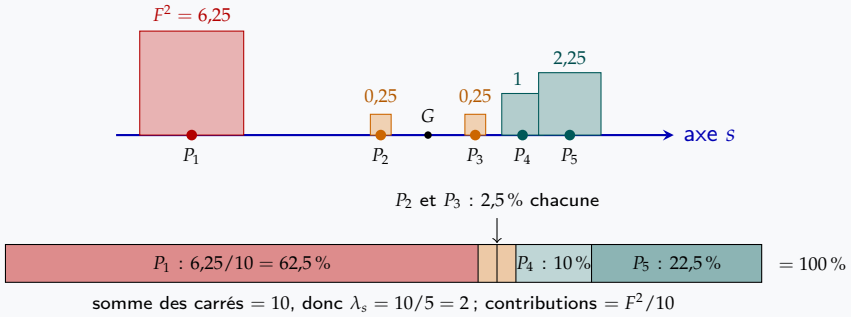
L'inertie d'un axe est une somme de morceaux, un par parcelle. On a vu à la question 5 c) que l'inertie portée par l'axe s est la variance des coordonnées sur cet axe :

$$\lambda_s = \mathrm{Var}(F_s) = \frac{1}{I}\sum_{i=1}^{I} F_s(i)^2 = \underbrace{\tfrac{1}{I}F_s(1)^2}_{\text{dû à } P_1} + \underbrace{\tfrac{1}{I}F_s(2)^2}_{\text{dû à } P_2} + \cdots + \underbrace{\tfrac{1}{I}F_s(I)^2}_{\text{dû à } P_I}.$$

La **contribution** de la parcelle P_i à l'axe s est la part de son morceau dans ce total :

$$\mathrm{contrib}_s(P_i) = \frac{\frac{1}{I}F_s(i)^2}{\lambda_s} \times 100 = \frac{F_s(i)^2}{I\,\lambda_s} \times 100 \quad \text{(en \%)}.$$

Dans l'exemple fictif ci-dessous, cinq parcelles ont sur un axe les coordonnées $-2{,}5$; $-0{,}5$; $0{,}5$; 1 et $1{,}5$. Au-dessus de chacune, on dessine un carré de côté $|F_s(i)|$: son *aire* $F_s(i)^2$ est son morceau d'inertie (à $\frac{1}{I}$ près).



- Ce qui compte est le *carré* de la coordonnée : une parcelle deux fois plus éloignée du centre le long de l'axe pèse quatre fois plus. Ici, P_1, seule à gauche et loin du centre, fait à elle seule près des deux tiers de l'inertie de l'axe.
- Sur chaque axe, les contributions des I parcelles totalisent $100\,\%$; la contribution moyenne est $100/I = 20\,\%$. On repère les parcelles qui dépassent nettement cette moyenne.
- Pourquoi dit-on qu'une telle parcelle « tire » l'axe ? L'ACP choisit l'axe 1 de façon à rendre λ_1, donc la somme des carrés, la plus grande possible. Un point très éloigné du centre a un très grand carré : l'axe a intérêt à s'orienter dans sa direction pour le « récupérer » en entier. La direction de l'axe dépend alors beaucoup de cette seule parcelle.

a) Compléter le tableau suivant (on pourra utiliser les colonnes F_1^2 et F_2^2 de la question 5 et la colonne $d^2(P_i, G)$ de la question 1).

	Qualité de représentation				Contribution (%)	
Parcelle	$\cos^2_1$	$\cos^2_2$	$\cos^2$ plan	$\cos^2_3$	axe 1	axe 2
P_1						
P_2						
P_3						
P_4						
P_5						
Total	—	—	—	—		

b) **Vérification sur un exemple.** Pour la parcelle P_4, calculer séparément ses trois qualités de représentation

$$\cos^2_1(P_4) = \frac{F_1(4)^2}{d^2(P_4,G)}, \qquad \cos^2_2(P_4) = \frac{F_2(4)^2}{d^2(P_4,G)}, \qquad \cos^2_3(P_4) = \frac{F_3(4)^2}{d^2(P_4,G)},$$

en prenant les carrés $F_s(4)^2$ dans le tableau de la question 5 b) et $d^2(P_4,G)$ dans celui de la question 1. Vérifier que leur somme vaut 1, puis expliquer pourquoi il en est de même pour toutes les parcelles.

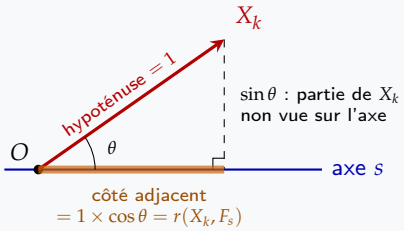
c) Quelles parcelles sont parfaitement représentées sur le plan ? Où se trouvent-elles par rapport au plan (F_1,F_2) ? Quelles parcelles le sont moins bien, et quelle part de leur position est « cachée » sur l'axe 3 ? Quelle précaution cela impose-t-il pour lire la carte de la question 5 ?

d) Quelle parcelle contribue le plus à l'axe 1 ? à l'axe 2 ? Comparer à la contribution moyenne.

e) Comparer P_1 et P_2 : toutes deux sont parfaitement représentées sur le plan, mais leurs contributions sont très différentes. Pourquoi ? En déduire que le $\cos^2$ et la contribution ne mesurent pas la même chose.

Et pour les variables ? La qualité de représentation

Les mêmes notions s'appliquent aux variables, lues sur le cercle des corrélations. On a vu à la question 6 d) que chaque variable centrée réduite est une flèche de **longueur** 1, et que sa coordonnée sur l'axe s est son produit scalaire avec le vecteur unitaire de l'axe, c'est-à-dire $r(X_k,F_s)$.

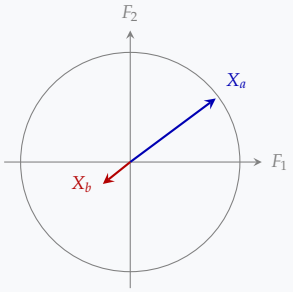


Dans ce triangle rectangle d'hypoténuse 1, le côté adjacent vaut $\cos\theta$. Or ce côté est la coordonnée de la variable sur l'axe : $\cos\theta = r(X_k,F_s)$, d'où

$$\cos^2_s(X_k) = \cos^2\theta = r^2(X_k,F_s).$$

C'est la part de la longueur^2 de la flèche (donc de la variance de X_k) vue sur l'axe.

La flèche ayant une longueur 1 dans l'espace complet, Pythagore donne $r^2(X_k,F_1) + r^2(X_k,F_2) + r^2(X_k,F_3) = 1$ (question 6 c)), et la qualité de représentation sur le plan, $r^2(X_k,F_1) + r^2(X_k,F_2)$, est le **carré de la longueur de la flèche dessinée sur le cercle**.



X_a : flèche qui touche presque le cercle, **bien représentée** (longueur $\approx 0{,}97$, $\cos^2$ sur le plan $\approx 0{,}94$)

X_b : flèche courte, **mal représentée** (longueur $\approx 0{,}32$, $\cos^2 \approx 0{,}1$) : son information est sur d'autres axes

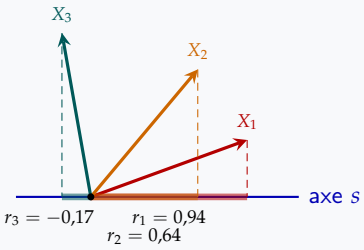
Règle de lecture. On ne lit l'angle entre deux flèches (donc leur corrélation) que si les deux flèches arrivent près du cercle.

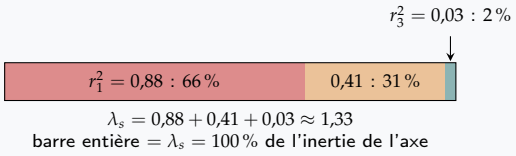
Et pour les variables ? La contribution

La valeur propre λ_s est la somme des carrés des corrélations des variables avec l'axe :

$$\lambda_s = \sum_{k=1}^{K} r^2(X_k,F_s).$$

On l'a démontré à la question 6 c) : $\sum_k r^2(X_k,F_s) = \sum_k \lambda_s u^2_{sk} = \lambda_s \|U_s\|^2 = \lambda_s$, car U_s est unitaire. Géométriquement, $r(X_k,F_s)$ est l'ombre de la flèche X_k (de longueur 1) sur l'axe s : λ_s est la *somme des carrés des ombres* des K flèches sur l'axe. Exemple fictif avec trois flèches :





Les deux points de vue donnent donc la même valeur propre, ce qui est un excellent contrôle de calcul :

côté parcelles (question 5)	**côté variables** (question 6)
$\lambda_s = \dfrac{1}{I}\sum_{i=1}^{I} F_s(i)^2$	$\lambda_s = \sum_{k=1}^{K} r^2(X_k, F_s)$

Comme pour les parcelles, la contribution de la variable X_k à l'axe s est sa part dans cette somme :

$$\text{contrib}_s(X_k) = \frac{r^2(X_k, F_s)}{\lambda_s} \times 100 \quad \text{(en \%)}.$$

Sur chaque axe, les contributions des K variables totalisent $100\,\%$; la contribution moyenne est $100/K \approx 33{,}3\,\%$. Une variable de forte contribution est une variable qui a *construit* l'axe : c'est elle qu'il faut regarder pour nommer l'axe.

f) Compléter le tableau suivant à l'aide des coordonnées des variables (question 6).

	Coordonnées		Qualité de représentation				Contrib. (%)	
Variable	$r(X_k, F_1)$	$r(X_k, F_2)$	$\cos^2_1$	$\cos^2_2$	$\cos^2$ plan	$\cos^2_3$	axe 1	axe 2
X_1 (azote)								
X_2 (rendement)								
X_3 (argile)								
Total	—	—			—			

g) Vérifier que la colonne $\cos^2_1$ totalise λ_1, que la colonne $\cos^2_2$ totalise λ_2 et que la colonne $\cos^2_3$ totalise λ_3, puis que, sur chaque ligne,

$$\cos^2_1 + \cos^2_2 + \cos^2_3 = 1.$$

h) Quelles variables sont parfaitement représentées sur le plan ? Lesquelles le sont moins bien, et où se trouve le reste de leur information ? Quelles variables ont construit l'axe 1 ? l'axe 2 ? En déduire un nom pour chaque axe.

8. Interprétation

a) En faisant des allers-retours entre la carte des parcelles et le cercle des corrélations, décrire les parcelles situées aux extrémités de chaque axe : qu'ont-elles de particulier ? Les noms d'axes proposés à la question 7 h) sont-ils confirmés ?

b) *Proche sur le plan veut-il dire semblable ?* P_4 et P_5 sont au même point du plan. Comparer leurs données brutes. Quel indicateur de la question 7 permettait de prévoir que leur ressemblance n'est qu'apparente ? Sur quel axe diffèrent-elles ?

c) *Deux flèches confondues veulent-elles dire* $r = 1$ *?* Sur le cercle, les flèches de l'azote et du rendement sont superposées, alors que $r_{12} = 0{,}8$. Expliquer ce paradoxe en vérifiant que

$$r_{12} = \sum_{s=1}^{3} r(X_1, F_s)\, r(X_2, F_s).$$

Quelle règle de lecture du cercle en déduisez-vous ?

d) *Un seul individu peut-il fabriquer un axe ?* Commenter la contribution de P_3 à l'axe 2. Que faudrait-il faire en pratique face à un tel axe ?

9. Pour aller plus loin : les parcelles et les variables sur un même graphique

Le problème

La carte des parcelles (question 5) et le cercle des corrélations (question 6) n'ont pas la même échelle. Sur la carte, les coordonnées $F_s(i)$ ont pour variance λ_s ; sur le cercle, les coordonnées des variables sont des corrélations, comprises entre -1 et 1. Superposer les deux graphiques tels quels n'aurait pas de sens. On veut pourtant les réunir sur un seul graphique, appelé **biplot**, en gardant comme référence *le cercle des corrélations de rayon* 1.

L'idée : standardiser les coordonnées des parcelles

On ne touche pas aux flèches des variables ; on divise chaque coordonnée de parcelle par l'écart-type de l'axe, $\sigma_{F_s} = \sqrt{\lambda_s}$. On obtient les **scores standardisés**

$$\left(\frac{F_1(i)}{\sqrt{\lambda_1}} \, ; \, \frac{F_2(i)}{\sqrt{\lambda_2}} \right),$$

qui ont une variance égale à 1 sur chaque axe : ils sont exprimés « en écarts-types ».

Attention. Une variance égale à 1 ne signifie pas que les valeurs sont comprises entre -1 et 1. Les flèches des variables (des corrélations) restent à l'intérieur du cercle ou sur le cercle, mais **les parcelles atypiques peuvent sortir du cercle**, ce qui traduit visuellement leur singularité.

Pourquoi ce choix est le bon (1) : revenir des $F_s(i)$ aux z_{ik}

À la question 5 a), on a calculé les nouvelles coordonnées $F_s(i)$ à partir des anciennes z_{ik}. On peut faire le chemin inverse :

$$z_{ik} = \sum_{s=1}^{K} u_{sk} \, F_s(i).$$

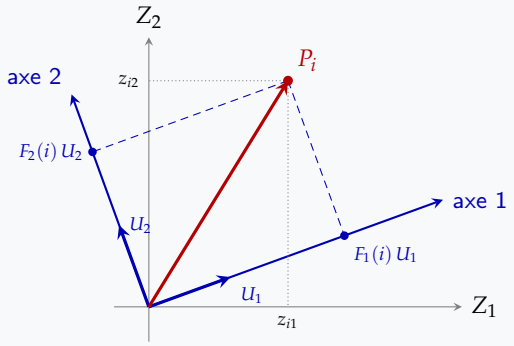
Démonstration. Les vecteurs U_1, U_2, U_3 forment une *base orthonormée* (trois vecteurs unitaires, deux à deux orthogonaux, question 4). Tout vecteur s'écrit donc comme une combinaison de ces trois vecteurs : $\overrightarrow{GP_i} = a_1 \, U_1 + a_2 \, U_2 + a_3 \, U_3$ pour certains nombres a_1, a_2, a_3. Pour trouver a_t, on fait le produit scalaire des deux membres avec U_t (règles de calcul du rappel de la question 4) :

$$\overrightarrow{GP_i} \cdot U_t = a_1 \, (U_1 \cdot U_t) + a_2 \, (U_2 \cdot U_t) + a_3 \, (U_3 \cdot U_t) = a_t,$$

car $U_s \cdot U_t = 0$ si $s \neq t$ (orthogonalité) et $U_t \cdot U_t = 1$ (vecteur unitaire). Or $\overrightarrow{GP_i} \cdot U_t = F_t(i)$ (question 5 a)). Donc $a_t = F_t(i)$, et

$$\overrightarrow{GP_i} = F_1(i) \, U_1 + F_2(i) \, U_2 + F_3(i) \, U_3.$$

Il suffit de lire la k-ième composante des deux membres : à gauche z_{ik}, à droite $F_1(i) \, u_{1k} + F_2(i) \, u_{2k} + F_3(i) \, u_{3k}$. C'est la formule annoncée.



Le *même* vecteur $\overrightarrow{GP_i}$ se décompose dans les deux repères (vue en deux dimensions pour simplifier) :

$$\overrightarrow{GP_i} = z_{i1}\,\vec{e}_1 + z_{i2}\,\vec{e}_2 = F_1(i)\,U_1 + F_2(i)\,U_2.$$

Passer des z aux F (question 5) ou des F aux z (ici), c'est seulement changer de repère.

Pourquoi ce choix est le bon (2) : le produit scalaire reconstitue les données

Avec le raccourci de la question 6 b), $u_{sk} = r(X_k, F_s)/\sqrt{\lambda_s}$, la formule précédente devient

$$z_{ik} = \sum_{s=1}^{K} \underbrace{\frac{F_s(i)}{\sqrt{\lambda_s}}}_{\text{score standardisé}} \times \underbrace{r(X_k, F_s)}_{\text{coordonnée de la flèche}}.$$

Sur le biplot, on ne garde que les axes 1 et 2 (les termes $s = 1$ et $s = 2$) : le **produit scalaire** entre le point d'une parcelle et la flèche d'une variable *reconstitue* donc, au moins approximativement, la valeur centrée réduite z_{ik} de la parcelle pour cette variable. Concrètement, une parcelle située dans la direction d'une flèche a une forte valeur de cette variable ; une parcelle située à l'opposé en a une faible.

a) Compléter le tableau des scores standardisés.

Parcelle	$F_1(i)$	$\frac{F_1(i)}{\sqrt{\lambda_1}}$	$F_2(i)$	$\frac{F_2(i)}{\sqrt{\lambda_2}}$	distance^2 à l'origine sur le biplot
P_1					
P_2					
P_3					
P_4					
P_5					

Tracer ensuite le **biplot** : sur un même graphique, le cercle des corrélations de rayon 1 avec les flèches des trois variables (question 6 d)) *et* les cinq parcelles, placées en leurs scores standardisés. Certaines parcelles sont-elles hors du cercle ? Est-ce gênant ?

b) Calculer le produit scalaire entre le point P_3 et la flèche de l'argile, puis entre P_1 et la flèche de l'azote, puis entre P_4 et la flèche de l'azote, en complétant le tableau. Comparer aux valeurs centrées réduites z_{ik} de la question 1.

Parcelle	**Variable**	point de la parcelle $(a\,;b)$	flèche de la variable $(c\,;d)$	produit scalaire $ac+bd$	z_{ik}	exact ou approché ?
P_3	argile					
P_1	azote					
P_4	azote					

Pourquoi la reconstitution est-elle exacte dans deux cas et seulement approchée dans le troisième ? (Penser à la qualité de représentation.)

Indication ▼ Correction ▼ [stat-0108]

Exercice 2 ACP à la main : le climat de cinq villes du Cameroun

Dans le cadre d'une étude agro-climatologique, un chercheur veut résumer la variabilité du climat de cinq villes camerounaises situées dans des zones différentes : **Yaoundé, Douala, Édéa, Kribi et Ngaoundéré**. Il retient trois variables ($K = 3$) :

- X_1 **(pluie) :** pluviométrie annuelle moyenne (en centaines de mm) ;
- X_2 **(température) :** température moyenne annuelle (°C) ;
- X_3 **(ensoleillement) :** durée moyenne d'ensoleillement (heures par jour).

Ville	X_1 : pluie (centaines de mm)	X_2 : température (°C)	X_3 : ensoleillement (h/j)
Yaoundé	18	25	6
Douala	32	27	3
Édéa	24	28	4
Kribi	26	24	5
Ngaoundéré	20	21	7

À la main, en autonomie

- On réalise une ACP normée de ce tableau ($I = 5$ villes, $K = 3$ variables), avec des variances « population » (division par I).
- Les valeurs ont été arrondies et simplifiées pour que les calculs se fassent à la main. On gardera les valeurs exactes (racines carrées) le plus longtemps possible.
- Les questions sont moins détaillées que dans l'exercice précédent : on s'inspirera de sa démarche, et l'on vérifiera chaque résultat avant de passer au suivant.

1. Préparer les données

a) Calculer les moyennes $\overline{X}_k$ et les écarts-types σ_k des trois variables.

b) Calculer le tableau des données centrées réduites z_{ik} et, pour chaque ville, le carré de sa distance au centre de gravité $d^2(i, G)$. Vérifier que l'inertie totale vaut K.

c) Quelle ville est la plus atypique ? Laquelle est la plus proche de la « ville moyenne » ?

2. Les corrélations

a) Calculer la matrice des corrélations R. Commenter les liaisons obtenues : sont-elles plausibles du point de vue du climat ?

b) On note Z la matrice 5×3 des données centrées réduites, obtenue à la question 1 b). Vérifier, sur un coefficient diagonal et un coefficient hors diagonale, que

$$R = \frac{1}{I} Z^{\mathrm{T}} Z.$$

3. Les valeurs propres

a) Calculer et factoriser le polynôme caractéristique $\det(R - \lambda \mathrm{I})$; en déduire les valeurs propres $\lambda_1 \geq \lambda_2 \geq \lambda_3$.

b) Donner le pourcentage d'inertie de chaque axe et le pourcentage cumulé.

c) Combien d'axes faut-il retenir selon le critère de Kaiser ? selon le critère du coude ? Faut-il pour autant renoncer à interpréter le troisième axe ?

4. Les vecteurs propres

a) Déterminer les vecteurs propres unitaires U_1, U_2, U_3 de R (on choisira le signe de sorte que la première coordonnée soit positive) et vérifier qu'ils sont deux à deux orthogonaux.

b) *Une valeur propre nulle.* On a trouvé $\lambda_3 = 0$.

i. Que valent la moyenne et la variance de la composante F_3 (la colonne des coordonnées des cinq villes sur l'axe 3) ? En déduire la valeur de $F_3(i)$ pour chacune des cinq villes.

ii. Écrire $F_3(i)$ à l'aide des coordonnées de U_3 et des z_{ik}, puis remplacer chaque z_{ik} par $\left(x_{ik} - \overline{X}_k\right)/\sigma_k$. En déduire que les cinq villes vérifient une même relation

$$a\,x_{i1} + b\,x_{i2} + c\,x_{i3} = d,$$

où a, b, c et d sont des entiers à déterminer. Vérifier cette relation sur les données brutes des cinq villes.

iii. Que permet de calculer cette relation pour une ville dont on connaît la pluie et la température ? Le *sens* de la relation (signes des coefficients) est-il plausible pour le climat ? Son *exactitude* l'est-elle ?

iv. Conséquences pour l'ACP : où se trouve le nuage des cinq villes dans l'espace à trois dimensions ? Que vaudront les qualités de représentation des villes et des variables sur le plan (F_1, F_2) ? Combien de variables « indépendantes » le tableau contient-il réellement ?

5. Les variables

a) Calculer les coordonnées $r(X_k, F_1)$ et $r(X_k, F_2)$ des trois variables, leurs qualités de représentation sur le plan et leurs contributions aux axes 1 et 2.

b) Tracer le cercle des corrélations. Nommer les deux axes.

c) Retrouver r_{12} et r_{13} à partir des angles entre les flèches. Pourquoi cette lecture est-elle exacte ici ?

6. Les villes

a) Calculer les coordonnées $F_1(i)$ et $F_2(i)$ des cinq villes et vérifier que $\mathrm{Var}(F_1) = \lambda_1$ et $\mathrm{Var}(F_2) = \lambda_2$.

b) Tracer la carte des villes dans le plan (F_1, F_2).

c) Calculer les qualités de représentation et les contributions des villes aux axes 1 et 2. Quelles villes construisent chaque axe ?

7. Interprétation et esprit critique

a) En confrontant la carte des villes et le cercle des corrélations, décrire le climat de chaque ville. Retrouver ces descriptions dans les données brutes.

b) Proposer une typologie climatique des cinq villes. Quelle explication géographique peut-on avancer pour l'axe 1 ?

c) Quelles sont les limites de cette analyse ?

Indication ▼ [stat-0062]

Indications

Indication pour l'exercice 1 ▲

1. Écarts à la moyenne de l'azote : -30, -10, 0, 10, 30 ; en déduire σ_1^2. En ACP normée, chaque variable a une variance égale à 1, donc l'inertie totale vaut K.

2. Deux des trois corrélations sont nulles.

3. $\det(R-\lambda\,\mathrm{I}) = (1-\lambda)\big[(1-\lambda)^2 - r_{12}^2\big]$; résoudre $1-\lambda=0$, puis $1-\lambda=\pm r_{12}$. Pour l'éboulis, comparer les deux écarts $\lambda_1-\lambda_2$ et $\lambda_2-\lambda_3$.

4. En notant $U=(u_1\,;u_2\,;u_3)$, pour $\lambda=1{,}8$ les équations donnent $u_1=u_2$ et $u_3=0$; ne pas oublier de normer $(2u_1^2=1)$.

5. Garder les carrés F_s^2 : ce sont des nombres simples (4,5 ; 0,5 ; 0 ; 2 ; 2 pour F_1^2). Remarquer que $F_2=Z_3$.

6. Avec la colonne $W=\sqrt{2}\,F_1=(-3\,;-1\,;0\,;2\,;2)$: $\mathrm{Var}(W)=3{,}6$, donc $r(X_k,F_1)=\frac{1}{\sqrt{3{,}6}}\times\frac{1}{5}\sum_i z_{ik}W_i$. Comme $F_2=Z_3$, $r(X_k,F_2)=r_{k3}$.

7. Les contributions (en %) à l'axe 1 se calculent par $100\times F_1^2/9$, celles à l'axe 2 par $100\times F_2^2/5$; pour une variable, $\cos_s^2(X_k)=r^2(X_k,F_s)$ et $\mathrm{contrib}_s(X_k)=100\times r^2(X_k,F_s)/\lambda_s$.

8. Regarder le $\cos^2$ sur le plan de P_4 et P_5, et leurs coordonnées sur F_3. Pour c), reprendre les $r(X_k,F_3)$ calculés à la question 6 c).

9. $\sqrt{1{,}8}\approx 1{,}34$; la distance² à l'origine vaut $\frac{F_1(i)^2}{\lambda_1}+\frac{F_2(i)^2}{\lambda_2}$. Le produit scalaire de $(a\,;b)$ et $(c\,;d)$ vaut $ac+bd$.

Indication pour l'exercice 2 ▲

1. $\sigma_1^2=24$, $\sigma_2^2=6$, $\sigma_3^2=2$. Les colonnes $\sqrt{6}\,Z_1$, $\sqrt{6}\,Z_2$ et $\sqrt{2}\,Z_3$ sont formées d'entiers : calculer avec elles.

2. Deux corrélations sont égales ; on peut écrire $\frac{\sqrt{3}}{2}\approx 0{,}866$.

3. En posant $\mu=1-\lambda$, la règle de Sarrus donne $\mu^3-\big(r_{12}^2+r_{13}^2+r_{23}^2\big)\mu+2\,r_{12}r_{13}r_{23}=0$; chercher une racine évidente. Pour le critère de Kaiser, comparer λ_s à 1.

4. Pour λ_1, les deux premières équations donnent $u_1=u_2$. U_3 est orthogonal à U_1 et à U_2. Une variable de variance nulle est constante ; multiplier la relation $z_{i1}+z_{i2}+\sqrt{3}\,z_{i3}=0$ par $2\sqrt{6}$.

5. $r(X_k,F_s)=\sqrt{\lambda_s}\,u_{sk}$; $\cos_s^2(X_k)=r^2(X_k,F_s)$; $\mathrm{contrib}_s(X_k)=100\times r^2(X_k,F_s)/\lambda_s$. Le cosinus de l'angle entre deux flèches de longueur 1 est leur produit scalaire.

6. Avec les colonnes entières de l'indication 1, $2\sqrt{5}\,F_1$ et $2\sqrt{3}\,F_2$ sont des entiers. $\mathrm{contrib}_s(i)=100\times F_s(i)^2/(I\,\lambda_s)$.

7. Regarder l'altitude des villes.

Corrections

Correction de l'exercice 1 ▲

1. Préparer le tableau : centrer et réduire

a) **Moyennes et écarts-types.** Pour chaque variable, on calcule la moyenne sur les $I = 5$ parcelles, puis les écarts à la moyenne, leurs carrés, et l'écart-type « population » $\sigma_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ik} - \overline{X}_k)^2}$.

Parcelle	Azote X_1			Rendement X_2			Argile X_3		
	x_{i1}	$x_{i1}-60$	$(\cdot)^2$	x_{i2}	$x_{i2}-4$	$(\cdot)^2$	x_{i3}	$x_{i3}-20$	$(\cdot)^2$
P_1	30	−30	900	2,5	−1,5	2,25	17	−3	9
P_2	50	−10	100	3,5	−0,5	0,25	17	−3	9
P_3	60	0	0	4	0	0	32	12	144
P_4	70	10	100	5,5	1,5	2,25	17	−3	9
P_5	90	30	900	4,5	0,5	0,25	17	−3	9
Somme	300	0	2 000	20	0	5	100	0	180
Somme/5	**60**		**400**	**4**		**1**	**20**		**36**

La ligne « Somme/5 » donne la moyenne (colonnes x_{ik}) et la variance (colonnes des carrés) :

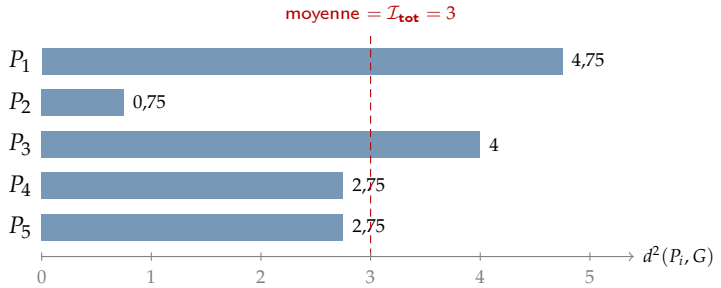
$$\overline{X}_1 = 60,\ \sigma_1 = \sqrt{400} = \boxed{20}\ ; \qquad \overline{X}_2 = 4,\ \sigma_2 = \sqrt{1} = \boxed{1}\ ; \qquad \overline{X}_3 = 20,\ \sigma_3 = \sqrt{36} = \boxed{6}.$$

✓ **Contrôle** les trois écarts-types sont bien entiers, et chaque colonne d'écarts a une somme nulle (c'est toujours le cas : les écarts à la moyenne se compensent).

b) **Données centrées réduites et distances au centre.** On divise chaque écart par l'écart-type de sa variable. Par exemple, $z_{11} = \frac{30-60}{20} = \frac{-30}{20} = -1{,}5$, $z_{42} = \frac{5{,}5-4}{1} = 1{,}5$ et $z_{33} = \frac{32-20}{6} = \frac{12}{6} = 2$.

Parcelle	Z_1 (azote)	Z_2 (rendement)	Z_3 (argile)	$d^2(P_i, G) = z_{i1}^2 + z_{i2}^2 + z_{i3}^2$
P_1	−1,5	−1,5	−0,5	$2{,}25 + 2{,}25 + 0{,}25 = \mathbf{4{,}75}$
P_2	−0,5	−0,5	−0,5	$0{,}25 + 0{,}25 + 0{,}25 = \mathbf{0{,}75}$
P_3	0	0	2	$0 + 0 + 4 = \mathbf{4}$
P_4	0,5	1,5	−0,5	$0{,}25 + 2{,}25 + 0{,}25 = \mathbf{2{,}75}$
P_5	1,5	0,5	−0,5	$2{,}25 + 0{,}25 + 0{,}25 = \mathbf{2{,}75}$
Moyenne	0	0	0	$15/5 = 3$

✓ **Contrôle** toutes les valeurs sont des multiples de 0,5 et chaque colonne Z_k a une moyenne nulle.



c) **Inertie totale.**

$$\mathcal{I}_{\text{tot}} = \frac{1}{5}\sum_{i=1}^{5} d^2(P_i, G) = \frac{4{,}75 + 0{,}75 + 4 + 2{,}75 + 2{,}75}{5} = \frac{15}{5} = \boxed{3} = K.$$

Pourquoi était-ce prévisible ? La somme des d^2 est la somme de *tous* les carrés z_{ik}^2 du tableau. On peut l'additionner ligne par ligne (parcelle par parcelle, ce que l'on vient de faire) ou colonne par colonne (variable par variable) :

$$\mathcal{I}_{\text{tot}} = \frac{1}{I}\sum_i (z_{i1}^2 + z_{i2}^2 + z_{i3}^2) = \underbrace{\tfrac{1}{I}\textstyle\sum_i z_{i1}^2}_{\text{Var}(Z_1)=1} + \underbrace{\tfrac{1}{I}\textstyle\sum_i z_{i2}^2}_{\text{Var}(Z_2)=1} + \underbrace{\tfrac{1}{I}\textstyle\sum_i z_{i3}^2}_{\text{Var}(Z_3)=1} = 3.$$

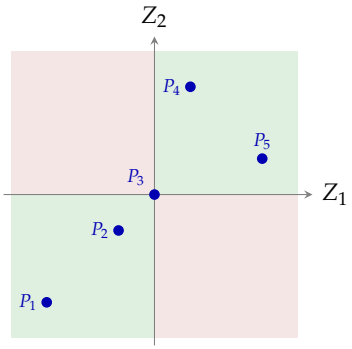
Ici, chaque colonne de carrés a pour somme 5 (par exemple $2{,}25 + 0{,}25 + 0 + 0{,}25 + 2{,}25 = 5$ pour Z_1). Chaque variable centrée réduite a une variance égale à 1 : l'inertie totale d'une ACP normée vaut toujours K, quelles que soient les données. Chaque variable « pèse » 1.

d) **Parcelle la plus atypique.** P_1 est la plus éloignée de la parcelle moyenne ($d^2 = 4{,}75$, la plus longue barre ci-dessus) : elle s'écarte d'un écart-type et demi *à la fois* sur l'azote et sur le rendement (peu d'azote, faible rendement). P_3 vient juste après ($d^2 = 4$), mais à cause d'*une seule* variable : son taux d'argile (32 % contre 17 % partout ailleurs) est à deux écarts-types de la moyenne ($z_{33} = 2$), la valeur $|z_{ik}|$ la plus forte du tableau. À l'opposé, P_2 ($d^2 = 0{,}75$) est la parcelle la plus « ordinaire ».

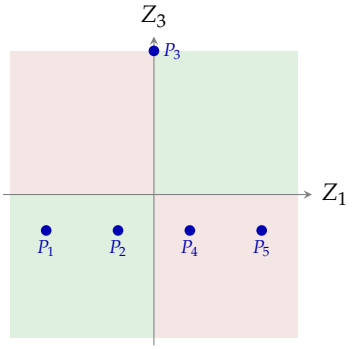
2. La matrice des corrélations

a) **Coefficients de corrélation.** On calcule les produits colonne à colonne, puis leur moyenne :

Parcelle	z_{i1}	z_{i2}	z_{i3}	$z_{i1}\,z_{i2}$	$z_{i1}\,z_{i3}$	$z_{i2}\,z_{i3}$
P_1	$-1{,}5$	$-1{,}5$	$-0{,}5$	2,25	0,75	0,75
P_2	$-0{,}5$	$-0{,}5$	$-0{,}5$	0,25	0,25	0,25
P_3	0	0	2	0	0	0
P_4	0,5	1,5	$-0{,}5$	0,75	$-0{,}25$	$-0{,}75$
P_5	1,5	0,5	$-0{,}5$	0,75	$-0{,}75$	$-0{,}25$
Somme				4	0	0
Moyenne $= r$				$\mathbf{r_{12} = 0{,}8}$	$\mathbf{r_{13} = 0}$	$\mathbf{r_{23} = 0}$



azote et rendement : tous les produits sont ≥ 0, donc $r_{12} > 0$



azote et argile : produits positifs (P_1, P_2) et négatifs (P_4, P_5) qui se compensent : $r_{13} = 0$

b) **Matrice des corrélations** (symétrique, avec des 1 sur la diagonale, car $r_{kk} = \frac{1}{5}\sum_i z_{ik}^2 = \text{Var}(Z_k) = 1$) :

$$R = \begin{pmatrix} 1 & 0{,}8 & 0 \\ 0{,}8 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

c) **Interprétation.** L'azote et le rendement sont fortement et positivement corrélés ($r_{12} = 0{,}8$) : les parcelles qui ont reçu plus d'azote produisent en général davantage, ce qui est agronomiquement plausible (l'azote est souvent le premier facteur limitant du maïs). La corrélation n'est pas égale à 1 : P_5, la plus fertilisée, n'a pas le meilleur rendement (rendement décroissant à forte dose, autres facteurs limitants…). L'argile n'est liée ni à l'azote ni au rendement ($r_{13} = r_{23} = 0$) ; c'est un choix des données fictives : en réalité, la texture du sol influence souvent le rendement.

3. Les valeurs propres : combien d'information par axe ?

a) **Polynôme caractéristique.** La dernière ligne de $R - \lambda\,\mathrm{I}$ est $(0 \;\; 0 \;\; 1-\lambda)$: en développant selon cette ligne, seul le dernier terme reste (signe $+$, car $(-1)^{3+3} = 1$), multiplié par le déterminant 2×2 obtenu en barrant la ligne 3 et la colonne 3 :

$$P(\lambda) = \det \begin{pmatrix} 1-\lambda & 0{,}8 & 0 \\ 0{,}8 & 1-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{pmatrix} = (1-\lambda) \times \det \begin{pmatrix} 1-\lambda & 0{,}8 \\ 0{,}8 & 1-\lambda \end{pmatrix} = (1-\lambda)\left[(1-\lambda)^2 - 0{,}8^2\right].$$

Avec $a^2 - b^2 = (a-b)(a+b)$, où $a = 1-\lambda$ et $b = 0{,}8$:

$$P(\lambda) = (1-\lambda)(1-\lambda-0{,}8)(1-\lambda+0{,}8) = (1-\lambda)(0{,}2-\lambda)(1{,}8-\lambda).$$

b) **Valeurs propres.** Un produit est nul si et seulement si l'un de ses facteurs est nul : $P(\lambda) = 0$ pour $\lambda = 1$, $\lambda = 0{,}2$ ou $\lambda = 1{,}8$. Rangées par ordre décroissant :

$$\boxed{\lambda_1 = 1{,}8}, \qquad \boxed{\lambda_2 = 1}, \qquad \boxed{\lambda_3 = 0{,}2}.$$

✓ **Contrôle** $1{,}8 + 1 + 0{,}2 = 3 = \mathrm{Tr}(R) = 1 + 1 + 1$.

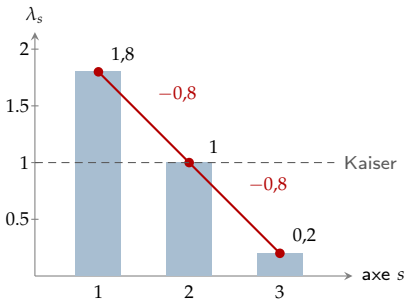
> **Autre méthode.** Développer $(1-\lambda)^2 - 0{,}64 = \lambda^2 - 2\lambda + 0{,}36$, de discriminant $\Delta = 4 - 1{,}44 = 2{,}56 = 1{,}6^2$, d'où $\lambda = \frac{2 \pm 1{,}6}{2}$, soit 1,8 ou 0,2.

c) **Pourcentages d'inertie et critère de Kaiser.**

	λ_s	% d'inertie $= 100\,\lambda_s/3$	% cumulé	Kaiser ($\lambda_s \geq 1$?)
Axe 1	1,8	60 %	60 %	oui
Axe 2	1	33,3 %	93,3 %	oui (de justesse)
Axe 3	0,2	6,7 %	100 %	non

Le plan (F_1, F_2) résume 93,3 % de l'inertie : la carte à deux dimensions sera très fidèle. Le critère de Kaiser retient les axes 1 et 2 et écarte l'axe 3. On remarque que $\lambda_2 = 1$ est exactement au seuil : l'axe 2 ne résume ni plus ni moins qu'une variable de départ (on comprendra pourquoi à la question 7).

d) **Éboulis et critère du coude.**



Les valeurs propres diminuent de 0,8 entre les axes 1 et 2, puis encore de 0,8 entre les axes 2 et 3 : la ligne brisée est une droite, **il n'y a pas de coude**. Avec seulement $K = 3$ valeurs propres, le critère du coude n'est pas concluant. On s'appuie donc sur le critère de Kaiser et sur le pourcentage cumulé (93,3 % pour le plan) : on retient **deux axes**.

4. Les vecteurs propres : la direction des axes

a) **Résolution de** $R\,U = \lambda\,U$, c'est-à-dire $(R - \lambda\,\mathrm{I})\,U = 0$, avec $U = (u_1\,;u_2\,;u_3)$.

- $\lambda_1 = 1{,}8$: le système s'écrit

$$\begin{cases} -0{,}8\,u_1 + 0{,}8\,u_2 = 0 \\ 0{,}8\,u_1 - 0{,}8\,u_2 = 0 \\ -0{,}8\,u_3 = 0 \end{cases} \qquad \text{donc} \qquad u_1 = u_2 \quad \text{et} \quad u_3 = 0.$$

On norme : $u_1^2 + u_1^2 = 1$, soit $u_1^2 = \frac{1}{2}$, d'où $u_1 = \frac{1}{\sqrt{2}}$ et

$$U_1 = \left(\tfrac{1}{\sqrt{2}}\,;\tfrac{1}{\sqrt{2}}\,;0\right) \approx (0{,}71\,;0{,}71\,;0).$$

- $\lambda_2 = 1$: le système s'écrit

$$\begin{cases} 0{,}8\,u_2 = 0 \\ 0{,}8\,u_1 = 0 \\ 0 \times u_3 = 0 \end{cases} \qquad \text{donc} \qquad u_1 = u_2 = 0 \quad \text{et} \quad u_3 \text{ quelconque;}$$

on norme $(u_3^2 = 1)$:

$$U_2 = (0\,;0\,;1).$$

- $\lambda_3 = 0{,}2$: le système s'écrit

$$\begin{cases} 0{,}8\,u_1 + 0{,}8\,u_2 = 0 \\ 0{,}8\,u_1 + 0{,}8\,u_2 = 0 \\ 0{,}8\,u_3 = 0 \end{cases} \qquad \text{donc} \qquad u_2 = -u_1 \quad \text{et} \quad u_3 = 0;$$

on norme $(2u_1^2 = 1)$:

$$U_3 = \left(\tfrac{1}{\sqrt{2}}\,;-\tfrac{1}{\sqrt{2}}\,;0\right) \approx (0{,}71\,;-0{,}71\,;0).$$

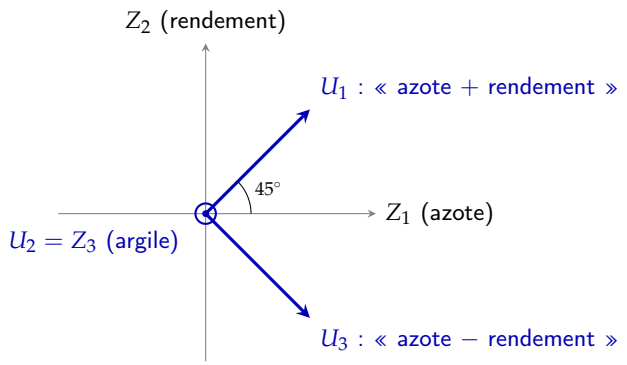
✓ **Contrôle** chaque vecteur a bien une coordonnée nulle ; on peut aussi vérifier $R\,U_1 = 1{,}8\,U_1$: la première ligne donne $1 \times \frac{1}{\sqrt{2}} + 0{,}8 \times \frac{1}{\sqrt{2}} + 0 = \frac{1{,}8}{\sqrt{2}} = 1{,}8 \times \frac{1}{\sqrt{2}}$.

> **Remarque.** $-U_s$ est aussi un vecteur propre unitaire. Le sens d'un axe est donc arbitraire ; un logiciel peut donner des signes opposés, ce qui retourne la carte sans changer l'interprétation.

b) **Orthogonalité.**

$$U_1 \cdot U_2 = \tfrac{1}{\sqrt{2}} \times 0 + \tfrac{1}{\sqrt{2}} \times 0 + 0 \times 1 = 0, \quad U_1 \cdot U_3 = \tfrac{1}{2} - \tfrac{1}{2} + 0 = 0, \quad U_2 \cdot U_3 = 0 + 0 + 0 = 0 :$$

les trois axes sont deux à deux perpendiculaires. Géométriquement, le nouveau repère est l'ancien **tourné de $45°$ autour de l'axe Z_3** :



vue de dessus : $U_2 = Z_3$ « sort » de la feuille, vers le lecteur

5. Les individus : coordonnées et carte des parcelles

a) **Formules des coordonnées.** On applique $F_s(i) = u_{s1}\,z_{i1} + u_{s2}\,z_{i2} + u_{s3}\,z_{i3}$ avec les coordonnées des vecteurs propres :

$$F_1(i) = \tfrac{1}{\sqrt{2}}\,z_{i1} + \tfrac{1}{\sqrt{2}}\,z_{i2} + 0 \times z_{i3} = \frac{z_{i1}+z_{i2}}{\sqrt{2}},$$
$$F_2(i) = 0 \times z_{i1} + 0 \times z_{i2} + 1 \times z_{i3} = z_{i3},$$
$$F_3(i) = \tfrac{1}{\sqrt{2}}\,z_{i1} - \tfrac{1}{\sqrt{2}}\,z_{i2} + 0 \times z_{i3} = \frac{z_{i1}-z_{i2}}{\sqrt{2}}.$$

L'axe 1 fait la somme (à $\sqrt{2}$ près) de l'azote et du rendement ; l'axe 2 est la variable argile elle-même ; l'axe 3 mesure l'écart entre azote et rendement.

b) **Tableau complété.** On calcule $\sqrt{2}\,F_1 = z_{i1} + z_{i2}$ et $\sqrt{2}\,F_3 = z_{i1} - z_{i2}$, puis $F_1^2 = (\sqrt{2}\,F_1)^2/2$ et $F_3^2 = (\sqrt{2}\,F_3)^2/2$. Par exemple, pour P_1 : $\sqrt{2}\,F_1 = -1{,}5 + (-1{,}5) = -3$, donc $F_1 = -3/\sqrt{2} \approx -2{,}12$ et $F_1^2 = 9/2 = 4{,}5$; pour P_4 : $\sqrt{2}\,F_3 = 0{,}5 - 1{,}5 = -1$ et $F_3^2 = 1/2 = 0{,}5$.

	Coordonnées				Carrés			Contrôle
Parcelle	$\sqrt{2}\,F_1$	F_1 (approx.)	F_2	$\sqrt{2}\,F_3$	F_1^2	F_2^2	F_3^2	somme $= d^2$
P_1	-3	$-2{,}12$	$-0{,}5$	0	$4{,}5$	$0{,}25$	0	$4{,}75$
P_2	-1	$-0{,}71$	$-0{,}5$	0	$0{,}5$	$0{,}25$	0	$0{,}75$
P_3	0	0	2	0	0	4	0	4
P_4	2	$1{,}41$	$-0{,}5$	-1	2	$0{,}25$	$0{,}5$	$2{,}75$
P_5	2	$1{,}41$	$-0{,}5$	1	2	$0{,}25$	$0{,}5$	$2{,}75$
Somme	0	0	0	0	9	5	1	15

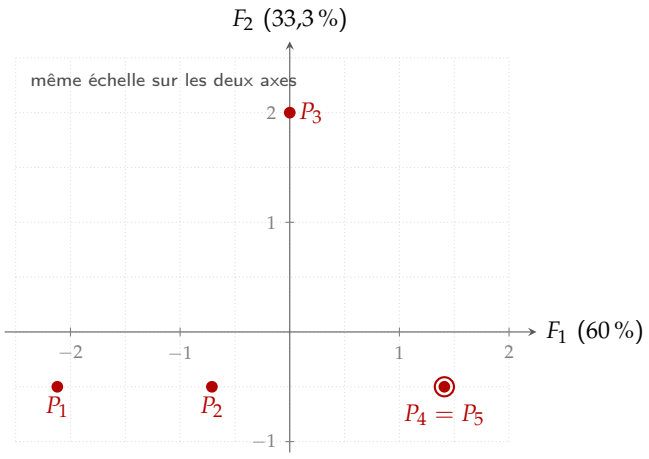
✓ **Contrôle** sur chaque ligne, $F_1^2 + F_2^2 + F_3^2 = d^2(P_i, G)$ (dernière colonne, à comparer à la question 1) : changer d'axes ne change pas les distances. Chaque colonne de coordonnées a une somme nulle : les F_s sont centrées.

c) **Variances.** Les colonnes F_s sont centrées, donc leur variance est la moyenne des carrés (ligne « Somme » du tableau, divisée par 5) :

$$\mathrm{Var}(F_1) = \frac{4{,}5 + 0{,}5 + 0 + 2 + 2}{5} = \frac{9}{5} = 1{,}8 = \lambda_1,$$
$$\mathrm{Var}(F_2) = \frac{0{,}25 + 0{,}25 + 4 + 0{,}25 + 0{,}25}{5} = \frac{5}{5} = 1 = \lambda_2,$$

et de même $\mathrm{Var}(F_3) = \frac{1}{5} = 0{,}2 = \lambda_3$. La variance des coordonnées sur un axe est exactement la valeur propre : c'est en ce sens que λ_s est « l'inertie portée par l'axe s ». Et l'on retrouve $9 + 5 + 1 = 15$, soit $\mathcal{I}_{\text{tot}} = 15/5 = 3$: l'inertie totale est découpée axe par axe.

d) **Carte des parcelles.**



P_4 et P_5 ont exactement les mêmes coordonnées $(1{,}41\,;-0{,}5)$: elles sont **superposées** sur la carte. Pourtant, leurs données brutes sont différentes (70 et 90 kg N/ha ; 5,5 et 4,5 t/ha) : elles ne diffèrent que selon l'axe 3 ($\sqrt{2}\,F_3 = -1$ et 1), qui n'est pas visible sur la carte. Les questions 7 et 8 reviennent sur ce point.

6. Les variables : le cercle des corrélations

a) **Calcul par la définition.** On utilise la colonne $W = \sqrt{2}\,F_1 = (-3\,;-1\,;0\,;2\,;2)$, centrée, de variance

$$\mathrm{Var}(W) = \frac{(-3)^2+(-1)^2+0^2+2^2+2^2}{5} = \frac{18}{5} = 3{,}6 \qquad (\text{soit } 2\times\lambda_1), \qquad \sigma_W = \sqrt{3{,}6}.$$

Comme W et F_1 sont proportionnelles avec un coefficient positif, $r(X_k,F_1) = r(Z_k,W) = \dfrac{\frac{1}{5}\sum_i z_{ik}\,W_i}{1\times\sqrt{3{,}6}}$.

Parcelle	$W_i = \sqrt{2}\,F_1(i)$	$z_{i1}\,W_i$	$z_{i2}\,W_i$	$z_{i3}\,W_i$
P_1	-3	$(-1{,}5)(-3) = 4{,}5$	$(-1{,}5)(-3) = 4{,}5$	$(-0{,}5)(-3) = 1{,}5$
P_2	-1	0,5	0,5	0,5
P_3	0	0	0	0
P_4	2	1	3	-1
P_5	2	3	1	-1
Somme/5 $= \mathrm{Cov}(Z_k,W)$		$9/5 = 1{,}8$	1,8	0
$r(X_k,F_1) = \mathrm{Cov}/\sqrt{3{,}6}$		$\frac{1{,}8}{\sqrt{3{,}6}} = \sqrt{0{,}9} \approx$ **0,95**	**0,95**	**0**

(On a utilisé $\frac{1{,}8}{\sqrt{3{,}6}} = \sqrt{\frac{1{,}8^2}{3{,}6}} = \sqrt{\frac{3{,}24}{3{,}6}} = \sqrt{0{,}9}$.)

Pour l'axe 2, $F_2 = Z_3$ a pour écart-type $\sqrt{\lambda_2} = 1$, donc $r(X_k,F_2) = \frac{1}{5}\sum_i z_{ik}\,z_{i3} = r_{k3}$: $r(X_1,F_2) = r_{13} = 0$, $r(X_2,F_2) = r_{23} = 0$ et $r(X_3,F_2) = r_{33} = 1$.

Variable	$r(X_k,F_1)$	$r(X_k,F_2)$
X_1 (azote)	$\sqrt{0{,}9} \approx 0{,}95$	0
X_2 (rendement)	$\sqrt{0{,}9} \approx 0{,}95$	0
X_3 (argile)	0	1

b) **Avec le raccourci $r(X_k,F_s) = \sqrt{\lambda_s}\,u_{sk}$.** Vérification de l'étape (5) pour $k=1$, $s=1$: la première ligne de R est $(1 \;\; 0{,}8 \;\; 0)$, d'où

$$(R\,U_1)_1 = 1\times\tfrac{1}{\sqrt{2}} + 0{,}8\times\tfrac{1}{\sqrt{2}} + 0\times 0 = \tfrac{1{,}8}{\sqrt{2}} = \lambda_1\,u_{11}.$$

Les six coordonnées :

- axe 1 : $r(X_1,F_1) = r(X_2,F_1) = \sqrt{1{,}8}\times\frac{1}{\sqrt{2}} = \sqrt{\frac{1{,}8}{2}} = \sqrt{0{,}9} \approx 0{,}95$ et $r(X_3,F_1) = \sqrt{1{,}8}\times 0 = 0$;
- axe 2 : $r(X_1,F_2) = r(X_2,F_2) = \sqrt{1}\times 0 = 0$ et $r(X_3,F_2) = \sqrt{1}\times 1 = 1$.

On retrouve les mêmes valeurs, en une ligne au lieu d'un tableau de produits : c'est la formule qu'utilisent les logiciels.

c) **Vérifications et lecture.** On calcule aussi l'axe 3 avec le raccourci :

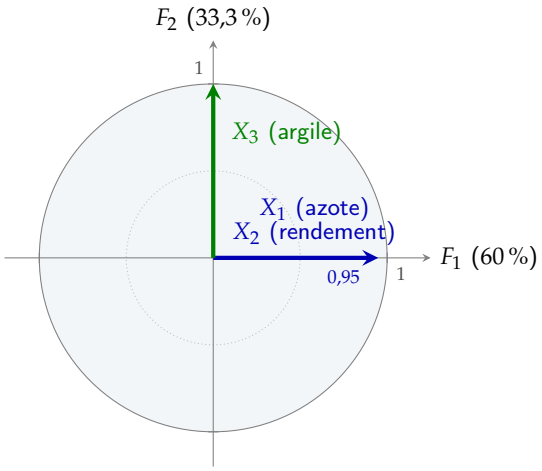
$$r(X_1,F_3) = \sqrt{0{,}2}\times\tfrac{1}{\sqrt{2}} = \sqrt{0{,}1} \approx 0{,}32, \qquad r(X_2,F_3) = \sqrt{0{,}2}\times\left(-\tfrac{1}{\sqrt{2}}\right) = -\sqrt{0{,}1} \approx -0{,}32,$$
$$r(X_3,F_3) = \sqrt{0{,}2}\times 0 = 0.$$

Le tableau des r^2 se remplit alors entièrement :

	axe 1	axe 2	axe 3	**total de la ligne**
X_1 (azote)	0,9	0	0,1	**1**
X_2 (rendement)	0,9	0	0,1	**1**
X_3 (argile)	0	1	0	**1**
total de la colonne	$1{,}8 = \lambda_1$	$1 = \lambda_2$	$0{,}2 = \lambda_3$	$3 = K$

- En colonne : l'axe 1 capte 90 % de la variance de l'azote et 90 % de celle du rendement, soit l'équivalent de 1,8 variable ; l'axe 2 capte toute l'argile, soit exactement une variable ; l'axe 3 ne capte que les 10 % restants de l'azote et du rendement (0,2 variable) : il est logique de l'abandonner.
- En ligne : rien n'est perdu. L'argile est entièrement sur l'axe 2 ; l'azote et le rendement sont à 90 % sur l'axe 1 et à 10 % sur l'axe 3, donc à 90 % visibles sur le plan (F_1, F_2).

d) **Cercle des corrélations.**



Les flèches de l'azote et du rendement sont **superposées**, horizontales et presque de longueur 1 (0,95) ; celle de l'argile est verticale et touche le cercle. L'axe 1 est donc l'axe « azote–rendement » et l'axe 2 l'axe « argile ». La superposition de X_1 et X_2 semble dire $r_{12} = 1$ (angle nul, cosinus égal à 1), alors que $r_{12} = 0{,}8$: voir la question 8.

7. Les aides à l'interprétation

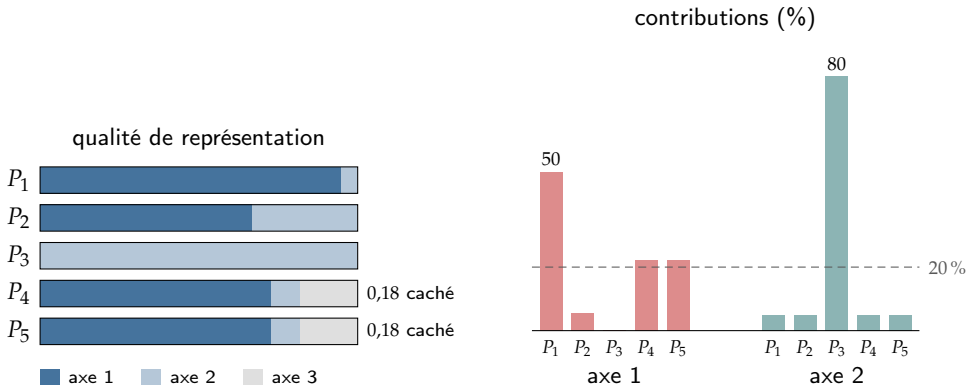
a) **Tableau des parcelles.** On divise chaque carré $F_s(i)^2$ par $d^2(P_i, G)$ pour les $\cos^2$, et l'on utilise $\text{contrib}_1 = \frac{100\,F_1^2}{5\times 1{,}8} = \frac{100\,F_1^2}{9}$ et $\text{contrib}_2 = \frac{100\,F_2^2}{5\times 1} = \frac{100\,F_2^2}{5}$. Détail des calculs :

Parcelle	$\cos^2_1 = \frac{F_1^2}{d^2}$	$\cos^2_2 = \frac{F_2^2}{d^2}$	$\cos^2_3 = \frac{F_3^2}{d^2}$	$\text{contrib}_1 = \frac{100\,F_1^2}{9}$	$\text{contrib}_2 = \frac{100\,F_2^2}{5}$
P_1	$\frac{4{,}5}{4{,}75} \approx 0{,}95$	$\frac{0{,}25}{4{,}75} \approx 0{,}05$	$\frac{0}{4{,}75} = 0$	$\frac{450}{9} = 50$	$\frac{25}{5} = 5$
P_2	$\frac{0{,}5}{0{,}75} \approx 0{,}67$	$\frac{0{,}25}{0{,}75} \approx 0{,}33$	0	$\frac{50}{9} \approx 5{,}6$	5
P_3	$\frac{0}{4} = 0$	$\frac{4}{4} = 1$	0	0	$\frac{400}{5} = 80$
P_4	$\frac{2}{2{,}75} \approx 0{,}73$	$\frac{0{,}25}{2{,}75} \approx 0{,}09$	$\frac{0{,}5}{2{,}75} \approx 0{,}18$	$\frac{200}{9} \approx 22{,}2$	5
P_5	$\approx 0{,}73$	$\approx 0{,}09$	$\approx 0{,}18$	$\approx 22{,}2$	5

D'où le tableau demandé :

	Qualité de représentation				Contribution (%)	
Parcelle	$\cos^2_1$	$\cos^2_2$	$\cos^2$ plan	$\cos^2_3$	axe 1	axe 2
P_1	0,95	0,05	**1**	0	**50**	5
P_2	0,67	0,33	**1**	0	5,6	5
P_3	0	1	**1**	0	0	**80**
P_4	0,73	0,09	**0,82**	0,18	22,2	5
P_5	0,73	0,09	**0,82**	0,18	22,2	5
Total	—	—	—	—	100	100

Les mêmes résultats en images : à gauche, la part de d^2 visible sur chaque axe (chaque barre fait 100 %) ; à droite, les contributions, à comparer à la moyenne 20 %.



b) **Vérification pour P_4.** On lit $F_1(4)^2 = 2$, $F_2(4)^2 = 0{,}25$, $F_3(4)^2 = 0{,}5$ (tableau de la question 5 b) et $d^2(P_4, G) = 2{,}75$ (question 1) :

$$\cos^2_1(P_4) + \cos^2_2(P_4) + \cos^2_3(P_4) = \frac{2}{2{,}75} + \frac{0{,}25}{2{,}75} + \frac{0{,}5}{2{,}75} = \frac{2 + 0{,}25 + 0{,}5}{2{,}75} = \frac{2{,}75}{2{,}75} = 1.$$

C'est vrai pour toutes les parcelles : les trois carrés ont toujours pour somme $d^2(P_i, G)$ (théorème de Pythagore dans le repère orthonormé (U_1, U_2, U_3), contrôle de la question 5), et l'on divise cette somme par elle-même.

c) **Qualité de représentation.** P_1, P_2 et P_3 ont un $\cos^2$ égal à 1 sur le plan : elles sont *dans* le plan (F_1, F_2) ($F_3 = 0$, car $z_{i1} = z_{i2}$), et la carte les montre sans déformation. P_4 et P_5 ont un $\cos^2$ de 0,82 : 18 % du carré de leur distance au centre est « caché » sur l'axe 3, l'une au-dessous du plan, l'autre au-dessus (comme A et B sur la figure de l'énoncé). Précaution : ne pas conclure que P_4 et P_5 sont semblables parce qu'elles sont confondues sur la carte.

d) **Contributions.** La contribution moyenne est $100/5 = 20\,\%$ (ligne pointillée ci-dessus). Axe 1 : P_1 contribue le plus (50 %, deux fois et demie la moyenne), suivie de P_4 et P_5 (22,2 % chacune) ; l'axe 1 oppose donc P_1 à P_4 et P_5. Axe 2 : P_3 contribue à elle seule pour 80 % (quatre fois la moyenne) ; l'axe 2 est presque entièrement « fabriqué » par P_3.

e) **P_1 et P_2.** Toutes deux sont parfaitement représentées ($\cos^2 = 1$), mais P_1 est loin du centre ($d^2 = 4{,}75$) alors que P_2 en est proche ($d^2 = 0{,}75$).

- Le $\cos^2$ est un *rapport* F_s^2/d^2 (le cosinus d'un angle) : il ne dépend que de la *direction* de $\overrightarrow{GP_i}$, pas de sa longueur. Les deux parcelles sont dans le plan, donc toutes deux ont $\cos^2 = 1$.
- La contribution dépend de $F_s(i)^2$, donc de la *distance* au centre : P_1 ($F_1^2 = 4{,}5$) pèse neuf fois plus que P_2 ($F_1^2 = 0{,}5$) sur l'axe 1.

Le $\cos^2$ dit si un point est *bien vu* sur la carte, la contribution dit s'il a *construit* l'axe : ce sont deux questions différentes.

f) **Tableau des variables.** Avec $\cos^2_s(X_k) = r^2(X_k, F_s)$ et $\text{contrib}_s(X_k) = 100 \times r^2(X_k, F_s)/\lambda_s$:

Variable	Coordonnées		Qualité de représentation				Contrib. (%)	
	$r(X_k, F_1)$	$r(X_k, F_2)$	$\cos^2_1$	$\cos^2_2$	$\cos^2$ plan	$\cos^2_3$	axe 1	axe 2
X_1 (azote)	0,95	0	0,9	0	**0,9**	0,1	**50**	0
X_2 (rendement)	0,95	0	0,9	0	**0,9**	0,1	**50**	0
X_3 (argile)	0	1	0	1	**1**	0	0	**100**
Total	—	—	1,8	1	—	0,2	100	100

Par exemple, $\text{contrib}_1(X_1) = 100 \times \frac{0,9}{1,8} = 50\,\%$ et $\text{contrib}_2(X_3) = 100 \times \frac{1}{1} = 100\,\%$.

g) **Vérifications.** Les colonnes $\cos^2_1$, $\cos^2_2$, $\cos^2_3$ totalisent $0,9+0,9+0 = 1,8 = \lambda_1$, $0+0+1 = 1 = \lambda_2$ et $0,1+0,1+0 = 0,2 = \lambda_3$. Sur chaque ligne, $0,9+0+0,1 = 1$ (azote et rendement) et $0+1+0 = 1$ (argile). On retrouve le tableau des r^2 de la question 6 c : qualité de représentation des variables et carrés des corrélations sont la même chose.

h) **Bilan pour les variables.** L'argile est parfaitement représentée ($\cos^2 = 1$, flèche qui touche le cercle). L'azote et le rendement sont bien représentés ($\cos^2 = 0,9$, flèche de longueur $\sqrt{0,9} \approx 0,95$) ; les 10 % restants de leur information sont sur l'axe 3. L'axe 1 est construit à parts égales par l'azote et le rendement (50 % chacun, contre 33,3 % en moyenne) ; l'axe 2 est construit uniquement par l'argile (100 %).

- Axe 1 : **niveau de fertilisation azotée et de production** (des parcelles peu fertilisées et peu productives, à gauche, aux parcelles fertilisées et productives, à droite).
- Axe 2 : **teneur du sol en argile**.

On comprend aussi pourquoi $\lambda_2 = 1$: l'argile n'est corrélée à aucune autre variable, elle forme donc à elle seule un axe, qui résume exactement une variable.

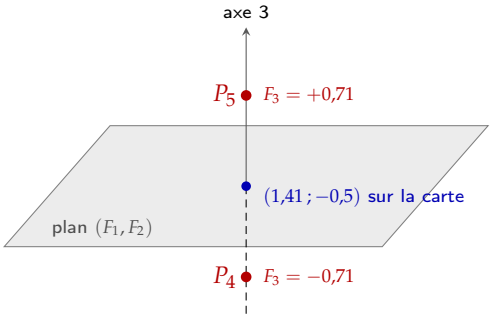
8. Interprétation

a) **Extrémités des axes.**

- Axe 1, côté négatif : P_1 (30 kg N/ha, 2,5 t/ha), la parcelle la moins fertilisée et la moins productive. Côté positif : P_4 et P_5 (70 et 90 kg N/ha, 5,5 et 4,5 t/ha), les plus fertilisées et les plus productives. Sur le cercle, les flèches de l'azote et du rendement pointent vers la droite : le nom « fertilisation et production » est confirmé.
- Axe 2, côté positif : P_3, seule parcelle à sol argileux (32 % contre 17 %) ; les quatre autres sont toutes à $F_2 = -0,5$. La flèche de l'argile pointe vers le haut : le nom « argile » est confirmé.

P_2, proche du centre, est une parcelle « ordinaire », un peu en dessous de la moyenne en azote et en rendement.

b) P_4 **et** P_5**.** Données brutes : $P_4 = (70\,;5,5\,;17)$ et $P_5 = (90\,;4,5\,;17)$. P_5 reçoit plus d'azote que P_4 mais produit moins : elles ne se ressemblent pas. Le $\cos^2$ sur le plan (0,82, soit $\cos^2_3 = 0,18$) permettait de prévoir que leur ressemblance n'est qu'apparente. Elles diffèrent sur l'axe 3, $F_3 = (z_{i1} - z_{i2})/\sqrt{2}$, qui oppose l'azote au rendement : $F_3(4) = -\frac{1}{\sqrt{2}} \approx -0,71$ et $F_3(5) = +\frac{1}{\sqrt{2}} \approx 0,71$.



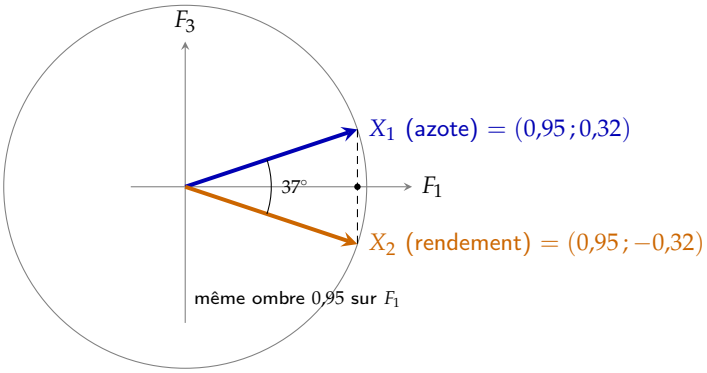
P_4 (sous le plan) : beaucoup de rendement pour la dose d'azote reçue ;
P_5 (au-dessus) : peu de rendement pour la dose reçue.
Sur la carte, toutes deux tombent au même point.

L'axe 3 est donc un axe d'*efficience de l'azote* (rendement obtenu par kilo d'azote apporté). Dans l'espace complet, $d^2(P_4, P_5) = (0,5-1,5)^2 + (1,5-0,5)^2 + (-0,5+0,5)^2 = 2$, soit aussi $\big(F_3(4) - F_3(5)\big)^2 = (\sqrt{2})^2 = 2$, alors que leur distance sur la carte est nulle.

c) **Les flèches superposées.**

$$\sum_{s=1}^{3} r(X_1,F_s)\,r(X_2,F_s) = \underbrace{\sqrt{0{,}9}\times\sqrt{0{,}9}}_{\text{axe 1}} + \underbrace{0\times 0}_{\text{axe 2}} + \underbrace{\sqrt{0{,}1}\times\left(-\sqrt{0{,}1}\right)}_{\text{axe 3}} = 0{,}9+0-0{,}1 = 0{,}8 = r_{12}.$$

C'est le produit scalaire des deux flèches complètes (de longueur 1), qui vaut bien leur corrélation (question 6 d) de l'énoncé) ; avec $k = l$, la même formule redonne $\sum_s r^2(X_k,F_s) = 1$. Sur le plan, les deux flèches sont confondues (angle nul), ce qui suggère $r = 1$. Mais une partie de l'information (ici 10 % pour chaque variable) est sur l'axe 3, où les deux flèches sont *opposées* ($+0{,}32$ et $-0{,}32$) : cette partie cachée fait baisser la corrélation de 0,9 à 0,8. Vu dans le plan (F_1,F_3), l'angle réel apparaît :



L'angle entre les deux variables vaut $\text{Arccos}(0{,}8) \approx 37°$, pas $0°$; projetées sur le plan (F_1,F_2), les deux flèches ont la même ombre et se superposent.

> **Règle de lecture.** L'angle entre deux flèches ne donne la corrélation que si les deux flèches touchent (presque) le cercle ; sinon, il ne donne qu'une indication, d'autant moins fiable que les flèches sont courtes.

d) P_3 **et l'axe 2.** Avec 80 % de contribution, l'axe 2 est presque entièrement dû à une seule parcelle : c'est la seule à sol argileux, toute la variabilité de l'argile vient d'elle. Un tel axe décrit une parcelle particulière plus qu'une structure générale. En pratique, il faut :

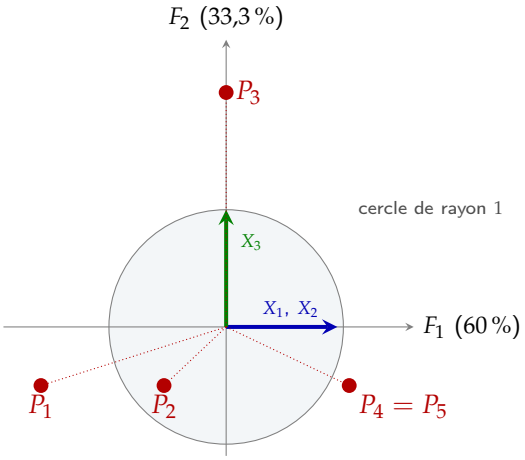
- vérifier la donnée (erreur de saisie, de mesure, de prélèvement ?) ;
- refaire l'ACP sans P_3, en la plaçant en *individu supplémentaire*, pour voir si les autres axes restent stables ;
- rester prudent, avec seulement cinq parcelles, sur toute généralisation.

9. Pour aller plus loin : le biplot

a) **Scores standardisés.** On divise F_1 par $\sqrt{1{,}8} \approx 1{,}34$ et F_2 par $\sqrt{1} = 1$. De façon exacte, $\frac{F_1(i)^2}{\lambda_1} = \frac{F_1(i)^2}{1{,}8}$ (par exemple $\frac{4{,}5}{1{,}8} = 2{,}5$ pour P_1), ce qui donne directement la distance² à l'origine $\frac{F_1(i)^2}{\lambda_1} + \frac{F_2(i)^2}{\lambda_2}$.

Parcelle	$F_1(i)$	$\frac{F_1(i)}{\sqrt{\lambda_1}}$	$F_2(i)$	$\frac{F_2(i)}{\sqrt{\lambda_2}}$	distance² à l'origine sur le biplot
P_1	$-2{,}12$	$-1{,}58$	$-0{,}5$	$-0{,}5$	$2{,}5+0{,}25 = \mathbf{2{,}75}$
P_2	$-0{,}71$	$-0{,}53$	$-0{,}5$	$-0{,}5$	$0{,}28+0{,}25 \approx \mathbf{0{,}53}$
P_3	0	0	2	2	$0+4 = \mathbf{4}$
P_4	1,41	1,05	$-0{,}5$	$-0{,}5$	$1{,}11+0{,}25 \approx \mathbf{1{,}36}$
P_5	1,41	1,05	$-0{,}5$	$-0{,}5$	$\approx \mathbf{1{,}36}$

✓ **Contrôle** chaque colonne de scores standardisés a une variance égale à 1 : par exemple $\frac{2{,}5+0{,}28+0+1{,}11+1{,}11}{5} = 1$.



P_1, P_3, P_4 et P_5 sont hors du cercle (distance^2 supérieure à 1) ; seule P_2, la parcelle la plus « ordinaire », est à l'intérieur. Ce n'est pas gênant : le cercle de rayon 1 borne les *corrélations*, pas les scores des parcelles. Une parcelle hors du cercle est simplement à plus d'un écart-type du centre dans le plan.

Contrôle de la formule inverse. Pour P_4 : $F_1(4)\,U_1 + F_2(4)\,U_2 + F_3(4)\,U_3 = \sqrt{2}\,(\frac{1}{\sqrt{2}}\,;\frac{1}{\sqrt{2}}\,;0) - 0{,}5\,(0\,;0\,;1) - \frac{1}{\sqrt{2}}(\frac{1}{\sqrt{2}}\,;-\frac{1}{\sqrt{2}}\,;0) = (1\,;1\,;0) + (0\,;0\,;-0{,}5) + (-0{,}5\,;0{,}5\,;0) = (0{,}5\,;1{,}5\,;-0{,}5)$: on retrouve bien la ligne P_4 du tableau centré réduit.

b) **Produits scalaires.** On travaille avec les coordonnées exactes :

$$\frac{F_1(1)}{\sqrt{\lambda_1}} = -\frac{3/\sqrt{2}}{\sqrt{1{,}8}} = -\frac{3}{\sqrt{3{,}6}} = -\frac{\sqrt{10}}{2}, \qquad \frac{F_1(4)}{\sqrt{\lambda_1}} = \frac{2/\sqrt{2}}{\sqrt{1{,}8}} = \frac{2}{\sqrt{3{,}6}} = \frac{\sqrt{10}}{3},$$

la flèche de l'azote étant $(\sqrt{0{,}9}\,;0)$ et celle de l'argile $(0\,;1)$.

Parcelle	**Variable**	point de la parcelle $(a\,;b)$	flèche de la variable $(c\,;d)$	produit scalaire $ac+bd$	z_{ik}	exact ou approché ?
P_3	argile	$(0\,;2)$	$(0\,;1)$	$0\times 0 + 2\times 1 = 2$	$z_{33} = 2$	exact
P_1	azote	$(-\frac{\sqrt{10}}{2}\,;-0{,}5)$	$(\sqrt{0{,}9}\,;0)$	$-\frac{\sqrt{10}\sqrt{0{,}9}}{2} = -\frac{\sqrt{9}}{2} = -1{,}5$	$z_{11} = -1{,}5$	exact
P_4	azote	$(\frac{\sqrt{10}}{3}\,;-0{,}5)$	$(\sqrt{0{,}9}\,;0)$	$\frac{\sqrt{10}\sqrt{0{,}9}}{3} = \frac{\sqrt{9}}{3} = 1$	$z_{41} = 0{,}5$	approché

Le biplot ne garde que les termes $s = 1$ et $s = 2$ de la formule $z_{ik} = \sum_s \frac{F_s(i)}{\sqrt{\lambda_s}}\, r(X_k, F_s)$; le terme oublié est $\frac{F_3(i)}{\sqrt{\lambda_3}}\, r(X_k, F_3)$.

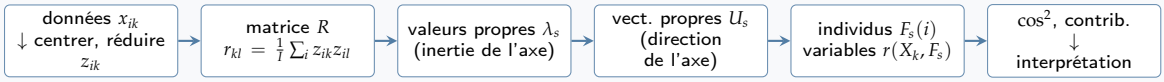
- Il est nul si la parcelle est dans le plan ($F_3(i) = 0$: c'est le cas de P_1 et P_3) *ou* si la variable est parfaitement représentée ($r(X_k, F_3) = 0$: c'est le cas de l'argile). La reconstitution est alors exacte.
- Pour P_4 et l'azote, aucune des deux conditions n'est remplie ($\cos^2(P_4) = 0{,}82$ et $\cos^2(X_1) = 0{,}9$ sur le plan) :

$$\frac{F_3(4)}{\sqrt{\lambda_3}} \times r(X_1, F_3) = \frac{-1/\sqrt{2}}{\sqrt{0{,}2}} \times \sqrt{0{,}1} = -\frac{\sqrt{0{,}1}}{\sqrt{0{,}4}} = -0{,}5,$$

et l'on retrouve bien $1 + (-0{,}5) = 0{,}5 = z_{41}$ en ajoutant le terme oublié.

Conclusion. Le biplot se lit correctement pour les parcelles et les variables bien représentées ; pour les autres, il ne donne qu'une approximation, et ce sont encore les $\cos^2$ qui disent jusqu'où on peut lui faire confiance.

À retenir : le fil d'une ACP normée



- L'inertie totale vaut K ; elle se découpe axe par axe : $\lambda_s = \frac{1}{I}\sum_i F_s(i)^2 = \sum_k r^2(X_k, F_s)$.
- Le $\cos^2$ dit si un point est *bien vu* ; la contribution dit s'il a *fait* l'axe.
- Deux points proches sur la carte, ou deux flèches confondues, ne se ressemblent que s'ils sont bien représentés.

Étudiant(e) (Nom, Prénom) : .....

Promotion, groupe : .....

Email : .....