La famille des statistiques multivariées

On distingue plusieurs grandes approches :

1. Méthodes de réduction dimensionnelle

- **ACP (Analyse en Composantes Principales)** : réduit un grand nombre de variables quantitatives en quelques axes principaux.
- **Analyse Factorielle des Correspondances (AFC)** : adaptée aux tableaux de contingence (variables qualitatives).
- **Analyse Factorielle Multiple (AFM)** : extension de l'ACP/AFC pour plusieurs groupes de variables.

2. Méthodes de classification

- **Classification ascendante hiérarchique (CAH)** : construit un dendrogramme pour regrouper les individus.
- **K-means** : partitionne les individus en k groupes homogènes.
- **Analyse discriminante** : cherche à séparer des groupes connus à partir de variables explicatives.

3. Méthodes de liaison entre ensembles de variables

- **Analyse Canonique** : étudie les relations entre deux ensembles de variables.
- **Analyse en Corrélation Canonique (ACC)** : maximise la corrélation entre combinaisons linéaires de deux groupes de variables.
- **Analyse de la redondance** : proche de l'ACC mais orientée vers la prédiction.

4. Méthodes exploratoires mixtes

- **Analyse des Correspondances Multiples (ACM)** : pour plusieurs variables qualitatives.
- **Analyse Factorielle Mixte (AFM)** : combine variables quantitatives et qualitatives.
- **Analyse conjointe** : utilisée en marketing pour comprendre les préférences.

Intérêt de l'analyse factorielle

- **Réduction de la dimensionnalité** : simplifie les données en quelques axes tout en conservant l'essentiel de l'information.
- **Visualisation** : permet de représenter graphiquement les individus et les variables.
- **Détection de structures** : met en évidence des regroupements, oppositions ou corrélations.
- **Interprétation** : aide à comprendre les relations entre variables et à identifier les variables les plus importantes.
- **Prétraitement** : utile avant d'appliquer des méthodes de classification ou de régression.

En résumé

La famille des statistiques multivariées pour l'analyse factorielle comprend des méthodes comme **ACP, AFC, ACM, AFM**, qui visent à **explorer, réduire et interpréter** des données multidimensionnelles. Elles sont essentielles en sciences, économie, marketing, chimie des matériaux, biologie, etc., dès qu'on manipule des jeux de données complexes.

1. Fondement géométrique

- On considère les données comme des **points dans un espace multidimensionnel** (chaque variable = une dimension).
- L'ACP cherche à trouver de **nouveaux axes** (appelés composantes principales) qui résument au mieux la dispersion des points.
- Ces axes sont des **combinaisons linéaires des variables initiales** et sont **orthogonaux** (perpendiculaires entre eux).

2. Fondement statistique (variance)

- L'idée est de **maximiser la variance expliquée** par chaque axe.
- Le premier axe (PC1) est celui qui capture la plus grande partie de la variance totale des données.
- Le deuxième axe (PC2) capture la plus grande variance restante, sous la contrainte d'être orthogonal au premier.
- Et ainsi de suite pour les autres axes.

3. Extraction des axes principaux

- On calcule la **matrice de covariance ou de corrélation** des variables.
- On effectue une **décomposition en valeurs propres** :
  - **Valeurs propres** → quantité de variance expliquée par chaque axe.
  - **Vecteurs propres** → direction des axes principaux (combinaisons des variables).
- Les axes principaux sont donc les **vecteurs propres** de la matrice de covariance/corrélation.

4. Résultat

- Chaque individu (point) peut être projeté sur ces nouveaux axes → coordonnées factorielles.
- Chaque variable peut être représentée par sa corrélation avec les axes → cercle des corrélations.
- On obtient une **représentation simplifiée** des données, souvent en 2D ou 3D, tout en conservant l'essentiel de l'information.

Définition Valeurs propres et vecteurs propres

λ est une valeur propre de A si seulement s'il existe un vecteur x non nul tel que : $Ax = \lambda x$ on dit alors que x est le vecteur propre de A associé à la valeur propre .λ.

Exemple

$$A = \begin{bmatrix} 1 & 2 & 2 \\ 0 & 2 & 1 \\ -1 & 2 & 2 \end{bmatrix} \; et \; x = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$$

$$\begin{bmatrix} 1 & 2 & 2 \\ 0 & 2 & 1 \\ -1 & 2 & 2 \end{bmatrix} \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = (2) = 2 \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \; est \; un \; vecteur \; propre \; de \; A \; de \; valeur \; propre \; \lambda = 2$$

Pour trouver les valeurs et vecteurs propres on utilisant la définition $Ax = \lambda x$

$$(A - \lambda)x = 0 \; puisque \; x \neq 0 \; alors \; on \; a$$
$$(A - \lambda) = 0$$

On multiplie λ par la matrice identité I ; on obtient la formule suivante :

$$(A - I\lambda) = 0$$

Pour déterminer les valeurs il faut développer le $\det(A - I\lambda) = 0$, on obtient le polynôme caractéristique de A dont les racines sont ses valeurs propres.

Exemple :

Prenons l'exemple précédent :

$$A = \begin{bmatrix} 1 & 2 & 2 \\ 0 & 2 & 1 \\ -1 & 2 & 2 \end{bmatrix}$$

$$(A - I\lambda) = \begin{bmatrix} 1-\lambda & 2 & 2 \\ 0 & 2-\lambda & 1 \\ -1 & 2 & 2-\lambda \end{bmatrix}$$

$$\det(A - I\lambda) = \det \begin{bmatrix} 1-\lambda & 2 & 2 \\ 0 & 2-\lambda & 1 \\ -1 & 2 & 2-\lambda \end{bmatrix} = 0$$

On choisi la colonne ou il y a des zéros poour simplifiés le calcul, ici la première colonne

$$\det \begin{bmatrix} 1-\lambda & 2 & 2 \\ 0 & 2-\lambda & 1 \\ -1 & 2 & 2-\lambda \end{bmatrix} = (1-\lambda) \begin{vmatrix} 2-\lambda & 1 \\ 2 & 2-\lambda \end{vmatrix} + 0 * \begin{vmatrix} 2 & 2 \\ 2 & 2-\lambda \end{vmatrix} - 1 * \begin{vmatrix} 2 & 2 \\ 2-\lambda & 1 \end{vmatrix}$$
$$= (1-\lambda) * [(2-\lambda)^2 - 1 * 2] + 0 - [2 * 1 - 2(2-\lambda)]$$
$$= 4 - 8\lambda + 5\lambda^2 - \lambda^3$$

On détermine les racines du polynôme du 3^{iéme} degré, on trouve une racine initile pour simpliefier le polynôme, on prend par exemple la racine =1 et on remplace dans l'équation $4 - 8 * 1 + 5 * 1^2 - 1^3 = 0$, donc 1 et la première racine, on va diviviser $\frac{4-8\lambda+5\lambda^2-\lambda^3}{(\lambda-1)} = (\lambda - 1)(\lambda^2 - 4\lambda + 4) = (\lambda - 1)(\lambda - 2)^2 = 0$on a deux racines doubles $(\lambda = 2)$ et$(\lambda = 1)$

1) $(A - I\lambda)x = \begin{bmatrix} 1-\lambda & 2 & 2 \\ 0 & 2-\lambda & 1 \\ -1 & 2 & 2-\lambda \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, on prend $(\lambda = 1)$ et on remplace dans la matrice

$\begin{bmatrix} 0 & 2 & 2 \\ 0 & 1 & 1 \\ -1 & 2 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, maintenant on le système d'équation linéaire suivant

$$\begin{cases} 0 * x + 2 * y + 2z = 0 \\ 0 * x + y + z = 0 \\ -x + 2y + z = 0 \end{cases} \quad \begin{matrix} \\ y = -z \\ x = -z \end{matrix}$$

Posant z=t, on obtient un espace propre engendré par v_1=[−1,−1,1].

2) $(A - I\lambda)x = \begin{bmatrix} 1-\lambda & 2 & 2 \\ 0 & 2-\lambda & 1 \\ -1 & 2 & 2-\lambda \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, on prend $(\lambda = 2)$ et on remplace dans la matrice

$$\begin{bmatrix} 1-2 & 2 & 2 \\ 0 & 2-2 & 1 \\ -1 & 2 & 2-2 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad \begin{bmatrix} -1 & 2 & 2 \\ 0 & 0 & 1 \\ -1 & 2 & 0 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad z = 0,, x = 2y$$

Posant y=s, on obtient l'espace propre (de dimension 1) engendré par v_2=[2,1,0].

Synthèse et remarque

- Valeurs propres: λ=1 et λ=2 (double).
- Vecteurs propres:
  - pour λ=1: tout multiple de [−1,−1,1],
  - pour λ=2: tout multiple de [2,1,0].

La multiplicité géométrique de λ=2 est 1 (un seul vecteur propre indépendant), donc A n'est pas diagonalisable sur R

La diagonalisation

Une matrice carré A est diagonale, s'il existe une matrice inversible P(appelée matrice de passage) et une matrice diagonale D satisfaisant la relation : $A=PDP^{-1}$

Les $\lambda_i$sont les valeurs propres de la matrice A et aussi la diagonale de D, et les colonnes de P sont les vecteurs propres associés

Exemple

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 2 & -1 & 2 \end{bmatrix}$$

Solution

Les valeurs propres et veteurs propres sont :

Polynôme caractéristique et valeurs propres

$$(A - I\lambda) = \begin{bmatrix} 1-\lambda & 2 & 0 \\ 0 & 3-\lambda & 0 \\ 2 & -1 & 2-\lambda \end{bmatrix} = (1-\lambda)\det\begin{bmatrix} 3-\lambda & 0 \\ -1 & 2-\lambda \end{bmatrix} - 2 * det\begin{bmatrix} 0 & 0 \\ 2 & 1-\lambda \end{bmatrix}$$
$$= (1-\lambda)(3-\lambda)(2-\lambda)$$

La matrice est diagonalisable (base de vecteurs propres).

$$\lambda_1 = 3 : V_1\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \lambda_2 = 2 : V_2\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \lambda_3 = 1 : V_3\begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}$$

$$D = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad P = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & -2 \end{bmatrix}, \qquad P^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 2 & -3 & 1 \\ 1 & -1 & 0 \end{bmatrix}$$

Vérification :$P * P^{-1} = I = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & -2 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ 2 & -3 & 1 \\ 1 & -1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Diagonalisation de A=PDP^{-1} :

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 2 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & -2 \end{bmatrix}\begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ 2 & -3 & 1 \\ 1 & -1 & 0 \end{bmatrix}$$

Les cinq étapes du déroulement de l'ACP

Pour interpréter les résultats produits par ce processus, des critères tels que la distance des points par rapport à l'axe sont utilisés.

Résumé en 5 points

1. **Réduire la complexité des données** : l'ACP simplifie un grand nombre de variables sans perte d'information.
2. **Mieux visualiser les corrélations** : elle révèle les liens entre variables et individus dans un espace graphique clair.
3. **Optimiser les modèles de données** : indispensable en data science et intelligence artificielle.
4. **Appliquer une méthode statistique puissante** : fondée sur la géométrie et la variance pour extraire les axes principaux.
5. **Un outil clé pour les métiers du digital et de la data** : utilisé en marketing, apprentissage automatique, biostatistique, et cybersécurité.

Exercie sur l'ACP

Dans cet exercice on va faire l'anlyse des préciptations (en cm) , la température maximale et minimale (en °C), pour quelques villes dans notre pays, les données colléctées sont présentées dans le tableau suivant (Données fictifs) :

Ville	Précipitation	$Température_{max}$	$Température_{min}$
Alger	12,04	23,7	5,9
Oran	17,18	15,5	-1,8
Constantine	11,83	13,1	2,8
Biskra	6,23	13,5	-2,4
Jijel	16,99	21,1	7,2
Bejaia	3,87	20,3	-0,9

Question :

1) Centarge et réduction des données (normalisation)
2) Déterminer la matrice de corrélation
3) Calcul de valeurs propres et vectrices propres
4) Déterminer les axes factoriels et composants principales

Solution

1) Centarge et réduction des données (normalisation)

Clcul de la moyenne, variance et ecarttype				
Ville	Précipitation	$Température_{max}$	$Température_{min}$	
Alger	12,04	23,7	5,9	
Oran	17,18	15,5	-1,8	
Constantine	11,83	13,1	2,8	
Biskra	6,23	13,5	-2,4	
Jijel	16,99	21,1	7,2	
Bejaia	3,87	20,3	-0,9	
Moy	11,3566667	17,8666667	1,8	
Variance	24,7782556	16,2988889	14,1433333	
σ	4,97777617	4,03718824	3,76076233	

La normalisation d'une composant est données la formule : $\frac{x_i - moy}{\sigma}$

Matrice réduite

Ville	Précipitation	$Température_{max}$	$Température_{min}$
Alger	0,13727683	1,44490001	1,09020449
Oran	1,16986645	-0,58621658	-0,95725273
Constantine	0,09508932	-1,18068972	0,26590354
Biskra	-1,02991105	-1,08161087	-1,11679485
Jijel	1,1316968	0,80088743	1,43587909
Bejaia	-1,50401834	0,60272972	-0,71793955

Note : X_{cr} : $matrice\ des\ données\ centrées\ reduites$

2) **Matrice de corrélation appelée R**

Elle donne les coefficients de corrélation linéaire des variables prises deux à deux. C'est une succession d'analyses bivariées.

$$R = \frac{1}{nbre\_individu} * X_{cr}^{t} * X_{cr}$$

On obtient une matrice symétrique

R=

	P	Tmax	Tmin
P	1	Corél(P, Tmax)	Corél(P, Tmin)
Tmax	Corél(Tmax, P)	1	Corél(Tmax, Tmin)
Tmin	Corél(Tmin, P)	Corél(Tmin, Tmax)	1

$$R=\begin{pmatrix} 0{,}14 & 1{,}17 & 0{,}09 & -1{,}03 & 1{,}131 & -1{,}50 \\ 1{,}44 & -0{,}60 & -1{,}20 & -1{,}1 & 0{,}80 & 0.60 \\ 1.09 & -0.95 & 0.26 & -1.11 & 1.43 & -0.71 \end{pmatrix} \begin{pmatrix} 0.14 & 1.44 & 1.09 \\ 1.17 & -060 & -095 \\ 0.09 & -1.20 & 0.26 \\ -1.03 & -1.10 & 1.11 \\ 1.31 & 0.80 & -1.43 \\ -1.5 & 0.60 & -0.71 \end{pmatrix} = \begin{pmatrix} 1 & 0.085 & 0.485 \\ 0.085 & 1 & 0.624 \\ 0.485 & 0.624 & 1 \end{pmatrix}$$

Calcul de valeurs propres et vectrices propres de la matrice de corrélation
Les valeurs propres classées en ordre décroissant

$$\lambda_1 = 1.83 : V_1 \begin{pmatrix} 0.46 \\ 0.56 \\ 0.69 \end{pmatrix}, \qquad \lambda_2 = 0.92 : V_2 \begin{pmatrix} 0.79 \\ -0.61 \\ -0.08 \end{pmatrix}, \qquad \lambda_3 = 0.025 : V_3 \begin{pmatrix} 0.41 \\ 0.56 \\ 0.72 \end{pmatrix}$$

Définition de l'Inertie expliquée
En Analyse en Composantes Principales **(ACP)**, l'**inertie expliquée** est une notion centrale qui permet de mesurer combien d'information (ou de variabilité) des données originales est conservée par les axes principaux.
Les valeurs propres permettent de déterminer l'inertié epliquée par chaque axe factoriel (principal).

Définition

- L'**inertie totale** correspond à la somme des variances de toutes les variables (ou, en termes matriciels, à la somme des valeurs propres de la matrice de covariance/corrélation). $Inertie\ total = \sum_{j=}^{N} \lambda_j$ $N\ est\ le\ nombre\ de\ valeurs\ propres$Dans l'exemple

$$\lambda_1 + \lambda_2 + \lambda_3 = 1.83 + 0.92 + 0.025 \approx 3$$

- Chaque **composante principale** est associée à une valeur propre λ_i, qui mesure la part de variance expliquée par cet axe.
- L'**inertie expliquée par une composante** est donc :

$$inertie\ expliquée\ par\ l'axe\ \text{i} = \frac{\lambda_i}{\sum_{j=}^{N} \lambda_j}$$

- L'**inertie cumulée** après k axes est :

$$inertie\ cumulé = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=}^{N} \lambda_j}$$

	Valeur propre	Inertié explique pour chaque composante(%)	Inertié cumulé(%)
λ_1	1.83	61.17	61.17
λ_2	0.92	30.51	91.68
λ_3	0.025	8.32	100

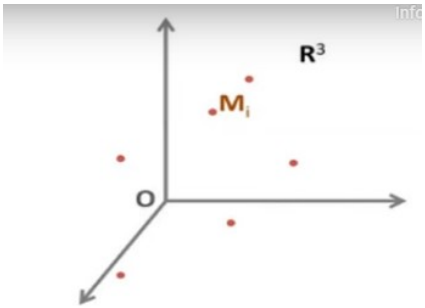
- Le premuer axe contient 61.17 de l'information
- Le deuxième contient 30.51 de l'information.
  Les deux premiers axes expliquent 91.68 % de l'inertie, cela signifie que projeter les données sur ces deux axes conserve 91.68 % de l'information initiale.
- Plus l'inertie expliquée est élevée, plus la représentation en dimensions réduites est fidèle.

Projection des données centrée réduites

1er étaoe :

On commence par déssiner les points de la matrice de données réduite (M1, M2, M3, M4, M5, M6) dans un graphe 3D.

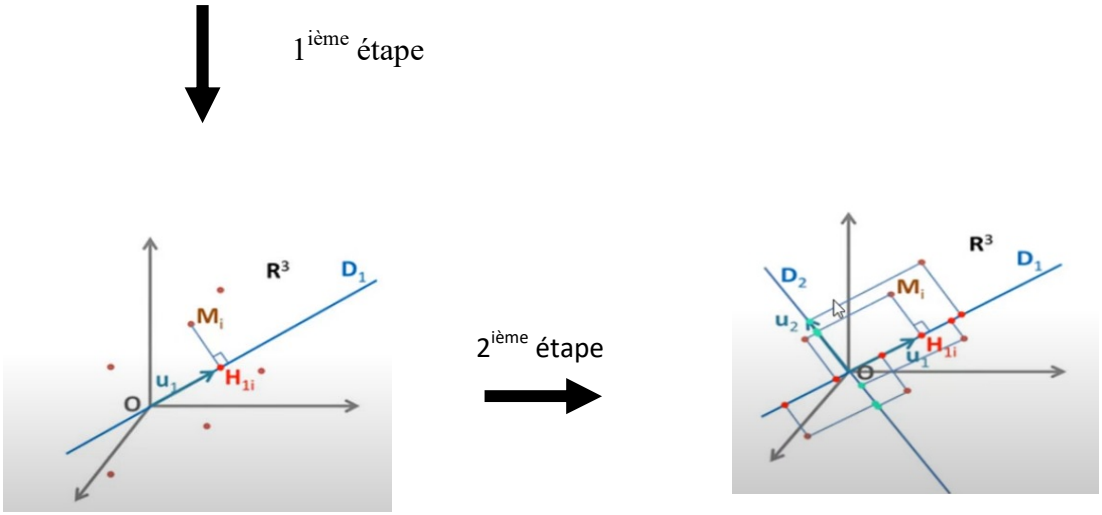
$$\boldsymbol{x} = \begin{bmatrix} 0.14 & 1.44 & 1.09 \\ 1.17 & -060 & -095 \\ 0.09 & -1.20 & 0.26 \\ -1.03 & -1.10 & 1.11 \\ 1.31 & 0.80 & -1.43 \\ -1.5 & 0.60 & -0.71 \end{bmatrix} \begin{bmatrix} \boldsymbol{M_1} \\ \boldsymbol{M_2} \\ \boldsymbol{M_3} \\ \boldsymbol{M_4} \\ \boldsymbol{M_5} \\ \boldsymbol{M_6} \end{bmatrix}$$



2ième étape

La 2ième étape est de passer à la projection dans un espace réduit 2D en utilisant les axes d'inerties, pour cela on dessine en premier lieu sur le graphe 3D l'axe à la plus grande valeur d'inertie qui explique l'information à 61.17% vient l'étape de la projection de ces 6 points dur cet axe (bien sur avec une perte d'une petite information sur ces données), puis on passe au

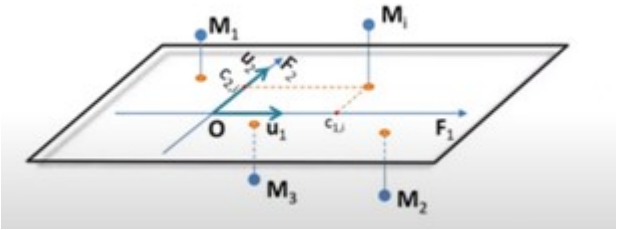
deuxième axe d'inertie qui explique l'information à 30.51% et on fait la projection des 6 points(voir la figure au dessous)



Plan factoruei

Le plan défini par le couple de vecteurs propres (V_1, V_2) est appelé plan factoriel :

- Le plus proche des points représentants les individus
- Sur lequels ces points se déforment le moins possible
- Explique le mieux l'inertie globale



Les composantes principales

$$\begin{pmatrix} 0.14 & 1.44 & 1.09 \\ 1.17 & -060 & -095 \\ 0.09 & -1.20 & 0.26 \\ -1.03 & -1.10 & 1.11 \\ 1.31 & 0.80 & -1.43 \\ -1.5 & 0.60 & -0.71 \end{pmatrix} \quad V_1\begin{pmatrix} 0.46 \\ 0.56 \\ 0.69 \end{pmatrix}, V_2\begin{pmatrix} 0.79 \\ -0.61 \\ -0.08 \end{pmatrix}, V_3\begin{pmatrix} 0.41 \\ 0.56 \\ 0.72 \end{pmatrix}$$

On va calculer les composantes principales en multipliant le s vecteur propres par les donnéeds centrées réduitents

- La projection sur l'axe 1 : $X_{cr} * V_1$

$$\begin{pmatrix} 0.14 & 1.44 & 1.09 \\ 1.17 & -060 & -095 \\ 0.09 & -1.20 & 0.26 \\ -1.03 & -1.10 & 1.11 \\ 1.31 & 0.80 & -1.43 \\ -1.5 & 0.60 & -0.71 \end{pmatrix} * \begin{pmatrix} 0.46 \\ 0.56 \\ 0.69 \end{pmatrix} = \begin{pmatrix} 1.63 \\ -0.45 \\ -0.43 \\ -1.85 \\ 1.96 \\ -0.85 \end{pmatrix}$$

- La projection sur l'axe 2 : $X_{cr} * V_2$

$$\begin{pmatrix} 0.14 & 1.44 & 1.09 \\ 1.17 & -060 & -095 \\ 0.09 & -1.20 & 0.26 \\ -1.03 & -1.10 & 1.11 \\ 1.31 & 0.80 & -1.43 \\ -1.5 & 0.60 & -0.71 \end{pmatrix} * \begin{pmatrix} 0.79 \\ -0.61 \\ -0.08 \end{pmatrix} = \begin{pmatrix} -081 \\ 1.31 \\ 0.79 \\ -0.12 \\ 0.37 \\ -1.54 \end{pmatrix}$$

- La projection sur l'axe 3 : $X_{cr} * V_3$

$$\begin{pmatrix} 0.14 & 1.44 & 1.09 \\ 1.17 & -060 & -095 \\ 0.09 & -1.20 & 0.26 \\ -1.03 & -1.10 & 1.11 \\ 1.31 & 0.80 & -1.43 \\ -1.5 & 0.60 & -0.71 \end{pmatrix} * \begin{pmatrix} 0.41 \\ 0.56 \\ 0.72 \end{pmatrix} = \begin{pmatrix} 0.07 \\ 0.84 \\ -0.81 \\ -0.21 \\ -0.13 \\ 0.24 \end{pmatrix}$$

Composantes principales

λ	1.83	0.92	0.025
Nouvelles variables	F_1	F_2	F_3
Alger	1.63	-0.81	0.07
Oran	-0.45	1.31	0.84
Constantine	-0.43	0.79	-0.81
Biskra	-1.85	-0.12	-0.21
Jijel	1.96	0.37	-0.13
Bejaia	-0.85	-1.54	0.24

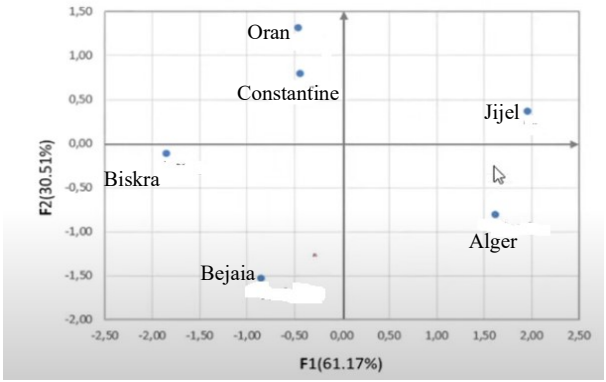
Propriétés :

- Les composantes principales sont centrées c'est-à-dire la moyenne de chaque composante est nulle
- La variance de chaque composante est égale à la valeur propre associée à l'axe

Nuage des indivudus

La projection des oints sur le plan factoriel défini par (v_1, v_2) permet d'obtenir le nuage des indivudus qui capture le maximu d'information possible.

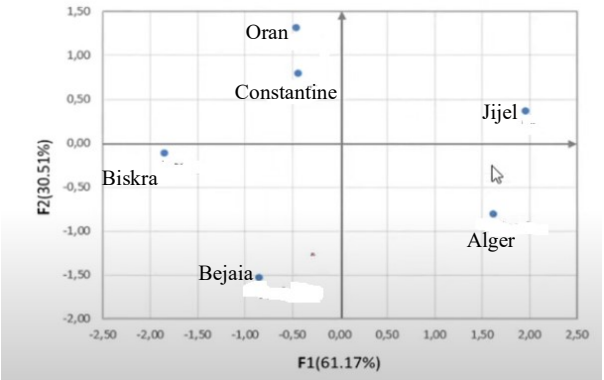
Nouvelles variables	F_1	F_2
Alger	1.63	-0.81
Oran	-0.45	1.31
Constantine	-0.43	0.79
Biskra	-1.85	-0.12
Jijel	1.96	0.37
Bejaia	-0.85	-1.54



L'inertié expliquéepar le plan factoriel : 91.68%

Nuage des individus : interprétation

Maintenant comment on va interpréter ce nuage de points



Ville	P	T_{max}	T_{min}
Alger	12,04	23,7	5,9
Oran	17,18	15,5	-1,8
Constantine	11,83	13,1	2,8
Biskra	6,23	13,5	-2,4
Jijel	16,99	21,1	7,2
Bejaia	3,87	20,3	-0,9

En utilisant le tableau d'induvidus et et le graphe associé on va essayer de cherhcer :

1) Quels sont les individus qui sont proches et les plus éloignés par rapport à l'axe 1 n, l'axe 2.
2) On voit que Jijel et Biskra sont éloignés par rapport à l'axe 1 (précipitation) ont deux comportements différends
3) De même Béjaia et Oran ont deux comportements différends par rapport à l'axee 2 (température maximale)

Pour mieux comprendre les comportements par rapport aux deux axes, alors il faut calculer le cercle de corrélation entre les variables initiales est les composantes principales

$$V_1\begin{pmatrix}0.46\\0.56\\0.69\end{pmatrix}, V_2\begin{pmatrix}0.79\\-0.61\\-0.08\end{pmatrix}, V_3\begin{pmatrix}0.41\\0.56\\0.72\end{pmatrix}$$

La corrélation entre une variable initiale K_i et une composante principale F_j est donnée par :

$$r(K_i, F_j) = élément\ \ i\ \ du\ vecteur\ propre\ V_{ij} * \sqrt{\lambda_j}$$

Où :
V_{ij} est la i-ème coordonnée du vecteur propre V_j,
λ_j est l'écart-type de la composante principale F_j.

	F1	F2	F3
P	0.62	0.76	0.21
T_{max}	0.76	-0.59	0.28
T_{min}	0.93	-0.03	-0.36

λ	1.83	0.92	0.025
Nouvelles variables	F_1	F_2	F_3
Alger	1.63	-0.81	0.07
Oran	-0.45	1.31	0.84
Constantine	-0.43	0.79	-0.81
Biskra	-1.85	-0.12	-0.21
Jijel	1.96	0.37	-0.13
Bejaia	-0.85	-1.54	0.24

$$r(P, F_j) = élément\ \ 1\ \ du\ vecteur\ propre\ V_1 * \sqrt{\lambda_1}$$
$$= 0.46 * \sqrt{1.83} = 0.62$$

$$r(T_{min}, F_j) = élément\ \ 3\ \ du\ vecteur\ propre\ V_1$$
$$* \sqrt{\lambda_1} = 0.69 * \sqrt{1.83} = 0.93$$

Calculs manuels

Étape 1 :

- **Pour F_1 ($\sqrt{\lambda_1} = \sqrt{1.83} \approx 1.353$)**
  - r (P,F_1)=0.46*1.353≈0.62
  - r (T_{max},F_1)=0.56*1.353≈0.76
  - r (T_{min},F_1)=0.69*1.353≈0.93
- **Pour F_2 ($\sqrt{\lambda_2} = \sqrt{0.92} \approx 0.959$)**
  - r(P,F_2)=0.79*0.959≈0.76
  - r(T_{max},F_2)=−0.61*0.959≈−0.59
  - r(T_{min},F_2)=−0.08*0.959≈−0.03
- **Pour F3 ($\sqrt{\lambda_3} = \sqrt{0.025} \approx 0.158$)**
  - r(P,F_3)=0.41*0.158≈0.21
  - r(T_{max},F_3)=0.56*0.158≈0.28
  - r(T_{min},F_3)=0.72*0.158≈−0.36 (attention au signe selon le vecteur propre)

Étape 2 : Résultats obtenus

	F1	F2	F3
P	0.62	0.76	0.21
T_{max}	0.76	-0.59	0.28
T_{min}	0.93	-0.03	-0.36

Étape 3 : Cercle de corrélation

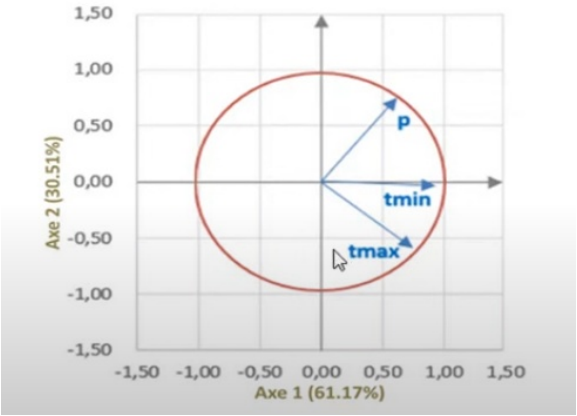
- On trace un cercle unité (rayon = 1).
- Chaque variable est représentée par ses coordonnées (r(K_i,F_1),r(K_i,F_2)).
- **Exemple :**
  - **P** → (0.62, 0.76)
  - **T_{max}** → (0.76, -0.59)
  - **T_{min}** → (0.93, -0.03)

Ces points montrent comment les variables originales se projettent sur le plan formé par les deux premiers axes (F_1–F_2).

- Plus un point n'est proche du cercle, mieux la variable est représentée par ce plan.
- Ici, **T_{min}** est très bien représentée par F_1, **P** est bien corrélé aux deux axes, et **T_{max}** est corrélée positivement à F_1 mais négativement à F_2.

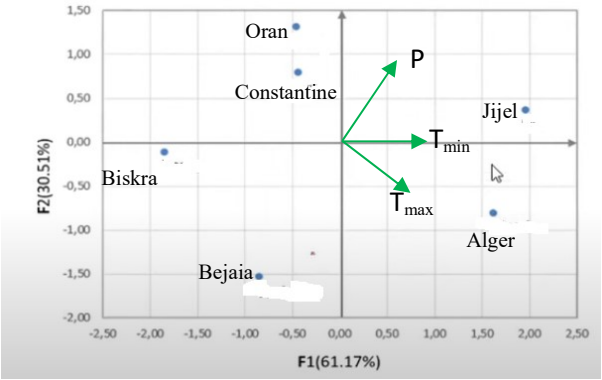
En résumé : le cercle de corrélation est obtenu en multipliant les coordonnées des vecteurs propres par la racine carrée des valeurs propres, puis en projetant les variables sur le plan des deux premiers axes.

Génèration du schéma du cercle de corrélation avec ces trois variables (V_1, V_2, V_3) pour visualiser leur position par rapport aux axes F_1 et F_2



Rappel : étudier les axes d'inertie des individus est équivalent à étudier les axes d'inertié des variables

En ACPquand on analyse un échantillon, étudier les indivudus ou étudier les variables, c'est étudier deux facettes d'une même chose



Le schéma du cercle de corrélation montre les trois variables **P**, $\mathbf{T_{max}}$ et $\mathbf{T_{min}}$ projetées sur les axes principaux F_1 et F_2, avec leurs coordonnées respectives (P : 0.62, 0.76 ; T_{max} : 0.76, -0.59 ; T_{min} : 0.93, -0.03).

On peut l'examiner ci-dessous pour visualiser clairement la position des variables par rapport au cercle unité et aux deux axes :

Ce graphique permet de voir que :

- **P** est corrélé positivement avec F_1 et F_2 (en haut à droite).
- $\mathbf{T_{max}}$ est corrélée positivement avec F_1 mais négativement avec F_2 (en bas à droite).

- **T_{min}** est fortement corrélée avec F_1 et quasi indépendante de F_2 (à droite, proche de l'axe F_1).

Interprétation de la qualité de représentation (cos^2) en ACP

Le **cos^2** (cosinus carré) mesure à quel point une variable initiale est bien représentée par un plan formé par les composantes principales (souvent F_1 et F_2). C'est un indicateur clé dans le **cercle de corrélation**.

Définition mathématique

Pour une variable K_i, et les deux premières composantes principales F_1 et F_2, on calcule :

$$cos^2(K_i) = r(K_i, F_1)^2 + r(K_i, F_2)^2$$

Où $r(K_i,F_j)$ est la **corrélation** entre la variable K_i et l'axe F_j.

Interprétation

- **Si** $cos^2 \approx 1$: La variable est **très bien représentée** par le plan F_1–F_2. Sa position dans le cercle est proche du bord.
- **Si** $cos^2 \approx 0$: La variable est **mal représentée** par ce plan. Elle est projetée près du centre du cercle.
- **Si** $cos^2 = 0.7$: Cela signifie que **70 % de la variance** de cette variable est capturée par les deux premiers axes.

Exemple avec tes données

Prenons les corrélations :

Application aux variables

- **Données des corrélations:**
  - **P:** $r(F_1)=0.62$, $r(F_2)=0.76$
  - **Tmax:** $r(F_1)=0.76$, $r(F_2)=-0.59$
  - **Tmin:** $r(F_1)=0.93$, $r(F_2)=-0.03$
- **Calculs:**
  - **P:** $cos^2=0.622+0.762=0.3844+0.5776=0.9620$
  - **Tmax:** $cos^2=0.762+(-0.59)2=0.5776+0.3481=0.9257$
  - **Tmin:** $cos^2=0.932+(-0.03)2=0.8649+0.0009=0.8658$

Variable	**F_1**	**F_2**	**cos^2**
P	0.62	0.76	0.962
Tmax	0.76	-0.59	0.9257
Tmin	0.93	-0.03	0.8658

Interprétation :

- **P** est très bien représentée (96 %) par le plan F_1–F_2.
- **Tmax** aussi (92 %).
- **Tmin** un peu moins (86 %), mais reste bien projetée.

Utilisation pratique

- On peut **filtrer les variables** avec un cos² élevé pour se concentrer sur celles bien représentées.
- On peut aussi **colorer les flèches** dans le cercle de corrélation selon leur cos² pour visualiser la qualité.

Application aux individus

Qualité de représentation d'un individu sur un axe

Concept

La qualité de représentation d'un individu sur un axe principal mesure la part de sa « variance » capturée par cet axe. C'est la proportion de son score au carré sur l'axe, rapportée à la somme des carrés de ses scores sur tous les axes (sa norme au carré). Elle correspond au cosinus carré de l'angle entre le vecteur des scores et l'axe.

Formule

Pour l'individu M_i et l'axe F_j, avec score C_{ij} sur l'axe F_j et norme $\|M_i\|^2 = \sum_k C_{iK}^2$, la qualité est:

$$Qualité(M_i, F_j) = Cos^2(\theta) = \frac{C_{ij}^2}{\|M_i\|^2}$$

Exemple détaillé sur Biskra

- **Scores:** $F_1 = -1.85$, $F_2 = -0.12$, $F_3 = -0.21$
- **Norme au carré:**

$\|\text{Biskra}\|^2 = (-1.85)^2 + (-0.12)^2 + (-0.21)^2 = 3.4225 + 0.0144 + 0.0441 = 3.4810$

- **Qualité par axe:**
  - Qualite´(Biskra,F1)=3.4225/3.4810≈0.983
  - Qualite´(Biskra,F2)=0.0144/3.4810≈0.004
  - Qualite´(Biskra,F3)=0.0441/3.4810≈0.013

Application aux individus donnés

Les scores fournis sont:

- **Alger:** F1=1.63, F2=−0.81, F3=0.07 → ‖·‖2=3.3179
- **Oran:** F1=−0.45, F2=1.31, F3=0.84 → ‖·‖2=2.6242

- **Constantine:** F1=−0.43, F2=0.79, F3=−0.81 → ‖·‖2=1.4651
- **Biskra:** F1=−1.85, F2=−0.12, F3=−0.21 → ‖·‖2=3.4810
- **Jijel:** F1=1.96, F2=0.37, F3=−0.13 → ‖·‖2=3.9954
- **Bejaia:** F1=−0.85, F2=−1.54, F3=0.24 → ‖·‖2=3.1517

Qualités par axe

Individu	Qualité(F1)	Qualité(F2)	Qualité(F3)
Alger	0.801	0.198	0.0015
Oran	0.077	0.654	0.269
Constantine	0.126	0.426	0.448
Biskra	0.983	0.004	0.013
Jijel	0.962	0.034	0.0042
Bejaia	0.229	0.753	0.018

Ces valeurs corrigent les incohérences du tableau initial où certaines entrées de F2/F3 semblaient être des scores bruts et non des qualités. La somme des trois qualités par individu vaut 1 (à l'arrondi près), ce qui vérifie la décomposition de la variance sur les axes.

Lecture et interprétation

- **Variables très bien sur F1–F2:** Alger, Jijel, Biskra, Bejaia (leurs qualités sur F1–F2 sont proches de 1).
- **Individus nécessitant F3 pour bien comprendre:** Oran et Constantine ont des qualités F3 non négligeables (≈ 0.27 et 0.45), donc leur position sur le plan F1–F2 perd de l'information.

Maintenant on calcule aussi les cos² pour les **individus** (villes comme Alger, Oran, etc.) sur le plan F_1–F_2 pour voir lesquelles sont bien représentées

Cos² des individus sur le plan F_1–F_2

Le cos² pour un individu mesure la qualité de sa représentation sur le plan F1–F2. On le calcule par:

$$cos^2 = \frac{F_1^2 + F_2^2}{F_1^2 + F_2^2 + F_3^2}$$

- **Scores donnés:**
  - **Alger:** F1=1.63, F2=−0.81, F3=0.07
  - **Oran:** F1=−0.45, F2=1.31, F3=0.84
  - **Constantine:** F1=−0.43, F2=0.79, F3=−0.81
  - **Biskra:** F1=−1.85, F2=−0.12, F3=−0.21
  - **Jijel:** F1=1.96, F2=0.37, F3=−0.13
  - **Bejaia:** F1=−0.85, F2=−1.54, F3=0.24

- **Calculs:**
  - **Alger:** $\cos^2 = 1.632 + (-0.81)21.632 + (-0.81)2 + 0.072 = 2.6569 + 0.65612.6569 + 0.6561 + 0.0049 \approx 3.31303.3179 = 0.9985$
  - **Oran:** $\cos^2 = 0.2025 + 1.71610.2025 + 1.7161 + 0.7056 = 1.91862.6242 = 0.7315$
  - **Constantine:** $\cos^2 = 0.1849 + 0.62410.1849 + 0.6241 + 0.6561 = 0.80901.4651 = 0.5522$
  - **Biskra:** $\cos^2 = 3.4225 + 0.01443.4225 + 0.0144 + 0.0441 = 3.43693.4810 = 0.9873$
  - **Jijel:** $\cos^2 = 3.8416 + 0.13693.8416 + 0.1369 + 0.0169 = 3.97853.9954 = 0.9960$
  - **Bejaia:** $\cos^2 = 0.7225 + 2.37160.7225 + 2.3716 + 0.0576 = 3.09413.1517 = 0.9827$

Résultats

Individu	F_1	F_2	F_3	$F_1^2+F_2^2$	Total ($F_1^2+F_2^2+F_3^2$)	$\cos^2$ F_1–F_2
Alger	1.63	-0.81	0.07	3.3130	3.3179	0.9985
Oran	-0.45	1.31	0.84	1.9186	2.6242	0.7315
Constantine	-0.43	0.79	-0.81	0.8090	1.4651	0.5522
Biskra	-1.85	-0.12	-0.21	3.4369	3.4810	0.9873
Jijel	1.96	0.37	-0.13	3.9785	3.9954	0.9960
Bejaia	-0.85	-1.54	0.24	3.0941	3.1517	0.9827

Interprétation

- **Seuil pratique:** un $\cos^2$ supérieur à 0.8 indique une très bonne représentation sur le plan F_1–F_2.
- **Très bien représentés:**
  - **Alger (0.9985), Jijel (0.9960), Biskra (0.9873), Bejaia (0.9827)** — leurs positions dans le plan F_1–F_2 reflètent presque toute leur variance.
- **Bien à moyennement représentés:**
  - **Oran (0.7315)** — une part notable est captée par F_3, donc sa position sur F_1–F_2 perd un peu d'information.
  - **Constantine (0.5522)** — fortement influencée par F_3; il faut regarder le plan incluant F_3 pour une interprétation complète.