TD : Analyse en Composantes Principales (ACP) – Énoncé

1. Objectif du TD

Ce TD a pour but de comprendre et d'appliquer pas à pas les calculs de base de l'Analyse en Composantes Principales (ACP), à partir d'un petit jeu de données, en utilisant les formules fondamentales. Les étudiants devront effectuer tous les calculs à la main (ou sur tableur) pour bien assimiler la démarche.

2. Jeu de données

On dispose de trois variables quantitatives observées sur quatre individus :

Individu	**X_1 (Poids)**	**X_2 (Taille)**	**X_3 (Âge)**
A	60	165	25
B	72	180	30
C	57	170	22
D	80	190	28

3. Étape préliminaire : Standardisation des données

Avant de commencer l'ACP, il faut centrer et réduire les variables.

Formule : $z_{ij} = (x_{ij} - \bar{x}_j) / s_j$

avec : $\bar{x}_j$ = moyenne de la variable X_j et s_j = écart-type de la variable X_j.

- **<u>Travail demandé :</u>**

1. Calculez les moyennes et écarts-types de chaque variable.
2. Déduisez les valeurs centrées et réduites (z_{ij}) pour chaque individu.

4. Calcul de la matrice de corrélation

Formules :

$r_{ij} = \mathrm{Cov}(X_i, X_j) / (s_i \times s_j)$

$\mathrm{Cov}(X_i, X_j) = (1/(n-1))\, \Sigma(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$

- **<u>Travail demandé :</u>**

1. Construisez la matrice de corrélation R.
2. Vérifiez que R est symétrique et que ses coefficients sont compris entre −1 et 1.

5. Calcul des valeurs propres et vecteurs propres

Formules :

$|R - \lambda I| = 0$

$(R - \lambda_k I)u_k = 0$

$||u_k|| = 1$

- **Travail demandé :**

1. Résolvez l'équation caractéristique pour obtenir les valeurs propres (λ_1, λ_2, λ_3).
2. Pour chaque λ_k, trouvez le vecteur propre u_k associé.
3. Normalisez chaque vecteur propre.
4. Calculez la part de variance expliquée par chaque axe : $Var(F_k) = \lambda_k / \Sigma\lambda_i$.

6. Calcul des composantes principales

Formules :

$F_{ik} = z_{i1}u_{1k} + z_{i2}u_{2k} + z_{i3}u_{3k}$

ou matriciellement : $F = ZU$

- **Travail demandé :**

1. Calculez les coordonnées des individus sur les deux premiers axes (F_1 et F_2).
2. Représentez les individus sur le plan factoriel (F_1, F_2).

7. Qualité de représentation des variables (cos²)

Formules :

$corr(X_j, F_k) = \sqrt{\lambda_k} \times u_{jk}$

$\cos^2(X_j, F_k) = \lambda_k \times u_{jk}^2$

- **Travail demandé :**

1. Calculez les corrélations entre chaque variable et les axes F_1 et F_2.
2. En déduisez la qualité de représentation ($\cos^2$) des variables sur chaque axe.
3. Interprétez : quelles variables sont bien représentées sur le premier plan ?

8. Contributions des variables aux axes

Formules :

$Ctr(X_j, F_k) = u_{jk}^2$

ou en pourcentage : $Ctr(X_j, F_k) = (u_{jk}^2 / \Sigma u_{mk}^2) \times 100$

- **Travail demandé :**

1. Calculez les contributions des variables à la formation des axes F_1 et F_2.
2. Interprétez : quelles variables déterminent le plus chaque axe ?

9. Interprétation finale

1. Quelle proportion de la variance totale est expliquée par les deux premiers axes ?
2. Que représente le premier axe ? (un facteur global, une opposition, etc.)
3. Quelle variable semble indépendante ou spécifique ?
4. Quelle réduction de dimension est raisonnable ici ?