ACP : Analyse en Composantes principales

Introduction :

ACP est une approche qui est employée pour résumer l'information contenue dans une matrice (tableau) de données comportant un nombre important de variables numériques.

⚠ Si le table de données ne contient que deux variables, alors c'est facile d'étudier **le lien** entre ces deux variables en appliquant un analyse bivariée : donc on calcule la corrélation entre ces variables.

Mais Si le tableau contient beaucoup des variables cela devient impossible à appréhender.

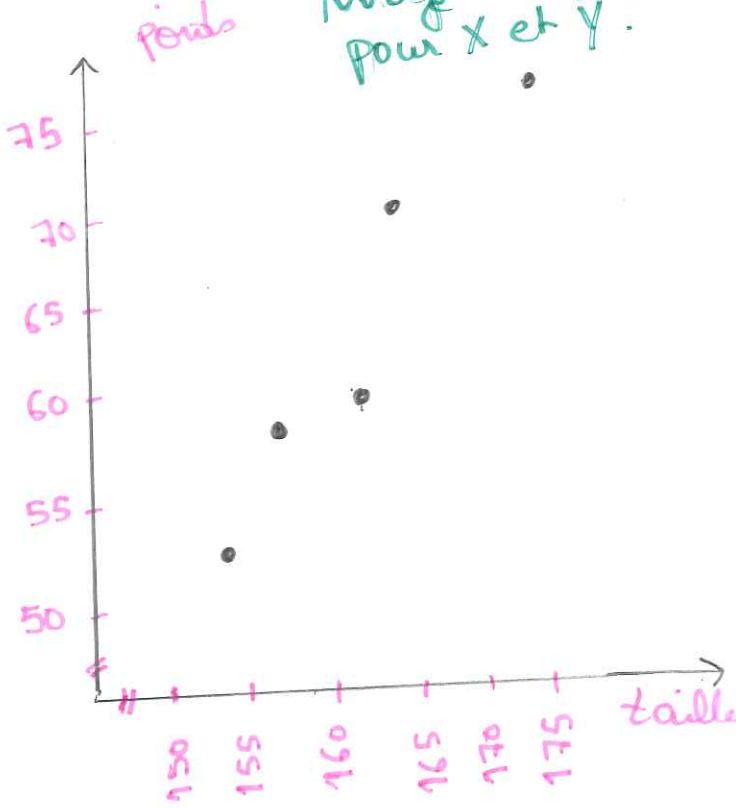
Alors, l'ACP permet alors de passer d'un nuage de points qui évolue dans un espace à $\underline{p}$ dimensions (p = nb de variables ou nb de colonnes) à une représentation en deux dimensions (ou éventuellement plusieurs représentations à 2 dimensions)

- Exp:
- On donne le tableau suivant :

① Etude de deux variables :

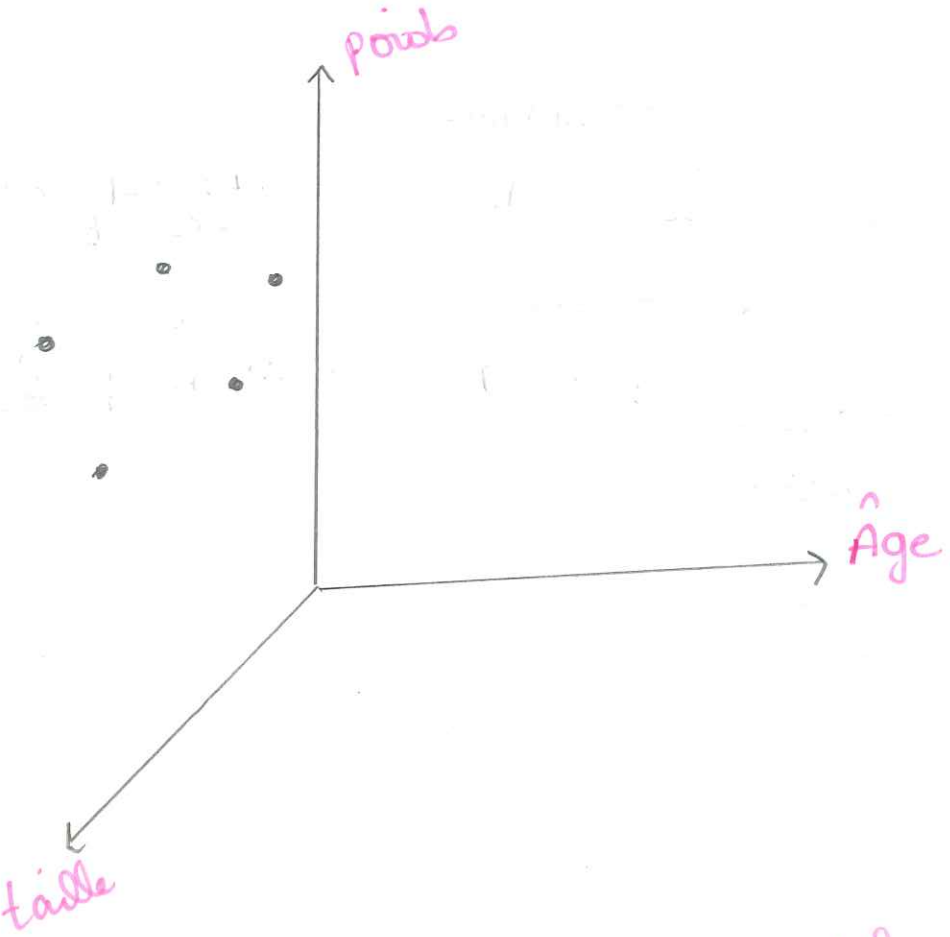
Etudiant	taille X	Poids Y
Sarah	158	57
Olivier	169	70
Paul	174	75
Manon	155	52
Serena	163	58

Nuage de point pour X et Y.



② Si on a 3 variables : On a ajouté une 3e variable Âge.

Etudiant	taille X	Poids Y	Âge Z
Sarah	158	57	21
Olivia	169	70	19
Paul	174	75	18
Manon	155	52	22
Serena	163	58	23



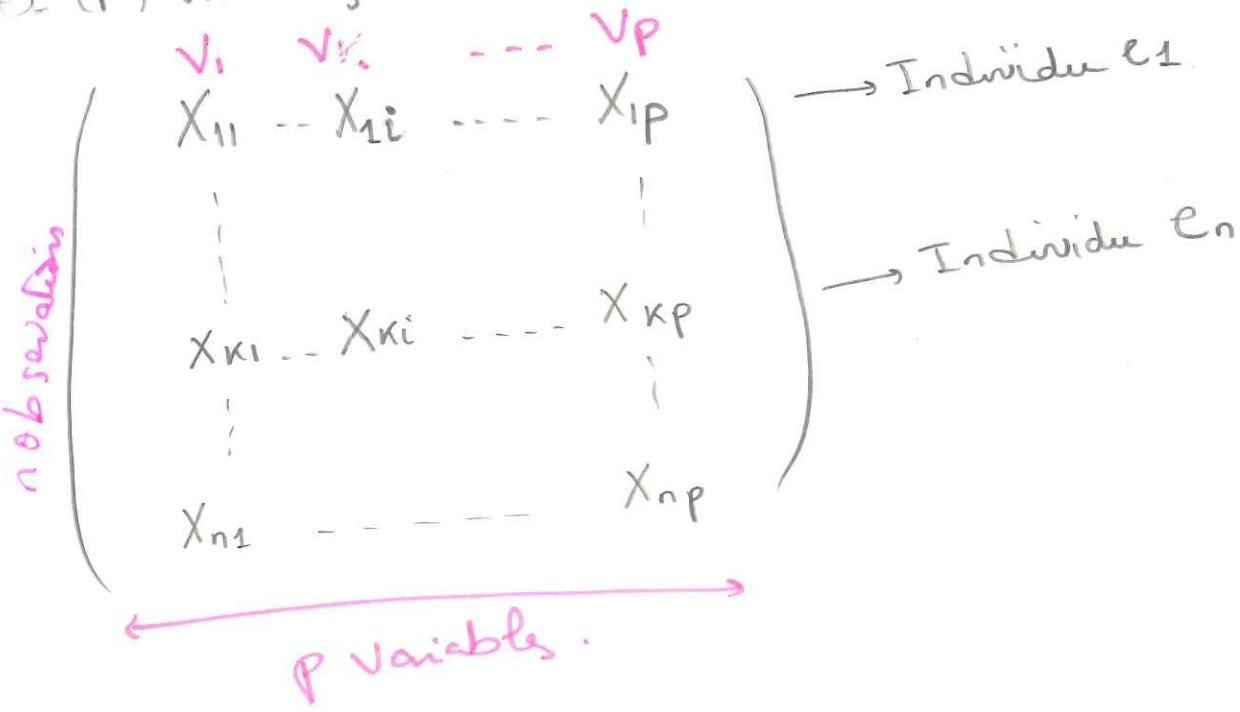
③ Si on ajoute une 4ème variable exemple : l'année d'obtention de Bac : ça sera difficile de représenter ces donner, car on a 4 variables alors 4 axes (X, Y, Z, t) ; dans ce cas on applique un ACP :

⚠ BUT : ACP va rechercher l'axe dans lequel les observations sont le plus dispersées. On l'appelle cet axe : première composante principale C_1

Maintenant

- Ensuite ; IL va chercher l'axe perpendiculaire à C_1 et qui va être le deuxième composante principale C_2. (Les axes C_1 et C_2 sont orthogonal $\rightarrow [\cos(\widehat{C_1, C_2}) = \text{corelat}(C_1, C_2) = 0]$

On considère le jeu de données avec (n) individus et (P) variables :



Centre de gravité : est le point dont les coordonnées sont les valeurs moyennes des variables :

$$G = (\bar{X}_1, \ldots, \bar{X}_p)$$

Inertie :

$$I = \sum_{K=1}^{n} \|e_K - G\|^2$$

(e_K : observations ; G : centre de gravité ; $\|\cdot\|$: Norme euclidienne)

Mesure l'information contenue dans un nuage de points. C'est la somme des distances au carré entre les observations et le centre de gravité.

Propriétés :

- Norme euclidienne pour vecteur $U(x,y)$: $\|u\| = \sqrt{x^2+y^2}$
- Norme euclidienne entre 2 pts A et B : $\|AB\| = \sqrt{(x_B-x_A)^2 + (y_B-y_A)^2}$

Propriétés de l'inertie :

$$I = tr(V) = \sum_{k=1}^{p} s_k^2 \qquad \left[s_k^2 = Var(X_k) \right]$$

V = matrice de var-cov de X

$$V = \begin{pmatrix} s_1^2 & \cdots & C_{X_1 X_p} \\ \vdots & \ddots & \vdots \\ C_{X_1 X_p} & \cdots & s_p^2 \end{pmatrix}$$

$C_{X_1 X_p}$ → covariance entre X_1 et X_p

matrice symétrique définie positive

$$I = \sum_{k=1}^{p} \lambda_k$$

λ_k : valeurs propres de matrice de var-cov $\underline{\underline{V}}$.

$\lambda_1 + \cdots + \lambda_p$

→ Si les variables sont centrées réduites alors :

$$I = p$$

p : nb variables

Interprétation (Inertie)

- Inertie : Mesure les informations expliquée par l'axe
- On fait la somme de % de variance donné pour C_1 et C_2 pour l'avoir

Ici : 96,5 % d'inertie expliqué par le plan (C_1, C_2)

Centre et réduire les variables:

① Distance entre 2 observations (individus) (k) et (k'):

$$\|e_k - e_{k'}\|^2 = \sum_{i=1}^{p} (x_{ki} - x_{k'i})^2$$

② Centré et réduite les variables:

but: Si les variables ont pas m̂ unité, certains variables à valeurs faible « disparaitons » de l'information au profit de celles ayant de fort valeurs.

Comment faire? Replacer chaque variable X_i par:

$$X_i \longrightarrow \frac{X_i - \bar{X}_i}{S_i} \quad ; i = 1, \dots, p$$

$\bar{X}_i$ → moyenne variable X_i

S_i → écart-type de variable X_i.

Produit scalaire:

→ Le produit scalaire entre 2 variables X_i et X_j est:

$$\langle X_i, X_j \rangle = \frac{1}{n} \sum_{k=1}^{n} x_{ki} x_{kj}$$

d'où La norme: $\|X_i\|^2 = \frac{1}{n} \sum_{k=1}^{n} (x_{ki})^2$

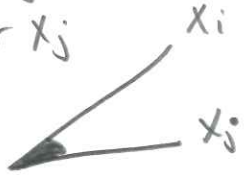
. Si les variables sont centré alors:

- $< X_i, X_j > = \cos(X_i, X_j) = C_{X_i X_j}$
- $\| X_i \|^2 = S_i^2$
- $\cos(\hat{X_i, X_j}) = \dfrac{< X_i, X_j >}{\| X_i \| . \| X_j \|} = \dfrac{C_{X_i . X_j}}{S_i^2 . S_j^2}$

$= r_{X_i X_j}$ → corrélation entre X_i et X_j

→ $\boxed{\cos(\hat{X_i, X_j}) = \text{cor}(X_i, X_j)}$

$r(X_i, X_j) > 0$, alors $\cos(\hat{X_i, X_j})$, donc il y a angle aigu (entre 0 et $\frac{\pi}{2}$) entre X_i et X_j



$r(X_i, X_j) < 0$ alors $\cos(\hat{X_i, X_j}) < 0$ donc il y a un angle obtu (entre $\frac{\pi}{2}$ et π) entre X_i et X_j

X_i et X_j corrélé négativement

$r(X_i, X_j) = 0$ alors $\cos(X_i, X_j) = \frac{\pi}{2}$ donc les variables sont orthogonal ↓ pas corrélé linéairement

Principe ACP:

. Trouver des espaces des petits dimensions, sur lesquel les projections des observat minimisent la défor de la réalité

. on cherche un sous espace F_q de $\mathbb{R}^p$ (de dimension $q = 2, 3, \ldots$) sur lequel on projete le nuage de points

→ f_K la projection de e_K dans nouveau plan F_q.
Il faut que la distance entre l'individu e_K et sa projecti f_K soit minimale

→ U_K → Le vecteur propre unitaire de la matri $\underline{\underline{V}}$ associé à La Keme plus grande valeur propre.

- L'inertie du nuage projeté sur U_K est λ_K.
- l'inertie du nuage projeté sur F_q est $\lambda_1 + \cdots + \lambda_q$
- l'inertie totale : $I = \lambda_1 + \cdots + \lambda_q$.

→ Les vecteurs propres $U_1, \ldots, U_K$ sont appellé axes principaux

« Le 1er axe principale U_1 est associe à la plus grand valeur propre λ_1. ; pareil U_2 pour λ_2.

→ La projection des individus sur un axe principale est une nouvelle variable appelli Composante principale.
→ La 1ère composante C_1 représente les coordonnés des projections des individus sur l'axe U_1

Interpretation :

. Representation des observations :

	eginvalue	% vaine	cumulative varia.
C_1	2.663	66.57	
C_2	1.19.	29.94	
C_3		3,180	
C_4			

C_1 contient 66,5% de l'inertie (des informations)

$\lambda_1 = 2.663$

$\lambda_2 = 1.19$

C_2 contient 29.94% des information

4 Variables

⇓

On obtient 4 Comp principales

alors : ces deux axes contient $66,5 + 29,94 = 96,5\%$ de l'information.

- % Variance de C_1 : $\dfrac{\lambda_1}{\lambda_1 + \dots + \lambda_p}$

- A'. Savoir :

$$C_i = \tilde{X} \cdot U_i$$

C_i : composante principale i ; $\tilde{X}$: Matrice des données centrés réduit ; U_i : axe principal

$$Var(C_i) = S^2_{C_i} = \lambda_i$$

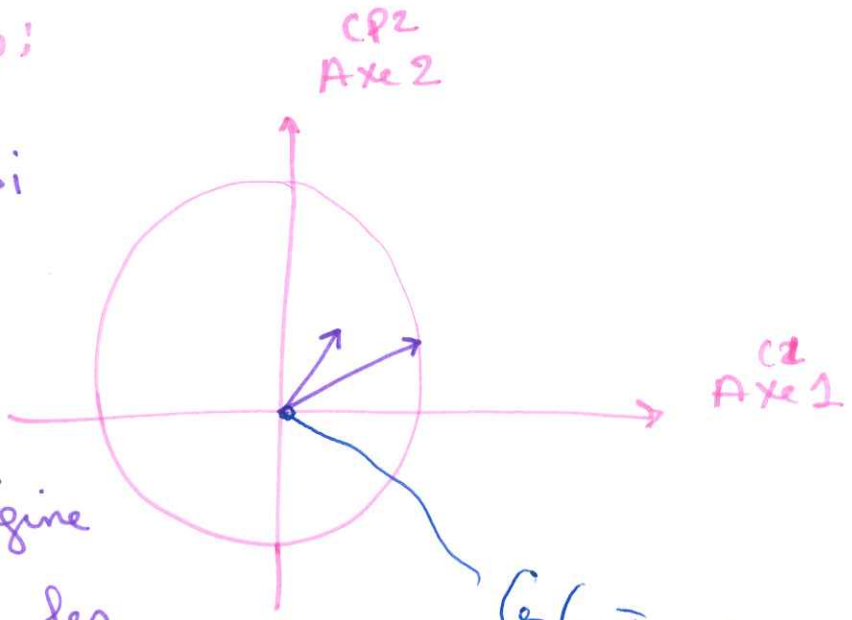
et

$$\sum \lambda_i = p$$

donc $\sum S^2_{C_i} = \sum S^2_{X_i} = \sum \lambda_i = p = I$ (nb variable)

(9)

Représentation des variables :



* Variables bien représenté si elles sont proches du cercle.

* Variables proche de l'origine sont peu corrélé avec les axes ; sont mals représenté alors : pas d'interpétation possible pour ces variables.

. Les individus sont bien représenté s'ils ne sont pas trop éloignés de l'axe sur lequel ont les projette.
→ Donc, il faut vérifier cos entre l'individu de l'axe (proche de 1).

→ valable si l'individu loin du centre de gravité.