I. Généralités :

L'Analyse en Composantes Principales (ACP) est une technique statistique utilisée pour réduire la dimensionnalité d'un jeu de données tout en conservant un maximum d'information. Voici comment elle fonctionne :

1. **Objectif** : L'ACP vise à transformer des variables potentiellement corrélées en un ensemble de nouvelles variables non corrélées appelées "composantes principales". Ces composantes sont des combinaisons linéaires des variables d'origine et capturent la variance maximale dans les données.
2. **Étapes clés** :
    - **Standardisation des données** : Si les variables ont des unités différentes, il est recommandé de standardiser les données pour les rendre comparables.
    - **Matrice de covariance** : L'ACP calcule ensuite la matrice de covariance ou de corrélation entre les variables pour comprendre leur relation.
    - **Valeurs propres et vecteurs propres** : Les valeurs propres (qui mesurent la quantité de variance expliquée par chaque composante) et les vecteurs propres (qui définissent la direction des composantes principales) sont extraits de la matrice de covariance.
    - **Projection des données** : Les données d'origine sont projetées dans un nouvel espace défini par les composantes principales. Les premières composantes principales expliquent le plus de variance dans les données.
3. **Interprétation** :
    - Les premières composantes principales contiennent l'essentiel de l'information du jeu de données.
    - Le nombre de composantes retenues dépend du pourcentage de variance totale que l'on souhaite conserver (souvent 80-90%).

L'ACP est donc un outil puissant pour simplifier des jeux de données complexes tout en minimisant la perte d'information.

Caractéristiques des composantes principales :

- **Composante principale 1 (CP1)** : C'est la direction qui explique le maximum de variance dans les données.
- **Composante principale 2 (CP2)** : C'est la direction perpendiculaire à CP1, qui explique la deuxième plus grande part de variance restante, et ainsi de suite.

Ces composantes sont **orthogonales** entre elles, c'est-à-dire non corrélées.

4. Interprétation :

- **Variance expliquée** : Chaque composante principale est associée à une part de variance expliquée. Par exemple, si la première composante explique 70 % de la variance, cela signifie qu'elle capture 70 % de l'information présente dans les données d'origine.
- **Projection des données** : Les données initiales peuvent être projetées sur les premières composantes principales pour visualiser les relations principales avec moins de dimensions.

Typiquement, on utilise souvent les **deux premières composantes** pour faire une représentation en deux dimensions des données.

5. Utilité de l'ACP :

1. **Réduction de la dimensionnalité** : Elle permet de réduire le nombre de variables tout en gardant l'essentiel de l'information.
2. **Visualisation** : En projetant les données sur les deux ou trois premières composantes principales, on peut facilement visualiser des données complexes en 2D ou 3D.
3. **Suppression de la redondance** : L'ACP élimine la **multicolinéarité** entre les variables (les variables redondantes qui véhiculent la même information).

II. Principe de l'ACP

L'inertie d'un nuage de points = inertie totale = $\sum_i \boldsymbol{P_i d^2}(\boldsymbol{x_i}, \boldsymbol{g})$ $(g\ est\ le\ centre\ de\ gravité)$

Comme déjà mentionné l'objectif principal de l'ACP est la recherche d'un ensemble réduit de variables non corrélés qui sont des combinaisons linéaires des variables initiales et qui résume avec précision les variables initiales. En d'autres termes, la recherche d'un sous espace représentant au mieux le nuage initial.

Nuage d'un points : Poids d'un individu en général égal a 1/N (N nombre d'individu), $\sum_{i=1}^{N} P_i = 1$

Inertie : $\boldsymbol{I_{A/B} = P_A d^2(A, B)}$,

$\boldsymbol{I_{A/\Delta} = P_A d^2(A, \Delta)}$

Inertie globale : $\boldsymbol{I_g = \sum_{i=1}^{n} d^2(e_i, g)}$

Rappel :

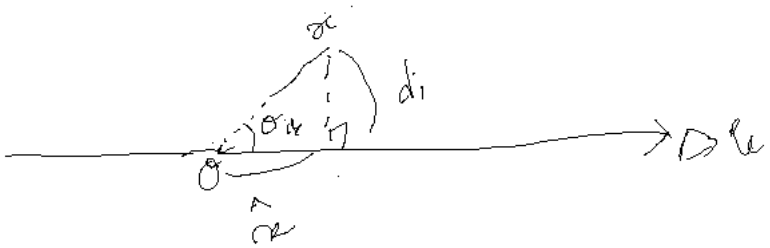
$\boldsymbol{x_1 = \begin{pmatrix} a \\ b \end{pmatrix}, x_2 = \begin{pmatrix} a' \\ b' \end{pmatrix}, M = I\ (matrice\ identite)}$

$\boldsymbol{d^2(x_1, x_2) = (a - a')^2 + (b - b')^2 = \|x_1 - x_2\|_M^2 = (x_1 - x_2)^t M (x_1 - x_2)}$

$\boldsymbol{v_1 \begin{pmatrix} x \\ y \end{pmatrix}, v_2 \begin{pmatrix} x' \\ y' \end{pmatrix}}$

Le produit scalaire de deux vecteurs $\boldsymbol{\langle v_1, v_2 \rangle = xx' + yy' = v_1^t \cdot v_2}$

Les axes principaux d'inertie :



$$\boldsymbol{I_{\Delta k} = \sum_{i=1}^{n} p_i d_M^2 x_{i/\Delta k} = \sum_{i=1}^{n} p_i \left(\|x_i\|_M^2 - \|\hat{x}_i\|_M^2 \right) = \sum_{i=1}^{n} P_i \|x_i\|_M^2 - \sum_{i=1}^{n} P_i \|\hat{x}_i\|_M^2}$$

$= \sum_{i=1}^{n} \boldsymbol{P_i \|x_i\|_M^2} - \boldsymbol{\sum p_i \langle x_{i'} u_k \rangle_M^2}$

L'inertie expliquée de l'axe u_k

A maximiser

$= \sum P_i \|x_i\|^2 - \sum P_i \langle x_{i'} u_k \rangle_M^t \cdot \langle x_{i'} u_k \rangle_M$

$= \sum P_i \|x_i\|^2 - \sum p_i (x_i M u_k)^t (x_i M u_k)$

$= \sum P_i \|x_i\|^2 - \sum P_i u_k^t M^t x^t x M u_k$

Matrice V de variance covariance

$= \sum P_i \|x_i\|^2 - u_k^t M^t (\sum P_i x^t x) M u_k$

$= \sum P_i \|x_i\|^2 - u_k^t M^t V M u_k$

Solution : u_k vecteur propre de VM associe aux valeurs propres λ_k, $VMu_k = \lambda_k u_k$

RQ : Les vecteurs propres forme une base orthonormée c-à-d :

$$\boldsymbol{\langle u_i, u_j \rangle_M = u_i^t M u_j = 0 \ \ \forall i \neq j, et \ , \|u_i\|^2 = \langle u_i^t, u_i \rangle_M = u_i^t M u_i = 1}$$

$$\boldsymbol{I_{\Delta k} = \sum P_i \|x_i\|^2 - u_k^t M^t \lambda_k u_k}$$

$$\boldsymbol{I_{\Delta k} = \sum P_i \|x_i\|^2 - (u_k^t M^t u_k) \lambda_k} \qquad (u_k^t M^t u_k = 1)$$

$$\boldsymbol{I_{\Delta k} = \sum P_i \|x_i\|^2 - \lambda_k}$$

M est la métrique :

Si les variables homogènes donc M est la matrice identité M=Id.

Si les données hétérogènes, M=D_{1/σ^2}

III. Les étapes d'une ACP

1/ Centrer le tableau :

$\boldsymbol{x = x - \bar{x}}$ (Données centrées dans le cas d'une ACP non normée)

Ou bien $\boldsymbol{x = \frac{x - \bar{x}}{\sigma_x}}$ (Données centrées-réduites dans le cas d'une ACP normée)

$$\bar{x} = g = \frac{\sum_{i=1}^{n} P_i x_{ij}}{\sum_{i=1}^{n} P_i}$$

Matrice de variance-covariance:

$v = \frac{1}{N} x^t x$ ($dans\ l'ACP\ normée\ est\ consideree\ comme\ matrice\ de\ correlation$)

2/ Déterminer les axes principaux d'inertie :

Recherche des valeurs propres λ_k et vecteurs propres u_k

$\mathrm{Det}(vM - \lambda I) = 0$

$VMU_k = \lambda_k u_k$

RQ : $\boldsymbol{TR(vM) = \sum \lambda_i}$

3/ Composantes principales :

$\boldsymbol{C_k = \langle x, u_k \rangle_M}$

$$C_k = \begin{pmatrix} C_k^1 \\ C_k^2 \\ . \\ . \\ . \\ C_k^n \end{pmatrix} \qquad C_k^i = x_i^t M u_k$$

Remarque :

$\boldsymbol{moy(C_k) = 0}$

$\boldsymbol{Var(C_k) = \lambda_k}$

$\boldsymbol{corr(C_i, C_j) = cov(C_i, C_j) = 0 \; \forall i, j}$

4/ Qualité de représentation :

λ_k *en ordre decroissant* $\lambda_1 > \lambda_2 > \cdots > \lambda_k$

$Q_1 = \dfrac{\lambda_1}{\sum \lambda_i} \geq 80\%$

Sinon $Q_2 = \dfrac{\lambda_1 + \lambda_2}{\sum \lambda_i} \geq 80\%$

5/ Contributions aux inerties :

a. **Part d'inertie de x_i prise en compte par $\boldsymbol{\Delta_k(u_k)}$**

$\|\hat{x}\|^2$ *la val de projection de* x_i Sur (Δ_k)

$$\boldsymbol{\cos^2 \theta_{ik} = \frac{\left(C_k^i\right)^2}{\|x_i\|_M^2}}$$

b. **Contribution relative d'un individu a l'inertie expliquée ($\boldsymbol{\lambda_k}$) de l'axe $\boldsymbol{\Delta_k(u_k)}$**

$\boldsymbol{cont_{ik} = \frac{P_i (C_k^i)^2}{\lambda_k}}$ A quel point un individu a servi à la création de l'axe

$$\lambda_k = \sum_{i=1}^{n} P_i \left(c_k^i\right)^2 = P_1 \left(C_k^1\right)^2 + P_2 \left(C_k^2\right)^2 + \cdots + P_n (C_k^n)^2$$

6/ Description des variables

Les composantes principales (C_k) forment une base.

La projection e la var (X^j) sur l'axe C_k est le coefficient de corrélation entre $x^j et\ C_k$

$$cor(x^j, C_k) = \frac{cov(x^j, C_k)}{\sigma_{x_j}\ \sigma_{C_k}} = \frac{\sum P_i x_{ij} C_{ik}}{\sqrt{\sum P_i x_{ij}^2} . \sqrt{\lambda_k}}$$

$$-1 \leq cor \leq 1$$

IV. Exemple pratique :

Supposons que vous ayez un jeu de données avec trois variables : **hauteur**, **poids** et **âge** pour un groupe d'individus. Si hauteur et poids sont fortement corrélés, l'ACP pourrait identifier une composante principale qui combine ces deux variables pour capturer leur contribution commune à la variance. La deuxième composante principale pourrait alors capturer la variance liée à l'âge, qui est moins corrélée aux deux autres variables.

- ✓ **Avantages et limites de l'ACP :**

Avantages :

- Réduction de la complexité des données.
- Identification des relations entre les variables.
- Aide à la visualisation de grands jeux de données.

Limites :

- Les résultats sont parfois difficiles à interpréter, surtout si les composantes principales sont des combinaisons complexes des variables d'origine.
- L'ACP est une méthode linéaire, elle peut ne pas bien capturer des relations non linéaires dans les données.

En résumé, l'ACP est un outil puissant pour réduire la dimension des données, en maximisant la quantité d'information conservée dans un nombre réduit de variables (les composantes principales).

Exemple d'ACP

Imaginons que nous avons un petit ensemble de données sur trois étudiants avec des notes dans trois matières : **Mathématiques**, **Physique** et **Informatique**. Voici les notes de chaque étudiant (sur 20) :

Étudiant	Mathématiques	Physique	Informatique
Étudiant 1	15	14	16
Étudiant 2	12	10	11
Étudiant 3	18	17	19

Nous souhaitons appliquer l'ACP pour comprendre la relation entre ces matières et réduire potentiellement le nombre de variables à analyser.

Étape 1 : Standardisation des données

Les notes des trois matières ont des échelles similaires, mais il est tout de même possible de les **standardiser (ACP normée)** (moyenne = 0, écart-type = 1). Cela permet de rendre les variables comparables. Pour standardiser les données, on soustrait la moyenne et on divise par l'écart-type pour chaque variable.

Étudiant	Mathématiques (standardisé)	Physique (standardisé)	Informatique (standardisé)
Étudiant 1	0.27	0.29	0.30
Étudiant 2	-1.09	-1.17	-1.16
Étudiant 3	0.82	0.88	0.87

Étape 2 : Calcul de la matrice de covariance

Nous calculons ensuite la **matrice de covariance** entre les variables (les matières). La covariance indique comment deux variables varient ensemble. Si la covariance est positive, cela signifie que les deux variables augmentent ou diminuent ensemble.

Voici la matrice de covariance pour nos données :

	Mathématiques	Physique	Informatique
Mathématiques	1.0	0.99	0.99
Physique	0.99	1.0	0.99
Informatique	0.99	0.99	1.0

On voit ici que les trois matières sont très corrélées entre elles (toutes les covariances sont proches de 1).

Étape 3 : Calcul des vecteurs propres et valeurs propres

Les **valeurs propres** et **vecteurs propres** sont ensuite calculés à partir de la matrice de covariance. Ces vecteurs propres représentent les directions des composantes principales, et les valeurs propres nous disent combien de variance chaque composante principale explique.

Voici les valeurs propres et les pourcentages de variance expliqués :

- **Composante principale 1 (CP1)** : Valeur propre = 2.98 (explique environ 99.3 % de la variance totale).
- **Composante principale 2 (CP2)** : Valeur propre = 0.02 (explique environ 0.6 % de la variance totale).
- **Composante principale 3 (CP3)** : Valeur propre ≈ 0.00 (explique environ 0.1 % de la variance totale).

Étape 4 : Interprétation des résultats

- **CP1** explique 99.3 % de la variance totale, ce qui signifie qu'elle capture presque toute l'information contenue dans les trois matières. Autrement dit, la majeure partie de la variation des notes des étudiants est résumée par une seule composante principale, ce qui est une combinaison linéaire des trois matières.
- **CP2** et **CP3** n'ajoutent presque rien en termes de variance expliquée (0.6 % et 0.1 % respectivement).

Étape 5 : Définition des composantes principales

La **première composante principale (CP1)** est une combinaison des trois matières. Cela signifie que CP1 est une sorte de **moyenne pondérée** des trois matières.

Étape 6 : Projection des données

Maintenant, on peut **projeter** les notes des étudiants sur la première composante principale (CP1). Comme CP1 capture presque toute la variance, nous n'avons plus besoin de trois variables pour analyser les données, mais seulement une.

Par exemple :

Étudiant	**Projection sur CP1**
Étudiant 1	15.26
Étudiant 2	10.51
Étudiant 3	18.74

Cela signifie que, même si nous avons trois matières initialement, une seule composante (CP1) capture presque toute l'information, ce qui simplifie l'analyse.

Étape 7 : Visualisation

En visualisant les étudiants sur l'axe de CP1, on pourrait voir à quel point ils se situent les uns par rapport aux autres en fonction de leur performance globale en mathématiques, physique et informatique. Il serait également possible de tracer un graphique avec CP1 et CP2, mais comme CP2 n'explique que 0.6 % de la variance, il n'ajouterait pas beaucoup d'information.

- ✓ Dans cet exemple, l'ACP a montré que les trois matières sont fortement corrélées, et qu'une seule composante principale suffit à expliquer presque toute la variance. Cela nous permettrait de simplifier notre analyse, en remplaçant les trois matières par une unique mesure synthétique qui résume les performances globales des étudiants.
- ✓ L'ACP est ainsi un outil très utile pour **réduire la dimensionnalité** des jeux de données tout en conservant un maximum d'information.