Réduction de Dimension

Analyse en Composantes Principales (ACP)

APEKE Séna / BABA Tchao

EPL

Juillet 2024

Plan

1. Introduction à la Réduction de Dimension
2. Principe de l'Analyse en Composantes Principales (PCA)
3. Rappels mathématiques
4. Les éléments de l'ACP
5. Aspect pratique de l'ACP
6. Notion de variables et d'individus suplémentaire
7. Exemple d'application de l'ACP

Introduction à la Réduction de Dimension

- **Définition :** La réduction de dimension consiste à transformer des données de haute dimension en une représentation de dimension inférieure tout en préservant autant que possible l'information essentielle.
- **Importance :** Elle permet d'éliminer la redondance, de simplifier l'analyse et d'améliorer les performances des algorithmes.
- **Problèmes liés à la haute dimensionnalité :** Surajustement (overfitting), difficulté de visualisation en dimension supérieure à 3.

Quelque méthodes réduction de dimension

Supervision	**Linéarité**	**Exemples**
Non supervisé	Linéaire	Analyse en Composantes Principales (PCA)
Supervisé	Linéaire	Analyse Discriminante Linéaire (LDA)
Non supervisé	Non linéaire	Isometric Mapping (ISOMAP)
Supervisé	Non linéaire	Autoencodeurs supervisés
Non supervisé	Non linéaire	Réseaux de neurones autoencodeurs

Table: Algorithmes de réduction de dimension

Principe de l'Analyse en Composantes Principales (PCA)

Définition de PCA

L'Analyse en Composantes Principales (PCA) est une méthode statistique permettant de transformer un jeu de données en identifiant les axes de plus grande variance. Ces axes sont appelés composantes principales, et sont obtenus par décomposition des données en directions orthogonales maximisant la variance.

Un petit conseil !

L'ACP est cruciale et stratégique pour un data analyst. Il est donc très important de bien la comprendre. Elle nécessite un minimum de pratique pour analyser les données correctement. Si la compréhension est approximative, elle mène alors très facilement à des analyses erronées, imprécises.

Objectifs de l'ACP

L'ACP a deux objectifs principaux. Retenez-les bien, nous y ferons référence tout au long de la partie. Elle permet d'étudier :

- la **variabilité entre les individus**, c'est-à-dire quelles sont les différences et les ressemblances entre individus
- les **liaisons entre les variables** : y a-t-il des groupes de variables très corrélées entre elles, qui peuvent être regroupées en de nouvelles variables synthétiques ?

L'enjeu de l'ACP

Pour analyser des données, il est souhaitable de les visualiser dans un repère. La représentation des données s'appelle **nuage de points**. C'est ce nuage que nous souhaitons étudier, et donc visualiser. Mais comme nous l'avons déjà évoqué, nous avons un petit souci pour visualiser ce nuage de points. En effet, il a souvent plus de 2 dimensions, alors que nos écrans sont en 2 dimensions. Un écran ou une feuille de papier sont des plans.

L'enjeu de l'ACP

Pour analyser des données, il est souhaitable de les visualiser dans un repère. La représentation des données s'appelle **nuage de points**. C'est ce nuage que nous souhaitons étudier, et donc visualiser. Mais comme nous l'avons déjà évoqué, nous avons un petit souci pour visualiser ce nuage de points. En effet, il a souvent plus de 2 dimensions, alors que nos écrans sont en 2 dimensions. Un écran ou une feuille de papier sont des plans.

Pour visualiser des points dans un espace à n dimensions (avec $n > 2$) sur un plan, la solution est d'effectuer une **projection orthogonale**. Le mot **projection** vous évoque peut-être une projection de cinéma, ou bien le fait de lancer un objet sur un mur, comme le font certains artistes quand ils projettent de la peinture sur une toile. Dans les 2 cas, il s'agit de la même chose : pour le cinéma, on projette une image (l'image d'un acteur).

Dans la réalité, cet acteur est à 3 dimensions (non, un être humain n'est jamais plat !), mais l'image projetée sur l'écran est en 2 dimensions, car l'écran de cinéma est un plan. De même, la toile du peintre est aussi plane, alors que les gouttes de peinture qui y sont projetées sont à peu près sphériques, donc en 3D. L'espace dans lequel nous vivons est à 3 dimensions, que l'on a coutume d'appeler **largeur, hauteur et profondeur.**

La projection mathématique, c'est la même chose : c'est représenter des points à p dimensions dans un espace plus petit , c'est-à-dire à q dimensions, avec q¡p

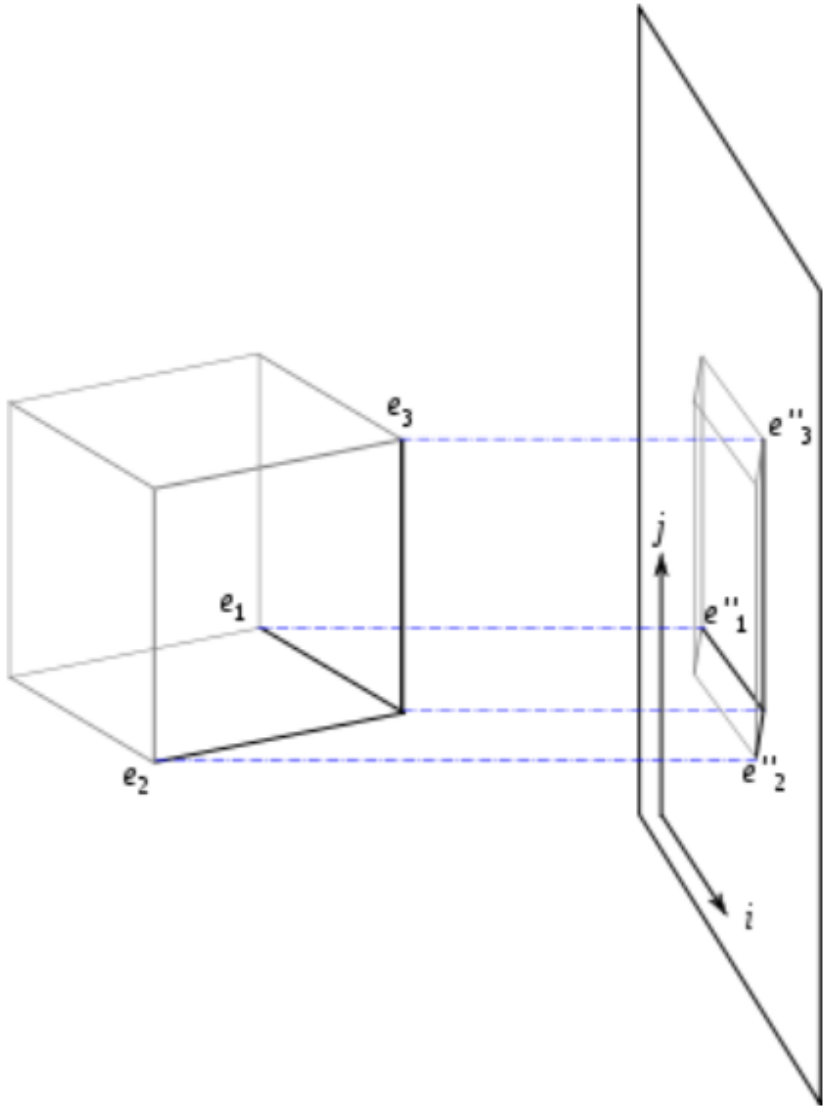


Figure: Exemple de la projection d'un cube d'un espace de dimension 3 dans un espace de dimension 2

Malheureusement, quand on projette des points, on perd de l'information.

Eh oui ! Reprenons l'exemple du cinéma. Imaginons que l'acteur soit face à la caméra et qu'il tienne dans ses mains une règle graduée en centimètres, face à la caméra. La règle est à la même distance de la caméra que l'acteur. Sur l'écran de cinéma, vous pouvez facilement savoir quelle distance sépare les 2 yeux de l'acteur grâce à la règle graduée. Vous pouvez également mesurer l'espacement entre ses 2 oreilles ou même la taille de l'acteur. Vous avez donc une bonne appréhension d'au moins 2 dimensions : la largeur et la hauteur.

Cependant, pour la 3e dimension (la profondeur), c'est plus compliqué. En effet, vous ne pouvez pas mesurer avec précision la distance entre le ventre de l'acteur et son dos, ni même la distance entre le bout de son nez et le reste de son visage. Pour mesurer ces 2 longueurs, il faut placer la règle dans le sens de la profondeur, et il vous est impossible d'appréhender avec précision cette profondeur sur un écran de cinéma. C'est pour cela que l'on dit qu'il y a une perte d'information : vous perdez l'information de la profondeur.

Perdre de l'information, c'est frustrant !

PS : Il faut essayer de perdre le moins d'information possible.

Cela tombe bien, car pour un même objet (l'acteur, par exemple), il y a plusieurs projections possibles : en fonction de là où se place la caméra. La caméra peut se placer au-dessus de l'acteur, ou de profil, de face, ou même en contre-plongée, un peu en biais, etc.

Mais avez-vous déjà vu un film dans lequel les acteurs sont filmés constamment du dessus, ou même du dessous ?

Vue de dessous

Vue de dessus

Non , me direz-vous, car on ne verrait rien ! Rien, ou en tout cas, pas grand-chose. Effectivement, on verrait moins bien que quand les acteurs sont filmés de face. Dire on ne voit pas grand-chose équivaut à dire on perd beaucoup d'information . Vu de dessus, on perd par exemple l'information de l'expression du visage : l'acteur est-il heureux ? Triste ? Vues de face, les expressions du visage sont bien plus visibles.

Il y aurait donc des projections qui seraient meilleures que d'autres ?

Tout à fait, il y a des projections pour lesquelles on voit moins bien que d'autres, avec lesquelles on perd plus d'information que d'autres.

Tout l'enjeu est donc de trouver une projection pour nos données qui perde le moins d'information possible.

Pourquoi voit-on mieux un acteur de face que de dessus ?

La principale raison est que la forme de l'être humain est allongée : notre hauteur est plus grande que notre largeur. Ainsi, l'image d'un acteur sera plus allongée s'il est filmé de face plutôt que d'en haut. Son image sera plus étalée à l'écran.

Une image étalée ! De quoi parle t-on ?

Ah Oui !! : c'est la notion d'inertie ! Des points très étalés ont une grande inertie.

C'est ici la clé de l'ACP : rechercher la projection pour laquelle l'inertie des points est maximale.

Principe de l'A.C.P.

On cherche une représentation des n individus , dans un sous-espace F_k de $\mathbb{R}^p$ de dimension k (k petit 2, 3 . . . ; par exemple un plan).

Autrement dit, on cherche à définir k nouvelles variables combinaisons linéaires des p variables initiales qui feront perdre le moins d'information possible.

Ces variables seront appelées **composantes principales** ,
les axes qu'elles déterminent : **axes principaux**
les formes linéaires associées : **facteurs principaux** .

Représentation statistique d'un jeu de données

Variable

On appelle variable un vecteur v de taille n. Chaque coordonnée v_i correspond à un individu. Chaque individu peut avoir un poids w_i, tel que $w_1 + w_2 + \cdots + w_n = 1$. On a souvent $w = 1/n$

Moyenne arithmétique

La moyenne arithmétique est une mésure de tendance centrale qui dépend de toutes les observations et est sensibles aux valeurs extrêmes. Elle est donnée par:

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i \qquad \text{ou} \qquad \bar{v}_{\text{pondérée}} = \frac{\sum_{i=1}^{n} w_i v_i}{\sum_{i=1}^{n} w_i}$$

Quartiles :

Les quartiles divisent un ensemble de données en quatre parties égales. On note :

- Q_1 : le premier quartile, qui correspond à la valeur en dessous de laquelle se trouvent 25% des données.
- Q_2 : le deuxième quartile, qui est la médiane et divise les données en deux parties égales (50% des données en dessous).
- Q_3 : le troisième quartile, qui correspond à la valeur en dessous de laquelle se trouvent 75% des données.

Médiane :

La médiane est la valeur qui sépare un ensemble de données en deux parties égales. Si les données sont ordonnées, la médiane est définie comme :

$$\text{Médiane} = \begin{cases} \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{si } n \text{ est pair} \\ x_{((n+1)/2)} & \text{si } n \text{ est impair} \end{cases}$$

Déciles :

Les déciles divisent un ensemble de données en dix parties égales. Le k-ième décile, noté D_k, correspond à la valeur en dessous de laquelle se trouvent $10k\%$ des données. Par exemple, le premier décile (D_1) correspond à la valeur en dessous de laquelle se trouvent 10% des données, tandis que le cinquième décile (D_5) est la médiane (50% des données).

Percentiles :

Les percentiles divisent un ensemble de données en 100 parties égales. Le p-ième percentile, noté P_p, est la valeur en dessous de laquelle se trouvent $p\%$ des données. Par exemple, le $30^{\text{ème}}$ percentile (P_{30}) est la valeur qui sépare les 30% inférieurs des données des 70% supérieurs.

Variance et écart-type

La variance est la moyenne des carrées moins le carré de la moyenne.
L'écart-type, qui a la même unité que v est une mesure de dispersion.
Variance non pondéré:

$$\text{Var}(v) = \frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v})^2$$

Variance pondéré:

$$\text{Var}_{\text{pondérée}}(v) = \frac{\sum_{i=1}^{n} w_i(v_i - \bar{v}_{\text{pondérée}})^2}{\sum_{i=1}^{n} w_i}$$

Ecart-type non pondéré:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v})^2}$$

Ecart-type pondéré:

$$\sigma_{\text{pondérée}} = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(v_i - \bar{v}_{\text{pondérée}}\right)^2}{\sum_{i=1}^{n} w_i}}$$

Remarque :
La variance satisfait également la formule suivante:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} v_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} v_i\right)^2$$

$$\sigma_w^2 = \frac{\sum_{i=1}^{n} w_i v_i^2}{\sum_{i=1}^{n} w_i} - \left(\frac{\sum_{i=1}^{n} w_i v_i}{\sum_{i=1}^{n} w_i}\right)^2$$

Représentation matricielle d'un jeu de données

Un jeu de données peut être représenté sous forme de matrice, où chaque ligne correspond à un individu ou une observation, et chaque colonne représente une variable ou un attribut de ces observations. Cette représentation facilite l'analyse statistique et le traitement des données. Soit un jeu de données $X \in \mathbb{R}^{n \times p}$, où :

- n est le nombre d'échantillons (observations),
- p est le nombre de variables (caractéristiques).

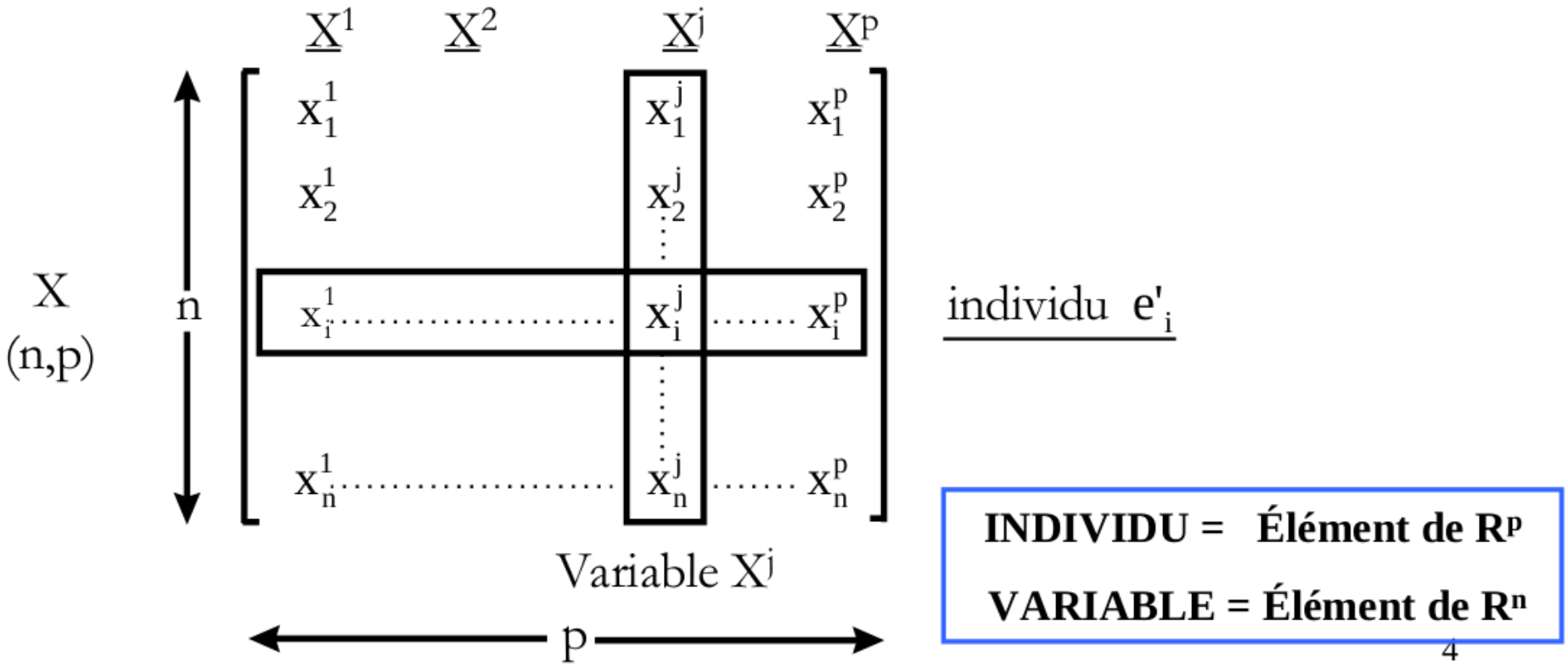


Figure: Représentation matricielle d'un jeu de données

Matrice de poids : On associe aux individus un poid w_i. Soit D_w la matrice de poids w_i telle que $w_1 + w_2 + \cdots + w_n = 1$ que l'on représente par la matrice diagonale de taille n

$$D_w = \begin{pmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & w_2 & 0 & \cdots & 0 \\ 0 & 0 & w_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_n \end{pmatrix}$$

Cas uniforme : Tous les individus ont le même poids $w_i = 1/n$ et $D_w = \frac{1}{n} I_n$

Point moyen et tableau centré :

Point moyen :

On appelle point moyen de la distribution de donées, le vecteur $\mathbf{g}$ des moyennes arithmétiques de chaque variable: $\mathbf{g}' = (\bar{\mathbf{x}}^{\mathbf{1}}, \ldots, \bar{\mathbf{x}}^{\mathbf{p}})$, où $\bar{\mathbf{x}}^{\mathbf{j}} = \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} \mathbf{w_i x_i^j}$

ici $\mathbf{g}'$ désigne la transposé du vecteur $\mathbf{g}$ des moyennes arithmétiques.

Sous forme matricielle on a $\mathbf{g} = \mathbf{X}'\mathbf{D_w}\mathbf{1_n}$; où $\mathbf{X}'$ désigne la transposé de la matrice de données $\mathbf{X}$; $\mathbf{D}_w$ est la matrice des poids et $\mathbf{1}_n$ est une matrices colonne remplie de 1

Tableau centré :

Le tableau centré est obtenu en centrant les variables autour de leur moyenne $\mathbf{y_i^j} = \mathbf{x_i^j} - \bar{\mathbf{x}}^{\mathbf{j}}$ en notation matricielle on a :

$\mathbf{Y} = \mathbf{X} - \mathbf{1_n g}' = (\mathbf{I_n} - \mathbf{1_n 1}'_{\mathbf{n}}\mathbf{D_w})\mathbf{X}$

Données réduites

La matrice de données réduites est obtenue après avoir centré puis réduit les variables, c'est-à-dire en soustrayant la moyenne de chaque variable (centrage) et en divisant chaque variable par son écart-type (réduction). Cela permet de mettre toutes les variables sur une échelle comparable, avec une moyenne nulle et une variance égale à 1.

$$z_i^j = \frac{x_i^j - \bar{x}^j}{\sigma_j}$$

où $\bar{x}^j$ est la moyenne de la j-ème variable et σ_j son écart-type.

Sous forme matricielle, les données réduites peuvent être obtenues en combinant la matrice centrée et une matrice de normalisation, avec les inverses des écarts-types sur la diagonale. Si $\mathbf{Y}$ est la matrice de données centrées, et $\mathbf{D}_{\frac{1}{\sigma}}$ la matrice diagonale des inverses des écarts-types, on peut écrire la transformation comme suit :

$$\mathbf{Z} = \mathbf{Y}\mathbf{D}_{\frac{1}{\sigma}}$$

où :

- $\mathbf{Z}$ est la matrice des données réduites,
- $\mathbf{Y}$ est la matrice des données centrées,
- $\mathbf{D}_{\frac{1}{\sigma}}$ est une matrice diagonale contenant les inverses des écarts-types des variables sur sa diagonale.

Cette transformation garantit que chaque colonne de $\mathbf{Z}$ a une moyenne nulle et une variance égale à 1, permettant une comparaison équitable entre les variables, même si elles sont mesurées sur des échelles différentes.

NB:

$$D_{\frac{1}{\sigma}} = D_{\sigma}^{-1}$$

Ci-dessous est illustrée la variation de la structure des données dans le plan, en fonction de leur centrage et de leur réduction.

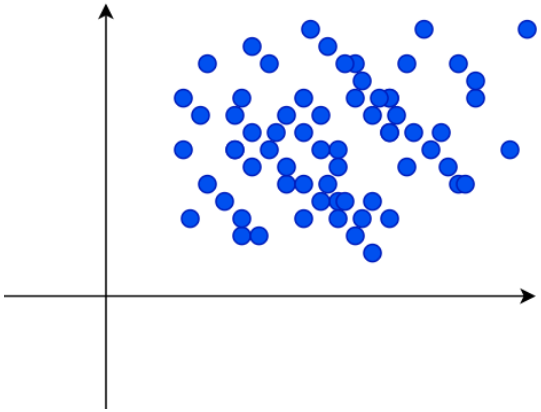


Figure: Données brute

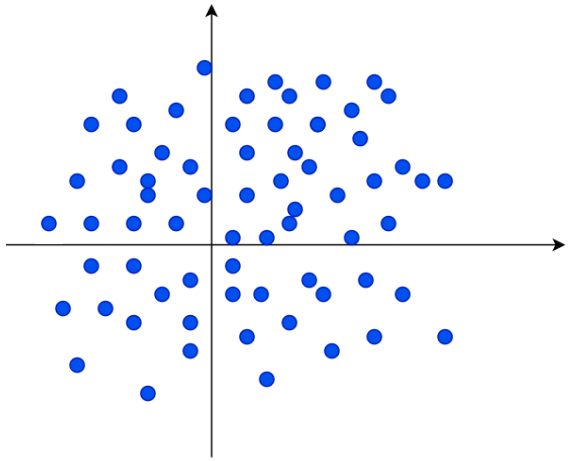


Figure: Données centrées

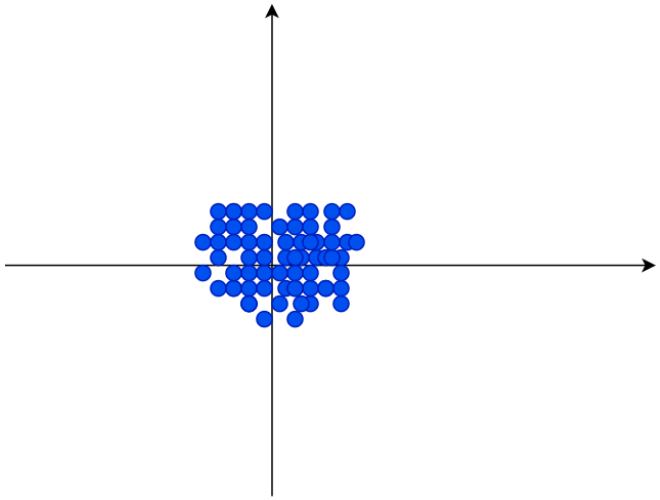


Figure: Données après centrage et réduction

La matrice de covariance

La covariance observée entre deux variables a et b est:

$$\text{Cov}(a, b) = \frac{1}{n} \sum_{i=1}^{n} (a_i - \bar{a})(b_i - \bar{b})$$

La matrice de covariance Σ est une matrice carrée de dimension p qui capture les relations linéaires entre les variables après centrage.

$$\Sigma = \begin{pmatrix} \sigma^2(a_1) & \sigma(a_1, a_2) & \dots & \sigma(a_1, a_p) \\ \sigma(a_2, a_1) & \sigma^2(a_2) & \dots & \sigma(a_2, a_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(a_p, a_1) & \sigma(a_p, a_2) & \dots & \sigma^2(a_p) \end{pmatrix}$$

où $\sigma^2(a_i)$ représente la variance de la variable a_i et $\sigma(a_i, a_j)$ représente la covariance entre les variables a_i et a_j.
La formule matricielle de la matrice de covariance est:

$$\mathbf{\Sigma} = \mathbf{X}'\mathbf{D_w}\mathbf{X} - \mathbf{g}\mathbf{g}' = \mathbf{Y}'\mathbf{D_w}\mathbf{Y}$$

Propriétés de la matrice de covariance

- Σ est une matrice symétrique $p \times p$.
- Les éléments diagonaux de Σ représentent la variance de chaque variable.
- Les éléments hors diagonaux de Σ représentent les covariances entre les différentes variables.

Matrice de correlation

Le coefficient r de Bravais-Pearson ou coefficient de correlation est donné par:

$$r_{a,b} = \frac{\text{Cov}(a,b)}{\sigma_a \sigma_b}$$

- $\text{Cov}_{a,a} = var(a)$ et $r_{a,a} = 1$
- $\text{Cov}_{a,b} = \text{Cov}_{b,a}$ et donc $r_{a,b} = r_{b,a}$

La matrice de correlation R est donnée par:

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

où r_{ij} représente le coefficient de corrélation entre la variable i et la variable j.

Propriétés du coefficient de correlation

- On a toujours l'inégalité $-1 \leq r_{v,w} \leq 1$ dite inégalité de Cauchy-Schwarz.
- $|r_{v,w}| = 1$ si et seulement si v et w sont linéairement liés.
- Si $r_{v,w} = 0$, on dit que les variables sont décorrélées. Cela ne veut pas dire qu'elles sont indépendantes.

Une corrélation linéaire mesure la force et la direction d'une relation linéaire entre deux variables, tandis que la dépendance indique simplement qu'un changement dans une variable est associé à un changement dans l'autre, sans nécessairement impliquer une relation linéaire.

La formule matricielle est:

$$\mathbf{R} = \mathbf{D}_{\frac{1}{\sigma}} \mathbf{\Sigma} \mathbf{D}_{\frac{1}{\sigma}}$$

où

$$\mathbf{D}_{\frac{1}{\sigma}} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ & \frac{1}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_p} \end{pmatrix}$$

Notion de métrique dans un espace vectoriel de dimension p

Afin de pouvoir considérer la structure du nuage des individus, on définit une distance, qui induira une géométrie.

Distance euclidienne classique

La distance la plus simple entre deux points $A = (a_1, a_2, \ldots, a_p)$ et $B = (b_1, b_2, \ldots, b_p)$ dans $\mathbb{R}^p$ est donnée par:

$$d(A, B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_p - b_p)^2}$$

$$= \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}$$

Notion de métrique dans un espace vectoriel de dimension p

Afin de pouvoir considérer la structure du nuage des individus, on définit une distance, qui induira une géométrie.

Distance euclidienne classique

La distance la plus simple entre deux points $A = (a_1, a_2, \ldots, a_p)$ et $B = (b_1, b_2, \ldots, b_p)$ dans $\mathbb{R}^p$ est donnée par:

$$d(A, B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_p - b_p)^2}$$

$$= \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}$$

Généralisation simple de la distance euclidienne

Si on donne un poids $w_i > 0$ à la variable i on a:

$$d(A, B) = \sqrt{w_1(a_1 - b_1)^2 + w_2(a_2 - b_2)^2 + \cdots + w_p(a_p - b_p)^2}$$

$$= \sqrt{\sum_{i=1}^{p} w_i (a_i - b_i)^2}$$

Métrique

Une métrique dans un espace vectoriel de dimension p, est une généralisation de la notion de distance qui peut être influencée par des poids ou des relations spécifiques entre les différentes dimensions. Cela peut se faire à travers un produit scalaire défini par une matrice symétrique définie positive.

Formellement, considérons une métrique définie par une matrice symétrique $M \in \mathbb{R}^{p \times p}$, où M est une matrice définie positive (i.e. $x^T M x > 0$ pour tout vecteur non nul $x \in \mathbb{R}^p$).

La distance entre deux points $A = (a_1, a_2, \ldots, a_p)$ et $B = (b_1, b_2, \ldots, b_p)$ dans $\mathbf{R}^p$ est donnée par :

$$d_M(A, B) = \sqrt{(A - B)^T M (A - B)}$$

où $(A - B)$ est le vecteur de différence entre A et B, et M est la matrice qui définit la métrique.

Interprétation avec les produits scalaires

Dans ce cadre, on peut interpréter la transformation de l'espace à l'aide du produit scalaire. Le produit scalaire entre deux vecteurs **A** et **B** dans un espace euclidien est défini par :

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i=1}^{p} A_i B_i$$

Quelques propriétés importantes du produit scalaire sont :

- **Linéarité** : $\langle \alpha \mathbf{A} + \beta \mathbf{B}, \mathbf{C} \rangle = \alpha \langle \mathbf{A}, \mathbf{C} \rangle + \beta \langle \mathbf{B}, \mathbf{C} \rangle$
- **Symétrie** : $\langle \mathbf{A}, \mathbf{B} \rangle = \langle \mathbf{B}, \mathbf{A} \rangle$
- **Positivité** : $\langle \mathbf{A}, \mathbf{A} \rangle \geq 0$ et $\langle \mathbf{A}, \mathbf{A} \rangle = 0$ si et seulement si $\mathbf{A} = 0$

La norme d'un vecteur **A** est liée au produit scalaire et est donnée par :

$$\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle}$$

Dans ce contexte, la matrice $\mathbf{M}$ modifie la structure de l'espace en "déformant" les distances selon les directions associées à ses valeurs propres. Si $\mathbf{M}$ est la matrice identité, on retrouve la distance euclidienne classique entre deux points $\mathbf{A}$ et $\mathbf{B}$:

$$\mathbf{d(A,B)} = \sqrt{(\mathbf{A}-\mathbf{B})^{\mathbf{T}}(\mathbf{A}-\mathbf{B})} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{p}}(\mathbf{a_i}-\mathbf{b_i})^2}$$

Cependant, si $\mathbf{M}$ est une matrice diagonale $\mathbf{D_w}$ avec des coefficients différents, cela signifie que certaines directions dans l'espace sont "plus importantes" que d'autres, et la distance devient "pondérée" par les coefficients de $\mathbf{D_w}$. Ainsi, la nouvelle distance devient :

$$\mathbf{d_M(A,B)} = \sqrt{(\mathbf{A}-\mathbf{B})^{\mathbf{T}}\mathbf{D_w}(\mathbf{A}-\mathbf{B})}$$

où les poids dans $\mathbf{D_w}$ influencent la contribution de chaque direction à la distance globale.

Existence de la M-Orthogonalité

Deux points **A** et **B** sont dits M-**orthogonaux** si leur produit scalaire selon la métrique **M** est nul :

$$(\mathbf{A} - \mathbf{G})^{\mathsf{T}} \mathbf{M} (\mathbf{B} - \mathbf{G}) = \mathbf{0}$$

où **G** est le centre de gravité du nuage de points dans un espace vectoriel de dimension **p**.

Remarque :
Utiliser une métrique revient à "tordre" les données par exemple pour les rendre comparable.

Métrique Réduite :
La notion de métrique réduite est une généralisation de la métrique classique appliquée à des données qui ont été prétraitées par réduction. Rappelons que pour effectuer cette réduction, nous avons multiplier la matrice de données par une matrice de normalisation $\mathbf{D}_{\frac{1}{\sigma}}$, où $\mathbf{D}_{\frac{1}{\sigma}}$ est une matrice diagonale dont les éléments sur la diagonale sont les inverses des écarts-types des variables.
La matrice $\mathbf{D}_{\frac{1}{\sigma}}$ est de la forme suivante :

$$\mathbf{D}_{\frac{1}{\sigma}} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_p} \end{pmatrix}$$

Le fait de diviser les variables par leurs écarts-types σ_i peut être vu comme une opération de transformation matricielle qui revient à utiliser des poids $\mathbf{w_i} = \sigma_i^2$, ce qui nous conduit à la métrique suivante :

$$\mathbf{D}_{\frac{1}{\sigma^2}} = \mathbf{D}_{\frac{1}{\sigma}} \times \mathbf{D}_{\frac{1}{\sigma}}$$

où $\mathbf{D}_{\frac{1}{\sigma^2}}$ est une matrice diagonale contenant les inverses des variances des variables $\frac{1}{\sigma_i^2}$.

Interprétation :

Travailler avec la métrique $\mathbf{D}_{\frac{1}{\sigma^2}}$ revient à diviser chaque variable par son écart-type puis à utiliser la métrique classique (métrique Euclidienne).

Autrement dit, appliquer la métrique $\mathbf{D}_{\frac{1}{\sigma^2}}$ revient à normaliser les données puis à appliquer la métrique classique $\mathbf{I}$ (la matrice identité).

Cela permet d'analyser les données dans un espace où chaque variable a un impact équivalent, car les variables sont sur des échelles normalisées et comparables. Cela est particulièrement utile dans des analyses ACP, où la variance de chaque variable joue un rôle crucial.

Notions d'Inertie

L'inertie en un point **A** d'un nuage de points, lorsqu'on utilise une métrique définie par une matrice **M**, est une mesure de la dispersion ou de la variation des données autour de ce point. Contrairement au cas euclidien classique, cette inertie tient compte de la déformation introduite par la matrice **M**, qui peut pondérer ou influencer les différentes directions de l'espace.

Soit $\mathbf{E_i}$ un ensemble de points dans un espace vectoriel de dimension $\mathbf{p}$, avec un point **A** de coordonnées $\mathbf{A} = (\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_p})$ de ce nuage de points, et une matrice métrique $\mathbf{M} \in \mathbb{R}^{\mathbf{p}\times\mathbf{p}}$ donnée. L'inertie autour du point **A** est alors définie comme la distance quadratique entre **A** et tous les points $\mathbf{E_i}$ selon la métrique **M** :

$$\mathbf{I(A)} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i} \|\mathbf{E_i} - \mathbf{A}\|^2 = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i} (\mathbf{E_i} - \mathbf{A})' \mathbf{M} (\mathbf{E_i} - \mathbf{A})$$

où **A** est le vecteur des coordonnées du point **A**, $\mathbf{E_i}$ représente les autres points du nuage, $\mathbf{w_i}$ est un poids associé au point $\mathbf{E_i}$, et **M** est une matrice symétrique définie positive.
Si **M** est la matrice identité, cette inertie revient à la norme euclidienne classique du vecteur **A** :

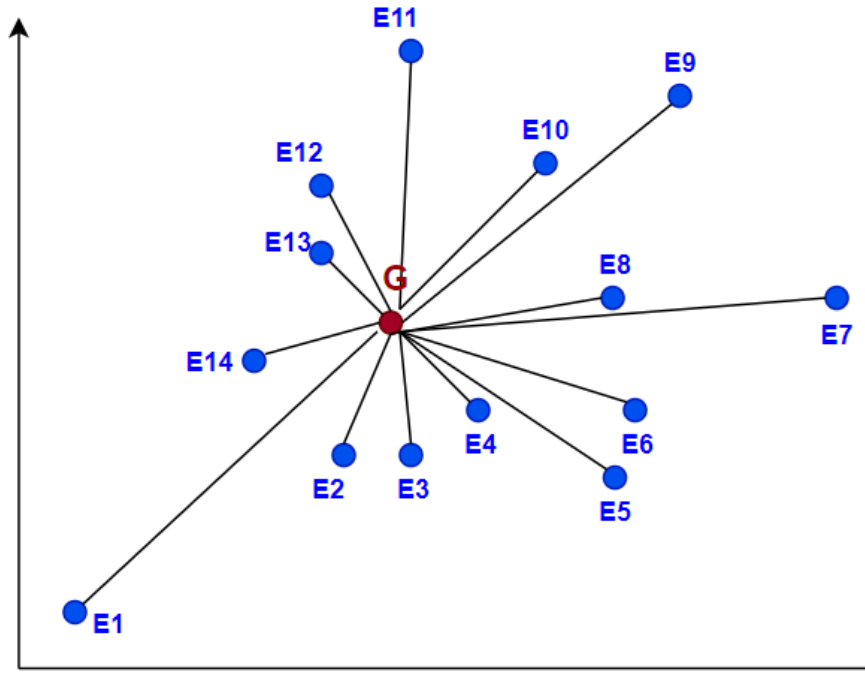
$$\mathbf{I(A)} = \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{a_j^2} = \|\mathbf{A}\|^2$$

Inertie totale autour du barycentre du nuage de points :
L'inertie totale autour du barycentre **G** d'un nuage de points est une mesure de la dispersion globale des points par rapport au barycentre **G**. Elle est donnée par la somme des distances quadratiques pondérées entre chaque point $\mathbf{E_i}$ et le barycentre **G**.
Formellement, si $\mathbf{E_1}, \mathbf{E_2}, \ldots, \mathbf{E_n}$ sont les points du nuage dans un espace de dimension **p**, alors l'inertie totale est définie par :

$$I_G = \sum_{i=1}^{n} w_i \|E_i - G\|^2$$

où :

- E_i est le vecteur des coordonnées du i-ème point,
- G est le barycentre, défini par $G = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i E_i$,
- w_i est un poids associé au point E_i,
- $\|E_i - G\|^2$ est la distance quadratique (ou norme euclidienne au carré) entre le point E_i et le barycentre G.



L'inertie totale autour du barycentre G I_G est la plus petite inertie possible d'un nuage de point. Elle mesure l'étalement du nuage de points

Autre relation

L'inertie en un point **A** d'un nuage de points peut être exprimée en fonction de l'inertie totale par rapport au barycentre **G** selon la relation suivante :

$$\mathbf{I(A)} = \mathbf{I(G)} + \mathbf{n} \cdot \|\mathbf{G} - \mathbf{A}\|^2$$

où :

- $\mathbf{I(A)}$ est l'inertie en **A**,
- $\mathbf{I(G)}$ est l'inertie totale par rapport au barycentre **G**,
- **n** est le nombre de points dans le nuage,
- $\|\mathbf{G} - \mathbf{A}\|^2$ est la distance quadratique entre le barycentre **G** et le point **A**.

Cette relation montre que l'inertie totale au point **A** se décompose en deux termes :

1. L'inertie par rapport au barycentre **G**, qui mesure la dispersion intrinsèque des points autour du barycentre.
2. Un terme supplémentaire qui est proportionnel à la distance entre le point **A** et le barycentre **G**, pondéré par **n**, le nombre de points du nuage.

$\mathbf{I_G}$ mésure également la moyenne des carées des distances entre les individus.

$$\mathbf{I_G} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \|E_i - E_j\|^2$$

le facteur $\frac{1}{2}$ est dû au fait que les distances entre les points sont comptées deux fois dans la double somme.

Métriques particulières :

Forme matricielle :

En termes de la matrice de covariance $\mathbf{\Sigma}$, l'inertie totale autour de $\mathbf{G}$ peut également être exprimée comme :

$$\mathbf{I_G} = \text{tr}(\mathbf{\Sigma M})$$

où tr désigne la trace de la matrice. Cette formule relie l'inertie totale à la covariance des points et à la métrique utilisée pour mesurer les distances.

Métriques classique :

$\mathbf{M} = \mathbf{I_p}$ correspond au produit scalaire usuel

$$\mathbf{I_G} = \text{tr}(\mathbf{\Sigma}) = \sum_{\mathbf{i=1}}^{\mathbf{p}} \sigma^2$$

Métriques réduite :

Cette métrique est obtenue lorsqu'on pose $\mathbf{M} = \mathbf{D}_{\frac{1}{\sigma^2}} = \mathbf{D}^2_{\frac{1}{\sigma}}$ On a donc

$$\mathbf{I_G} = \text{tr}(\mathbf{D}_{\frac{1}{\sigma^2}}\mathbf{\Sigma}) = \text{tr}(\mathbf{D}_{\frac{1}{\sigma}}\mathbf{\Sigma}\mathbf{D}_{\frac{1}{\sigma}}) = \text{tr}(\mathbf{R}) = \mathbf{p}$$

p étant le nombre total de variables dans le jeu de données.

Principe

Nous cherchons à projeter orthogonalement le nuage de points sur un espace $\mathbb{E}^k$ de dimension $\mathbf{k} < \mathbf{p}$, sous la forme :

$$\mathbf{E_i^*} - \mathbf{G} = \mathbf{c_{i,1}a_1} + \mathbf{c_{i,2}a_2} + \cdots + \mathbf{c_{i,k}a_k}$$

où les vecteurs $\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_k}$ définissent l'espace $\mathbb{E}^k$ et les $\mathbf{c_{i,l}}$ sont les coordonnées de $\mathbf{E_i^*}$.

Critère

Nous souhaitons minimiser la moyenne des carrés des distances entre les points $\mathbf{E_i}$ et leurs projections $\mathbf{E_i^*}$. Selon le théorème de Pythagore, nous avons toujours :

$$\|\mathbf{E_i} - \mathbf{G}\|^2 = \|\mathbf{E_i} - \mathbf{E_i^*}\|^2 + \|\mathbf{E_i^*} - \mathbf{G}\|^2$$

Cela revient donc à maximiser l'inertie du nuage projeté.

Nous cherchons ainsi $\mathbb{E}^k$, un sous-espace de dimension **k** de $\mathbb{E}^p$ tel que l'inertie du nuage projeté sur $\mathbb{E}^k$ soit maximale.

Théorème principal:
Le sous espace $\mathbb{E}^k$ de dimension k portant l'inertie maximale est engendré par les k vecteurs propres de ΣM associés aux k plus grandes valeurs propres.

Rappel sur quelques éléménts des matrices

Valeur propre et vecteur propre

Une **valeur propre** λ et un **vecteur propre u** de taille **p** associés à une matrice carrée **X** de taille $\mathbf{p} \times \mathbf{p}$ sont définis comme suit :
Un vecteur propre $\mathbf{u} \in \mathbb{R}^{\mathbf{p}}$ est un vecteur non nul tel que l'application de la matrice **X** à **u** redimensionne le vecteur sans changer sa direction. Cela est représenté par l'équation :

$$\mathbf{Xu} = \lambda\mathbf{u}$$

La PCA transforme les variables d'origine en nouvelles variables non corrélées (composantes principales) à l'aide de vecteurs propres $\mathbf{u_i}$ et de valeurs propres $\lambda_\mathbf{i}$. Les vecteurs propres représentent les directions de ces nouvelles variables, et les valeurs propres représentent la variance expliquée par chaque composante.

Problème aux valeurs propres

Décomposition de la matrice de covariance

Le cœur de la PCA repose sur la décomposition en valeurs propres de la matrice de covariance. Nous cherchons les vecteurs propres $\mathbf{u_i}$ et les valeurs propres $\lambda_\mathbf{i}$ qui satisfont l'équation :

$$\mathbf{\Sigma u_i} = \lambda_\mathbf{i} \mathbf{u_i}, \quad \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{p}$$

- $\mathbf{u_i}$ représente la direction des composantes principales (vecteurs propres).
- $\lambda_\mathbf{i}$ représente la quantité de variance expliquée par $\mathbf{u_i}$ (valeurs propres).

Maximisation de la variance La première composante principale $\mathbf{u_1}$ est celle qui maximise la variance des données projetées :

$$\mathbf{u_1} = \arg\max_{\mathbf{u}} \mathrm{Var}(\mathbf{Xu})$$

Matrice diagonalisable

Une matrice est dite **diagonalisable** si elle peut être exprimée sous la forme :

$$\mathbf{X} = \mathbf{PDP}^{-1}$$

où $\mathbf{P}$ est une matrice composée des vecteurs propres de $\mathbf{X}$, et $\mathbf{D}$ est une matrice diagonale contenant les valeurs propres de $\mathbf{X}$.

Une matrice $\mathbf{X}$ de,taille $\mathbf{p} \times \mathbf{p}$ qui a $\mathbf{p}$ valeurs propres distinctes est diagonalisable et on a:

$$\mathrm{Tr}(\mathbf{X}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \lambda_{\mathbf{i}}$$

Rappel sur quelques matrices diagonalisables

- **Matrice symétrique :** Une matrice symétrique réelle ($\mathbf{X}' = \mathbf{X}$) possède une base de vecteurs propres orthogonaux réèls et ses valeurs propres sont elle aussi réelles.
- **Matrice M-symétrique :** Une matrice M-symétrique réelle ($\mathbf{X}'\mathbf{M} = \mathbf{M}\mathbf{X}$) possède une base de vecteurs propres M-orthogonaux réèls et ses valeurs propres sont elles mêmes réelles.
- **Matrice définie positive :** Une matrice définie positive est une matrice symétrique dont les valeurs propres sont strictements positives.

Analyse de la matrice $\mathbf{\Sigma M}$

- **Valeurs propres :** La matrice $\mathbf{\Sigma M}$ est M-**symétrique**; elle est donc diagonalisable est ses valeurs propres $\lambda_1, \ldots, \lambda_p$ sont réelles.
- **Axes principaux d'inertie :** On appelle axe principaux d'inertie, les p vecteurs $a_1, a_2, \ldots, a_p$ tels que $\mathbf{\Sigma M a_k} = \lambda_k \mathbf{a_k}$ les a_k sont M-**orthogonaux** c'est à dire $\langle \mathbf{a_i}, \mathbf{a_j} \rangle_{\mathbf{M}} = \mathbf{1}$ si i = j, $\mathbf{0}$ sinon
- **Signe des valeurs propres :** Les valeurs propres de $\mathbf{\Sigma M}$ sont positives et on peut les classer par ordre décroissant:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

Axes principaux d'inertie

Dans le cadre de l'Analyse en Composantes Principales (ACP), uns axe principal d'inertie($\mathbf{a_k}$) est une direction définie par un vecteur propre de la matrice de covariance (ou de corrélation) des données, le long de laquelle la dispersion (ou inertie) des données est maximale.

Autrement dit, les $\mathbf{a_k}$ sont les vecteurs propres de la matrice de covariance (ou de corrélation). Ils correspondent aux directions dans l'espace des variables d'origine, le long desquelles la variance des données est maximale. Ces vecteurs définissent les axes des nouvelles variables (composantes principales).
Le vecteur $\mathbf{a_1}$ associé à la plus grande valeur propre λ_1 est celui qui maximise la variance des données projetées. Le vecteur $\mathbf{a_2}$, associé à λ_2, est orthogonal à $\mathbf{a_1}$ et maximise la variance restante, et ainsi de suite.

- Les Vecteurs propres représentent les directions des axes principaux d'inertie. Ce sont les directions le long desquelles la dispersion (inertie) des points de données est la plus grande.
- Les Valeurs propres indiquent la quantité de variance expliquée par chaque axe principal. Plus la valeur propre est grande, plus l'axe associé capture une part importante de la variance des données.

En ACP, l'axe principal d'inertie associé à la plus grande valeur propre capture la plus grande partie de la variance, suivi par les axes successifs, qui sont orthogonaux entre eux et qui capturent des parts décroissantes de la variance. Ces axes permettent ainsi de réduire la dimension des données tout en conservant l'information la plus importante.

Composantes principales

Une composante principale ($\mathbf{c_k}$), également appelée axe factoriel, est une nouvelle variable obtenue par une combinaison linéaire des variables d'origine. Elle représente la projection des données sur l'axe principal d'inertie, indiquant ainsi la position des points dans cette nouvelle direction.

Chaque composante principale est donc associée à un axe principal d'inertie, et la première composante principale est la projection sur l'axe qui capture la plus grande variance des données.

De façon formelle, les composantes principales sont les nouvelles variables $\mathbf{c_k} = (\mathbf{c_{1k}}, \mathbf{c_{2k}}, \ldots, \mathbf{c_{pk}})$ de taille $\mathbf{p}$ obtenues après la transformation des données. Elles sont définies comme suit:

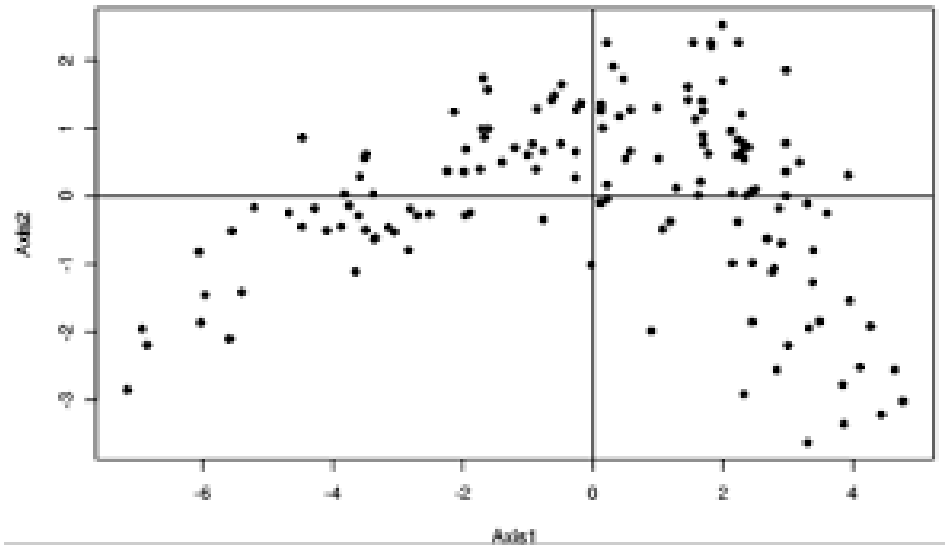
$$\mathbf{c_k} = \mathbf{YMa_k}$$

Chaque $\mathbf{c_k}$ contient les coordonnées des projections M-Orthogonales des individus centrés sur l'axe défini par les $\mathbf{a_k}$(axes principaux d'inertie)

Représentation des individus dans un plan principal

Elle permet de visualiser les données projetées dans un espace de dimension réduite, souvent à deux ou trois dimensions. Les individus sont projetés sur ce plan, défini par les deux premières composantes principales (premier plan factoriel), qui capturent la plus grande partie de la variance totale des données.

Chaque point représente un individu, et la proximité entre deux points reflète leur similarité dans l'espace des variables originales. Cette représentation permet ainsi de dégager des groupes d'individus similaires et d'identifier des tendances ou des structures sous-jacentes dans les données.



Coordonnées des individus dans l'espace des composantes principales

Les coordonnées des individus dans l'espace des composantes principales sont obtenues en projetant les observations sur les axes des composantes principales, qui maximisent la variance des données. En d'autres termes, ces coordonnées représentent les nouvelles positions des observations dans l'espace des composantes.

En supposant que $\mathbf{E_i} - \mathbf{G} = \sum_{\mathbf{l=1}}^{\mathbf{p}} \mathbf{c_{i,l}} \mathbf{a_l}$, alors

$$\langle \mathbf{E_i} - \mathbf{G}, \mathbf{a_k} \rangle_{\mathbf{M}} = \sum_{\mathbf{l=1}}^{\mathbf{p}} \mathbf{c_{i,l}} \langle \mathbf{a_l}, \mathbf{a_k} \rangle_{\mathbf{M}} = \mathbf{c_{i,k}}$$

La coordonnée de l'individu centré $\mathbf{E_i} - \mathbf{G}$ sur l'axe principal $\mathbf{a_k}$ est donnée par la projection M-Orthogonale :

$$\mathbf{c_{i,k}} = \langle \mathbf{E_i} - \mathbf{G}, \mathbf{a_k} \rangle_{\mathbf{M}} = (\mathbf{E_i} - \mathbf{G})' \mathbf{M} \mathbf{a_k}$$

Exemple : Supposons que nous ayons un ensemble de données centré avec trois individus $\mathbf{E_1}, \mathbf{E_2}, \mathbf{E_3}$ et deux composantes principales $\mathbf{a_1}, \mathbf{a_2}$, ainsi qu'une matrice métrique M. Si la projection de $\mathbf{E_1} - \mathbf{G}$ sur $\mathbf{a_1}$ donne la coordonnée $c_{1,1}$ et sur $\mathbf{a_2}$ donne $c_{1,2}$, alors les coordonnées de $\mathbf{E_1}$ dans l'espace des composantes principales sont $(c_{1,1}, c_{1,2})$. Par exemple, si $(E_1 - G)'Ma_1 = 2.5$ et $(E_1 - G)'Ma_2 = 1.2$, les nouvelles coordonnées de $\mathbf{E_1}$ seront $(2.5, 1.2)$ dans cet espace.

Propriétés des Composantes Principales

1. Moyenne des Composantes Principales

Les composantes principales sont centrées:

$$\bar{c_k} = \frac{1}{n} \sum_{i=1}^{n} w_i c_{i,k} = \frac{1}{n} 1_n{}' D_w C_k$$

la moyenne des composantes principales c_k est calculée en prenant la somme pondérée des valeurs des composantes principales $c_{i,k}$ de chaque individu, où chaque individu est pondéré par w_i. Le terme $1_n{}'$ correspond à un vecteur de dimension n contenant uniquement des 1, utilisé pour sommer les contributions de chaque individu. Cette somme est ensuite divisée par n, le nombre total d'individus, pour obtenir la moyenne pondérée.

Cela permet de refléter l'importance relative de chaque individu dans la moyenne, en fonction des poids assignés D_w.

2. Variance des Composantes Principales

La variance de $\mathbf{c_k}$ est $\lambda_\mathbf{k}$

Preuve

$$\text{Var}(\mathbf{c_k}) = \mathbf{c_k'D_wc_k} = \mathbf{a_k'MY'D_wYMa_k} = \mathbf{a_k'M\Sigma Ma_k} = \lambda_\mathbf{k}\mathbf{a_k'Ma_k} = \lambda_\mathbf{k}$$

Par conséquent, on a toujours $\lambda_\mathbf{k} \geq \mathbf{0}$

3. Covariance entre Composantes Principales

Pour $k \neq l$

$$\text{cov}(\mathbf{c_k}, \mathbf{c_l}) = \mathbf{c_k'D_wc_l} = \cdots = \lambda_\mathbf{l}\mathbf{a_k'Ma_l} = \mathbf{0}$$

Interprétation

Les résultats ci-dessus nous indiquent que :

- La moyenne des composantes principales est nulle, ce qui signifie qu'elles sont centrées autour de l'origine.
- La variance des composantes principales est égale aux valeurs propres associées, ce qui reflète la quantité de variance expliquée par chaque composante.
- La covariance entre les composantes principales est nulle, indiquant qu'elles sont orthogonales et non corrélées.
- **Réduction de dimension** : Les premières composantes principales capturent la majeure partie de la variance des données, permettant ainsi de réduire la dimensionnalité tout en préservant une grande partie de l'information. Typiquement, seules les premières composantes principales sont retenues pour l'analyse, en négligeant celles qui capturent peu de variance.

- **Interprétation dans l'espace des variables** : Chaque composante principale peut être interprétée comme une direction dans l'espace des variables initiales, avec des coefficients (poids) qui indiquent l'importance relative de chaque variable dans la construction de cette composante. Ces coefficients sont donnés par les vecteurs propres associés aux valeurs propres de la matrice de covariance ou de corrélation.
- **Sommation de la variance** : La somme des variances expliquées par les composantes principales est égale à la variance totale des variables initiales. Cela garantit que toute l'information de la matrice de covariance est répartie entre les composantes.

Ainsi, dans l'**espace des variables**, les composantes principales permettent de comprendre quelles variables contribuent le plus à la variation observée et comment elles sont liées entre elles à travers ces nouvelles directions orthogonales.

Facteurs principaux et formules de reconstitution

Les facteurs principaux sont les coordonnées des individus dans le nouvel espace des composantes principales. Chaque individu est projeté sur les axes définis par les composantes principales, et les facteurs principaux correspondent aux coordonnées des individus sur ces axes. Ces facteurs représentent la contribution de chaque composante principale à chaque individu.

Associons à $\mathbf{a_k}$ le facteur principal $\mathbf{u_k} = \mathbf{M a_k}$ de taille $\mathbf{p}$. C'est un vecteur propre de $\mathbf{M\Sigma}$.
Preuve:

$$\mathbf{M\Sigma u_k} = \mathbf{M\Sigma M a_k} = \lambda_\mathbf{k} \mathbf{M a_k} = \lambda_\mathbf{k} \mathbf{u_k}$$

En pratique, les $\mathbf{u_k}$ sont calculé par diagonalisation de la matrice $\mathbf{M\Sigma}$ on obtient donc $\mathbf{c_k} = \mathbf{Y u_k}$. Les $\mathbf{a_k}$ ne sont pas interessants.
En posant $\mathbf{u'_k} = (\mathbf{u_{1k}}, \mathbf{u_{1k}}, \ldots, \mathbf{u_{pk}})$, on voit que la matrice des $\mathbf{u_{kj}}$ sert de matrice de passage entre la nouvelle base et l'ancienne base:

$$\mathbf{c_{ik}} = \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{y_i^j u_{jk}} \qquad \mathbf{c_k} = \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{y^j u_{jk}}$$

Formules de reconstitution

La reconstitution des données initiales à partir des facteurs principaux peut être réalisée en utilisant les premières **k** composantes principales. Cela permet d'approcher les données initiales tout en réduisant la dimensionnalité.

Par définition des $\mathbf{c_k}$, on a $\mathbf{E_i - G} = \sum_{\mathbf{k=1}}^{\mathbf{p}} \mathbf{c_{ik} a_k}$, et donc

$$\mathbf{y_i^j} = \sum_{\mathbf{k=1}}^{\mathbf{p}} \mathbf{c_{ik} a_{kj}}, \qquad \mathbf{y^j} = \sum_{\mathbf{k=1}}^{\mathbf{p}} \mathbf{c_k a_{kj}}, \qquad \mathbf{Y} = \sum_{\mathbf{k=1}}^{\mathbf{p}} \mathbf{c_k a_k'}$$

Les $\mathbf{a_{kj}}$ forment la matrice de passage entre l'ancienne base et la nouvelle. D'après le théorème d'**Eckart-Young**, on peut affirmer que les **k** premiers termes fournissent la meilleur approximation de **Y** par une matrice de rang **k** au sens des moindres carrées.

Aspect pratique de l'ACP

ACP sur des données centrées et réduites

Dans ce cas, l'analyse est réalisée à partir de la matrice de corrélation **R**, car les données sont désormais sur une échelle comparable (moyenne nulle et variance égale à 1).

La relation entre les données centrées-réduites et la matrice de corrélation est exprimée de la manière suivante :

$$\mathbf{Z}'\mathbf{D_w}\mathbf{Z} = \mathbf{D}_{\frac{1}{\sigma}}\mathbf{Y}'\mathbf{D_w}\mathbf{Y}\mathbf{D}_{\frac{1}{\sigma}} = \mathbf{D}_{\frac{1}{\sigma}}\mathbf{\Sigma}\mathbf{D}_{\frac{1}{\sigma}} = \mathbf{R}$$

où :

- **Z** représente les données réduites,
- $\mathbf{D_w}$ est la matrice des poids (ou des masses) associée aux observations,
- **Y** est la matrice des données centrées,
- $\mathbf{D}_{\frac{1}{\sigma}}$ est la matrice de normalisation, une matrice diagonale dont les éléments sont les inverses des écarts-types $\frac{1}{\sigma_i}$ des variables,

- $\mathbf{\Sigma}$ est la matrice de covariance des données centrées,
- $\mathbf{R}$ est la matrice de corrélation.

Données centrées-réduites : Les données sont d'abord centrées en soustrayant la moyenne, puis réduites en divisant chaque variable par son écart-type. Cela nous donne la matrice $\mathbf{Z}$, qui représente les données après ces deux étapes de transformation.

Produit matriciel : Le produit $\mathbf{Z}'\mathbf{D_w}\mathbf{Z}$ représente la matrice de covariance pondérée des données réduites. En réécrivant $\mathbf{Z} = \mathbf{D}_{\frac{1}{\sigma}}\mathbf{Y}$, où $\mathbf{Y}$ est la matrice des données centrées, on obtient l'expression suivante :

$$\mathbf{Z}'\mathbf{D_w}\mathbf{Z} = \mathbf{D}_{\frac{1}{\sigma}}\mathbf{Y}'\mathbf{D_w}\mathbf{Y}\mathbf{D}_{\frac{1}{\sigma}}$$

Cette expression montre que les données centrées $\mathbf{Y}$ sont normalisées à l'aide de $\mathbf{D}_{\frac{1}{\sigma}}$, ce qui fait apparaître la matrice de covariance des données centrées $\mathbf{\Sigma} = \mathbf{Y}'\mathbf{D_w}\mathbf{Y}$.

Transformation en matrice de corrélation : Finalement, la matrice de covariance $\mathbf{\Sigma}$ est transformée en matrice de corrélation $\mathbf{R}$ en multipliant par $\mathbf{D}_{\frac{1}{\sigma}}$ à gauche et à droite :

$$\mathbf{R} = \mathbf{D}_{\frac{1}{\sigma}} \mathbf{\Sigma} \mathbf{D}_{\frac{1}{\sigma}}$$

Ce produit revient à normaliser la covariance pour obtenir la corrélation, qui est sans dimension et permet de comparer directement les variables.

Exemple illustratif

Supposons que nous ayons trois variables ($p = 3$) et que les écarts-types soient $\sigma_1 = 2$, $\sigma_2 = 3$, et $\sigma_3 = 4$. La matrice de normalisation $\mathbf{D}_{\frac{1}{\sigma}}$ est alors :

$$\mathbf{D}_{\frac{1}{\sigma}} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix}$$

Si $\mathbf{\Sigma}$ est la matrice de covariance des données centrées, par exemple :

$$\mathbf{\Sigma} = \begin{pmatrix} 4 & 1.5 & 2 \\ 1.5 & 9 & 3 \\ 2 & 3 & 16 \end{pmatrix}$$

La matrice de corrélation $\mathbf{R}$ est obtenue par :

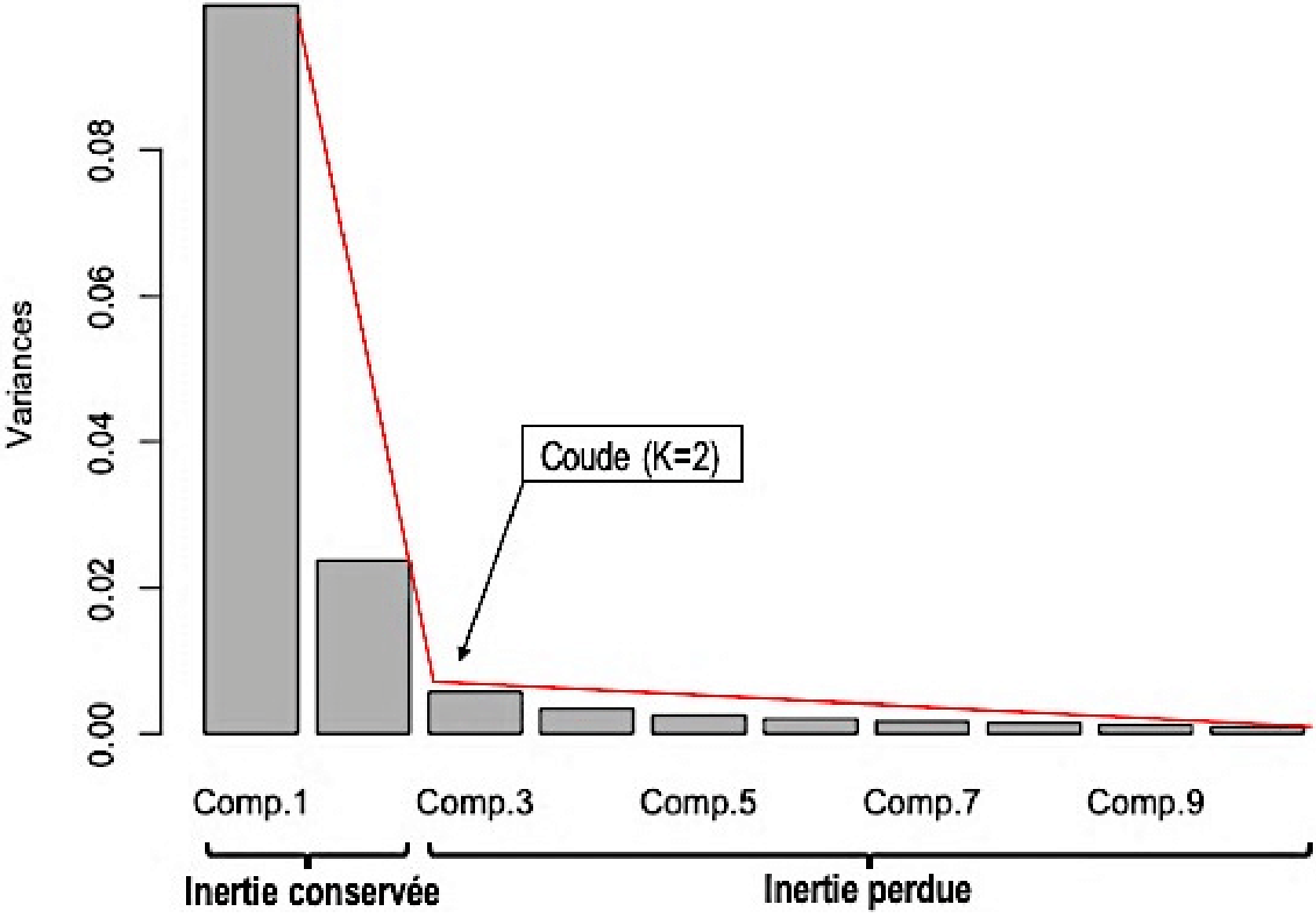
$$\mathbf{R} = \mathbf{D}_{\frac{1}{\sigma}} \mathbf{\Sigma} \mathbf{D}_{\frac{1}{\sigma}} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 4 & 1.5 & 2 \\ 1.5 & 9 & 3 \\ 2 & 3 & 16 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix}$$

Ce produit donne la matrice de corrélation, où les éléments diagonaux sont égaux à 1 et les autres éléments représentent les corrélations entre les variables.

Nombre d'axes à retenir

Lorsqu'on réalise une Analyse en Composantes Principales (ACP), la question du nombre d'axes à retenir est cruciale pour interpréter correctement les résultats. Les axes principaux correspondent aux directions de plus grande variance dans les données.

En pratique, il est conseillé de retenir les axes qui capturent la majeure partie de l'inertie (ou variance) totale des données. Un critère commun est de conserver les axes dont les valeurs propres sont supérieures à 1 (critère de Kaiser), ou encore de se baser sur la représentation graphique appelée "scree plot" (coudée), où l'on retient les axes jusqu'à ce que la décroissance des valeurs propres devienne moins significative. Il est également essentiel de prendre en compte la signification pratique des axes pour l'interprétation des résultats.
Il faut noter que les seuls critères utilisables pour la sélèction des axes sont empiriques.
Ci-dessous un scree-plot pour aider à une sélection optimale des axes lors d'une ACP


Variances
0.00
0.02
0.04
0.06
0.08
Coude (K=2)
Comp.1
Comp.3
Comp.5
Comp.7
Comp.9
Inertie conservée
Inertie perdue

Interprétation du Scree Plot

1. **Valeurs propres** : Les hauteurs des barres ou des points sur le graphique correspondent aux valeurs propres. Une valeur propre élevée indique que la composante principale associée capture une grande partie de la variance des données.
2. **Coudée du graphique** : Un des aspects les plus importants à rechercher est la "coudée" (ou "élan") dans le graphique. La coudée est le point où les valeurs propres commencent à diminuer de manière significative et où la pente devient moins abrupte. Ce point indique généralement le nombre optimal de composantes principales à retenir.

Critère de Kaiser : Une autre méthode d'interprétation est le critère de Kaiser, qui suggère de retenir les composantes ayant une valeur propre supérieure à 1. Cela signifie que ces composantes expliquent plus de variance qu'une variable originale standardisée.

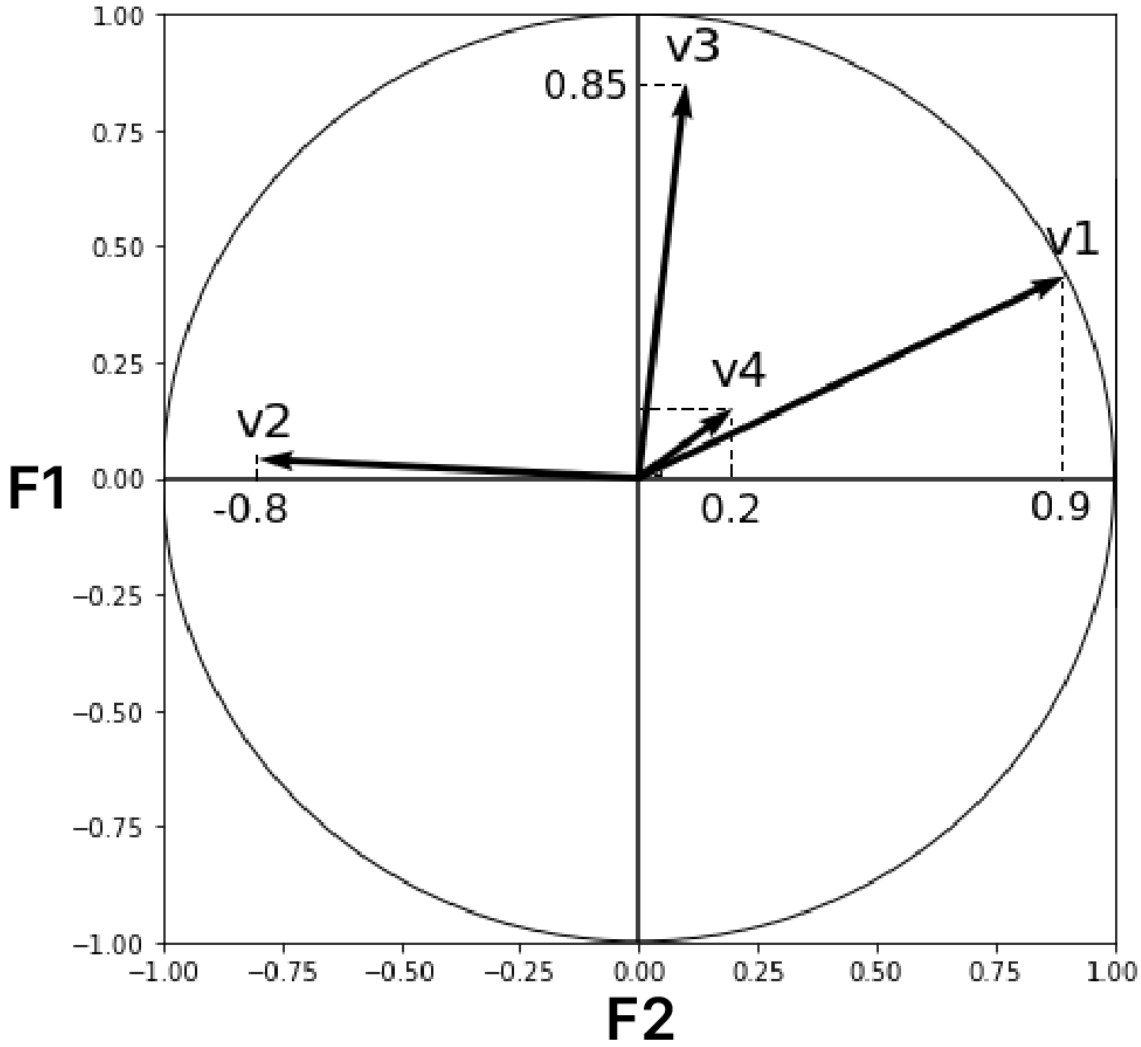
Interprétation qualitative : Il est également important d'interpréter les résultats dans le contexte de l'étude. Parfois, des composantes avec de faibles valeurs propres peuvent avoir des significations pratiques importantes qui ne devraient pas être négligées.

Le cercle des correlations

Le cercle des corrélations est un outil graphique utilisé dans le cadre de l'Analyse en Composantes Principales (ACP) pour visualiser les relations entre les variables initiales et les composantes principales. Dans un graphique, chaque variable est représentée par un vecteur dont la direction et la longueur indiquent respectivement la force et la direction de sa contribution aux composantes principales.

Dans le cercle de corrélation :

- Axes : Les axes correspondent aux premières composantes principales (généralement les deux premières) dans l'espace des composantes principales.
- Vecteurs : Chaque variable est représentée par un vecteur qui part de l'origine du cercle et pointe vers la position de la variable sur le cercle. La longueur du vecteur est proportionnelle à l'importance de la variable dans l'explication de la variance des données.
- Cercle unité : Le cercle est souvent normalisé pour avoir un rayon de 1, ce qui permet de visualiser facilement la contribution relative des variables.


v3
0.85
v1
v4
v2
-0.8
0.2
0.9
F1
F2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

Interprétation du cercle des corrélations

- Direction et corrélation :
  La direction du vecteur d'une variable indique sa corrélation avec les composantes principales. Si deux vecteurs sont proches l'un de l'autre, cela signifie que les variables qu'ils représentent sont fortement corrélées. Inversement, des vecteurs éloignés indiquent une corrélation faible ou négative.
- Longueur des vecteurs : La longueur d'un vecteur est indicative de l'importance de la variable dans le modèle. Des vecteurs plus longs suggèrent que la variable contribue de manière significative à l'inertie expliquée par les composantes principales, tandis que des vecteurs courts indiquent une contribution marginale.

- Interprétation des axes :
  Les axes du graphique (composantes principales) représentent les directions de la variance maximale dans les données. Les variables qui pointent dans la même direction que les axes des composantes principales ont une forte influence sur ces axes, et leur contribution est donc importante pour l'interprétation des résultats de l'ACP.
- Identification des variables influentes :
  Le cercle de corrélation aide à identifier les variables les plus influentes dans l'explication de la variance des données. En observant la position et la longueur des vecteurs, on peut déterminer quelles variables sont cruciales pour le modèle et lesquelles peuvent être considérées comme moins significatives.

En résumé, le cercle de corrélation est un outil puissant pour visualiser et interpréter les relations entre les variables initiales et les composantes principales dans une ACP. Il permet d'obtenir des insights sur la structure des données et d'orienter les décisions concernant la réduction de dimension et l'interprétation des résultats.

Contribution d'un Individu à une Composante Principale

La contribution d'un individu à une composante principale est un aspect essentiel de l'analyse en composantes principales (ACP), car elle permet d'évaluer dans quelle mesure chaque observation influence les directions principales de la variance des données.

On sait que $\text{var}(\mathbf{c_k}) = \lambda_\mathbf{k} = \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbf{w_i c_{ik}^2}$. La contribution de l'individu **i** à la composante principale **k** est donc:

$$\frac{\mathbf{w_i c_{ik}^2}}{\lambda_\mathbf{k}}$$

Ainsi, la contribution d'un individu **i** est importante si elle excède d'un facteur α le poids $\mathbf{w_i}$ de l'individu concerné, c'est à dire:

$$\frac{\mathbf{w_i c_{ik}^2}}{\lambda_\mathbf{k}} \geq \alpha \mathbf{w_i}$$

ou de manière équivalente:

$$|\mathbf{c_{ik}}| \geq \sqrt{\alpha \lambda_\mathbf{k}}$$

Choix de alpha

Selon les données, on se fixe en générale une valeur de l'ordre de 2 à 4, que l'on garde pour tous les axes.
Ainsi on dira qu'un idividu est sur-représenté lorsqu'il joue un rôle trop fort dans la définition d'un axe, par exemple:

$$\frac{\mathbf{w_i c_{ik}^2}}{\lambda_\mathbf{k}} > \mathbf{0,25}$$

En effet, cet individu tire à lui seul l'axe **k** et risque de perturber les représentations des autres points sur les axes de rang $\geq$ **k**. Il est surtout problématique sur les premiers axes. Un tel individu peut être le signe de données erronées.
Pour corriger ça, on peut tout simplement retirer cet individu de l'analyse et le mettre en individu supplémentaire.
Evaluer la contribution d'un individu à une composante principale est crucial pour interpréter les résultats de l'ACP et pour guider d'éventuelles décisions analytiques ultérieures.

Qualité de l'analyse

On sait que $\mathbf{I_G} = \text{tr}(\mathbf{\Sigma M})$. Comme la trace d'une matrice est la somme de ses valeurs propres, nous avons :

$$\mathbf{I_G} = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

Donc la qualité de la représentation obtenue par $\mathbf{k}$ valeurs propres est la proportion de l'inertie expliquée

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

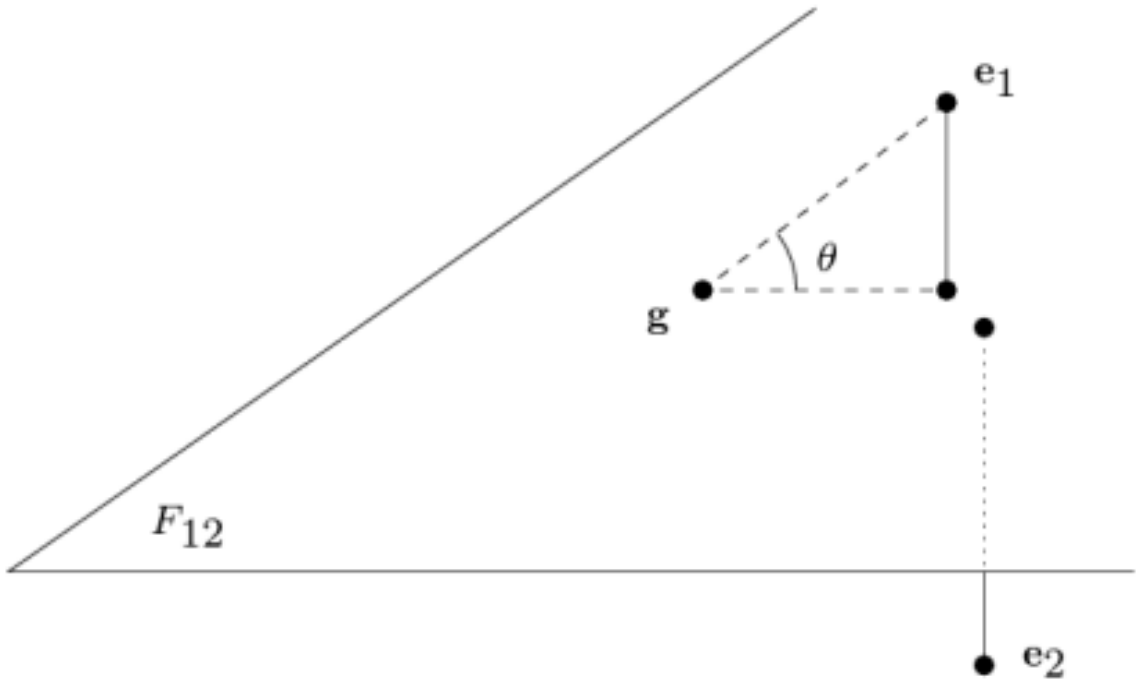
Par exemple, si $\lambda_1 + \lambda_2 = \mathbf{0,9 I_G}$, cela indique que le nuage de points est aplati autour du premier plan principal.

Pour des variables centrées et réduites, on a $\mathbf{I_G} = \text{tr}(\mathbf{R}) = \mathbf{p}$ ainsi, la somme des valeurs propres est le nombre de variables.
Notons également que cette valeur sert seulement à évaluer la projection retenue, pas à choisir le nombre d'axes à garder.

Qualité locale de la représentation

Ici, nous visons à évaluer si le nuage de points est fortement aplati en raison de la projection sur les sous-espaces principaux. Dans ce contexte, deux individus qui sont en réalité éloignés pourraient donner l'impression d'être proches l'un de l'autre.

La figure ci-dessous illustre la qualité locale de la représentation des points après projection.



Angle entre un individu et un axe principal

On le défini par son cosinus carré.

Le cosinus de l'angle entre l'individu centré $\mathbf{i}$ et l'axe principal $\mathbf{a_k}$ est

$$\cos(\mathbf{E_i}\hat{,}\mathbf{a_k}) = \frac{\langle \mathbf{E_i} - \mathbf{G}, \mathbf{a_k} \rangle_{\mathbf{M}}}{\|\mathbf{E_i} - \mathbf{G}\|_{\mathbf{M}}}$$

Puisque les $\mathbf{a_k}$ forment une base orthogonale.
Comme $\langle \mathbf{E_i} - \mathbf{G}, \mathbf{a_k} \rangle_{\mathbf{M}} = c_{ik}$,

$$\cos^2(\mathbf{E_i}\hat{,}\mathbf{a_k}) = \frac{\mathbf{c_{ik}^2}}{\sum_{\mathbf{l=1}}^{\mathbf{p}} \mathbf{c_{il}^2}}$$

qui est une grandeure qui mésure la qualité de la représentation de l'individu $\mathbf{i}$ sur l'axe principal $\mathbf{a_k}$.

Angle entre un individu et un sous espace principal

Il s'agit ici de l'ange que fait l'individu par rapport à sa projection orthogonale sur le sous espace.

La projection de $\mathbf{E_i} - \mathbf{G}$ sur le sous espace $\mathbb{E}_k$, $k \leq p$, est $\sum_{j=1}^{k} c_{ij} a_j$, et donc:

$$\mathbf{cos^2}(\mathbf{E_i}, \hat{\mathbb{E}}_\mathbf{k}) = \frac{\sum_{\mathbf{j=1}}^{\mathbf{k}} \mathbf{c_{ij}^2}}{\sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{c_{ij}^2}}$$

La qualité de la représentation de l'individu i sur le plan $\mathbb{E}_k$ est donc la somme des qualités de représentation sur les axes formants $\mathbb{E}_k$.
Notons qu'un $\cos^2$ égal à 0,9 correspond à un angle de 18°. Par conséquent, une valeur de 0,5 correspond à un angle de 45°.
On peut considérer que des valeurs supérieures à 0,80 indiquent une bonne qualité de projection, tandis que des valeurs inférieures à 0,5 signalent une mauvaise qualité.
Cependant, une mauvaise qualité de projection n'est significative que si le point projeté n'est pas trop proche de l'origine.

Notion de variables et d'individus suplémentaire

Variables supplémentaires qualitatives

Les composantes principales sont définies pour maximiser les contributions. Par conséquent, le fait que les corrélations obtenues soient proches de 1 peut ne pas être significatif. En revanche, une forte corrélation entre une composante principale et une variable qui n'a pas participé à l'analyse est très significative.

En pratique, certaines variables sont mises de côté afin qu'elles ne soient pas utilisées dans l'analyse (ce qui réduit la dimension de la matrice R en retirant des lignes et des colonnes). L'objectif est alors de déterminer si elles sont associées à un axe donné.
On calcule la corrélation de ces variables avec les composantes principales, puis on les place dans le cercle des corrélations. Si $\hat{\boldsymbol{z}}$ est le vecteur centré-réduit correspondant à cette variable, on calcule :

$$\text{cor}(\hat{\mathbf{z}}, \mathbf{c_k}) = \frac{\text{cov}(\hat{\mathbf{z}}, \mathbf{c_k})}{\sqrt{\text{var}(\mathbf{c_k})}} = \frac{\mathbf{1}}{\sqrt{\lambda_k}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i} \hat{\mathbf{z}}_\mathbf{i} \mathbf{c_{ik}}$$

Il est également possible d'utiliser un test statistique pour valider ces résultats.

Variables supplémentaires qualitatives

Il est possible de représenter les individus appartenant à différentes catégories (par exemple, homme, femme, etc.) à l'aide de symboles distincts sur les axes principaux.

Afin de déterminer si les étiquettes sont associées à l'axe principal $\mathbf{k}$, on peut calculer la coordonnée $\hat{\mathbf{c_k}}$ du barycentre de ces individus sur cet axe.

Valeur test

Soit $\hat{\mathbf{n}}$ le nombre d'individus parmi un total de n appartenant à une certaine catégorie, et $\hat{\mathbf{c_k}}$ la coordonnée de leur barycentre sur la k-ième composante principale. La valeur test s'exprime comme suit :

$$\hat{\mathbf{c_k}} \sqrt{\frac{\hat{\mathbf{n}}}{\lambda_\mathbf{k}}} \sqrt{\frac{\mathbf{n}-\mathbf{1}}{\mathbf{n}-\hat{\mathbf{n}}}}$$

Interprétation et usage

Cette valeur test est significative si les deux conditions suivantes sont remplies :

- $\hat{\mathbf{n}}$ et $\mathbf{n} - \hat{\mathbf{n}}$ sont suffisamment grands (généralement supérieurs à 30 pour que le théorème central limite s'applique) ;
- La valeur absolue de la statistique est supérieure à 2 ou 3.

Si ces conditions ne sont pas satisfaites, on conclura qu'il n'est pas possible d'affirmer avec certitude si la catégorie est liée à l'axe k.

Explication du calcul

L'idée sous-jacente est que, si les $\hat{\mathbf{n}}$ individus avaient été sélectionnés aléatoirement, $\hat{\mathbf{c_k}}$ suivrait une loi normale centrée (puisque les $\mathbf{z}$ ont une moyenne nulle) avec une variance égale à :

$$\frac{\lambda_{\mathbf{k}}}{\hat{\mathbf{n}}} \times \frac{\mathbf{n} - \hat{\mathbf{n}}}{\mathbf{n} - \mathbf{1}}$$

Cette variance découle du fait que l'échantillonnage se fait sans remise.

Individus supplémentaires

La méthode consiste à exclure certains individus de l'analyse, c'est-à-dire qu'ils ne sont pas pris en compte dans le calcul des covariances. Ensuite, l'objectif est de déterminer s'ils sont liés à un axe principal donné.

Cas des individus sur-représentés

On peut choisir de traiter ces points comme des individus supplémentaires, notamment lorsqu'ils font partie d'un échantillon qui, en lui-même, ne présente pas un grand intérêt. Cela peut être utile pour des cas où les points représentent des échantillons et non des entités d'intérêt direct.

Représentation

Ces individus supplémentaires sont ensuite ajoutés à la représentation graphique dans les plans principaux. Pour calculer leurs coordonnées sur un axe principal donné, on utilise la relation suivante :

$$\hat{c_k} = \sum_{j=1}^{p} \hat{z}^j u_{jk}$$

où les $\hat{\mathbf{z}}^{\mathbf{j}}$ représentent les coordonnées centrées-réduites d'un individu supplémentaire $\hat{\mathbf{z}}$.
Ces individus supplémentaires peuvent alors servir d'échantillons-tests pour valider les hypothèses obtenues à partir de l'ACP appliquée aux individus actifs.

Exemple d'application de l'ACP

Données

Notre jeu de données comprend les notes sur 20 de 29 élèves d'un lycée, en se concentrant spécifiquement sur cinq matières : les mathématiques, la physique, le français, l'anglais et le latin.

INDIVIDU	Maths	Physique	Français	Anglais	Latin
Basile	14	15	13	12	11
Thierry	4	3	6	4	12
Genevieve	16	15	5	6	10
Odilon	19	18	6	6	12
Edouard	15	15	14	16	11

Figure: Aperçu des données

Modélisation du jeu de données

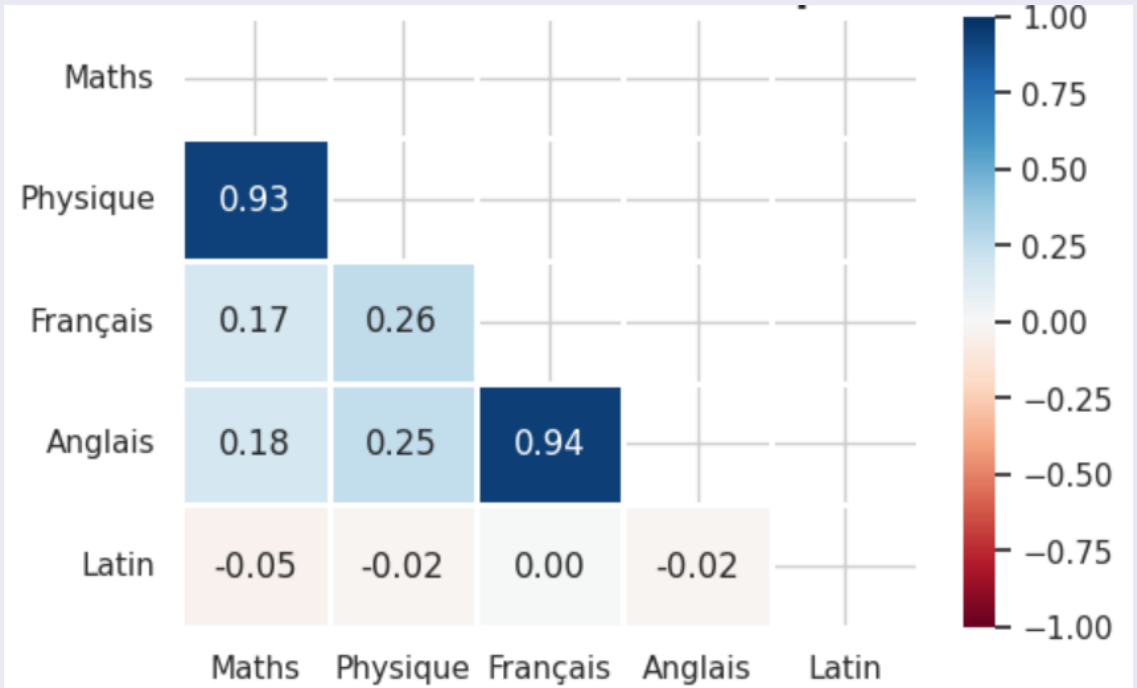
Soit $X \in \mathbb{R}^{29\times 5}$ notre jeu de données, où :

- 29 est le nombre d'élèves (observations),
- 5 est le nombre de matières concernées (caractéristiques).

Le but de la PCA est de trouver un ensemble de vecteurs propres orthogonaux $F_1, F_2, \ldots, F_k$ tels que les nouvelles variables $Y = XF$ maximisent la variance expliquée par chaque composante principale.

Matrice de correlation

La matrice de corrélation permet d'évaluer la force et la direction des relations linéaires entre plusieurs variables ; chaque coefficient de corrélation dans la matrice indique dans quelle mesure deux variables varient ensemble, avec des valeurs allant de -1 (corrélation négative parfaite représentée ici par la couleur rouge) à 1 (corrélation positive parfaite représentée ici par la couleur bleue), tandis qu'une valeur proche de 0 suggère une absence de relation linéaire.

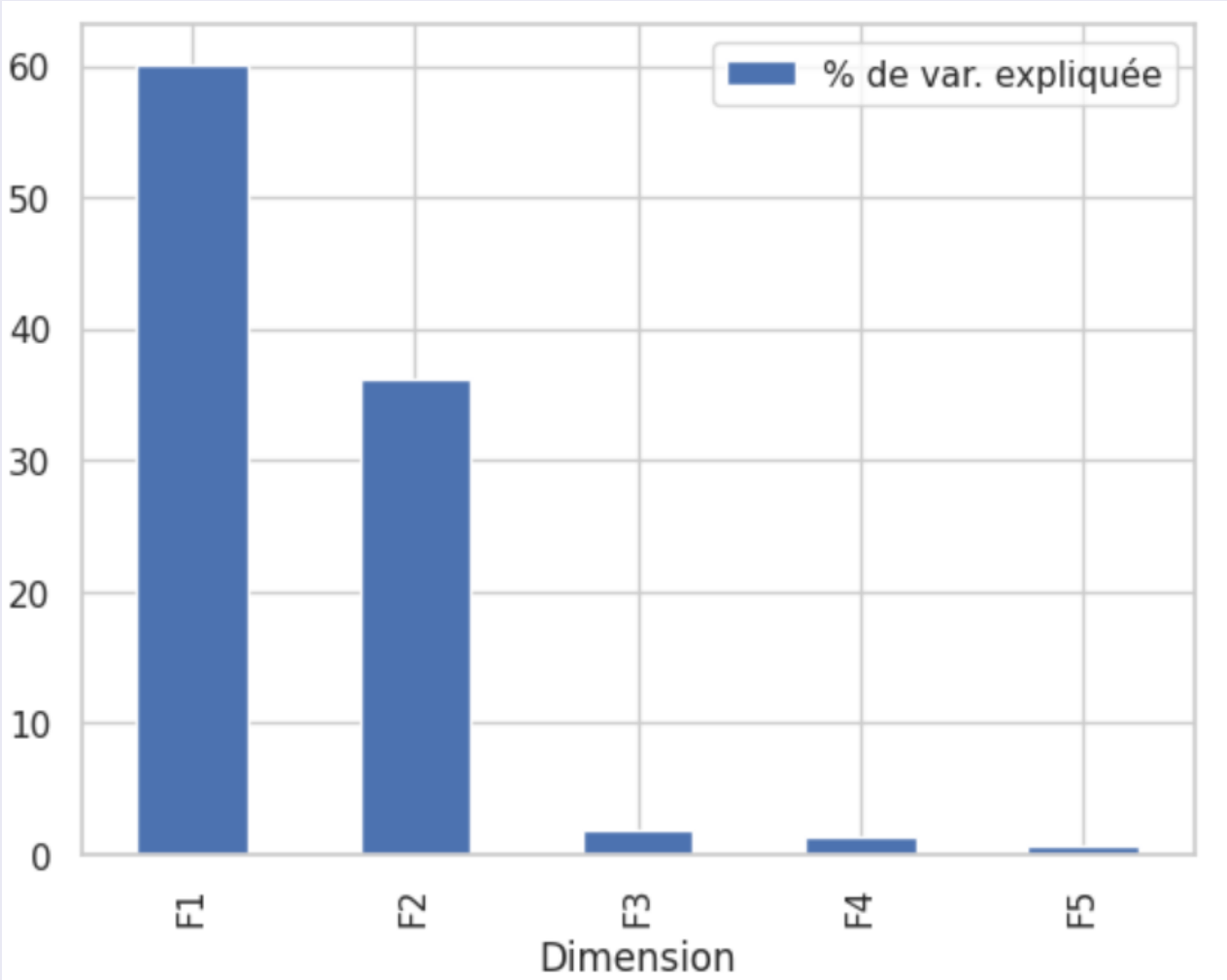


Eblouie des Valeurs propres

Lorsqu'on applique l'Analyse en Composantes Principales (ACP), les données sont projetées sur les axes principaux d'inertie, lesquels sont arrangés en ordre décroissant en fonction de l'inertie du nuage projeté, de la plus grande à la plus petite. En additionnant les inerties associées à tous les axes, on obtient l'inertie totale du nuage des individus.
Les inerties associées à chaque axe, notées F_i, sont équivalentes aux valeurs propres de la matrice de covariance des données.

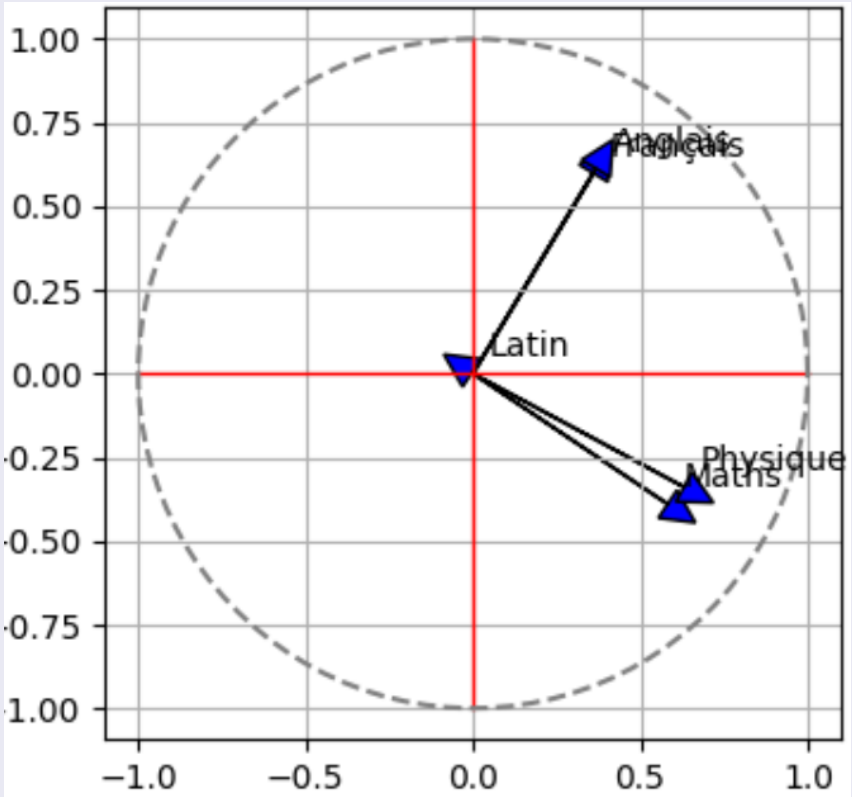
Dimension	Var. expliquée	% de var. expliquée	% cum. var. expliquée
F1	63.612576	60.130628	60.130628
F2	38.202932	36.111826	96.242454
F3	1.868437	1.766165	98.008619
F4	1.352795	1.278747	99.287366
F5	0.753900	0.712634	100.000000

Un moyen de représenter cette répartition de l'inertie consiste à afficher un diagramme décrivant le pourcentage d'inertie totale attribué à chaque axe. Ce diagramme est communément appelé l'éboulis des valeurs propres. Un exemple de ce type de représentation est illustré ci-dessous



Le cerle de correlation dans le premier plan factoriel

Le cercle de corrélation, permet de visualiser les relations entre les variables initiales et les axes factoriels ; plus une variable est proche du bord du cercle, plus elle est bien représentée par les composantes principales, tandis que les variables qui sont proches l'une de l'autre indiquent une forte corrélation positive, et celles en opposition sont corrélées négativement.

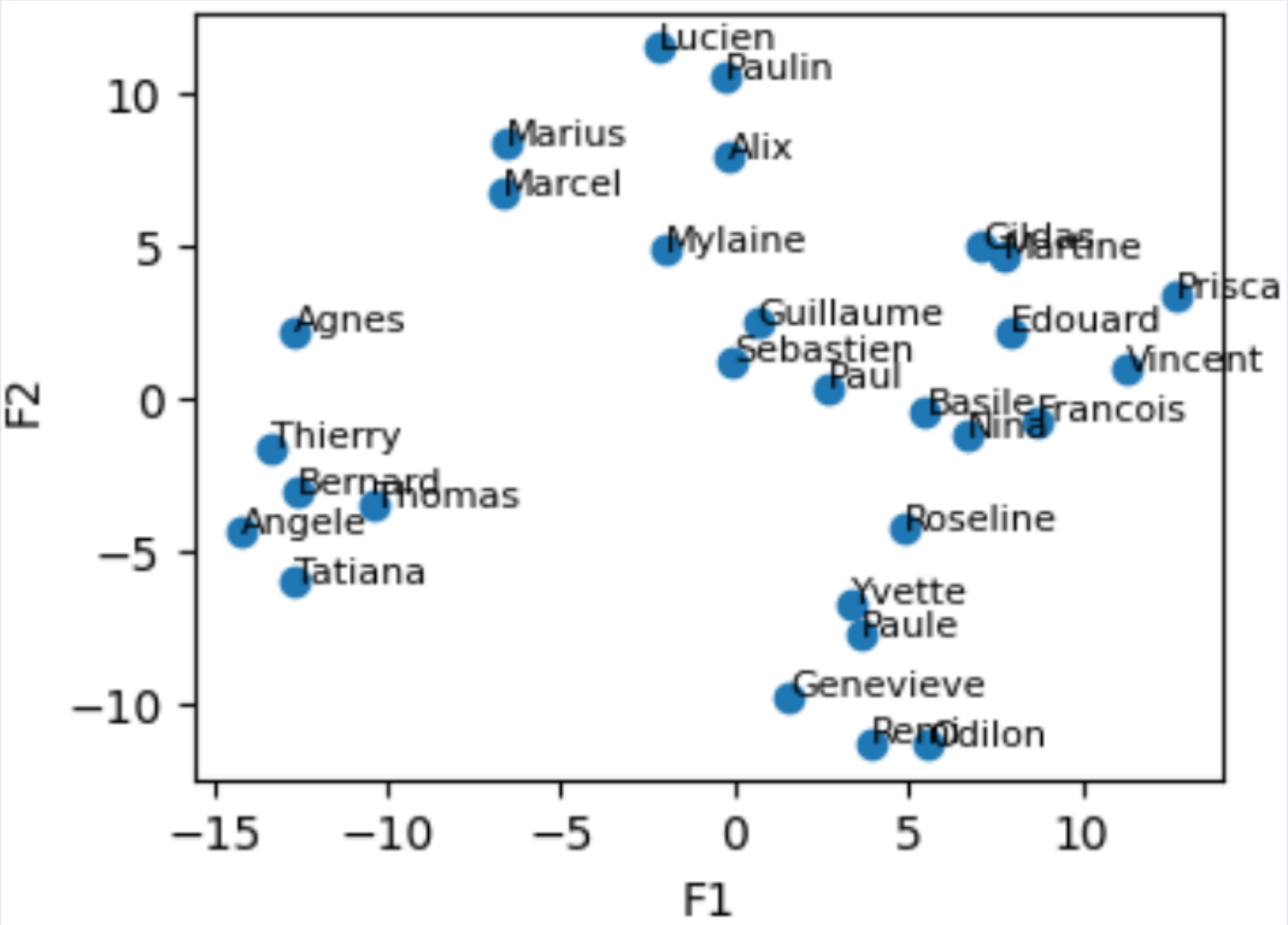


Interprétation du cercle de correlation

- **Position des variables :** Chaque variable originale(Math,Physique,Français,Anglais,Latin) est représentée par un vecteur dans le cercle de corrélation. La direction de ce vecteur indique la corrélation de la variable avec les composantes principales, et sa longueur indique la force de cette corrélation.
- **Proximité au cercle :** Les variables qui sont proches du cercle(Math,Physique,Français,Anglais) sont bien représentées par les composantes principales. Plus précisément, si un vecteur est proche du bord du cercle, cela signifie qu'il a une corrélation élevée avec la composante principale correspondante. Les variables situées près du centre du cercle ont une faible corrélation avec les composantes principales(Latin).

- **Angle entre les Variables :** L'angle entre deux vecteurs représente la corrélation entre les deux variables correspondantes. Si deux vecteurs pointent dans la même direction, les variables correspondantes sont fortement corrélées (Math et Physique sont fortement corrélées d'une part et Français et Anglais sont également fortement corrélées d'autre part). Si les vecteurs sont à 90 degrés l'un de l'autre, les variables sont non corrélées(Math et Français sont non corrélées).
- **Contribution à l'Inertie :** La variance expliquée par chaque composante principale est également représentée dans le cercle de corrélation. Les variables qui contribuent fortement à l'inertie de la composante principale sont situées près du bord du cercle.

Projection des données dans le premier plan factoriel



Interprétation

Dans le premier plan factoriel après PCA, les élèves sont projetés en fonction de leurs performances dans les cinq matières (mathématiques, physique, français, anglais, latin) en relation avec les composantes principales.

- Élèves proches des flèches des mathématiques et de la physique : Ces élèves ont des performances élevées en mathématiques et en physique, car ces matières sont fortement corrélées (flèches proches et un angle faible). Leur position indique qu'ils réussissent particulièrement bien dans ces matières.
- Élèves proches des flèches du français et de l'anglais : Les élèves qui se situent dans cette zone du plan ont de bonnes notes en français et en anglais. La forte corrélation entre ces deux matières (angle faible entre les flèches) indique que les élèves qui réussissent dans l'une ont tendance à réussir dans l'autre.

- Élèves éloignés des flèches de toutes les matières : Les élèves situés loin des flèches de toutes les matières n'ont pas de résultats particulièrement élevés dans aucune des matières. Ils peuvent avoir des notes moyennes ou faibles, et leurs performances sont globalement dispersées.
- Élèves proches du centre du plan : Les élèves positionnés près du centre du cercle ont des performances équilibrées dans toutes les matières, sans exceller particulièrement dans aucune d'elles. Leur position suggère une faible différenciation en fonction des composantes principales.
- Élèves proches de la flèche du latin : Comme la flèche du latin est près du centre du cercle, cela signifie que cette matière ne contribue pas fortement à la variabilité expliquée par les deux premières composantes principales. Les élèves proches de cette flèche ont probablement des performances relativement similaires dans cette matière, mais le latin n'est pas un facteur déterminant dans leur différenciation.