IPSA
ÉCOLE D'INGÉNIEUR EN AÉRONAUTIQUE

IPSA Toulouse
2024-2025
Promotion 2k27

Ma315

Introduction aux Sciences des Données

BE

UNE PREMIÈRE APPROCHE DE L'ANALYSE EN COMPOSANTES PRINCIPALES

Mathys HERBRETEAU
Tom KRENCKER
Florent KOHLMULLER
Adrian CALESTINI
Maxime GOT

TABLE DES MATIERES

INTRODUCTION

L'Analyse en Composantes Principales (ACP) est une méthode essentielle dans les sciences des données pour réduire la dimensionnalité tout en préservant l'essentiel de l'information. Ce rapport est structuré en deux grandes parties : la première présente l'analyse théorique des concepts fondamentaux de l'ACP, tandis que la seconde illustre leur mise en œuvre pratique à travers des scripts Python. Cette organisation permet de relier théorie et pratique, en mettant l'accent sur les points essentiels pour maîtriser l'ACP.

PARTIE 1 – ANALYSE THÉORIQUE

La première partie explore les fondements mathématiques de l'ACP. Nous y examinons les propriétés des vecteurs aléatoires, la diagonalisation des matrices de covariance, et les concepts de maximisation de la variance. Ces notions constituent la base théorique de l'ACP.

SECTION 1.1 : PROPRIETES DES VECTEURS ALEATOIRES ET VARIANCE

Nous débutons par l'étude des propriétés fondamentales des vecteurs aléatoires, comme leur centrage et leur variance. Ces notions sont essentielles pour comprendre comment maximiser la dispersion des données dans un espace de dimension réduite, objectif central de l'ACP.

QUESTION 1

Commençons par introduire X comme un vecteur aléatoire à n composantes et v un vecteur colonne dans R^{n}.

Nous posons $Y = Xv$ une combinaison linéaire de X, si X est centrée alors nous avons :

$E(X) = 0$ ainsi cela implique :

$$E(Y) = E(Xv) = vE(X) = 0$$

Nous avons donc Y qui est également centré.

Nous avons également la variance de Y :

$$Var(Y) = E(Y^2) - E(Y)^2 = E(Y^2)$$

Or nous avons posé que $Y = Xv$, on a donc :

$$Y^2 = (v^T X)^2 = (v^T X)(X^T v) = v^T (XX^T)v$$

La variance de Y est donc égale a :

$$Var(Y) = E(v^T (XX^T)v)$$

On applique la linéarité de l'espérance :

$$Var(Y) = v^T\, E(XX^T)v$$

Or nous savons que la matrice de covariance pour des X centrée est la suivante :

$$Cov(X) = E(XX^T)$$

Ainsi nous avons :

$$Var(Y) = v^T\, Cov(X)v$$

Or $Y = Xv$, donc finalement :

$$Var(Xv) = v^T\, Cov(X)v$$

Ainsi nous avons démontré que la transformation linéaire d'un vecteur aléatoire centré conserve son centrage et que la variance de la transformation dépend de la matrice de covariance. Ce résultat est fondamental, car il justifie l'intérêt d'étudier les directions dans lesquelles la variance est maximisée.

QUESTION 2

D'après les questions précédentes, nous devons donc résoudre le problème d'optimisation avec contrainte suivant :

$$(\mathcal{P}) : \arg \max_{\|v\|_2=1, v\in\mathbb{R}^n} v^T Cov(X)v$$

En notant :

$$\mathcal{L}(v,\lambda) \coloneqq v^T Cov(X)v - \lambda(\|v\|_2^2 - 1)$$

Nous cherchons maintenant à optimisation cette variance pour identifier les directions principales, qui capturent la majorité de l'information contenue dans les données.

Pour commencer nous voulons maximiser $v^T\, Cov(X)v$ avec la contrainte $\|v\|_2 = 1$, cette contrainte implique que $v^T v = 1$.

On introduit grâce au lagrangien la contrainte dans $\mathcal{L}(x,\lambda)$:

$$\mathcal{L}(x,\lambda) = v^T Cov(X)v - \lambda(v^T v - 1)$$

On a dans ce cas des conditions de premier ordre, ainsi pour maximiser $\mathcal{L}(x,\lambda)$ nous allons calculer les dérivées partielles de $\frac{\partial \mathcal{L}}{\partial v}(x,\lambda)$ et de $\frac{\partial \mathcal{L}}{\partial \lambda}(x,\lambda)$.

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial v}(x,\lambda) = 0 \\ \frac{\partial \mathcal{L}}{\partial v}(x,\lambda) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} 2Cov(X)v - 2\lambda v = 0 \\ -v^T v + 1 = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} Cov(X)v = \lambda v \\ \|v\|_2 - 1 = 0 \end{cases}$$

Nous obtenons :

$$Cov(X)v = \lambda v$$

Ainsi nous pouvons affirmer que les solutions v et λ, sont respectivement le vecteur propre et la valeur propre de X, vérifient bien $Cov(X)v = \lambda v$.

Donc les résultats montrent que maximiser la variance revient à résoudre un problème d'optimisation, où les solutions sont liées aux vecteurs propres de la matrice de covariance.

SECTION 1.2 : DIAGONALISATION ET PROPRIETES DES MATRICES DE COVARIANCE

Maintenant que nous avons établi l'importance de la variance dans l'ACP, nous allons examiner comment la matrice de covariance permet de formaliser cette dispersion et d'identifier les directions principales.

La diagonalisabilité de la matrice de covariance est un point clé, car elle permet de décomposer les données dans une base orthogonale. Cela facilite l'identification des composantes principales et leur interprétation.

QUESTION 3

Soit X un vecteur aléatoire dans $\mathbb{R}^n$, nous allons justifier que la matrice $Cov(X)$ se diagonalise dans une base orthonormée, c'est-à-dire qu'il existe une matrice orthogonale V et une matrice diagonale D telles que :

$$Cov(X) = VDV^T$$

Nous savons que la matrice de covariance de X s'exprimer comme suit :

$$Cov(X) = E\big((X - E[X])(X - E[X])^T\big)$$

Puisque $Cov(X)$ est une matrice symétrique, elle est donc diagonalisable d'après le théorème spectral. On a bien :

$$Cov(X) = VDV^T$$

Avec

$$D = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$$

En supposant que les valeurs propres sont ordonnées de manière décroissante :

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

La diagonalisation de $Cov(X)$ simplifie l'identification des directions principales et établit une structure claire pour les transformations utilisées en ACP.

QUESTION 4

Après avoir démontré que $Cov(X)$ est diagonalisable, nous pouvons relier cette propriété à la définition des composantes principales. Cela nous permet de formaliser comment les données peuvent être projetées dans un espace réduit tout en maximisant la variance conservée.

Ainsi nous avons $Cov(X) = VDV^T$ avec V matrice orthogonale avec les vecteurs propres $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n$.

On exprime $Y_k = Xv_k$ avec v_k le k vecteur propre de $Cov(X)$.

Donc en passant par la variance :

$$Var(Y_k) = Var(Xv_k)$$

On comprend alors que si on calcule la variance,

$$Var(Y_k) = Var(Xv_k)$$

Par la décomposition et les propriétés de la variance

$$Var(Y_k) = v_k^T Cov(Xv_k) v_k$$

Or on sait que v_k est un vecteur propre de $Cov(X)$, on a donc :

$$Cov(X)v_k = \lambda_k v_k$$

La variance de Y_k s'exprime donc comme suit :

$$Var(Y_k) = v_k^T(\lambda_k v_k) = \lambda_k\left(v_k^T v_k\right)$$

Or, on rappelle que $v_k^T v_k = 1$

Donc :

$$Var(Y_k) = \lambda_k$$

La trace de la matrice de covariance $Cov(X)$ est égale à la somme des variances des composantes principales :

$$Trace\big(Cov(X)\big) = \sum_{i=1}^{n} \lambda_i$$

Nous remarquons qu'au passage que les valeurs propres de la matrice de covariance $Cov(X)$ mesurent la dispersion du vecteur aléatoire X. Nous définissons le pourcentage de l'explication de la variance de la k-composante principale par :

$$\rho_k = \frac{\lambda_k}{Tace\big(Cov(X)\big)}$$

Ainsi nous avons établi que les composantes principales correspondent aux projections des données sur les vecteurs propres de la matrice de covariance et que la variance de chaque projection est donnée par la valeur propre associée. Ce résultat montre que l'ACP permet d'identifier les axes maximisant la dispersion des données, justifiant ainsi son utilisation pour la réduction de dimension.

SECTION 1.3 : ESTIMATION EMPIRIQUE ET COVARIANCE

Après avoir démontré que la matrice de covariance peut être diagonalisée, il reste à comprendre comment ces résultats théoriques peuvent être appliqués à des données réelles. Ainsi dans cette section nous abordons l'estimation empirique et les transformations nécessaires.

Dans la pratique, nous travaillons souvent avec des données échantillonnées. Cette section examine comment estimer la matrice de covariance et comment centrer et réduire les données pour garantir que l'ACP soit appliqué correctement.

QUESTION 5.A.

Nous avons établi que les vecteurs propres de $Cov(X)$ définissent les directions principales et leurs variances. Passons maintenant au cas pratique, nous calculons ici une approximation de la matrice de covariance à partir d'un échantillon E.

On note E_{cr} la matrice « centrée réduite » associée à E dont les coefficients sont donnés par la variable aléatoires :

$$(E_{cr})_{jk} = \frac{X_{jk} - \bar{X}_k}{\sqrt{m-1}\bar{\sigma}_k}$$

Nous souhaitons justifier que :

$$C := \frac{1}{m-1} E_{cr}^T E_{cr}$$

est une approximation d'un estimateur de la matrice de covariance de X.

Calculons maintenant les coefficiente de C :

$$C_{jk} = \frac{1}{m-1}(E_{cr})_{ij}(E_{cr})_{ik}$$

$$= \frac{1}{m-1}\sum_{i=1}^{m} \frac{X_{ij} - \bar{X}_{mj}}{\bar{\sigma}_{mj}\bar{\sigma}_{mk}} \cdot \frac{X_{ik} - \bar{X}_{mk}}{\sqrt{m-1}\bar{\sigma}_{mk}}$$

$$= \frac{1}{(m-1)^2}\sum_{i=1}^{m} \frac{(X_{ij} - \bar{X}_{mj})(X_{ik} - \bar{X}_{mk})}{\bar{\sigma}_{mj}\bar{\sigma}_{mk}}$$

On applique maintenant l'espérance aux coefficients de C pour $j \neq k$, soit :

$$E(C_{jk}) = E\left(\frac{1}{(m-1)^2}\sum_{i=1}^{m} \frac{(X_{ij} - \bar{X}_{mj})(X_{ik} - \bar{X}_{mk})}{\bar{\sigma}_{mj}\bar{\sigma}_{mk}}\right)$$

$$E(C_{jk}) = \frac{1}{(m-1)^2}\sum_{i=1}^{m} E\left(\frac{(X_{ij} - \bar{X}_{mj})(X_{ik} - \bar{X}_{mk})}{\bar{\sigma}_{mj}\bar{\sigma}_{mk}}\right)$$

$$E(C_{jk}) = \frac{1}{(m-1)^2 E(\bar{\sigma}_{mj}) E(\bar{\sigma}_{mk})} \sum_{i=1}^{m} E\left((X_{ij} - \bar{X}_{mj})(X_{ik} - \bar{X}_{mk})\right)$$

Or nous supposons que :

$$E\left(\left(\frac{X_{ik} - \bar{X}_{mk}}{\bar{\sigma}_{mk}}\right)\left(\frac{X_{jp} - \bar{X}_{mp}}{\bar{\sigma}_{mp}}\right)\right) \approx \frac{E\left((X_{ij} - \bar{X}_{mj})(X_{ik} - \bar{X}_{mk})\right)}{E(\bar{\sigma}_{mj}) E(\bar{\sigma}_{mk})}$$

Ainsi,

$$E(C_{jk}) = \frac{1}{(m-1) E(\bar{\sigma}_{mj}) E(\bar{\sigma}_{mk})} \sum_{i=1}^{m} Cov(X_{ij}, X_{ik})$$

Donc C est bien une approximation d'un estimateur de la matrice de covariance de X. Nous obtenons ainsi une matrice empirique qui permet d'appliquer l'ACP même sans connaître la distribution réelle des données.

QUESTION 5.B.

i.A.

Avec une estimation de la matrice de covariance en place, nous devons préparer les données pour garantir que l'ACP produise des résultats fiables. Cela passe par les étapes clés de centrage et de réduction des données.

Nous centrons et réduisons les données pour uniformiser leur échelle. Cela implique de rendre la moyenne des vecteurs lignes nulle (centrage) et de normaliser la variance des colonnes à 1 (réduction). Ces étapes garantissent que toutes les variables contribuent équitablement à l'analyse.

On définit le vecteur colonne u de la manière suivante :

$$u = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^m$$

Ainsi ce vecteur a m composantes qui sont toutes égales à 1, de plus u^T est un vecteur ligne de taille m.

Donc en effectuant la multiplication $u^T A$ nous obtenons un vecteur ligne a n composantes.

En cherchant la k-ième composante de $u^T A$, nous avons :

$$(u^T A)_k = \sum_{i=1}^{m} u_k^T A_{i,k} = \sum_{i=1}^{m} A_{i,k}$$

Donc $u^T A$ contient la somme des lignes de A.

Finalement en divisant par m, nous obtenons :

$$\frac{1}{m}u^T A = \left(\sum_{i=1}^{m} A_{i,1}, \ldots, \sum_{i=1}^{m} A_{i,n}\right)$$

Ceci correspond exactement au vecteur ligne moyen l_g. Nous pouvons ainsi écrire que :

$$l_g = \frac{1}{m}u^T A$$

Ce vecteur nous permet de calculer la moyenne de chaque variable, qui sera ensuite soustraite pour assurer un centrage correct.

QUESTION 5.B.

i.B.

Comme dis plus haut dans le but de centrer la matrice A par rapport a la moyenne de ses lignes on veut soustraire à chaque ligne la moyenne, c'est à dire l_g.

On pose donc :

$$ul_g = u\frac{1}{m}u^T A = \frac{1}{m}uu^T A$$

Ainsi nous obtenons une matrice $\frac{1}{m}uu^T$ qui est une matrice dont chaque ligne est égale a la ligne moyenne l_g.

Donc pour obtenir la matrice centrée A on pose :

$$A^{(c)} = A - \frac{1}{m}uu^T A$$

Donc chaque ligne de $A^{(c)}$ est la ligne d'origine de la matrice A moins la ligne moyenne l_g, ceci veut exactement dire que chaque ligne est une ligne centrée.

QUESTION 5.B.

i.C.

Ayant centré la matrice A, nous vérifions maintenant que le centrage ne modifie pas une matrice déjà centrée.

Pour centrer à nouveau la matrice $A^{(c)}$ nous appliquons la même opération de centrage :

$$\left(A^{(c)}\right)^{(c)} = A^{(c)} - \frac{1}{m}uu^T\left(A^{(c)}\right)$$

Nous cherchons à montrer que $\left(A^{(c)}\right)^{(c)} = A^{(c)}$.

Montrons donc que $\frac{1}{m}uu^T\left(A^{(c)}\right) = 0$.

Par définition nous avons $A^{(c)} = A - \frac{1}{m}uu^T A$.

Nous multiplions à gauche par u^T, nous obtenons donc ;

$$u^T A^{(c)} = u^T \left(A - \frac{1}{m}uu^T A\right) = u^T A - \frac{1}{m}u^T(uu^T A)$$

Or $u^T u = m$, donc :

$$u^T A^{(c)} = u^T A - \frac{1}{m}(mu^T A) = u^T A - u^T A = 0$$

Ainsi nous avons,

$$\begin{aligned} &\left(A^{(c)}\right)^{(c)} = A^{(c)} - \frac{1}{m}uu^T\left(A^{(c)}\right) \\ &\Leftrightarrow \left(A^{(c)}\right)^{(c)} = A^{(c)} - 0 \\ &\Leftrightarrow \left(A^{(c)}\right)^{(c)} = A^{(c)} \end{aligned}$$

La démonstration confirme que l'opération de centrage est idempotente, ce qui garantit la stabilité mathématique de la transformation, un aspect crucial pour les calculs numériques.

QUESTION 5.B.

ii.

Les variables peuvent avoir des échelles différentes, il est donc nécessaire de normaliser leurs amplitudes.

Ainsi nous introduisons une matrice $A \in \mathcal{M}_{mn}(\mathbb{R})$ que l'on pense comme un tableau de données. Les variables/paramètres mesurés représentés par les vecteurs colonnes, notons les colonnes de A : c_i, possèdent en général une dimension (des longueurs, vitesses, températures, etc..) et peuvent appartenir à des plages de valeurs difficilement comparables entre elles, c'est-à-dire les valeurs d'une colonne c_i peuvent être « très grandes » alors qu'une autre colonne c_j n'avoir que des valeurs « petites ».

L'opération de réduction consiste à adimensionner les valeurs des colonnes et à les normaliser entre -1 et 1. Concrètement cela revient à remplacer les colonnes c_i par :

$$c_i \rightsquigarrow \frac{1}{\|c_i\|_2} c_i$$

Notre but est d'obtenir une matrice diagonale $D \in \mathcal{M}_{mn}(\mathbb{R})$ à l'aide d'un produit :

$$A^{(r)} = AD$$

Pour modifier les colonnes d'une matrice A, on la multiplie par une matrice diagonale D. Afin de normaliser chaque colonne de A à une norme de 1, il suffit de diviser chaque colonne par sa norme. Ainsi, pour que toutes les colonnes aient une norme égale à 1, les éléments diagonaux k_i de la matrice doivent être définis comme suit :

$$k_i = \frac{1}{\|c_i\|_2}$$

Pour que la colonne i du produit AD devienne $c_i^{(r)} = \frac{1}{\|c_i\|_2}$, il suffit juste de fixer $D_{ii} = \frac{1}{\|c_i\|_2}$. De plus pour que toutes les autres cases hors de la diagonale de D soit nulles nous les fixons a 0. Ainsi nous obtenons pour la matrice D :

$$D = \begin{pmatrix} \frac{1}{\|c_1\|_2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\|c_2\|_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \frac{1}{\|c_n\|_2} \end{pmatrix}$$

Nous avons donc bien montré qu'il existe une matrice $D \in \mathcal{M}_{n,n}(\mathbb{R})$ diagonale telle que :

$$A^{(r)} = AD$$

Ainsi avec cette transformation cela nous assure que toutes les variables ont une variance comparable, empêchant les valeurs les plus grandes d'écraser l'analyse.

QUESTION 5.B.

iii.B.

Une fois la décomposition en valeurs propres effectuée, nous obtenons une liste de composantes principales ordonnées selon l'importance de leur contribution à la variance totale des données.

Cependant, conserver toutes les composantes n'est pas nécessaire. Nous devons déterminer combien de dimensions sont réellement utiles pour résumer l'information sans perte significative**.**

Nous avons :

$$\|A\|_F = \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{n} A_{i,j}^2} = \sqrt{\sum_{i=1}^{n} \|c_i\|_2^2}$$

Or nous avons établis précédemment que $\|c_i\|_2 = 1$.

On a :

$$\|A\|_F^2 = \sum_{i=1}^{n} \|c_i\|_2^2$$

$$\|A\|_F^2 = \sum_{i=1}^{n} 1 = n$$

$$\|A\|_F^2 = n$$

Nous obtenons donc :

$$\|A\|_F = \sqrt{n}$$

Nous pouvons noter la somme des valeurs propres de $A^T A$ de la manière suivante :

$$\sum_{i=1}^{n} \lambda_i = tr(A^T A)$$

Mais la trace de $A^T A$ se note également de la manière suivante :

$$tr(A^T A) = \sum_{k=1}^{n} (A^T A)_{k,k} = \sum_{k=1}^{n} \sum_{i=1}^{n} A_{i,k}^2 = \|A\|_F^2$$

Or nous venons juste d'établir que $\|A\|_F^2 = n$:

$$\sum_{i=1}^{rg(A)} \lambda_i = tr(A^T A) = \|A\|_F^2 = n$$

Ainsi choisir un nombre optimal de composantes permet d'éviter un surajustement et de réduire la complexité computationnelle, tout en maintenant une représentation fidèle des données.

Ainsi dans la suite nous nous intéresserons aux quantités $\frac{\lambda_i}{n}$ dont la somme vaut 1.

QUESTION 5.B.

iii.C.

Après avoir sélectionné le nombre de composantes principales, nous devons maintenant exprimer les données dans ce nouvel espace réduit.

Pour commencer introduisons R_{cr} la matrice « centrée-réduite » associée à la réalisation R et qui forme une réalisation de E_{cr}. Supposons que l'on a la diagonalisation suivante :

$$\sum = \frac{1}{m-1} R_{cr}^T R_{cr} = \frac{1}{m-1} V D^T V$$

Avec :

- V une matrice orthogonale où ses colonnes v_k sont les vecteurs propres de $\sum$.
- D une matrice diagonale où les coefficients sont rangés par ordre décroissant, tel que :

$$D := \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

Avec, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

Les directions principales v_k sont définis comme les vecteurs propres associés aux valeurs λ_k.

C'est-à-dire que les colonnes de V correspondent aux directions principales. De plus le vecteur v_k représente la direction dans laquelle la projection des données R_{cr} maximise la variance, avec la valeur λ_k représentant cette variance maximale.

D'autre part, notons Y_k les composantes principales obtenues en projetant les données R_{cr} sur les directions principales. Pour cela, notons :

$$Y_k = R_{cr} v_k$$

Nous pouvons écrire la variance de Y_k comme suit :

$$Var(Y_k) = \frac{1}{m-1} E\left(Y_k^2\right) = \frac{1}{m-1} \|Y_k\|_2^2 = \frac{1}{m-1} \|R_{cr} v_k\|_2^2$$

De plus, on sait que,

$$\|R_{cr} v_k\|_2^2 = v_k^T R_{cr}^T R_{cr} v_k$$

Nous obtenons,

$$Var(Y_k) = \frac{1}{m-1} \left(v_k^T R_{cr}^T R_{cr} v_k\right) = v_k^T \Sigma v_k$$

Nous savons aussi que v_k est un vecteur propre de Σ, tel que :

$$\Sigma v_k = \lambda\left(v_k^T v_k\right)$$

Donc,

$$Var(Y_k) = v_k^T (\lambda_k v_k) = \lambda_k \left(v_k^T v_k\right)$$

Or, nous savons que v_k est normalisé tel que $v_k^T v_k = 1$

Ainsi,

$$Var(Y_k) = \lambda_k$$

Centrer et réduire les données permet d'uniformiser leurs échelles et de préparer un terrain solide pour l'analyse en composantes principales.

SECTION 1.4 : PROJECTION DANS LA BASE DES DIRECTIONS PRINCIPALES

Une fois les données correctement transformées, il est possible de les exprimer dans la base des vecteurs propres, ce qui nous conduit à la prochaine section sur la projection dans l'espace des directions principales.

QUESTION 5.C.

Nous supposons dès à présents que nous avons calculé toutes les directions principales, c'est-à-dire que nous avons la matrice V à notre disposition. Prenons la base des directions principales $(v_i)_{i\in[\![1,n]\!]}$. Dans la notation d'ACP chaque ligne de R_{cr} est :

$$ligne\ i = (x_{i,1}, \dots, x_{i,n})$$

Avec $n \in \mathbb{N}$. Cette ligne est un vecteur dans l'espace $\mathbb{R}^n$.

Afin d'obtenir les coordonnées d'une ligne x_i dans une nouvelle base $\{v_1, \dots, v_n\}$ on effectue le produit suivant :

$$x_i V$$

Cela nous donne un nouveau vecteur-ligne qui contient les coordonnées de x_i dans la base $\{v_j\}$.

Ainsi pour regrouper l'ensemble de ces nouveaux vecteurs-lignes ça revient à faire le produit global au niveau matriciel :

$$R_{cr}V$$

Ainsi nous pouvons poser la matrice P qui sera l'expression de nos données dans la base $\{v_1, \dots, v_n\}$:

$$P = R_{cr}V$$

Ainsi nous supposons en entier $k \in [\![1, n]\!]$ fixé. Nous pouvons ainsi projetés dans l'espace vectoriel engendré par les vecteurs de directions principales les données centrées réduites R_{cr}. C'est-à-dire nous avons le vecteur $Vrc(v_1, \dots, v_k)$ en considérant les k-premieres colonnes de la matrice P précédente. Cette matrice se notera $proj_k(R_{cr})$.

La projection dans l'espace des directions principales est l'objectif final de l'ACP. Elle permet de simplifier l'interprétation des données tout en préservant leur structure fondamentale.

Pour « bien analyser » les données il nous reste à établir un critère sur le nombre de composantes principales. Pour cela nous reprenons notre variable Y_i qui est centrée-réduite, puis nous supposons que $m \gg n$.

Nous avons n variables aleatoires, $Y_i = (Y_1, \dots, Y_n)$.

Or comme Y_i est centrée-réduite, $\forall i \in [1, n], Var(Y_i) = 1$

Ainsi :

$$\sum_{i=1}^{n} Var(Y_i) = \sum_{i=1}^{n} 1 = n$$

Donc en divisant par n :

$$\frac{1}{n}\sum_{i=1}^{n} Var(Y_i) = \frac{n}{n} = 1$$

Ainsi d'un point de vue théorique nous avons :

$$\frac{1}{n}\sum_{i=1}^{n} Var(Y_i) = 1$$

Ce résultat est un « repère statistique », il nous permet de juger si une variable a une variance supérieure, inférieure ou égale à la moyenne globale des variances.

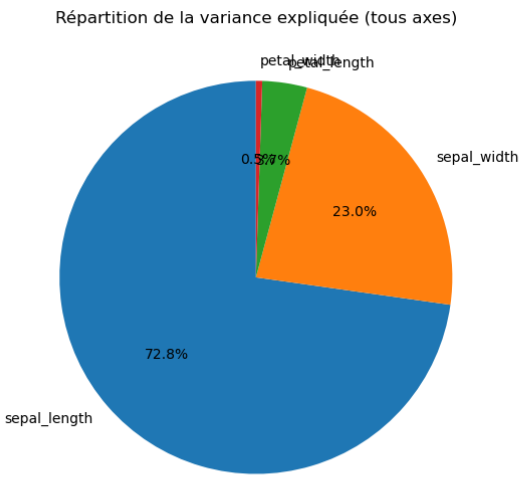
PARTIE 2 : APPLICATIONS PRATIQUES AVEC PYTHON

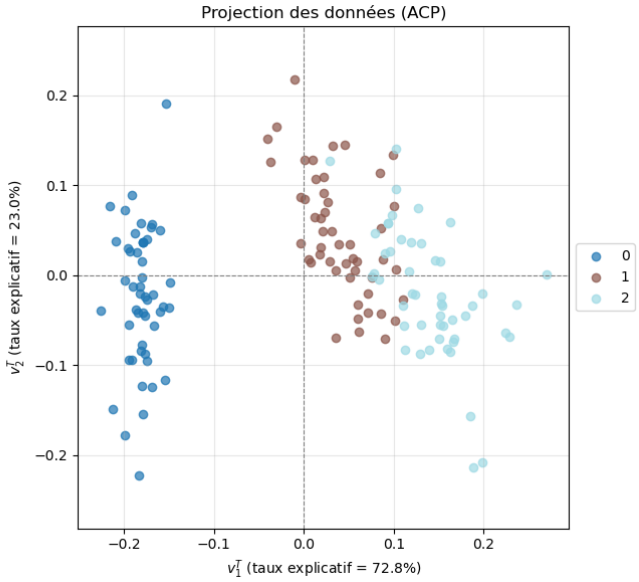
Après avoir établi toutes les bases théoriques, nous passons maintenant à la mise en œuvre pratique de ces concepts avec Python. Cela nous permettra d'analyser des jeux de données concrets et de visualiser les résultats.

Dans cette partie, nous mettons en œuvre les concepts théoriques développés dans la première partie. L'objectif est d'utiliser Python pour effectuer une ACP sur des jeux de données réels, en suivant les étapes théoriques : centrage, réduction, diagonalisation, et projection.

1. FICHIER IRIS.CSV

Cette ACP projette les données quadridimensionnelles telles que la longueur du sépale, la largeur du sépale, la longueur d'un pétale et la largeur d'un pétale sur un espace en 2 dimensions. Cette projection permet d'analyser ces données en détaille.



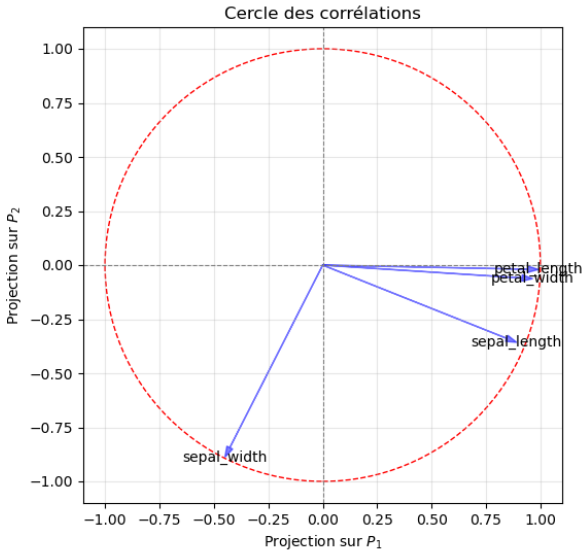


Dans le camembert de la variance nous observons plusieurs données intéressantes a relevés. La donnée $sepal\_length$ représente 72,8% de la variabilité sur le premier axe, cependant $sepal\_width$ exprime 23% de la variance selon le second axe. Les deux autres variables restantes, $petal\_length$ et $petal\_width$, représente très peu dans la construction des axes.

Ces axes sont représentés dans le schéma comportant le nuage des points, ce schéma est la projection ACP des données. Sur ce schéma sont représenté trois espèces différentes d'Iris, une espèce 0 (en bleue), une espèce 1(en marron) et une espèce 2 (en bleue ciel). Nous pouvons remarquer que l'axe 1 sépare nettement l'espèce bleu des autres espèces, pendant que l'axe 2 sépare médiocrement les espèces jaune et verte. Nous comprenons assez

rapidement qu'un individu ayant un $sepal\_length$ élevé se retrouve rapidement à droite du graphique, tandis que la variable $sepal\_width$ place les individus verticalement sur le graphique.

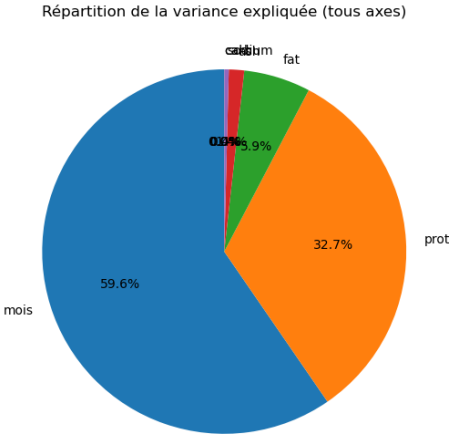
Concernant le cercle des corrélations nous pouvons dégager d'avantage des analyses sur la corrélation entre les différents facteurs qui dimensionne une plante d'Iris. Nous remarquons que $sepal\_length$ est très corrélé à l'axe 1 positivement tandis que $sepal\_width$ est corrélé négativement à l'axe 2. De plus $petal\_length$ et $petal\_width$ sont corrélé à l'axe 1 mais moins fortement que $sepal\_length$.

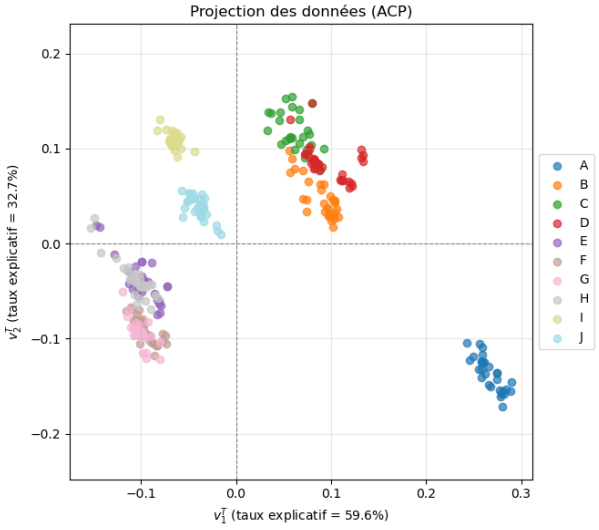


Nous pouvons également faire la remarque sur les angles qui séparent les différentes flèches. Un angle de 90 ° représente une faible corrélation, dans notre cas nous remarquons donc que $sepal\_width$ et $sepal\_length$ ont une faible corrélation, tandis que $petal\_length$ et $petal\_width$ ont une forte corrélation. Cela indique que pour avoir un pétale long il faut avoir une largeur de pétale grande et inversement. Au contraire un sépale long implique pas forcément un sépale large, de même inversement.

2. FICHIER PIZZAMOD

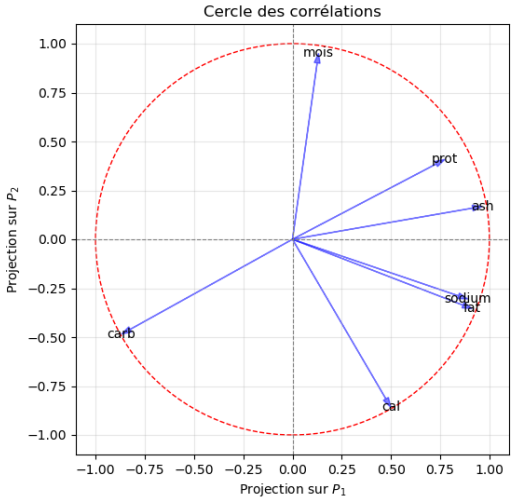
Dans ce cas nous utilisons les méthodes établies pour étudier les corrélations entre les nutriments présentes dans différentes marques de pizzas et leurs teneurs calorifiques. Les données utiliser pour effectuer cette analyse se trouvent dans le fichier $pizzamod.csv$.





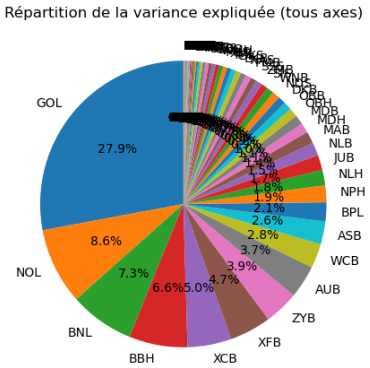
Dans ce camembert de variance nous remarquons que l'humidité ($mois$) domine le premier axe avec 59,6%, les protéines ($prot$) dominent le second axe avec 32,7% et finalement fat, ash, $sodium$, $calcium$ sont d'un rôle plus faible. Ces derniers représentent donc peu d'effet dans la construction des axes.

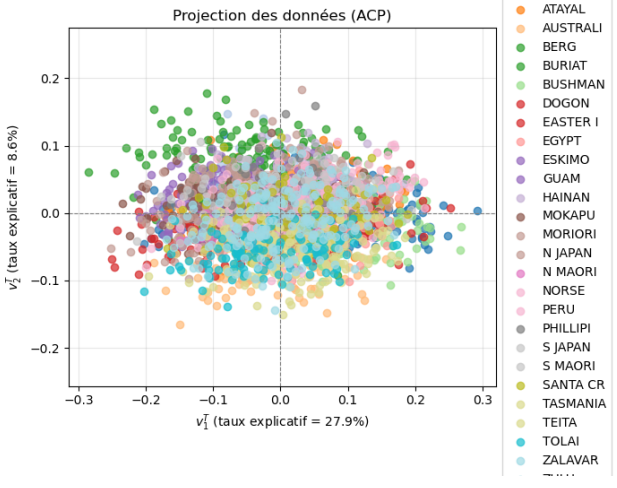
De la même manière dans le nuage des points nous avons plusieurs données, les différentes lettres correspondent à différentes marques de Pizza. Sur l'axe 1 les pizzas qui ont beaucoup d'humidité (variable $mois$ élevé) se placeront à gauche. Sur l'axe 2 les pizzas contenant beaucoup de protéines ($prot$) se positionneront de haut en bas. Nous pouvons donc observer plusieurs groupes bien séparés. Par exemple le groupe bleu se situe en bas à droite qui signifie peu d'humidité beaucoup de protéine.



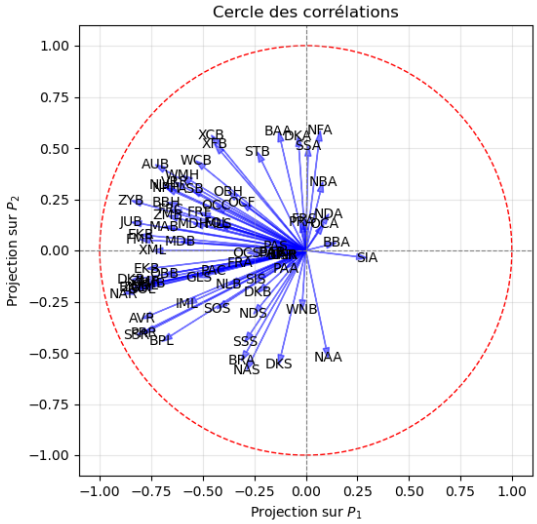
Avec le cercle des corrélations nous remarquons premièrement que l'humidité ($mois$) pointe quasiment vers le +1 de l'axe 1. Cela signifie son importance dans la construction de l'axe 1. Nous remarquons également après l'analyse des angles que le sel ($sodium$) et gras (fat) ont une forte corrélation entre elles. Tandis que les protéines ($prot$) ont une faible corrélation avec les calories (cal) car un angle de 90 ° les sépares, de même pour les calories (cal) et les glucides ($carb$). Ceci nous permet de dégager la conclusion suivante, les nutriments responsables a la teneur calorifique élevés sont le gras et le sel, tandis que les glucides, l'humidité et les protéines ne sont pas directement responsable a la teneur calorifique des pizzas. Nous remarquons également que la flèche représentant les protéines se situe à 180° de la flèche représentant les glucides, cela implique que les protéines sont inversement corrélées aux glucides, donc si on augmente les glucides nous baissons indirectement les protéines contenues dans les pizzas.

3. FICHIER HOWELLMOD





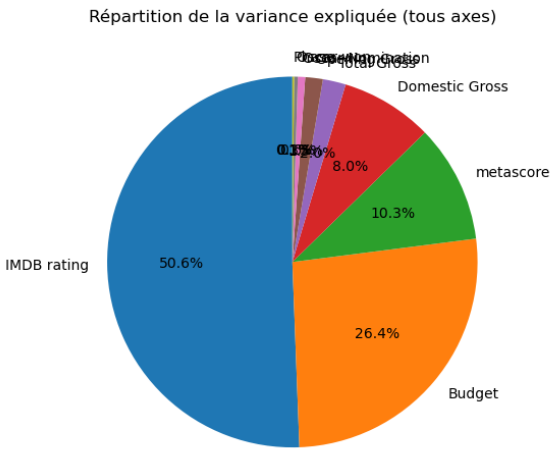
Le camembert révèle qu'aucune variable ne domine totalement, même si "GOL" explique presque 28 % de la variance. Les deux premiers axes de l'ACP (27,9 % et 8,6 %) ne totalisent qu'environ 36 % de la variance, on remarque donc que notre set de données est très partitionné et ne contient pas d'axe majeur en terme de variance. Sur le nuage de points, les groupes de populations (AINU, ANDAMAN...) se superposent.

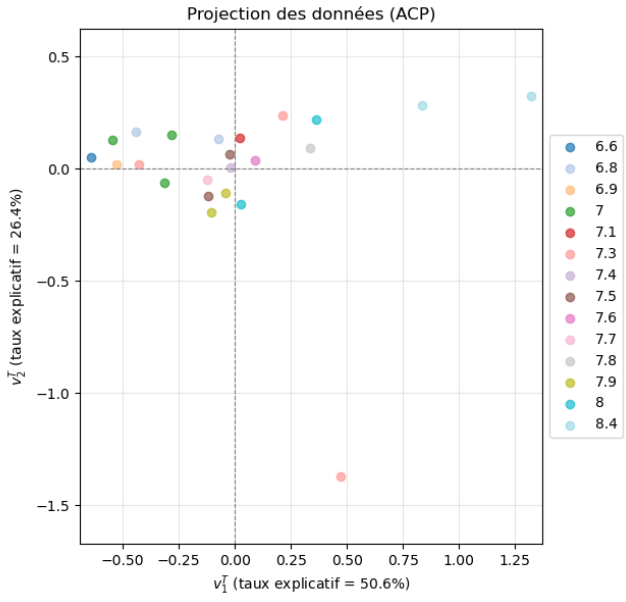


Concernant le cercle des corrélations, de nombreuses variables sont projetées, avec des flèches souvent regroupées au centre, signe qu'elles ne sont pas toutes bien exprimées sur les deux premiers axes et donc inutile ici. Néanmoins on remarque certains attroupement de données qui laissent voir qu'il existe des corrélations malgré tout cela.

Ainsi on en déduit que la variabilité est assez répartie sur plusieurs axes : on voit quelques tendances, mais il faudrait sans doute regarder les axes suivants ou isoler des groupes de données pour mieux comprendre la dispersion des populations et la corrélation de toutes ces variables. On atteint ici les limites de notre model bidimensionnel.

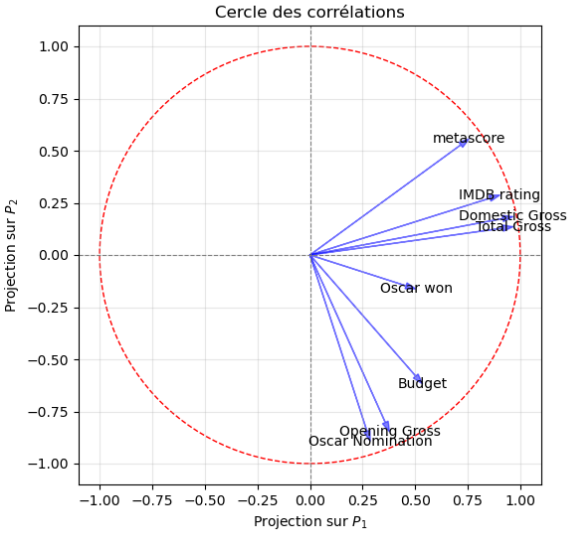
4. FICHIER MCU-DATASET





Dans la construction de ce camembert nous remarquons une répartition nette. Le IMDB rating contribue le plus dans la construction de l'axe 1 avec 50,3%. Le budget contribue majoritairement dans la construction du second axe avec 29,6%. Nous avons également d'autre données disponibles mais ces donnes ne contribuent pas assez dans la construction des axes.

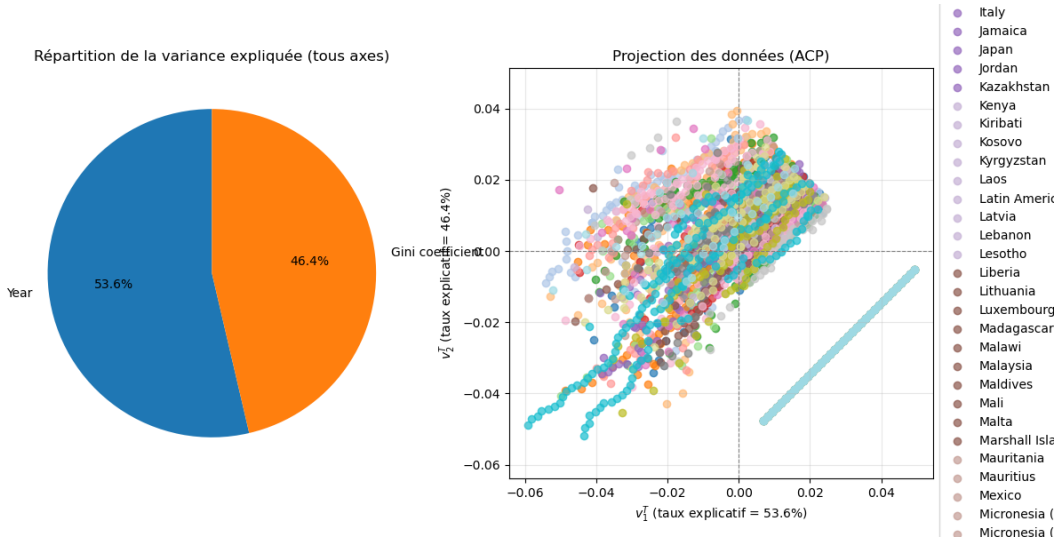
Les points colorés quant à eux correspondent aux différentes notes IMDB attribuées à différents films Marvel. On remarque que l'on ne peut pas réellement distinguer de groupe de données. Néanmoins on remarque une valeur qui est complétement hors du reste celle-ci est soit probablement très exceptionnel du à d'autres effets extérieurs à notre étude ou bien tout bonnement une erreur.



Sur le cercle des corrélations nous remarquons dès le début que IMDB rating, Total Gross, Domestic Gross et metascore sont groupée et pointes dans la même zone ainsi nous pouvons en déduire quelles sont positivement corrélés entre eux. Cependant nous remarquons des faits intéressants, le score donnés par le public (metascore et IMDB rating) n'est pas directement corrélé avec le budget (angle de 90 °) cela indique qu'un film recevant un bon score du public ne veut pas forcément dire que ce film a couté cher à produire.

5. FICHIER GINI-DATASET

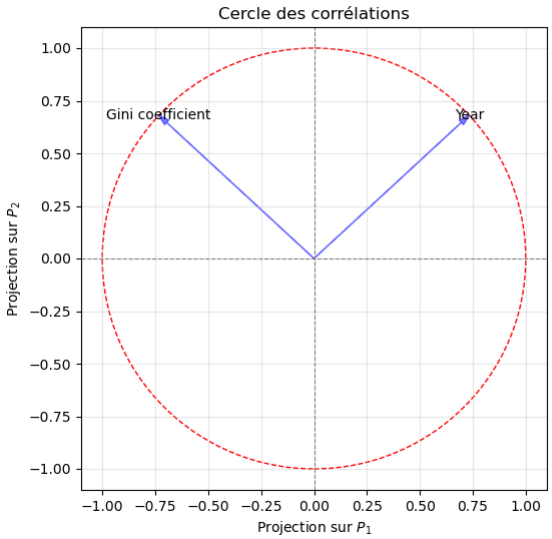
Dans ce cade d'étude nous allons nous intéresser au Gini coefficient en fonction de l'année pour différents pays. L'indice de Gini est un indicateur statistique qui évalue l'inégalité de la distribution des revenus au sein d'une population. Il varie de 0 (égalité parfaite) à 1 (inégalité extrême, où tous les revenus sont concentrés chez une seule personne ou un petit groupe).



Vis-à-vis du camembert, on observe deux variables : Year et Gini coefficient. Ce camembert indique une répartition en deux composantes principales, la variable Year qui est responsable de la construction de l'axe 1 avec une variance de 53,6% et la variable Gini coefficient qui de la construction de l'axe 2 avec une variance de 46,4%.

En s'intéressant maintenant au nuage de points nous remarquons que l'axe 1 sépare les observations en fonction des années, les années évolue de gauche à droite. L'axe 2 sépare les donnes selon leurs coefficients Gini, le coefficient évolue de bas en haut. Nous observons visuellement que le coefficient Gini évolue positivement en fonction des années qui passent.

Finalement le cercle des corrélations comporte deux flèches, une flèche qui représente le coefficient Gini et une autre représentant les années. Nous remarquons assez rapidement que les deux flèches sont séparées d'un angle se situant entre 90° et 120°, cela présente une très faible corrélation. Cela nous indique que l'évolution du coefficient Gini ne dépend pas directement des années qui passent.



En conclusion pour chacun de ces cinq cas on utilise globalement la même méthode d'analyse ACP. Nous étudions premièrement le camembert qui représente la répartition des variances expliquée. Ce camembert permet d'identifier les variables responsables a la construction des axes 1 et 2. Ces axes interviennent dans la prochaine étape de l'étude qui est l'analyse du nuage des points.

Ensuite, sur le graphique du nuage des points nous pouvons voir comment les différentes entrés se place par rapport aux axes établis dans l'étape précédente. Cela nous permet d'identifier différente entées qui appartiennent à la même famille ou de voir comment ces entrées se place par rapport aux axes paramétrés précédemment.

La dernière étape est la construction du cercle de corrélations qui indique comment les différentes variables se projette sur les axes et surtout comment les différentes variables se corrèle entre eux-mêmes. Nous pouvons identifier des corrélations entre les variables grâce aux angles qui séparent deux flèches représentatives, par exemple un angle de 90 ° indique une no corrélation entre variables tandis que si les flèches pointent globalement vers la même direction cela indique une corrélation entre variables.

Autrement dit, dans chaque cas, l'ACP fait ressortir les variables majeures et permet de visualiser rapidement la structure sous-jacente aux données : que ce soit l'effet de l'année vs. le Gini, la longueur des sépales chez l'iris, le rating IMDB d'un film, ou la teneur en eau d'un aliment.

CONCLUSION

L’analyse en composantes principales (ACP), explorée dans ce rapport, repose sur des fondements mathématiques rigoureux, étudiés en détail dans la première partie. Nous avons démontré que l’ACP cherche à maximiser la dispersion des données en identifiant les directions principales, c'est-à-dire les vecteurs propres de la matrice de covariance. Ces vecteurs sont obtenus en résolvant un problème d’optimisation où l’on cherche à maximiser la variance d’une projection linéaire sous contrainte de norme. Cette démarche nous a conduit à l'équation caractéristique $Cov(X)v = \lambda v$, liant vecteurs et valeurs propres.

Une étape essentielle de l’analyse a été de prouver que la matrice de covariance est symétrique et donc diagonalisable dans une base orthonormée. Cette propriété permet de projeter les données sur les axes définis par les vecteurs propres, simplifiant ainsi la compréhension des relations internes entre les variables. Par ailleurs, nous avons expliqué comment centrer et réduire les données pour garantir que les résultats de l’ACP ne soient pas biaisés par des différences d’échelles entre les variables.

Dans la partie pratique, nous avons appliqué ces concepts à différents jeux de données. Par exemple, dans le cas des données *iris.csv*, la projection sur les deux premières composantes principales a permis de distinguer clairement certaines espèces, illustrant la capacité de l’ACP à extraire des structures cachées. Le cercle des corrélations a montré que certaines variables, comme la longueur et la largeur des pétales, sont fortement corrélées et influencent de manière significative les premiers axes principaux.

Nous avons également étudié la réduction de dimension, démontrant qu’il est possible de conserver l'essentiel de l’information avec un nombre limité de composantes principales. Ce choix se justifie par la distribution décroissante des valeurs propres, qui reflète la part d’inertie expliquée par chaque composante.

Enfin, l’ACP s’est révélée être un outil puissant pour explorer des données complexes. Les résultats obtenus confirment que cette méthode permet de maintenir une grande part de la variance totale tout en simplifiant l’interprétation des données. Grâce à la mise en œuvre rigoureuse des étapes théoriques et pratiques, nous avons pu tirer des conclusions pertinentes sur la structure des jeux de données étudiés, démontrant ainsi l’efficacité de l’ACP dans les sciences des données.