Introduction

Purpose of PCA

L'Analyse en Composantes Principales (ACP) est une méthode classique d'analyse de données et de statistique multivariée, toujours largement utilisée aujourd'hui. Elle a été introduite par Karl Pearson en 1901 et formalisée par Harold Hotelling dans les années 1930 [**?**].

Objectifs principaux

- **Réduction de dimension** : L'ACP permet de transformer un tableau de données de dimension (n, p) en un tableau de dimension (n, q) avec $q < p$, en construisant q nouvelles variables, appelées composantes principales, qui capturent la majorité de l'information des p variables originales [**?**].
- **Élimination de la redondance** : Les composantes principales sont non corrélées, contrairement aux variables originales qui sont généralement corrélées [**?**].

Applications

L'ACP est utilisée dans divers domaines pour :

- **Résumer les données** : Identifier des similarités entre observations et relations entre variables.
- **Faciliter la visualisation** : Projeter les données dans un espace de dimension 2 ou 3 pour une représentation graphique [**?**].
- **Prétraiter les données** : Améliorer les performances des méthodes supervisées en réduisant le bruit et la redondance [**?**].
- **Réduire les coûts** : Diminuer les coûts de calcul, de stockage et d'acquisition des données [**?**].

Exemples de tableaux de données

Voici quelques exemples de tableaux de données où l'ACP est couramment appliquée :

- **Écologie** : Concentrations de polluants dans différents sites.
- **Économie** : Indicateurs économiques sur plusieurs années.
- **Génétique** : Expression de gènes chez des patients.
- **Biologie** : Mesures morphologiques pour différentes espèces ou variétés de plantes.
- **Marketing** : Indices de satisfaction client pour différentes marques.
- **Sociologie** : Répartition du temps selon les groupes socio-professionnels.
- **Analyse sensorielle** : Notes de descripteurs pour différents produits [**?**].

Toy example: athletic performance

Considérons un exemple concret avec un tableau de données représentant les performances athlétiques de 12 athlètes lors des Jeux Olympiques d'Athènes 2004, avec trois disciplines : 100 m, saut en longueur et lancer de disque.

Questions abordées par l'ACP

- **Étude des observations** : Identifier les profils d'athlètes. Peut-on regrouper les athlètes ayant des capacités physiques similaires (vitesse, endurance, force, etc.) ?
- **Étude des variables** : Analyser les relations (corrélations) entre les performances dans différentes disciplines. Certaines disciplines reposent-elles sur les mêmes qualités physiques ?
- **Création de nouvelles variables** : Peut-on créer de nouvelles variables qui résument plusieurs disciplines [**?**] ?

Principle of PCA

Projection orthogonale

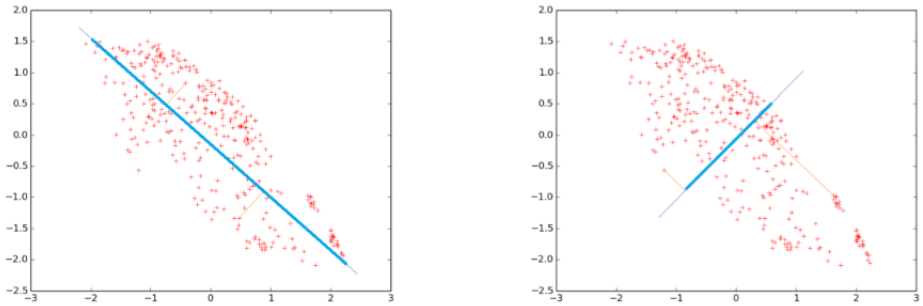
L'ACP consiste à définir une projection orthogonale sur un sous-espace E_q de $\mathbb{R}^p$ de dimension $q < p$. Pour ce faire, il est nécessaire de trouver une nouvelle base vectorielle de dimension q qui déforme le moins possible le nuage de points, c'est-à-dire qui préserve au maximum la dispersion des données (variance, inertie).

Composantes principales

Les q nouvelles variables obtenues par projection, appelées composantes principales, sont :

- Des combinaisons linéaires des p variables originales.
- Non corrélées, contrairement aux p variables originales qui sont généralement corrélées [**?**].

Example: Projection from $\mathbb{R}^2$ to $E_1 = \mathbb{R}$



Example: Projection from $\mathbb{R}^3$ to $E_2 = \mathbb{R}^2$

Data

Representation of data

Matrice X des données brutes

En ACP, les données sont quantitatives et organisées dans un tableau à n lignes et p colonnes, c'est-à-dire une matrice (n, p) :

- Chaque ligne correspond à une observation i, représentée par un vecteur $\mathbf{x}_i \in \mathbb{R}^p$.
- Chaque colonne correspond à une variable j, représentée par un vecteur $\mathbf{x}^j \in \mathbb{R}^n$.

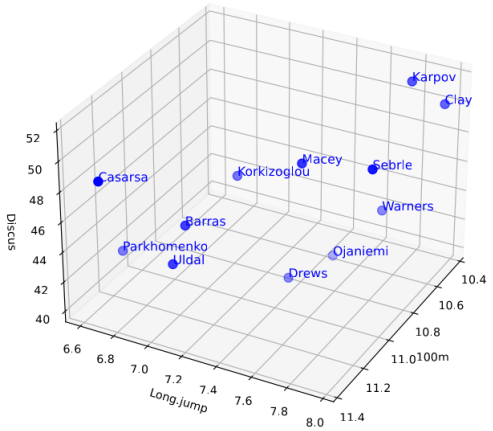
$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^j & \cdots & x_1^p \\ x_2^1 & x_2^2 & \cdots & x_2^j & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_i^1 & x_i^2 & \cdots & x_i^j & \cdots & x_i^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^1 & x_n^2 & \cdots & x_n^j & \cdots & x_n^p \end{pmatrix}$$

Nuage des observations

Chaque observation i est représentée par $\mathbf{x}_i = [x_i^1, x_i^2, \ldots, x_i^p]^T$:

- Vecteur des valeurs mesurées pour les p variables.
- Point dans $\mathbb{R}^p$, appelé espace des observations.

Les n observations forment un nuage de points, appelé **nuage des observations** (ne peut pas être représenté graphiquement si $p > 3$).



Nuage des variables

Chaque variable j est représentée par $\mathbf{x}^j = [x_1^j, x_2^j, \ldots, x_n^j]^T$:

- Vecteur des valeurs pour les n observations.
- Point dans $\mathbb{R}^n$, appelé espace des variables.

Les p variables forment un nuage de points, appelé **nuage des variables** (ne peut pas être visualisé si $n > 3$).

Pondération des observations

Dans certaines applications, il est utile de pondérer les observations selon leur importance relative :

- On assigne à chacune des n observations un poids $w_i > 0$ tel que $\sum_{i=1}^{n} w_i = 1$.
- La matrice des poids $\mathbf{D}_w$ est une matrice diagonale (n, n) :

$$\mathbf{D}_w = \text{diag}(w_1, w_2, \ldots, w_n) = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$$

- Cas le plus courant : poids uniformes → toutes les observations ont le même poids $w_i = \frac{1}{n}$, donc $\mathbf{D}_w = \frac{1}{n}\mathbf{I}_n$ où $\mathbf{I}_n$ est la matrice identité (n, n) [**?**].

Centering and scaling of data

Normalisation des données

L'ACP est toujours appliquée à des données centrées :

- Le centrage des variables déplace l'origine du système de coordonnées au centre du nuage de points sans changer sa forme.
- Les données peuvent également être réduites :
- La réduction des variables élimine l'effet des unités et standardise les magnitudes des différentes variables [**?**].

Moyenne empirique de la variable j

La moyenne empirique de la variable j est donnée par :

$$\bar{x}^j = \sum_{i=1}^{n} w_i x_i^j \quad \text{(somme pondérée de la colonne } j \text{ sur les } n \text{ observations)}$$

En notation matricielle : $\bar{x}^j = \mathbf{x}^j \mathbf{D}_w \mathbf{1}_n$ où $\mathbf{1}_n$ est le vecteur composé de n valeurs égales à 1.

Centroïde du nuage des observations

Le centroïde $\mathbf{g}$ du nuage des observations est le vecteur moyen des observations :

$$\mathbf{g} = \sum_{i=1}^{n} w_i \mathbf{x}_i = \begin{bmatrix} \bar{x}^1 \\ \vdots \\ \bar{x}^p \end{bmatrix}$$

En notation matricielle : $\mathbf{g} = \mathbf{X}^T \mathbf{D}_w \mathbf{1}_n$. $\mathbf{g}$ représente une observation moyenne de la population ; ses composantes sont les moyennes empiriques des variables.

Matrice Y des données centrées

La matrice $\mathbf{Y}$ des données centrées (colonnes de moyenne nulle) est donnée par :

$$\mathbf{Y} = \begin{pmatrix} y_1^1 & y_1^2 & \cdots & y_1^j & \cdots & y_1^p \\ y_2^1 & y_2^2 & \cdots & y_2^j & \cdots & y_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_i^1 & y_i^2 & \cdots & y_i^j & \cdots & y_i^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n^1 & y_n^2 & \cdots & y_n^j & \cdots & y_n^p \end{pmatrix}, \quad \text{avec} \quad y_i^j = x_i^j - \bar{x}^j$$

Variance empirique de la variable j

La variance empirique de la variable j est donnée par :

$$s_j^2 = \text{var}(\mathbf{x}^j) = \sum_{i=1}^{n} w_i (x_i^j - \bar{x}^j)^2 = \sum_{i=1}^{n} w_i (y_i^j)^2 = \text{var}(\mathbf{y}^j)$$

En notation matricielle : $s_j^2 = \mathbf{y}^j \mathbf{D}_w \mathbf{y}^j$.

Matrice Z des données standardisées

La matrice $\mathbf{Z}$ des données standardisées (centrées et réduites) est donnée par :

$$\mathbf{Z} = \begin{pmatrix} z_1^1 & z_1^2 & \cdots & z_1^j & \cdots & z_1^p \\ z_2^1 & z_2^2 & \cdots & z_2^j & \cdots & z_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_i^1 & z_i^2 & \cdots & z_i^j & \cdots & z_i^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_n^1 & z_n^2 & \cdots & z_n^j & \cdots & z_n^p \end{pmatrix}, \quad \text{avec} \quad z_i^j = \frac{y_i^j}{s_j} = \frac{x_i^j - \bar{x}^j}{s_j}$$

Covariance and correlation matrices

Covariance empirique des variables j et k

La covariance empirique entre les variables j et k est donnée par :

$$s_{jk} = \text{cov}(\mathbf{x}^j, \mathbf{x}^k) = \sum_{i=1}^{n} w_i (x_i^j - \bar{x}^j)(x_i^k - \bar{x}^k) = \sum_{i=1}^{n} w_i y_i^j y_i^k = \text{cov}(\mathbf{y}^j, \mathbf{y}^k)$$

En notation matricielle : $s_{jk} = \mathbf{y}^j \mathbf{D}_w \mathbf{y}^k$.

Matrice de covariance empirique V

La matrice de covariance empirique $\mathbf{V}$ est donnée par :

$$\mathbf{V} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1k} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2k} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{j1} & s_{j2} & \cdots & s_{jk} & \cdots & s_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pk} & \cdots & s_p^2 \end{pmatrix}$$

En notation matricielle : $\mathbf{V} = \mathbf{Y}^T \mathbf{D}_w \mathbf{Y}$.

Coefficient de corrélation empirique des variables j et k

Le coefficient de corrélation empirique entre les variables j et k est donné par :

$$r_{jk} = \text{cor}(\mathbf{x}^j, \mathbf{x}^k) = \frac{\text{cov}(\mathbf{x}^j, \mathbf{x}^k)}{\sqrt{\text{var}(\mathbf{x}^j)}\sqrt{\text{var}(\mathbf{x}^k)}} = \frac{s_{jk}}{s_j s_k} = \text{cor}(\mathbf{y}^j, \mathbf{y}^k)$$

Matrice de corrélation empirique R

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

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2k} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{j1} & r_{j2} & \cdots & r_{jk} & \cdots & r_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pk} & \cdots & 1 \end{pmatrix}$$

En notation matricielle : $\mathbf{R} = \mathbf{D}_{1/s}\mathbf{V}\mathbf{D}_{1/s} = \mathbf{Z}^T\mathbf{D}_w\mathbf{Z}$.

Propriétés des matrices de covariance et de corrélation

Les matrices de covariance et de corrélation sont des matrices carrées (p, p), symétriques et semi-définies positives. Une matrice $\mathbf{A}$ est dite semi-définie positive si pour tout vecteur non nul $\mathbf{u} \in \mathbb{R}^p$, $\mathbf{u}^T\mathbf{A}\mathbf{u} \geq 0$ [**?**].

Types d'ACP

On distingue deux types d'ACP :

- **ACP canonique (ou non standardisée)** : Analyse les données centrées $\mathbf{Y}$. Puisque $\mathbf{V}_Y = \mathbf{V}$, elle est basée sur la matrice de covariance.
- **ACP standardisée** : Analyse les données standardisées $\mathbf{Z}$. Puisque $\mathbf{V}_Z = \mathbf{R}$, elle est basée sur la matrice de corrélation.

Remarques

- La plupart du temps, l'ACP standardisée est préférée afin que chaque variable ait la même contribution. Cependant, si les variables partagent les mêmes unités, on peut préférer ne pas normaliser afin de tenir compte de leurs magnitudes relatives [**?**].
- L'ACP n'est pas invariante par changement d'échelle, donc les valeurs numériques diffèrent dans les deux cas.

Metric space - Reminders

Espace métrique

Pour caractériser la structure d'un nuage de points, il est nécessaire de pouvoir mesurer la proximité entre les points. Cela nécessite de définir un espace métrique, c'est-à-dire de définir une distance qui induit une géométrie.

Distance euclidienne

Soient $\mathbf{u}$ et $\mathbf{v}$ deux vecteurs non nuls de $\mathbb{R}^p$:

- Produit scalaire usuel : $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{j=1}^p u^j v^j = \mathbf{u}^T\mathbf{v}$.

- Norme euclidienne usuelle : $\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle} = \sqrt{\sum_{j=1}^{p} (u^j)^2} = \sqrt{\mathbf{u}^T \mathbf{u}}$.
- Distance euclidienne usuelle : $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\| = \sqrt{\sum_{j=1}^{p} (u^j - v^j)^2} = \sqrt{(\mathbf{u} - \mathbf{v})^T (\mathbf{u} - \mathbf{v})}$.

Cas général : $\mathbf{M} \neq \mathbf{I}_p$

Soit $\mathbf{M}$ une matrice carrée (p, p), symétrique, définie positive.

- Produit scalaire défini à partir de $\mathbf{M}$: $\langle \mathbf{u}, \mathbf{v} \rangle_M = \mathbf{u}^T \mathbf{M} \mathbf{v}$.
- Norme définie à partir de $\mathbf{M}$: $\|\mathbf{u}\|_M = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle_M} = \sqrt{\mathbf{u}^T \mathbf{M} \mathbf{u}}$.
- Distance définie à partir de $\mathbf{M}$: $d_M(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_M = \sqrt{(\mathbf{u} - \mathbf{v})^T \mathbf{M} (\mathbf{u} - \mathbf{v})}$.

Metric space of observations

Application au nuage des observations

On s'intéresse à la proximité entre les observations dans $\mathbb{R}^p$ afin de quantifier leur similarité. On choisit simplement la métrique $\mathbf{M} = \mathbf{I}_p$ (distance euclidienne).

Distance entre deux observations i et l

$$d(\mathbf{y}_i, \mathbf{y}_l) = \|\mathbf{y}_i - \mathbf{y}_l\| = \sqrt{(\mathbf{y}_i - \mathbf{y}_l)^T (\mathbf{y}_i - \mathbf{y}_l)}$$

Metric space of variables

Application au nuage des variables

On s'intéresse à la proximité entre les variables dans $\mathbb{R}^n$ afin de quantifier leur corrélation. On choisit la métrique $\mathbf{M} = \mathbf{D}_w$, la matrice diagonale des poids.

Produit scalaire et norme

Avec la métrique $\mathbf{M} = \mathbf{D}_w$:

- $\langle \mathbf{y}^j, \mathbf{y}^k \rangle_{D_w} = \mathbf{y}^j \mathbf{D}_w \mathbf{y}^k = \text{cov}(\mathbf{y}^j, \mathbf{y}^k) = s_{jk}$.
- $\|\mathbf{y}^j\|^2_{D_w} = \langle \mathbf{y}^j, \mathbf{y}^j \rangle_{D_w} = \mathbf{y}^j \mathbf{D}_w \mathbf{y}^j = \text{var}(\mathbf{y}^j) = s_j^2$.
- $d^2_{D_w}(\mathbf{y}^j, \mathbf{y}^k) = \|\mathbf{y}^j - \mathbf{y}^k\|^2_{D_w} = \text{var}(\mathbf{y}^j - \mathbf{y}^k)$.
- $\cos \theta_{D_w}(\mathbf{y}^j, \mathbf{y}^k) = \frac{\langle \mathbf{y}^j, \mathbf{y}^k \rangle_{D_w}}{\|\mathbf{y}^j\|_{D_w} \|\mathbf{y}^k\|_{D_w}} = \frac{s_{jk}}{s_j s_k} = \text{cor}(\mathbf{y}^j, \mathbf{y}^k) = r_{jk}$.

Inertia of the observation cloud

Inertie totale du nuage des observations

L'inertie totale du nuage des observations est définie comme la moyenne pondérée des carrés des distances entre les points et le centroïde :

$$I(\mathbf{X}) = \sum_{i=1}^{n} w_i d^2(\mathbf{x}_i, \mathbf{g}) = \sum_{i=1}^{n} w_i \|\mathbf{x}_i - \mathbf{g}\|^2 = \sum_{i=1}^{n} w_i \|\mathbf{y}_i\|^2 = I(\mathbf{Y})$$

Remarques

- L'inertie correspond à la généralisation de la variance au cas multivarié ; elle mesure la dispersion des données à travers les p variables.
- $I(\mathbf{Z}) = \sum_{i=1}^{n} w_i \|\mathbf{z}_i\|^2 \neq I(\mathbf{Y})$ car $\mathbf{z}_i = \mathbf{D}_{1/s} \mathbf{y}_i$.

Expression matricielle de l'inertie

$$I = \text{trace}(\mathbf{V})$$

Remarques sur la trace

La trace d'une matrice carrée est la somme de ses éléments diagonaux, qui est aussi égale à la somme de ses valeurs propres.

Inertie du nuage des observations projeté

L'objectif de l'ACP est de projeter les observations sur un sous-espace vectoriel de $\mathbb{R}^p \rightarrow$ On s'intéresse à l'inertie après projection, appelée inertie expliquée (ou portée) par le sous-espace.

Inertie expliquée par un sous-espace vectoriel

L'inertie du nuage des observations expliquée par un sous-espace F de $\mathbb{R}^p$ est l'inertie du nuage projeté sur F :

$$I_F(\mathbf{Y}) = \sum_{i=1}^{n} w_i \|P_F(\mathbf{y}_i)\|^2 = \sum_{i=1}^{n} w_i P_F(\mathbf{y}_i)^T P_F(\mathbf{y}_i)$$

où $P_F(\mathbf{y}_i)$ est la projection orthogonale de $\mathbf{y}_i$ sur F.

Cas particulier : Inertie expliquée par un sous-espace de dimension 1

Si le sous-espace F est une droite Δ_u engendrée par un vecteur normalisé $\mathbf{u}$ ($\|\mathbf{u}\| = 1$), alors l'inertie expliquée par $F = \Delta_u$ est donnée par :

$$I_{\Delta_u}(\mathbf{Y}) = \sum_{i=1}^{n} w_i \|P_u(\mathbf{y}_i)\|^2 = \mathbf{u}^T \mathbf{V} \mathbf{u}$$

PCA Method

Problem formulation

Objectif et principe de l'ACP

L'objectif principal de l'ACP est d'obtenir une représentation "la plus fidèle possible" du nuage de points dans un espace de dimension inférieure. Le critère de sélection du sous-espace est basé sur l'inertie.

Problème de l'ACP

On cherche le sous-espace E_q de dimension $q < p$ tel que l'inertie du nuage expliquée par E_q soit maximale :

$$E_q = \arg \max_{E, \dim(E)=q} I_E$$

E_q est appelé le sous-espace principal de dimension q. Cela revient à préserver au maximum la structure du nuage de points.

Solution to the problem

Théorème

Soit E_{k-1} un sous-espace de dimension $k-1 < p$ portant l'inertie maximale du nuage. Alors le sous-espace E_k de dimension k portant l'inertie maximale est obtenu de la manière suivante :

$$E_k = E_{k-1} \oplus \Delta_{u_k}$$

où Δ_{u_k} est la droite orthogonale à E_{k-1}, engendrée par le vecteur $\mathbf{u}_k$, portant l'énergie maximale.

Procédure de construction des sous-espaces principaux

1. $E_1 = \Delta_{u_1}$ où Δ_{u_1} est la droite qui maximise l'inertie expliquée $I_{\Delta_{u_1}}$.
2. $E_2 = E_1 \oplus \Delta_{u_2}$ où Δ_{u_2} est la droite orthogonale à E_1 qui maximise l'inertie expliquée $I_{\Delta_{u_2}}$.
3. ...
4. $E_q = E_{q-1} \oplus \Delta_{u_q}$ où Δ_{u_q} est la droite orthogonale à E_{q-1} qui maximise l'inertie expliquée $I_{\Delta_{u_q}}$.

Vecteurs et axes principaux

Les vecteurs $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ sont appelés vecteurs principaux, et les droites $\Delta_{u_1}, \Delta_{u_2}, \ldots, \Delta_{u_q}$ sont appelées axes principaux.

Computation of principal vectors

Calcul du premier vecteur principal $\mathbf{u}_1$

On cherche un vecteur normalisé $\mathbf{u}_1$ tel que l'inertie expliquée par la droite Δ_{u_1} engendrée par $\mathbf{u}_1$ soit maximale :

$$\mathbf{u}_1 = \arg \max_{\mathbf{u} \in \mathbb{R}^p, \|\mathbf{u}\|=1} I_{\Delta_u} = \arg \max_{\mathbf{u} \in \mathbb{R}^p, \mathbf{u}^T\mathbf{u}=1} \mathbf{u}^T \mathbf{V} \mathbf{u}$$

Ceci est un problème d'optimisation sous contrainte, résolu par les multiplicateurs de Lagrange :

$$(\mathbf{u}_1, \lambda_1) = \arg \max_{\mathbf{u} \in \mathbb{R}^p, \lambda \in \mathbb{R}} \mathbf{u}^T \mathbf{V} \mathbf{u} - \lambda(\mathbf{u}^T \mathbf{u} - 1)$$

Il peut être montré que $\mathbf{u}_1$ et λ_1 doivent satisfaire :

$$\mathbf{V}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1 \quad \text{et} \quad \mathbf{u}_1^T \mathbf{u}_1 = 1$$

Ainsi, λ_1 est une valeur propre de la matrice $\mathbf{V}$ et $\mathbf{u}_1$ est le vecteur propre normalisé associé à λ_1. L'inertie est alors :

$$I_{\Delta_{u_1}} = \mathbf{u}_1^T \mathbf{V} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1^T \mathbf{u}_1 = \lambda_1$$

Pour maximiser l'inertie, on choisit la plus grande valeur propre λ_1 de la matrice $\mathbf{V}$.

Premier vecteur principal $\mathbf{u}_1$ de l'ACP

Le premier vecteur principal $\mathbf{u}_1$ est le vecteur propre normalisé associé à la plus grande valeur propre λ_1 de la matrice de covariance $\mathbf{V}$. L'inertie expliquée par le premier axe principal Δ_{u_1} (engendré par $\mathbf{u}_1$) est :

$$I_{\Delta_{u_1}} = \lambda_1$$

Une procédure similaire est utilisée pour calculer les vecteurs principaux suivants $\mathbf{u}_k, k > 1$.

Théorème

Pour tout $q \leq p$, l'espace E_q de dimension q portant l'inertie maximale du nuage de points est engendré par les q vecteurs propres normalisés $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ associés aux q plus grandes valeurs propres de la matrice de covariance $\mathbf{V}$: $\lambda_1, \lambda_2, \ldots, \lambda_q$ (ordonnées par ordre décroissant). L'inertie expliquée par $E_q = \Delta_{u_1} \oplus \Delta_{u_2} \oplus \cdots \oplus \Delta_{u_q}$ est :

$$I_{E_q} = I_{\Delta_{u_1}} + I_{\Delta_{u_2}} + \cdots + I_{\Delta_{u_q}} = \lambda_1 + \lambda_2 + \cdots + \lambda_q$$

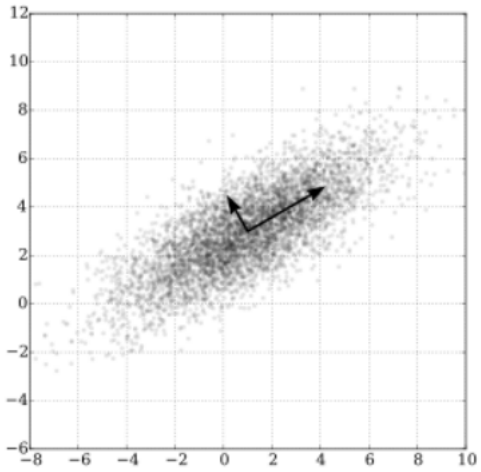
Remarques

- Si $q = p$, on a : $I_{E_p} = \sum_{j=1}^{p} \lambda_j = \text{trace}(\mathbf{V})$.
- $\mathbf{V}$ est symétrique et semi-définie positive. Par conséquent, toutes les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_p$ sont réelles et positives, et les vecteurs propres $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$ forment une base orthonormée de E_p.

Interprétation

Les vecteurs propres correspondent aux directions privilégiées de la matrice. Les valeurs propres représentent les facteurs multiplicatifs (étirement ou compression) dans ces directions.

- 1er vecteur propre : direction correspondant à la variance maximale.
- 2ème vecteur propre : direction orthogonale au 1er vecteur propre avec la plus grande variance restante.



Algorithme de l'ACP

Entrées

- Matrice de données (n, p).
- ACP canonique : données centrées $\mathbf{Y}$.
- ACP standardisée : données centrées et normalisées $\mathbf{Z}$.
- $\mathbf{D}_w$: matrice de poids (n, n) des observations.

Sorties

- $\lambda_1, \ldots, \lambda_p$: les valeurs propres (réelles et positives) de la matrice de covariance $\mathbf{V}$ classées par ordre décroissant.
- ACP canonique : $\mathbf{V} = \mathbf{V}_Y = \mathbf{Y}^T \mathbf{D}_w \mathbf{Y}$.
- ACP standardisée : $\mathbf{V} = \mathbf{V}_Z = \mathbf{R} = \mathbf{Z}^T \mathbf{D}_w \mathbf{Z}$.

- $\mathbf{u}_1, \ldots, \mathbf{u}_p$: les vecteurs principaux, c'est-à-dire les vecteurs propres normalisés associés aux valeurs propres.

PCA in practice

Nombre d'axes à retenir

Puisque l'objectif de l'ACP est la réduction de dimension, on souhaite conserver le moins d'axes possible. Plusieurs critères sont proposés dans la littérature pour choisir le nombre d'axes q :

- **Critère d'inertie** : Choisir le nombre d'axes q de sorte à retenir un ratio donné de l'inertie totale (ratio d'inertie expliquée) :

$$\frac{I_{E_q}}{I} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\sum_{j=1}^{p} \lambda_j} \geq \alpha$$

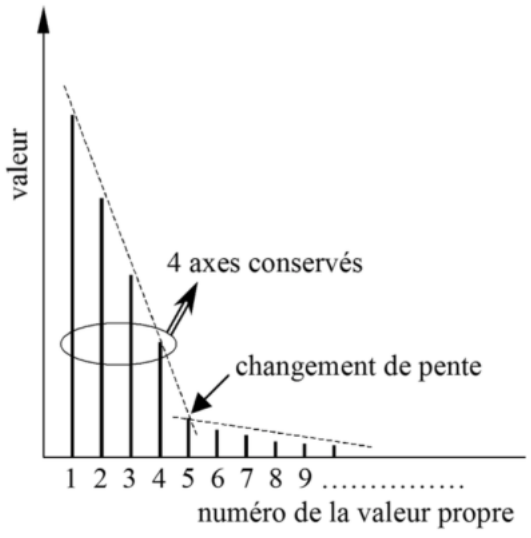
  où α est un seuil à fixer (le plus souvent entre 70% et 80%).

- **Règle de Kaiser** : Conserver les axes qui portent une inertie (valeur propre) supérieure à l'inertie moyenne par variable (moyenne des valeurs propres), c'est-à-dire choisir q tel que :

$$\lambda_q \geq \frac{I}{p} = \frac{\sum_{j=1}^{p} \lambda_j}{p} \quad \text{et} \quad \lambda_{q+1} < \frac{I}{p}$$

  Pour l'ACP standardisée, $\frac{I}{p} = 1$.

- **Courbe des valeurs propres (Scree plot)** : Tracer les valeurs propres par ordre décroissant (λ_j en fonction de l'indice j) et rechercher un "coude" dans la courbe. Conserver les axes associés aux valeurs propres situées avant le "coude" (le point où les valeurs propres commencent à se stabiliser, indiquant que les composantes suivantes expliquent peu de variance) [**?**].



Représentation graphique d'un nuage de points

Pour la visualisation, une représentation en 2 dimensions est généralement choisie :

- Pour le nuage des observations : projection sur un plan principal engendré par deux vecteurs principaux successifs $(\mathbf{u}_j, \mathbf{u}_{j+1}), j \leq p-1$.
- La meilleure représentation en deux dimensions est obtenue dans le premier plan principal $\Delta_{u_1} \oplus \Delta_{u_2}$.

- Pour le nuage des variables : projection sur un plan factoriel engendré par deux facteurs principaux successifs $(\mathbf{d}^k, \mathbf{d}^{k+1}), k \leq p-1$.
- La meilleure représentation en deux dimensions est obtenue dans le premier plan factoriel $\Delta_{d^1} \oplus \Delta_{d^2}$.

Qualité globale de la représentation

On s'intéresse à la qualité globale de la représentation obtenue par projection.

Qualité globale de la représentation d'un nuage de points sur un axe

La qualité globale de la représentation d'un nuage de points (observations ou variables) sur un axe Δ_i est mesurée par le ratio d'inertie expliquée par cet axe :

$$\frac{I_{\Delta_i}}{I} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

Plus la qualité est proche de 1, plus le nuage original (avant projection) est concentré autour de l'axe et moins il est déformé par la projection sur l'axe.

Qualité globale de la représentation sur un sous-espace de dimension $q > 1$

La qualité globale de la représentation du nuage sur le sous-espace principal (factoriel) E_q engendré par les q premiers vecteurs principaux (facteurs) est mesurée par le ratio d'inertie expliquée par E_q :

$$\frac{I_{E_q}}{I} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\sum_{j=1}^{p} \lambda_j}$$

Plus la qualité est proche de 1, plus le nuage original est concentré autour de E_q et moins il est déformé par la projection sur E_q.

Projection of observations and principal components

Coordinates of observations: principal components

Meilleure représentation en dimension $q < p$

La meilleure représentation en dimension $q < p$ du nuage des observations (en termes d'inertie) est obtenue par projection orthogonale sur le sous-espace principal E_q engendré par les q premiers vecteurs principaux $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$.

Coordonnée c_i^j de l'observation $\mathbf{y}_i$ sur l'axe principal Δ_{u_j} engendré par $\mathbf{u}_j$

$$c_i^j = \langle \mathbf{y}_i, \mathbf{u}_j \rangle = \mathbf{y}_i^T \mathbf{u}_j$$

La projection de l'observation i sur E_q est :

$$P_{E_q}(\mathbf{y}_i) = \sum_{j=1}^{q} c_i^j \mathbf{u}_j$$

Composante principale $\mathbf{c}^j$ sur l'axe principal Δ_{u_j}

La j-ème composante principale $\mathbf{c}^j$ est le vecteur dans $\mathbb{R}^n$ collectant les coordonnées des n observations sur l'axe Δ_{u_j} :

$$\mathbf{c}^j = \begin{bmatrix} c_1^j \\ \vdots \\ c_n^j \end{bmatrix} = \mathbf{Y}\mathbf{u}_j$$

Propriétés des composantes principales

- Les vecteurs $\mathbf{c}^j = \mathbf{Y}\mathbf{u}_j$ sont des combinaisons linéaires des variables originales $\mathbf{y}^j$.
- Les vecteurs $\mathbf{c}^j$ sont centrés, ont une variance λ_j, et sont deux à deux non corrélés (c'est-à-dire D_w-orthogonaux).
- Il est possible de normaliser les $\mathbf{c}^j$ pour former un système D_w-orthonormé dans l'espace des variables $\mathbb{R}^n$ (sur lequel le nuage des variables peut être projeté).

Facteur principal $\mathbf{d}^j$ sur l'axe Δ_{u_j}

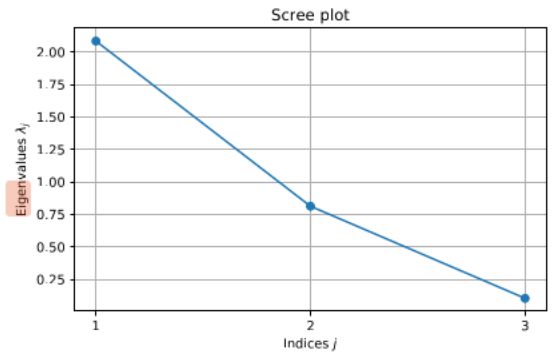
Le j-ème facteur principal est le vecteur dans $\mathbb{R}^n$ de variance 1 défini par :

$$\mathbf{d}^j = \frac{\mathbf{c}^j}{\sqrt{\lambda_j}}$$

Exemple de performances athlétiques (suite)

- Choice of the number of axes q :

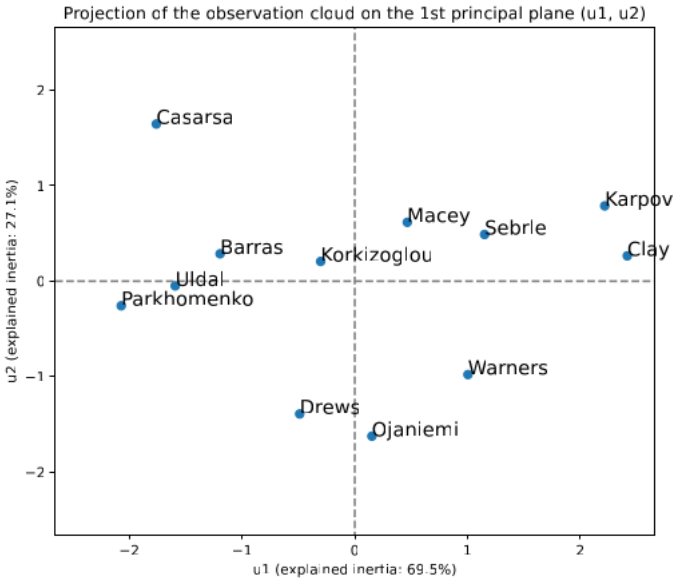
Indices j	Explained inertia: Eigenvalues λ_j	Explained inertia ratios: $\frac{\lambda_j}{\sum_{j=1}^{3}\lambda_j}$	Cumulative ratios: $\frac{\sum_{j=1}^{q}\lambda_j}{\sum_{j=1}^{3}\lambda_j}$
1	2.0838	0.6946	0.6946
2	0.8124	0.2708	0.9654
3	0.1038	0.0346	1



- Inertia criterion: $q = 2$ to retain more than 80% of total inertia.
- Kaiser rule: $q = 1$ (number of eigenvalues greater than 1).
- Scree plot: $q = 2$ (not very clear).

Next we retain the first 2 axes and project on the first principal plane spanned by the first 2 principal vectors.

Représentation des observations sur le premier plan principal



Projection of variables and correlation circle

Projection des variables

Coordonnées des variables

En ACP, les variables peuvent également être projetées sur les axes principaux. Cela permet de visualiser les relations entre les variables originales et les composantes principales.

Coordonnées de la variable $\mathbf{y}^j$ sur l'axe principal Δ_{u_k}

La coordonnée de la variable $\mathbf{y}^j$ sur l'axe principal Δ_{u_k} est donnée par le produit scalaire entre $\mathbf{y}^j$ et $\mathbf{u}_k$:

$$f_k^j = \langle \mathbf{y}^j, \mathbf{u}_k \rangle = {\mathbf{y}^j}^T \mathbf{u}_k$$

Interprétation

Les coordonnées f_k^j représentent la contribution de la variable $\mathbf{y}^j$ à la formation de la k-ème composante principale. Elles sont souvent utilisées pour interpréter les composantes principales en termes des variables originales.

Cercle des corrélations

Le cercle des corrélations est une représentation graphique des variables dans le plan principal. Chaque variable est représentée par un vecteur dont les coordonnées sont les corrélations entre la variable et les composantes principales.

Coordonnées des variables dans le cercle des corrélations

Les coordonnées des variables dans le cercle des corrélations sont données par :

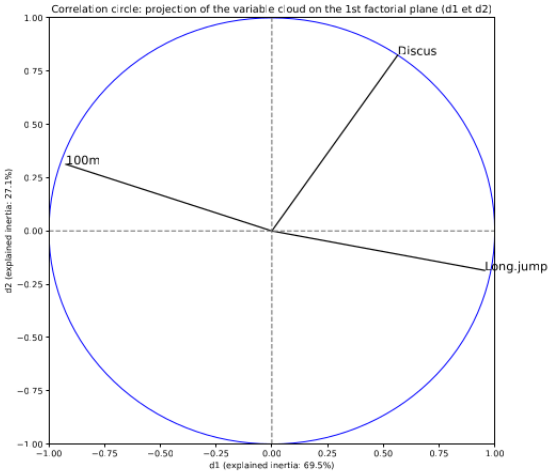
$$\text{cor}(\mathbf{y}^j, \mathbf{c}^k) = \frac{\text{cov}(\mathbf{y}^j, \mathbf{c}^k)}{\sqrt{\text{var}(\mathbf{y}^j)}\sqrt{\text{var}(\mathbf{c}^k)}} = \frac{f_k^j}{\sqrt{\lambda_k}}$$

où $\mathbf{c}^k$ est la k-ème composante principale.

Propriétés du cercle des corrélations

- Les variables sont représentées par des vecteurs dans le plan principal.
- La longueur d'un vecteur représente la qualité de la représentation de la variable sur le plan.
- L'angle entre deux vecteurs représente la corrélation entre les deux variables correspondantes.
- Si deux vecteurs sont proches, les variables correspondantes sont fortement corrélées.
- Si deux vecteurs sont orthogonaux, les variables correspondantes sont non corrélées.

Variables	on axis 1 $\tilde{\mathbf{c}}_1 = \sqrt{\lambda_1}\mathbf{u}_1$	on axis 2 $\tilde{\mathbf{c}}_2 = \sqrt{\lambda_2}\mathbf{u}_2$
100m	-0.9238	0.3129
Long jump	0.9549	-0.1849
Discus	0.5645	0.8248



Interprétation des résultats de l'ACP

Interprétation des composantes principales

Les composantes principales sont des combinaisons linéaires des variables originales. Chaque composante principale peut être interprétée en fonction des coefficients des variables originales.

Contribution des variables

La contribution d'une variable $\mathbf{y}^j$ à la formation de la k-ème composante principale est donnée par le carré de la coordonnée de la variable sur l'axe principal :

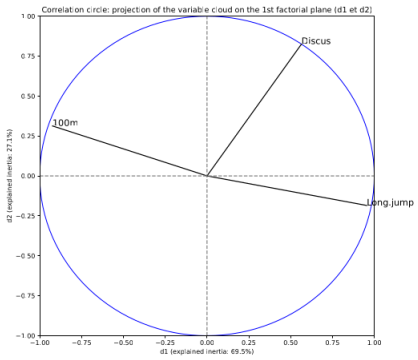
$$\mathrm{CTR}_k^j = \left(\frac{f_k^j}{\sqrt{\lambda_k}}\right)^2$$

Qualité de représentation des variables

La qualité de représentation d'une variable $\mathbf{y}^j$ sur le plan principal est donnée par la somme des carrés des coordonnées des variables sur les deux premiers axes principaux :

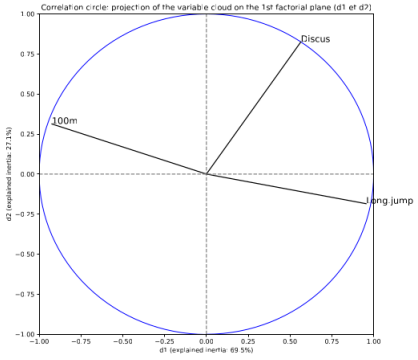
$$\mathrm{Qlt}^j = \left(\frac{f_1^j}{\sqrt{\lambda_1}}\right)^2 + \left(\frac{f_2^j}{\sqrt{\lambda_2}}\right)^2$$

Exemple de performances athlétiques (suite)



Variables contributing most to axis 1:

- (+) Long jump / (-) 100m
- Strong negative correlation between Long jump and 100m (opposite directions)



Variables contributing most to axis 2:

- (+) Discus
- Weak correlation between Discus and the 2 other variables (nearly orthogonal directions): different physical skills required by these disciplines

Exemple pratique : Interprétation des résultats

Analyse des composantes principales

Pour l'exemple des performances athlétiques, les deux premières composantes principales expliquent une grande partie de la variance totale. Voici comment interpréter ces résultats :

- **Première composante principale** : Elle est fortement corrélée avec les performances en saut en longueur et en 100 m. Cela suggère que cette composante représente principalement la vitesse et la puissance des athlètes.
- **Deuxième composante principale** : Elle est fortement corrélée avec les performances en lancer de disque. Cela suggère que cette composante représente principalement la force des athlètes.

Visualisation des résultats

La visualisation des observations et des variables sur le premier plan principal permet de voir les regroupements d'athlètes et les relations entre les variables.

Conclusion

Synthèse

L'ACP est une méthode puissante pour réduire la dimension des données tout en conservant l'information essentielle. Elle permet de :

- Visualiser les données multidimensionnelles en 2D ou 3D.
- Identifier les relations entre les variables.
- Résumer les données en un nombre réduit de composantes principales.

Limitations

Bien que l'ACP soit très utile, elle a certaines limitations :

- Elle suppose que les relations entre les variables sont linéaires.
- Elle est sensible à l'échelle des variables, d'où l'importance de standardiser les données si les variables ont des unités différentes.
- Elle peut être difficile à interpréter si les composantes principales ne sont pas clairement associées à des concepts physiques ou théoriques.

Références et lectures complémentaires

Pour approfondir vos connaissances sur l'ACP, voici quelques références utiles :

- **Livres** :
  - Jolliffe, I. T. (2002). Principal Component Analysis. Springer.
  - Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459.
- **Articles** :
  - Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, 2(11), 559-572.
  - Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology, 24(6), 417-441.
- **Ressources en ligne** :
  - `https://en.wikipedia.org/wiki/Principal_component_analysis`
  - `https://builtin.com/data-science/step-step-explanation-principal-component-analysis`
  - `https://www.geeksforgeeks.org/data-analysis/principal-component-analysis-pca/`