Analyse de données

Gwenlake

Présentation

Objectifs du cours

- Rappels sur le contexte et les fondements de l’Analyse en composante principale (ACP)
- Points clés mathématiques
- Choix du nombre d’axes principaux
- Interprétations

Analyse en composantes principales

Contexte

- Problème de départ : trop de variables, trop de corrélations

Exemple: Si vous mesurez la taille, le poids et la pointure d'une personne, vous avez trois informations, mais elles ne sont pas totalement différentes. Le poids et la pointure dépendent en grande partie de la taille. L'ACP est l'outil qui démêle ces informations redondantes pour n'en garder que l'essentiel.

Objectifs de l'ACP:

- Synthétiser l'information : Réduire la complexité des données en extrayant l'essentiel.
- Éliminer la redondance : Créer de nouvelles variables synthétiques et décorrélées (les composantes principales) à partir des variables d'origine.
- Réduire la dimensionnalité : Passer d'un grand nombre de variables à quelques composantes clés pour simplifier l'analyse.
- Faciliter la visualisation : Rendre les données facilement représentables sur un graphique en 2D ou 3D.

Analyse en composantes principales

Cadre théorique et rappels

- Soit **X** une matrice de données brutes, avec **n individus** (lignes) et **p variables** (colonnes). Ses dimensions sont **(n,p)**.
- Standardiser X consiste à centrer-réduire la matrice. *i.e*: Chacune de ses colonnes a une moyenne de 0 et un écart-type de 1.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

- Variance d'une variable X: quantifie la dispersion de la variable autour de sa moyenne $V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2$
- Corrélation entre 2 variables X et Y : Mesure la relation linéaire entre deux variables X et Y $\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$

Remarques

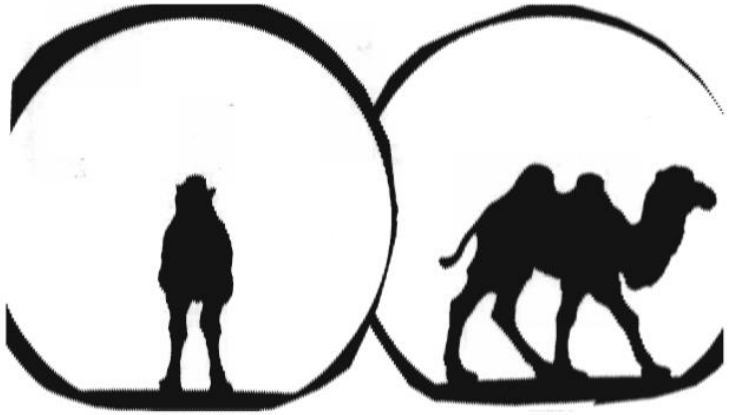
La corrélation entre deux variables est égale à la covariance de ces mêmes variables une fois centrées-réduites.

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{1 \cdot 1} = \mathrm{Cov}(X, Y)$$

Analyse en composantes principales

Modélisation du problème

- Le but de l'ACP est de réduire la dimensionnalité en projetant le nuage de points initial sur un nouvel axe, puis un deuxième, et ainsi de suite. L'enjeu est de choisir ces axes de manière à perdre le moins d'information possible lors de la projection.
- Exemple: trouver la meilleure photo 2D d’un objet 3D.
- En mathématiques, cette "photo" s'appelle une projection. C'est exactement comme projeter l'ombre de l'objet sur un mur. Notre question est : d'où devons-nous éclairer l'objet pour obtenir l'ombre la plus utile ?
- **Objectif** : trouver des nouveaux axes (variables) qui maximise la variance quand on projette nos données dessus

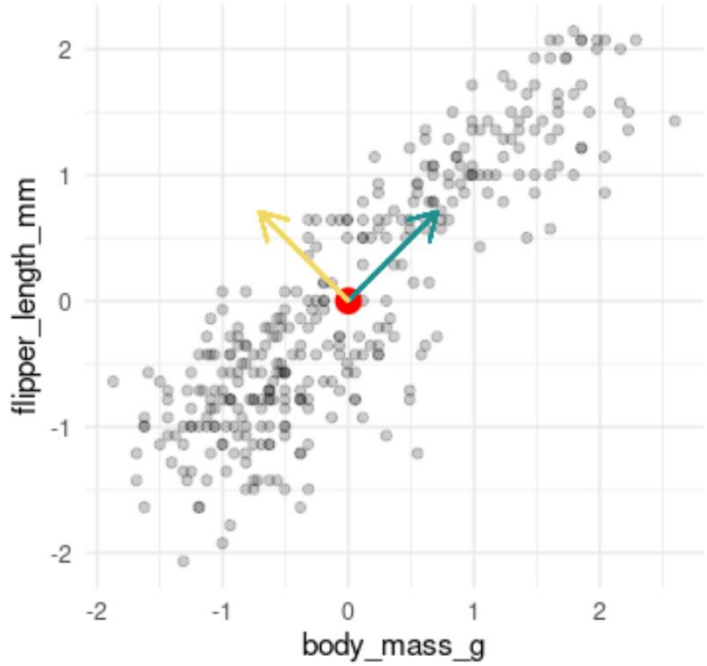


Quelle est la meilleure représentation du chameau?

Analyse en composantes principales

Modélisation du problème

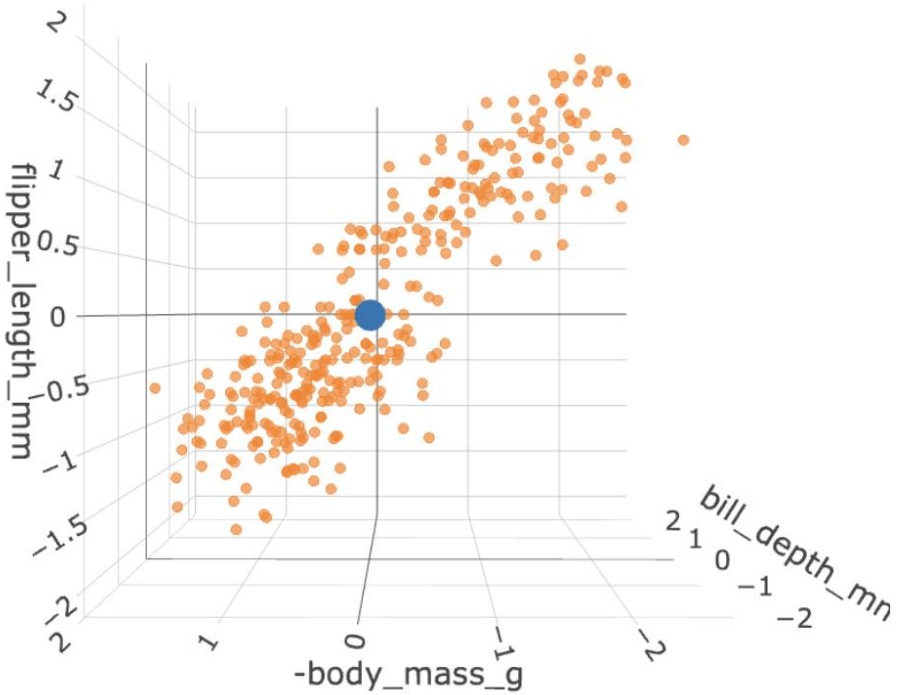
- La solution au problème est donnée par les vecteurs propres de la matrice de corrélation R.
- Qu'est-ce qu'un vecteur propre de R ? C'est une direction "spéciale" que la matrice R se contente d'étirer, sans jamais la faire tourner.
- Les nouveaux axes sont donc les vecteurs propres de R, une combinaison linéaire des variables.



Analyse en composantes principales

Une mesure de la quantité d'information

- Les données étant centrées, le point de coordonnées (0, …, 0) est le centre du nuage de points.
- **L'inertie** quantifie l'information/la variance portée par le nuage.
- En 1 dimension, l'inertie est la variance. En p dimension, c'est la somme des variances.



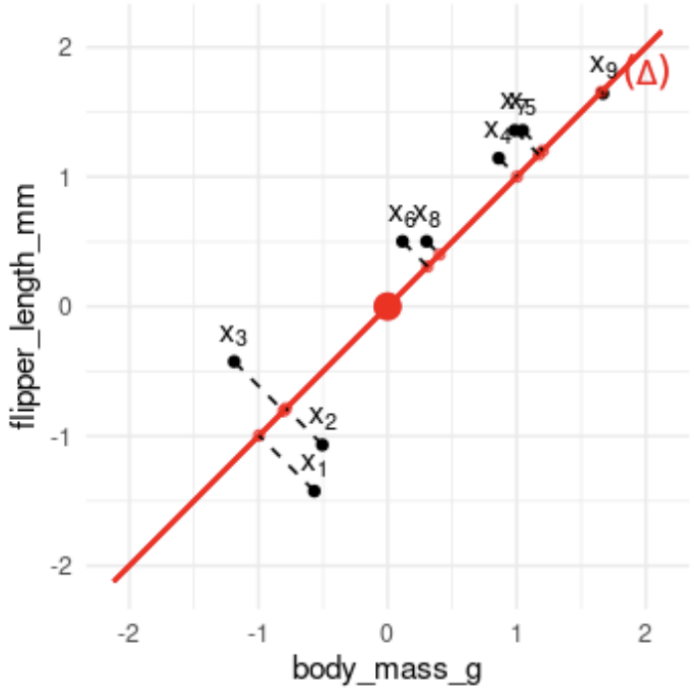
Remarque

Dans le cas de variables réduites (variance =1), **I=p**

Analyse en composantes principales

Une mesure de la quantité d'information

- La quantité d’information gardée lorsqu’on projette sur notre nouvel axe u est donnée par: I = Var(Projection sur u).
- Dans le cas où l’axe est le vecteur propre de la matrice de corrélation, la variance = la valeur propre.



Analyse en composantes principales

Sélection du nombre d'axes

- Le nombre de vecteurs propres lorsque nous faisons est en général égal au nombre de variables.
- Mais tous les axes ne sont pas utiles, car ne gardent pas beaucoup d'information (valeur propre proche de 0).

Le critère de Kaiser (simple mais approximatif):
On ne garde que les axes avec une valeur propre > 1

Le critère de la variance expliquée (méthode la plus courante)
On choisit le plus petit k tel que : Variance cumulée > seuil choisi (ex : 80 % ou 90 %)

Le critère du coude:
On trace la courbe :

- Axe des x : numéro des composantes
- Axe des y : valeurs propres (ou pourcentages de variance expliquée)

On cherche le "coude" de la courbe, là où la décroissance ralentit fortement.

Analyse en composantes principales

Interprétation des variables

- Une fois nos axes choisis (par exemple CP1 et CP2), il faut leur donner un sens. Pour cela, on interprète toujours deux axes en même temps, avec le **cercle des corrélations.**
- Le cercle est la carte des variables. Chaque flèche représente une variable.
- Les coordonnées (x,y) d'une flèche correspondent à sa corrélation avec l'Axe 1 et l'Axe 2 (valeur entre -1 et 1).

Comment le lire ?

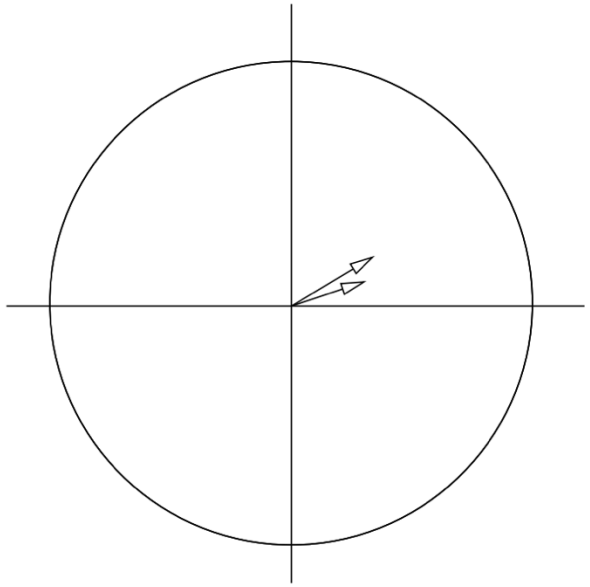
- **Flèche longue (proche du bord)** : la variable est bien représentée.
- **Angle faible entre deux flèches** : variables corrélées positivement.
- **Angle de 180°** : variables corrélées négativement.

Remarque

Le nom d'un axe est le concept commun aux variables qui lui sont les plus alignées.
Exemple : Poids et Taille alignés sur l'axe horizontal -> l'Axe 1 représente la "Corpulence".

Analyse en composantes principales

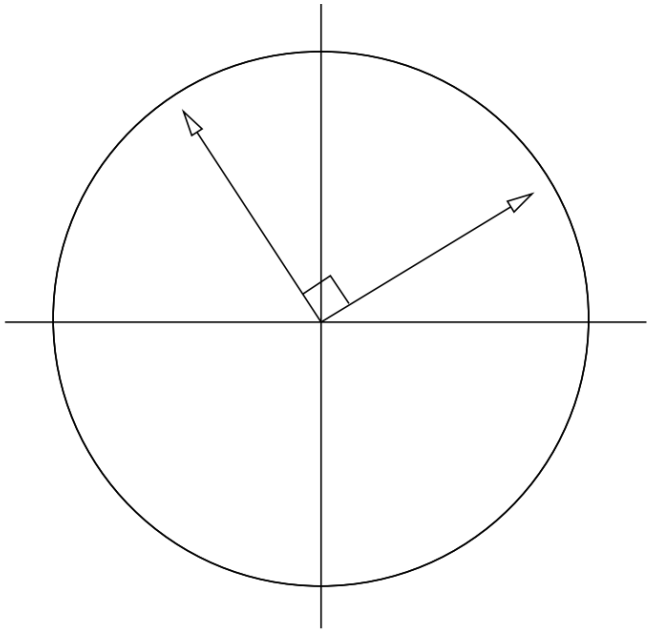
Interprétation des variables : cas particulier

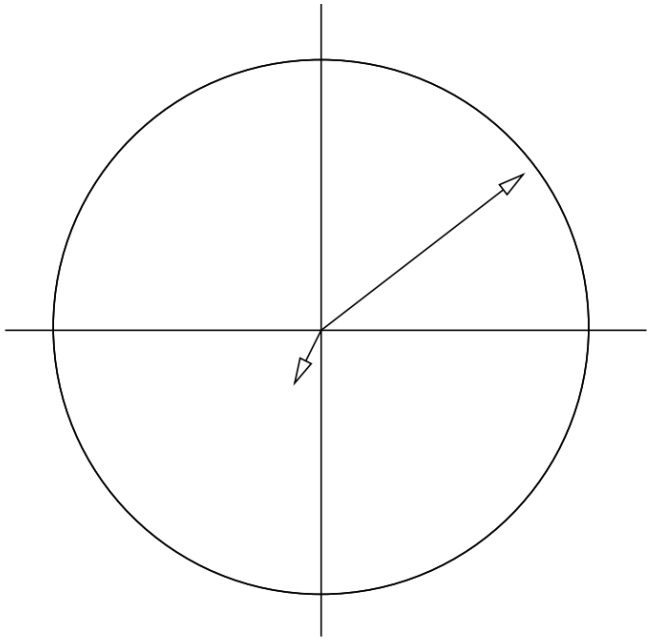


Pas d’interprétation, les variables de ne sont
Pas assez corrélées avec les nouveaux axes

Analyse en composantes principales

Interprétation des variables : cas particulier

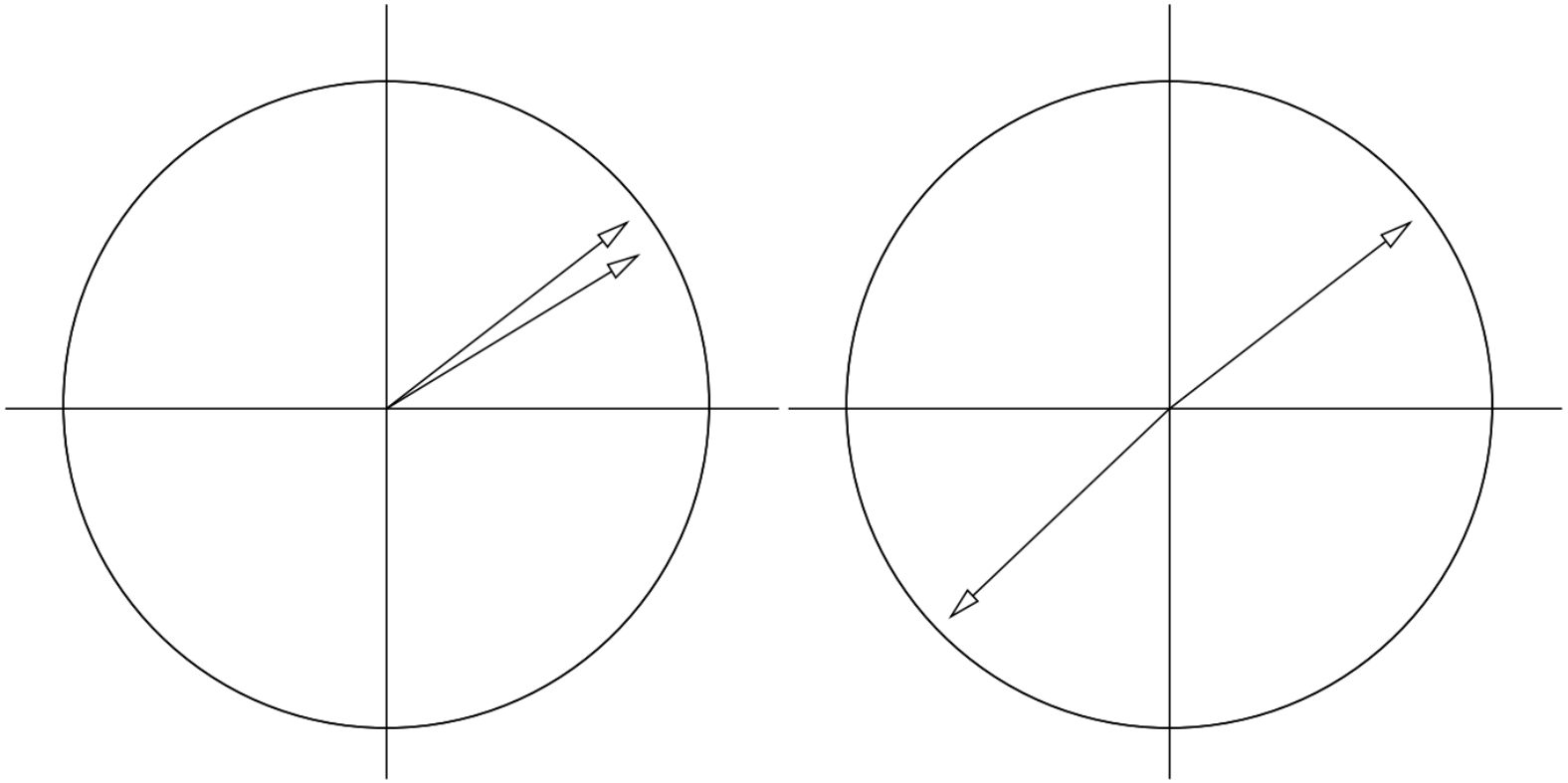




Les variables sont complétement décorrélées.

Analyse en composantes principales

Interprétation des variables : cas particulier

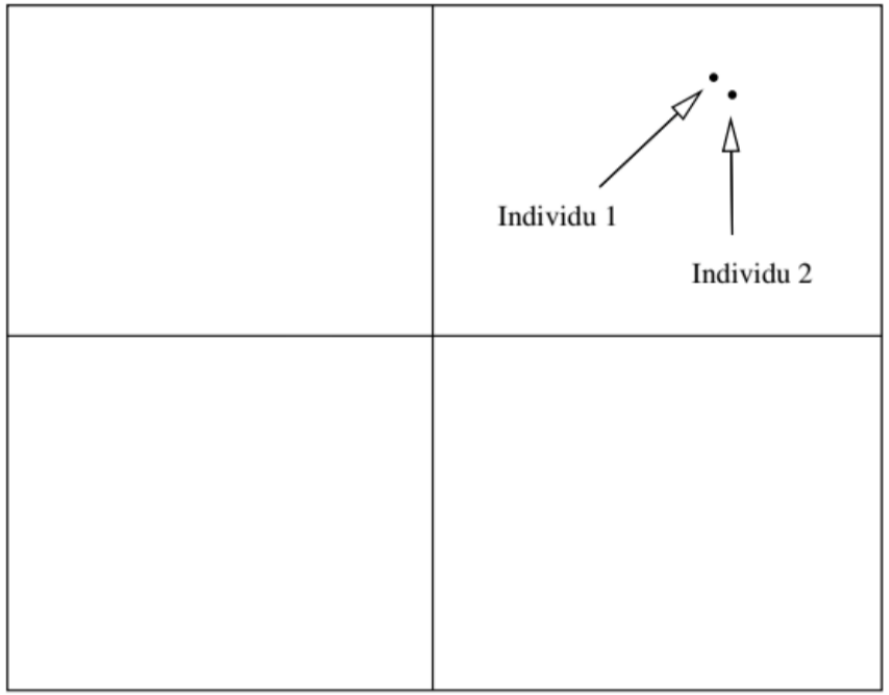


Les variables sont corrélées positivement dans le même sens et négativement dans le sens opposé.

Analyse en composantes principales

Interprétation des individus

On projette les individus dans le plan de deux axes de l'ACP.
Proximité = Similarité : Deux individus proches sur le graphique sont similaires.



Analyse en composantes principales

Interprétation des individus superposés aux variables

Un individu situé en haut à droite sur le graphe des individus sera fortement caractérisé par les variables qui se trouvent en haut à droite sur le cercle des corrélations.