Jour 34
#100JoursDeStat&ML
Isabelle LACMAGO

STAT&ML
POUR TOUS

L'ACP:

ANALYSE EN COMPOSANTES PRINCIPALES

DÉFINITION
UTILITÉ
OBJECTIF
CONSTRUCTION DES VARIABLES PERTINENTES

L'ACP:

- **Données non étiquetées:**

  → p variables explicatives quantitatives, n observations

- **Méthode d'apprentissage statistique non supervisée**

  ☞ Découvrir une structure

  ☞ Trouver un moyen simple et clair de représenter les données quand p est élevé

  ☞ Difficulté à vérifier les résultats

UTILITÉ ACP:

- **Data visualisation: r**eprésentation dans un espace de dimension plus faible tout en conservant le maximum d'informations
  - ✓Variables
  - ✓Observations

- **Data pré-processing**

→ Description des données: *comprendre la base de données et détecter les valeurs anormales*

→ Réduire les variables (regrouper) avant de mettre en œuvre certaines méthodes d'apprentissage supervisée sensible au fléau de la dimensionnalité: Modèle de régression, KNN

→ Imputation des valeurs manquantes (de type aléatoires)

Chaque individu de la population est caractérisé par p valeurs.

Les variables n'ont pas la même importance et peuvent être corrélées.

Principe de l'ACP

Construire un faible nombre de variables dérivées appelées **composantes principales**. À partir des combinaisons linéaires pondérées des p variables initiales.

CONSTRUCTION SUCCESSIVE DES COMPOSANTES PRINCIPALES :

- Avant tout:

→**centrer** les variables explicatives

  ✓ Pour avoir des variables de moyenne nulle

→**standardiser** (centrer et réduire) si elles n'ont pas la même unité de mesure.

  ✓ Pour avoir des variables de moyenne nulle et de variance 1

- Construire les composantes principales de la plus pertinente à la moins pertinente.

→ On note min(p,n-1) composantes principales

- **Première composante principale:** $Z_1 = \sum_{j=1}^{P} \varphi_{j1} X_j ; \sum_{j=1}^{P} \varphi_{j1}^2 = 1$

→c'est la combinaison linéaire des variables explicatives ayant la plus grande variance

- **Deuxième composante principale:** $Z_2 = \sum_{j=1}^{P} \varphi_{j2} X_j ; \sum_{j=1}^{P} \varphi_{j2}^2 = 1$

→ C'est la combinaison linéaire des variables explicative *qui est décorrélé de la première* composante principale et ayant la plus grande variance.

- **k ième composante principale:** $Z_{\mathrm{k}} = \sum_{j=1}^{P} \varphi_{jk} X_j ; \sum_{j=1}^{P} \varphi_{jk}^2 = 1$

→ C'est la combinaison linéaire des variables explicative qui est *décorrélé des (k-1) premières* composantes principales et ayant la plus grande variance.

- Une composante principale k sera donc ***complètement caractérisée*** par
  - ✓les coefficients $\varphi_{\mathrm{k}} = (\varphi_{1k}, \ldots, \varphi_{pk})$ et
  - ✓sa variance (proportion de la variance expliquée).

- La première composante principale est vu comme:
  - ✓*une ligne de dimension p, la plus proche des observations.*

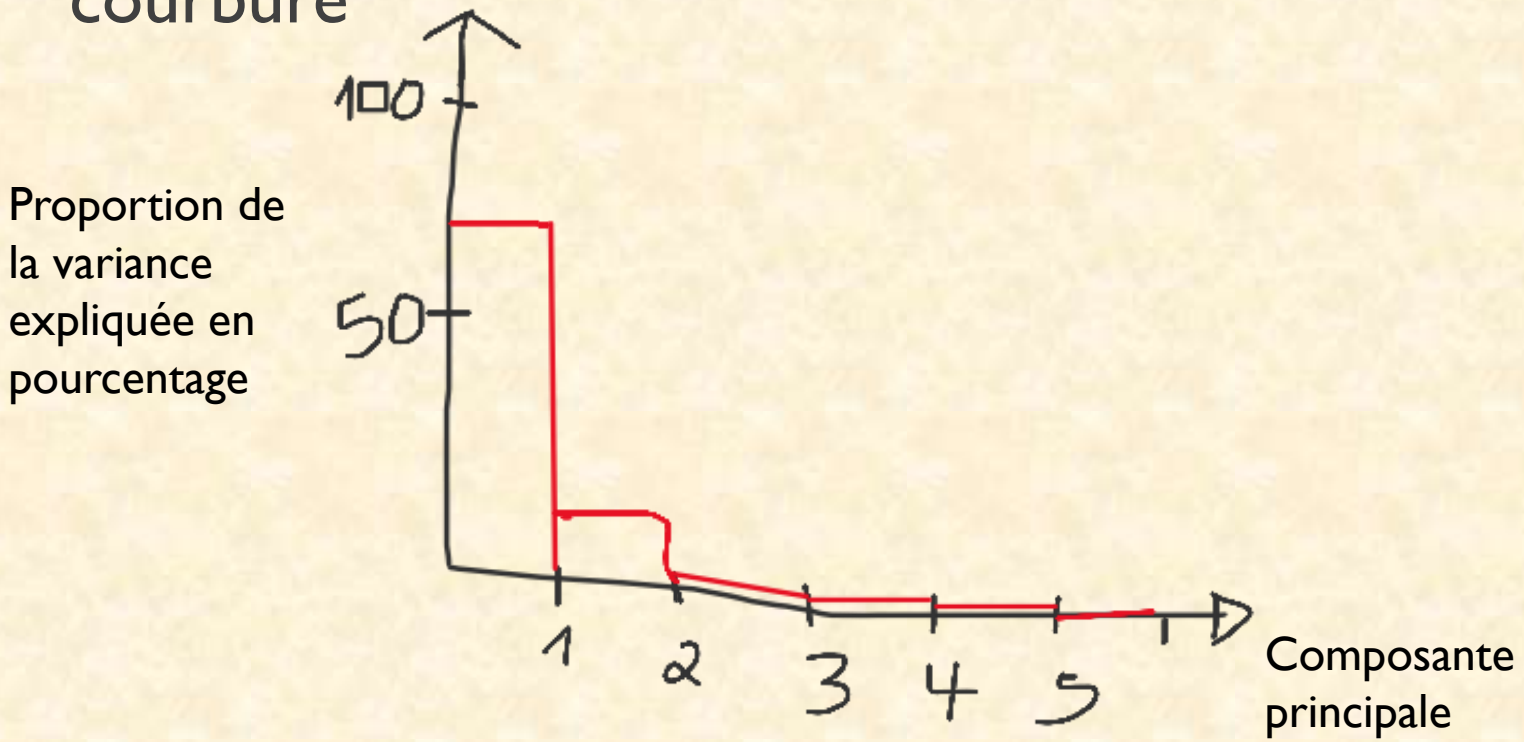
- **En pratique:** $M = \min(n-1, \boldsymbol{p})$

→ $\varphi_1, \ldots, \varphi_{\mathrm{M}}$ sont les vecteurs propres d'une matrice $X^T X$, et les variances sont les valeurs propres associées.

→ *La première composante principale est le vecteur propre associé à la valeur propre la plus élevée.*

CHOIX DU NOMBRE DE COMPOSANTES UTILES

- Choisir le plus petit nombre de composantes principales permettant d'interpréter, comprendre, résumer les données en utilisant la méthode du coude.
- Méthode du coude: Analyser les points d'inflexion, changement de courbure



STAT&ML
POUR TOUS

7 POINTS CLÉS SUR L'ACP?

ISABELLE LACMAGO

L'ACP est une méthode **d'apprentissage statistique non supervisée.**

C'est l'une des méthodes de référence en **réduction de la dimensionnalité des données**.

L'ACP est utilisée principalement lorsque l'on souhaite comprendre les individus d'une population caractérisée par un **grand nombre de variables**.

L'ACP nous permettra de comprendre la majorité des informations disponibles dans la base de données en *représentant les observations et les variables dans un espace de dimensions très réduites*.

Son principe consiste à construire un faible nombre de variables dérivées appelées **composantes principales, à partir des combinaisons linéaires pondérées des p variables initiales.**

Ces composantes sont construites de manière successive, en commençant par la plus pertinente. *La première composante est la plus proche des observations et contient le plus d'informations sur les données* (proportion de variance expliquée la plus élevée).

Afin de réaliser des analyses et de répondre à la question d'intérêt, on choisira le nombre de composantes en utilisant la **méthode du coude**.

STAT&ML
POUR TOUS

As-tu des questions?

As-tu des remarques?

Un like, un partage

Un commentaire!

ISABELLE LACMAGO