

Introduction et premiers pas dans l'ACP

Angelina Roche

Executive Master Statistique et Big Data

Plan du chapitre

Introduction au cours d'analyse de données

Tableaux de données

Réduction de la dimension

Plan

Introduction au cours d'analyse de données

Tableaux de données

Réduction de la dimension

Objectifs du cours

- ▶ Apprendre à extraire de l'information provenant de tableaux de données :
 - ▶ quantitatives (numériques) : **ACP** (Analyse en Composantes Principales),
 - ▶ qualitatives (données issues de questionnaires, données textuelles,...) : **AFC** (Analyse Factorielle des Correspondances), **ACM** (Analyse des Correspondances Multiples).
- ▶ Réduire la dimension des données comme première étape pour d'autres méthodes statistique (détection d'outliers, classification,...).
- ▶ Représenter graphiquement des données de grande dimension ou qualitatives.

Déroulement du cours

- ▶ **Cours 1** : introduction et premiers pas dans l'ACP.
- ▶ **Cours 2** : mise en oeuvre de l'ACP, étude des individus et des variables, aide à l'interprétation.
- ▶ **Cours 3** : AFC et ACM.

Quelques références

- ▶ **Page web de François Husson** : <http://math.agrocampus-ouest.fr/infoglueDeliverLive/membres/Francois.Husson/enseignement> incluant des vidéos et des références bibliographiques.
- ▶ Lebart, L., Morineau, A. et Piron, M. (2002). *Statistique exploratoire multidimensionnelle*, Dunod.
- ▶ Escofier, B. et Pagès ; J. (1998). *Analyses factorielles simples et multiples*, Dunod.
- ▶ Saporta, G. (1990). *Probabilités, Analyse de Données et Statistique*, Technip, Paris.

Plan

Introduction au cours d'analyse de données

Tableaux de données

Réduction de la dimension

Notations

- ▶ L'objectif est de décrire la distribution de plusieurs variables numériques observées sur les mêmes individus.
- ▶ Nous notons :
 - ▶ x_i^j l'observation de la j -ème variable sur l'individu i ,
 - ▶ p nombre de variables
 - ▶ n nombre d'individus.
- ▶ Les données sont donc représentées sous la forme d'une matrice à n lignes et p colonnes

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^p \\ \vdots & \ddots & \vdots \\ x_n^1 & \dots & x_n^p \end{pmatrix}.$$

- ▶ Ici, p est grand voire très grand.

Centrer, réduire, standardiser

- ▶ Centrer, c'est enlever la valeur de la moyenne de la **variable** :

$$x_i^j \leftarrow x_i^j - \bar{x}^j \text{ où } \bar{x}^j = \frac{1}{n} \sum_{i=1}^n x_i^j.$$

- ▶ Réduire, c'est diviser par l'écart-type de la variable :

$$x_i^j \leftarrow x_i^j / \sigma_j \text{ où } \sigma_j^2 = \frac{1}{n} \sum_{i=1}^n (x_i^j - \bar{x}^j)^2.$$

- ▶ Standardiser, c'est centrer et réduire :

$$x_i^j \leftarrow \frac{x_i^j - \bar{x}^j}{\sigma_j}.$$

Quand faut-il standardiser ou réduire les données ?

- ▶ **Indispensable** lorsque les variables ne sont pas exprimées dans la même unité.
- ▶ **Généralement conseillé** : permet d'accorder la même importance à chaque variable.
- ▶ Grande influence sur le résultat de l'étude.
- ▶ Mise en pratique : fonction `scale()` de R.

Pondération des individus

- ▶ Il peut être utile de pondérer les individus.
- ▶ On associe à chaque individu i un point p_i tel que

$$p_i \geq 0 \text{ pour tout } i \text{ et } \sum_{i=1}^n p_i = 1.$$

- ▶ Habituellement (c'est-à-dire sans pondération), $p_i = 1/n$.

Plan

Introduction au cours d'analyse de données

Tableaux de données

Réduction de la dimension

Nuage des individus

- ▶ Individu : $x_i = (x_i^1, \dots, x_i^p)$.
- ▶ Nuage des individus $N_I \subset \mathbb{R}^p$.
- ▶ ACP normée : les données sont standardisées,

$$N_I = \left\{ \left(\frac{x_i^1 - \bar{x}^1}{\sigma_1}, \dots, \frac{x_i^p - \bar{x}^p}{\sigma_p} \right), i = 1, \dots, n \right\}$$

- ▶ ACP non normée : les données sont juste centrées

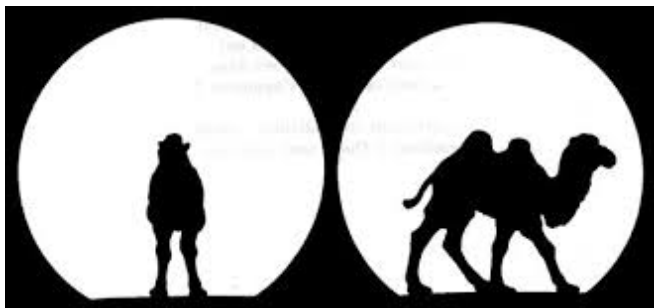
$$N_I = \left\{ (x_i^1 - \bar{x}^1, \dots, x_i^p - \bar{x}^p), i = 1, \dots, n \right\}$$

- ▶ **Objectif** : fournir une représentation simplifiée de N_I la plus fidèle possible.

Meilleure représentation plane d'un nuage de points N_i



Meilleure représentation plane d'un nuage de points N_i



Meilleure représentation d'un nuage de points N_I

- ▶ Inertie totale (= variance empirique) du nuage de point N_I :

$$I = \frac{1}{n} \sum_{i=1}^n \|x_i - \bar{x}\|^2,$$

avec $\bar{x} = (\bar{x}^1, \dots, \bar{x}^p)$.

- ▶ Version pondérée :

$$I = \sum_{i=1}^n p_i \|x_i - \bar{x}\|^2.$$

avec $\bar{x} = (\bar{x}^1, \dots, \bar{x}^p)$ où $\bar{x}^j = \frac{1}{n} \sum_{i=1}^n p_i x_i^j$.

Meilleure représentation d'un nuage de points N_I

- ▶ Inertie de la projection sur un sous-espace E où les données sont projetées (= variance expliquée) :

$$I_E = \frac{1}{n} \sum_{i=1}^n \|p_E(x_i) - \bar{x}\|^2,$$

où $p_E(x_i)$ est la projection orthogonale du point x_i sur le sous-espace E .

- ▶ Nous cherchons le sous-espace E_K de \mathbb{R}^n de dimension K d'inertie maximale.

Matrice de variance-covariance et matrice de corrélation

- La matrice de variance-covariance associée à X est la matrice

$$V = \begin{pmatrix} \sigma_1^2 & \text{Cov}(x^1, x^2) & \dots & \text{Cov}(x^1, x^p) \\ \text{Cov}(x^1, x^2) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \text{Cov}(x^1, x^p) & \dots & \dots & \sigma_p^2 \end{pmatrix},$$

où $\text{Cov}(x^j, x^{j'}) = \frac{1}{n} \sum_{i=1}^n (x_i^j - \bar{x}^j)(x_i^{j'} - \bar{x}^{j'})$.

Matrice de variance-covariance et matrice de corrélation

- La matrice de corrélation associée à X est la matrice

$$C = \begin{pmatrix} 1 & \frac{\text{Cov}(x^1, x^2)}{\sigma_1 \sigma_2} & \dots & \frac{\text{Cov}(x^1, x^p)}{\sigma_1 \sigma_p} \\ \frac{\text{Cov}(x^1, x^2)}{\sigma_1 \sigma_2} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\text{Cov}(x^1, x^p)}{\sigma_1 \sigma_p} & \dots & \dots & 1 \end{pmatrix}.$$

ACP et vecteurs propres

- ▶ Soient v^1, \dots, v^p les vecteurs propres de la matrice de corrélation C et $\lambda_1, \dots, \lambda_p$ les valeurs propres associées comptées avec multiplicité et numérotées telles que :

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0.$$

- ▶ En ACP normée, l'espace E_K de dimension K d'inertie maximale est

$$E_K = \text{Vect} \{v^1, \dots, v^K\}.$$

- ▶ En ACP non normée, nous considérons les éléments propres de la matrice de variance-covariance V .
- ▶ Mise en pratique : (par exemple) fonction `PCA()` du package `FactoMineR`.

Variance expliquée et valeurs propres

- ▶ λ_j : inertie du nuage de points N_I projetée sur l'axe j = variance expliquée par le j -ème axe.
- ▶ $I_{E_K} = \lambda_1 + \dots + \lambda_K$: inertie du nuage de points N_I projetée sur l'espace E_K = variance expliquée par les K premiers axes de l'ACP.
- ▶ $I = \lambda_1 + \dots + \lambda_p$: inertie totale.
- ▶ Proportion d'inertie expliquée par les K premiers axes :

$$\frac{I_{E_K}}{I}$$

Choix du nombre d'axes

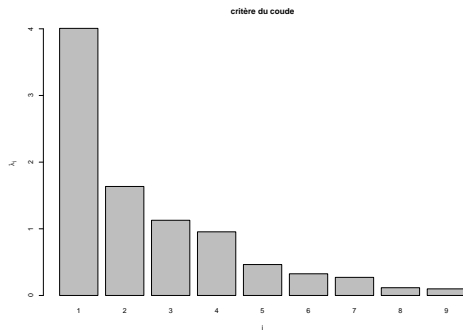
- ▶ **Critère du coude** : existence d'un coude dans le tracé de $j \mapsto \lambda_j$ (ébouli des valeurs propres) \leftrightarrow on garde les axes avant le coude.
- ▶ **Critère de Kaiser** : K le plus grand entier tel que $\lambda_K \geq I/p$ (entropie moyenne). En ACP normée $I/p = 1$.
- ▶ On garde les axes que l'on sait interpréter.
- ▶ Autre critère (très) répandu lorsque l'on souhaite réduire la dimension avant d'utiliser une autre méthode : K le plus grand entier tel que $I_{E_K}/I \geq s$ (souvent $s = 80\%$ ou $s = 90\%$).

Choix du nombre d'axes

- ▶ **Critère du coude** : existence d'un coude dans le tracé de $j \mapsto \lambda_j$ (*ébouli* des valeurs propres) \leftrightarrow on garde les axes avant le coude.
- ▶ **Critère de Kaiser** : K le plus grand entier tel que $\lambda_K \geq I/p$ (entropie moyenne). En ACP normée $I/p = 1$.
- ▶ **On garde les axes que l'on sait interpréter.**
- ▶ Autre critère (très) répandu lorsque l'on souhaite réduire la dimension avant d'utiliser une autre méthode : K le plus grand entier tel que $I_{E_K}/I \geq s$ (souvent $s = 80\%$ ou $s = 90\%$).

Retour aux données sur la consommation de protéines

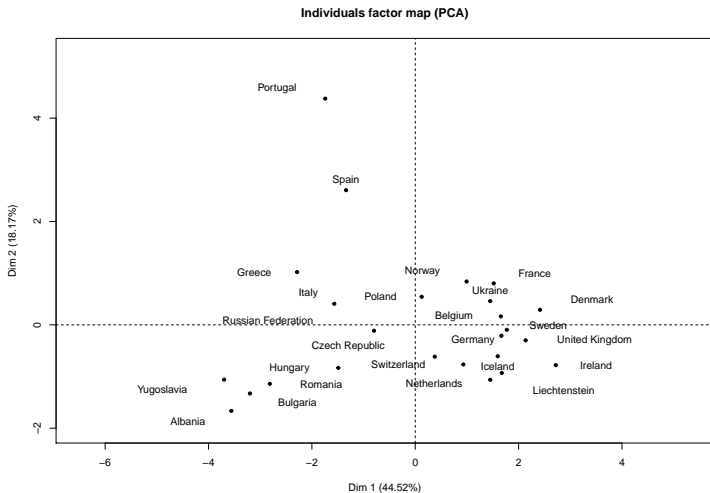
Choix du nombre d'axes



- ▶ Critère du coude : 1 (?) axe,
- ▶ Critère de Kaiser : on garde 3 axes,
- ▶ Critère des 80% : 4 axes.

Retour aux données sur la consommation de protéines

Représentation des individus sur le premier plan de l'ACP



Take-home message

- ▶ L'ACP est un **outil de visualisation de données multidimensionnelles** (utilisable en grande dimension) et de **réduction de la dimension**.
- ▶ L'espace de petite dimension permettant de représenter au mieux les données est déterminé par les **vecteurs propres** de la matrice de corrélation (ACP normée) ou variance-covariance (ACP non normée).
- ▶ Les **valeurs propres** représentent la part d'information (variance) apportée par chaque axe.
- ▶ **Problème** : les axes ne sont pas vraiment des variables \leftrightarrow difficiles à interpréter (des outils seront donnés au prochain cours).