Introduction et premiers pas dans l'ACP

Angelina Roche

Executive Master Statistique et Big Data

2018–2019

Plan du chapitre

Plan

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 (4 septembre) : introduction et premiers pas dans l'ACP.
- Cours 2 (11 septembre) : mise en oeuvre de l'ACP, étude des individus et des variables, aide à l'interprétation.
- Cours 3 (18 septembre) : 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

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'individus 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.

Exemple de données

Consommation de protéines en Europe : pour chacun des 25 pays de l'union européenne, relevé de la consommation moyenne journalière des 9 types de protéines.

Les individus (en ligne) sont les pays de l'union européenne et les variables sont la consommation journalière (en colonne) des 9 types de protéines.

Prot. / Pays	viandr	viandb	oeuf	lait	poisson	céréale	féculent	...
Bulgaria	7,8	6,0	1,6	8,3	1,2	56,7	1,1	...
Yugoslavia	4,4	5,0	1,2	9,5	0,6	55,9	3,0	...
Romania	6,2	6,3	1,5	11,1	1,0	49,6	3,1	...
Germany	11,4	12,5	4,1	18,8	3,4	18,6	5,2	...
France	18,0	9,9	3,3	19,5	5,7	28,1	4,8	...
Norway	9,4	4,7	2,7	23,3	9,7	23,0	4,6	...
Greece	10,2	3,0	2,8	17,6	5,9	41,7	2,2	...
...	...	...	...	...	...	...	...	...

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

Nuage des individus

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

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

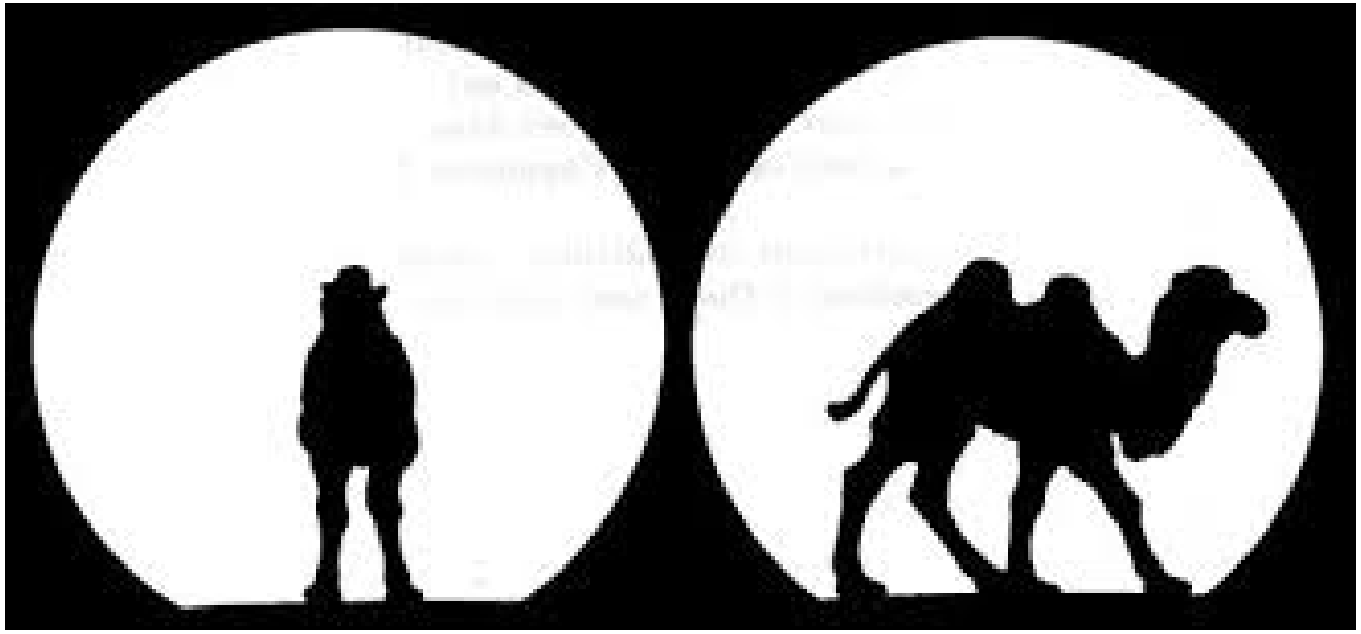
- ACP non normée : les données sont juste centrées

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

- **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 $\bar{x} = (\bar{x}^1, ..., \bar{x}^p)$.

- Version pondérée :

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

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

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

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

Matrice de variance-covariance et matrice de corrélation

- La matrice de corrélation associée à X est la matrice

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

ACP et vecteurs propres

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

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

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

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

- 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.
- $I_{E_K} = \lambda_1 + ... + \lambda_K$: inertie du nuage de points N_I projetée sur l'espace E_K = variance expliquée par les K premiers axes de l'ACP.
- $I = \lambda_1 + ... + \lambda_p$: 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) $\hookrightarrow$ 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 : K le plus grand entier tel que $I_{E_K}/I \geq s$ (souvent $s = 80\%$ ou $s = 90\%$).

Plan

Méthode générale

1. Analyser le jeu de données : fonction `summary()`, fonction `aggr()` du package VIM,...

2. Traiter les données manquantes :

   - Si la proportion d'individus/variables présentant des données manquantes est faible : supprimer les individus/variables présentant des données manquantes.
   - Sinon, utiliser une méthode d'imputation : remplacement par la moyenne, k-plus proches voisins, méthodes basées sur l'ACP (cf package missMDA),....

3. Une fois obtenu un tableau de données complet, réaliser l'ACP.