Introduction et ACP

Angelina Roche

Executive Master Statistique et Big Data

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

- 3 séances de 3h de cours et TP sous R.
- Plan du cours :
  1. Analyse en Composantes Principales (ACP).
  2. Analyse Factorielle des Correspondances (AFC).
  3. Suivant le temps :
     - 3.1 Analyse Factorielle des Correspondances Multiples (AFCM),
     - 3.2 ACP sur données mixtes
     - 3.3 classification sur composantes principales
     - 3.4 classification ascendante hiérarchique (CAH)
     - 3.5 ACP parcimonieuse

  Si vous avez une préférence entre ces différents thèmes n'hésitez pas à me le faire savoir.

Validation du cours

- **Deux mini-projets :**

  Projet 1 (P1) : à rendre avant le **mardi 28 mars**, application directe du cours d'aujourd'hui.

  Projet 2 (P2) : à rendre avant le **vendredi 5 mai**.

- **Note finale = (P1+2*P2)/3.**

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'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.

Exemple de données

Consommation de protéines en Europe : pour chacun des pays, 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,

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

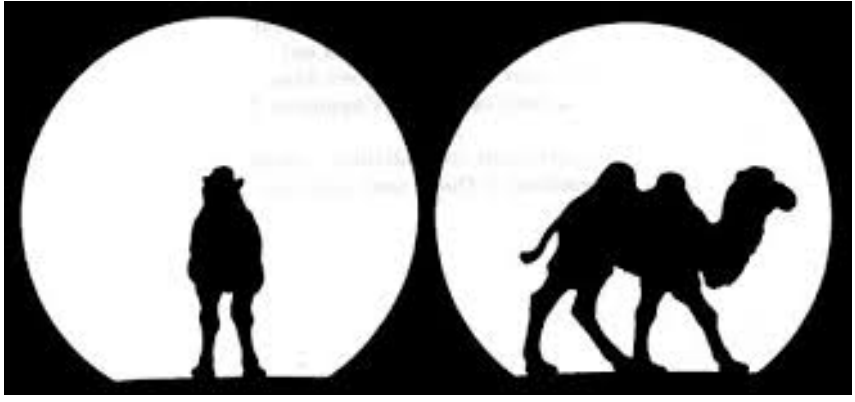
- ACP non normée : les données sont juste centrées

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

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

- Représente la **quantité d'information** apportée par le tableau de données.
- 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 1 : premiers pas dans l'ACP.

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 *empirique* : 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 $s = 90\%$).

Plan

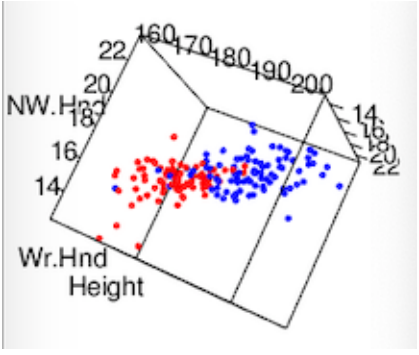
Projection du nuage des individus

- Les axes de l'ACP $v_1, ..., v_K$ sont des éléments de $\mathbb{R}^p$
- k-ème axe de l'ACP :

$$v_k = \begin{pmatrix} v_k^1 \\ \vdots \\ v_k^p \end{pmatrix}.$$

- $s_i^k = \tilde{x}_i v_k = \sum_{j=1}^p \tilde{x}_i^j v_k^j$: coordonnée du i-ème individu par rapport à l'axe k, où $\tilde{x}_i^j = (x_i^j - \bar{x}^j)/\sigma_j$ (ACP normée) ou $\tilde{x}_i^j = x_i^j - \bar{x}^j$ (ACP non normée).

Illustration données Mise en pratique 1



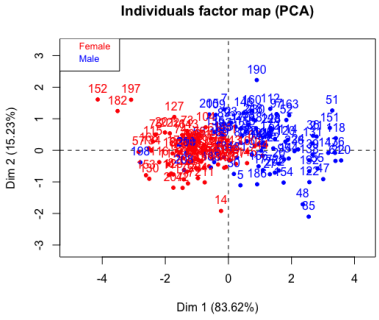


Figure – À droite : représentation du nuage N_I (tracé des points $(\tilde{x}_i^1, \tilde{x}_i^2, \tilde{x}_i^3)$, $i = 1, ..., n$). À gauche : représentation du nuage projeté sur le premier plan (tracé des points (s_i^1, s_i^2), $i = 1, ..., n$).

Composantes principales

- $s^k = (s_1^k, ..., s_n^k)$: composante principale $\hookrightarrow$ assimilable à une variable.

$$\sum_{i=1}^{n} s_i^k = \sum_{i=1}^{n} \tilde{x}_i v_k = \underbrace{\left(\sum_{i=1}^{n} \tilde{x}_i \right)}_{=0} v_k = 0$$

$\Rightarrow$ les composantes principales sont centrées.

Composantes principales (II)

- Soient

$$S = \begin{pmatrix} s_1^1 & \dots & s_1^p \\ \vdots & \ddots & \vdots \\ s_n^1 & \dots & s_n^p \end{pmatrix}, P = \begin{pmatrix} v_1^1 & \dots & v_1^p \\ \vdots & \ddots & \vdots \\ v_p^1 & \dots & v_p^p \end{pmatrix}, \tilde{X} = \begin{pmatrix} \tilde{x}_1^1 & \dots & \tilde{x}_1^p \\ \vdots & \ddots & \vdots \\ \tilde{x}_n^1 & \dots & \tilde{x}_n^p \end{pmatrix}.$$

- Par définition : $S = \tilde{X}P$, d'où

$$S^t S = P^t \tilde{X}^t \tilde{X} P = P^t C P = \text{diag}(\lambda_1, ..., \lambda_p).$$

$\Rightarrow \lambda_k = \sum_{i=1}^n (s_i^k)^2, \quad \sum_{i=1}^n s_i^j s_i^k = 0$ si $j \neq k$.

- **La variance de la k-ème composante est égale à λ_k.**
- **Les composantes principales sont décorrelées.**

Représentation des variables

- Corrélation de la variable $\tilde{x}^j$ par rapport à la k-ème composante principale s^k :

$$\text{cor}(\tilde{x}^j, s^k) = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i^j \frac{s_i^k}{\sigma(s^k)} \text{ où } \sigma^2(s^k) = \sum_{i=1}^{n}(s_i^k)^2 = \lambda_k.$$

- Rappels :
  - $-1 \leq \text{cor}(\tilde{x}^j, s^k) \leq 1$,
  - Plus $|\text{cor}(\tilde{x}^j, s^k)|$ proche de 1, plus on considèrera que la variable j est liée à l'axe k.
  - $\text{cor}(\tilde{x}^j, s^k) < 0$: corrélation négative,
  - $\text{cor}(\tilde{x}^j, s^k) > 0$: corrélation positive.

Cercles des corrélations

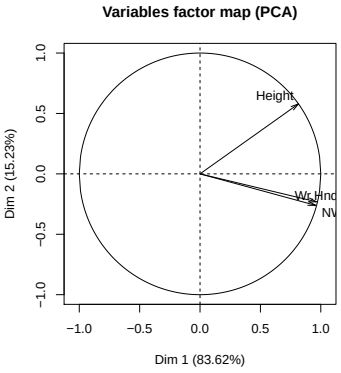
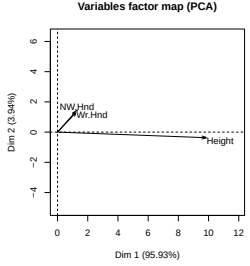


Figure – Représentation des corrélations sous la forme d'un cercle. Chaque flèche pointe sur le point de coordonnées $(\text{cor}(\tilde{x}^j, s^1), \text{cor}(\tilde{x}^j, s^2))$, $j = 1, ..., p$.

Cas de l'ACP non normée

- Dans une ACP non normée : on représente les covariances des variables par rapport aux axes et non les corrélations.
- Elles n'apparaissent donc plus sur un cercle.



Mise en pratique 2 : données températures/données protéines

Plan

Variables supplémentaires

- Utilité : variables construites à partir d'autres variables mais pouvant aider à l'interprétation ou variables quantitatives supplémentaires.
- Variables quantitatives : ajout sur le cercle des corrélations.
- Variables qualitatives : ajout dans le nuage des individus (coloration des individus en fonction des modalités par exemple).

Individus supplémentaires

- Utilité : individus ayant une contribution trop importante, ou dont on doute de la fiabilité, nouvelle étude,....
- Ajout dans le nuage des individus.

Contribution d'un individu à l'inertie d'un axe

- Rappel :

$$\lambda_k = \sum_{i=1}^{n} (s_k^i)^2.$$

- Contribution de l'individu i à l'inertie de l'axe k :

$$\text{ctr}(i,k) = \frac{(s_k^i)^2}{\lambda_k}.$$

- Lorsque les individus ne sont pas anonymes, ceux ayant une contribution importante (par exemple $> 1/n$) peuvent aider à l'interprétation des axes.
- Attention aux individus ayant une contribution trop importante ($> 25\%$).

Qualité de représentation d'un individu

- Nous avons : $\text{dist}(0, \tilde{x}_i)^2 = \sum_{j=1}^{p} (s_i^j)^2$.
- Qualité de représentation de l'individu i sur l'axe j :

$$Q(i,j) = \frac{(s_i^j)^2}{\text{dist}(0, \tilde{x}_i)^2}.$$

- On appelle parfois cet indice *cosinus carré*.