Analyse de données

Contents

Part I
ACP

1 Principe

On cherche à réduire le nombre de variable (réduction de la dimensionnalité) tout en gardant la même quantité d'information. On va donc garder les variables qui permettent de "départager" les individus.

2 Notations

Matrice $X = (x_{ij})_{i\leq n, j\leq p} = (e_i^T)_{j\leq p}^T$, pouvant dériver d'un tableau $(r_{ij})_{i\leq n, j\leq p}$.
X peut être vu sous 2 angles distincts:

- X est le nuage des n points-individus e_i plongé dans l'espace $\mathbb{R}^p$ des variables.
- X est le nuage des n points-variables e_i plongé dans l'espace $\mathbb{R}^n$ des variables.

Deux types d'ACP :

- ACP centrée : $x_{ij} = \frac{r_{ij}-\mu_j}{\sqrt{n}}$ où μ_j moyenne des $(r_{ij})_i$
- ACP normée : $x_{ij} = \frac{r_{ij}-\mu_j}{\sqrt{n}\sigma_j}$ où $\sigma_j^2 = \frac{1}{n}\sum_{i=1}^n (r_{ij}-\mu_j)^2$

matrice de corrélation $X^T X$
Les valeurs propres de la matrice sont notées λ_i. Les vecteurs propres associés sont les u_i.

Taux d'inertie $\tau_q = \frac{\sum_{i=1}^q \lambda_i}{\sum_{i=1}^p \lambda_i}$ On rappelle que $\sum_{i=1}^p \lambda_i = trace(X^T X)$

Choix du nombre d'axes :

- Règle de la part d'inertie : On cherche q, $\tau_q \geq 0.9$
- Règle de Kaiser : On ne garde que les axes correspondants aux valeurs propres supérieures à la moyenne des valeures propres
- Changement de pente : Si il y a un changement de pente lors de la représentation décroissante des valeures propres, on prend celles qui sont avant (coude, dérivée on continue)

Les valeurs propres ayant le plus d'inertie sont celles qui contiennent le plus d'information.

Axes factoriels Les nouveaux axes (correspondants aux nouvelles variables) sont les u_α. Si on fait l'analyse sur les individus, on obtient les v_α.

Importance d'un axe factoriel Contribution relative $Cr_\alpha(i) = \cos^2(e_i, u_\alpha) = (\frac{e_i^T u_\alpha}{\|e_i\|})^2 = \frac{\psi_{\alpha i}^2}{\|e_i\|^2}$

Contribution des e_i Contribution absolue $Cr_\alpha(i) = \frac{(e_i^T u_\alpha)^2}{\lambda_\alpha} = \frac{\psi_{\alpha i}^2}{\lambda_\alpha}$ où λ_α variance

Coefficient de corrélation $\rho_{j\alpha} = \cos(\beta_j) = \frac{x_j^T v_\alpha}{\|x_j\|}$

Part II
AFC

3 Principe

Elle est utilisée dans le cas de variables quantitatives. Elle peut être simple (2 questions) ou multiple (2 questions et plus).

4 Notations

Données Tableau des n_{ij}. De plus, $\sum_i \sum_j n_{ij} = n$

Tableau des fréquences relatives $f_{ij} = \frac{n_{ij}}{n}$

Fréquences marginales $\sum_i f_{ij} = f_j$ et $\sum f_j = 1$, de même pour f_i

Corrélation

- Si $f_{ij} = f_i f_j$, alors indépendance
- Si $f_{ij} > f_i f_j$, elles s'attirent
- Si $f_{ij} < f_i f_j$, il y a répulsion

Profils-Lignes $i = \frac{1}{f_i}(f_{ij})_j^T$, de même pour les profils-colonnes

Pondération des profils centre de gravité du nuage $g_L = (f_j)_j^T$ (pour profils-lignes)

Matrice L' des profils modifiés $l_{ij} = \frac{f_{ij}}{f_i\sqrt{f_j}}$ (lignes) et $i' = (\frac{f_{ij}}{f_i\sqrt{f_j}})_j^T$, $g'_L = (\sqrt{f_j})_j^T$

Inertie On la calcule grâce à la distance entre les profils $= u^T S'_L u = \sum_{j=1}^q \frac{f_{ij}}{f_i\sqrt{f_j}-\sqrt{f_j}}$ où $S'_L = \sum_{i=1}^p f_i(i' - g'_L)(i' - g'_L)^T$

Valeurs propres de S'_L g'_L est le vecteur propre correspondant à la valeur propre 0. L'espace cherché (des valeurs propres non nulles) est donc $\min(p-1, q-1)$

Part III
Régression

5 Principe

Elle consiste à analyser la relation linéaire ou non, entre des variables. Elle est dit simple lorsqu'il y a seulement 2 variables.

6 Notations

Equation de base $y = b_0 + b_1 x$

Méthode des moindre carrés On veut trouver une droite telle que la distance de chacuns des points à cette droite est minimale. $minZ = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$.
On dérive Z par rapport à b_0 et b_1, donne deux équations pour deux inconnues, on trouve b_0 et b_1.

Qualité de l'approximation $SSE = \sum(y_i - b_0 - b_1 x_i)^2$ et $SST = \sum(y_i - \frac{\sum y_i}{n})^2$
$SSE = SST + SSR$
On note $r^2 = \frac{SSR}{SST}$ le coefficent de détermination

- r faible mauvais ajustement
- r = 1 meilleur ajustement

Coefficient de corrélation $r_{xy} = (signe de b_1)|r|$ mesure la force de la relation entre deux variables

- +1 les deux variables sont liées à travers une pente positive
- -1 liées par une droite de pente négative
- 0 pas liées linéairement

Part IV
AFD

7 Principe

On a encore n individus, p variables mais aussi C classes.Elle peut être à but descriptif (recherche de facteurs déterminants) ou à but décisionnel (étant donné un nouvel individu de classe inconnue). Elle cherche à réduire la dimensionnalité tout en conservant la discrimination entre les classes.

8 Notations

Mêmes notations que l'ACP pour la plupart

Matrice de covariance (dispersion) intra-classes $S_W = \frac{1}{N}\sum_{k=1}^{C}\sum_{i=1}^{N_k}(e_i^{(k)} - \mu^{(k)})(e_i^{(k)} - \mu^{(k)})^T$
inter-classes $S_b = \sum_{k=1}^{C}\frac{N_k}{N}(\mu^{(k)} - \mu)(\mu^{(k)} - \mu)^T$
totale $S = S_W + S_b = \frac{1}{N}\sum_{i=1}^{N}(e_i - \mu)(e_i - \mu)^T$

Critère de discrimination $\lambda = \frac{u^T S_b u}{u^T S u}$ est la valeur propre correspondant au vecteur propre u de $S_W^{-1} S_b$. On cherche la plus grande.