Chazal
Baptiste

Analyse de données 1/6

Recherche des valeurs propres :

$$A = \begin{bmatrix} 3 & -2 & 2 \\ -1 & 1 & 3 \\ 2 & 3 & 1 \end{bmatrix} \quad \text{On pose} \quad \begin{vmatrix} 3-\lambda & -2 & 2 \\ -1 & 1-\lambda & 3 \\ 2 & 3 & 1-\lambda \end{vmatrix} = 0$$

$$\Leftrightarrow \begin{vmatrix} 3-\lambda & 0 & 2 \\ -1 & 4-\lambda & 3 \\ 2 & 4-\lambda & 1-\lambda \end{vmatrix} = 0 \qquad \Leftrightarrow \begin{vmatrix} 3-\lambda & 0 & 2 \\ -3 & 0 & 2+\lambda \\ 2 & 4-\lambda & 1-\lambda \end{vmatrix} = 0$$

$$\Leftrightarrow -(4-\lambda) \begin{vmatrix} 3-\lambda & 2 \\ -3 & 2+\lambda \end{vmatrix} = 0$$

...

$$\Leftrightarrow (\lambda-4)(-\lambda^2+\lambda+12) = 0$$

$$\Leftrightarrow \lambda-4 = 0 \quad \text{ou} \quad -\lambda^2+\lambda+12 = 0$$

$$\Leftrightarrow \lambda_1 = 4, \ \lambda_2 = -3, \ \lambda_3 = 4.$$

$$S = \{4; -3\}.$$

Recherche des vecteurs propres :

On part du même exemple que précédemment.

Pour $\lambda = -3$:

$$A\vec{u} = \lambda \vec{u}$$

$$\Leftrightarrow \begin{bmatrix} 3 & -2 & 2 \\ -1 & 1 & 3 \\ 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = -3 \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} 3x - 2y + 2z \\ -x + y + 3z \\ 2x + 3y + z \end{bmatrix} = \begin{bmatrix} -3x \\ -3y \\ -3z \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} 6 & -2 & 2 & 0 \\ -1 & 4 & 3 & 0 \\ 2 & 3 & 4 & 0 \end{bmatrix}$$ matrice augmentée

$$\Leftrightarrow \begin{bmatrix} 6 & -2 & 2 & 0 \\ 0 & 22 & 20 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \Leftrightarrow \begin{cases} 6x - 2y + 2z = 0 \\ 22y + 20z = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} x = \frac{7}{10} y \\ z = -\frac{11}{10} y \end{cases} \quad \Rightarrow \vec{u} = \begin{bmatrix} \frac{7}{10} y \\ y \\ \frac{11}{10} y \end{bmatrix} y \in \mathbb{R}.$$

Faire de même pour $\lambda = 4$.

Analyse de données 2/6

Recherche des vecteurs propres unitaires :

On part du même exemple que précédemment.

On a $\vec{u} = \begin{bmatrix} 7/10\, y \\ y \\ 11/10\, y \end{bmatrix}$ $y \in \mathbb{R}$.

On pose $\|\vec{u}\|^2 = 1$ (1 car unitaire)

$\Leftrightarrow (7/10\, y)^2 + y^2 + (11/10\, y)^2 = 1$

$\Leftrightarrow \dfrac{270\, y^2}{100} = 1$

$\Leftrightarrow 270 y^2 = 100$

$\Leftrightarrow y^2 = 10/27$

$\Leftrightarrow y = \sqrt{10/27}$ ou $-\sqrt{10/27}$

$$\rightarrow \begin{bmatrix} 7/10\sqrt{10/27} \\ \sqrt{10/27} \\ 11/10\sqrt{10/27} \end{bmatrix} \text{ ou } \begin{bmatrix} -7/10\sqrt{10/27} \\ -\sqrt{10/27} \\ -11/10\sqrt{10/27} \end{bmatrix}$$

Distance :

individu	poids	taille	âge	note
Bob	60	175	15	9
Billy	60	170	14	10
Brian	60	170	14	7

$$d^2(Bob, Billy) = (60-60)^2 + (175-170)^2 + (15-14)^2 + (9-10)^2$$
$$= 27.$$
$$d^2(Bob, Brian) = (60-60)^2 + (175-170)^2 + (15-14)^2 + (9-7)^2$$
$$= 30.$$
$$d^2(Billy, Brian) = (60-60)^2 + (170-170)^2 + (14-14)^2 + (10-7)^2$$
$$= 9.$$

Si l'on considère la taille en mètre, celle-ci devient très négligeable dans le calcul de la distance (par exemple $(1,75-1,70)^2 = 0,0025$). La distance entre Billy et Brian devient alors la plus élevée.

On appelle cela l'effet d'échelle.

Centrer les données ne modifie pas les distances.

Calcul du χ^2 (khi-deux) :

	Caractère 1	Caractère 2	Total
Classe 1	68	60	128
Classe 2	37	14	51
Total	105	74	179

On établit le tableau des k'. Pour chaque case, on multiplie le total ligne par le total colonne, et on divise par l'effectif total.

	Caractère 1	Caractère 2	Total
Classe 1	75,1	52,9	128
Classe 2	29,9	21,1	51
Total	105	74	179

tableau des effectifs théoriques

Ensuite on utilise $\chi^2_{obs} = \sum_i \sum_j \frac{(k_{ij} - k'_{ij})^2}{k'_{ij}}$.

$$\chi^2_{obs} = \frac{(68-75,1)^2}{75,1} + \frac{(60-52,9)^2}{52,9} + \frac{(37-29,9)^2}{29,9} + \frac{(14-21,1)^2}{21,1}$$

$$= 0,671 + 0,953 + 1,686 + 2,389$$

$$= 5,699.$$

Ensuite, on cherche le χ^2_{th}.

On calcule le degré de liberté $\nu = (n-1)(p-1)$
$= (2-1)(2-1)$
$= 1$.

On va donc regarder la ligne 1 de la table des χ^2_{th}.

Selon le seuil demandé, on choisit sa colonne :

Seuil de 5% → colonne $\chi^2_{0,95}$

Seuil de 10% → colonne $\chi^2_{0,90}$

etc...

Prenons ici 5%, on regarde donc la colonne $\chi^2_{0,95}$ et la ligne 1.

On a $\chi^2_{th} = 3,84$.

On a ici $\chi^2_{obs} > \chi^2_{th}$ donc une dépendance entre nos classes et nos caractères.

Si c'était l'inverse, il y aurait eu indépendance.

Analyse de données 4/6

Quelle méthode choisir ?

ACP ← n variables quantitatives de même importance

AFC ← 2 variables seulement, quelconques

ACM ← n variables quelconques

ACP

On utilise des données centrées réduites si :
- les unités sont différentes
- les ordres de grandeurs sont différents
- les variances/écart types sont trop différents

① Choisir le type d'ACP (centrée ? centrée réduite ?)

② Analyser la matrice des corrélations
- Relever les fortes corrélations et anti-corrélations

③ Choisir les axes

On relève les plus importantes valeurs propres λ et on donne le QGE (Qualité Globale d'Information).

$$\text{Qualité}_{\lambda_i} = \frac{\lambda_i}{\text{trace}} \cdot 100.$$

La trace est égale à la somme des valeurs propres.

On a ainsi quelque chose comme :

Axe 1 : $\frac{\lambda_1}{\text{trace}} \cdot 100 = 47{,}25\ \%$
Axe 2 : $\frac{\lambda_2}{\text{trace}} \cdot 100 = 19{,}27\ \%$
Axe 3 : $\frac{\lambda_3}{\text{trace}} \cdot 100 = 12{,}49\ \%$

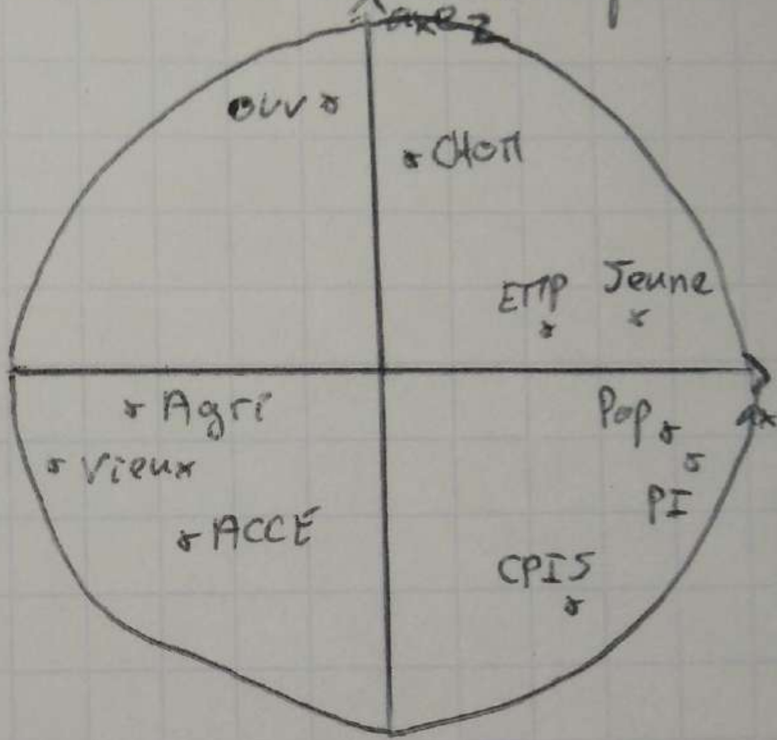
QGE $= 47{,}25 + 19{,}27 + 12{,}49 = 79{,}01\ \%$

④ Cercles de corrélations

On trace un ou plusieurs cercles de corrélation en fonction du nombre d'axes choisis. On y place les variables grâce aux coordonnées de la matrice des saturations.

Analyse de données 5/6

On relève pour chaque variable l'axe où la saturation est la plus forte (en valeur absolue). Si elle n'est pas notable, on ne la relève pas.

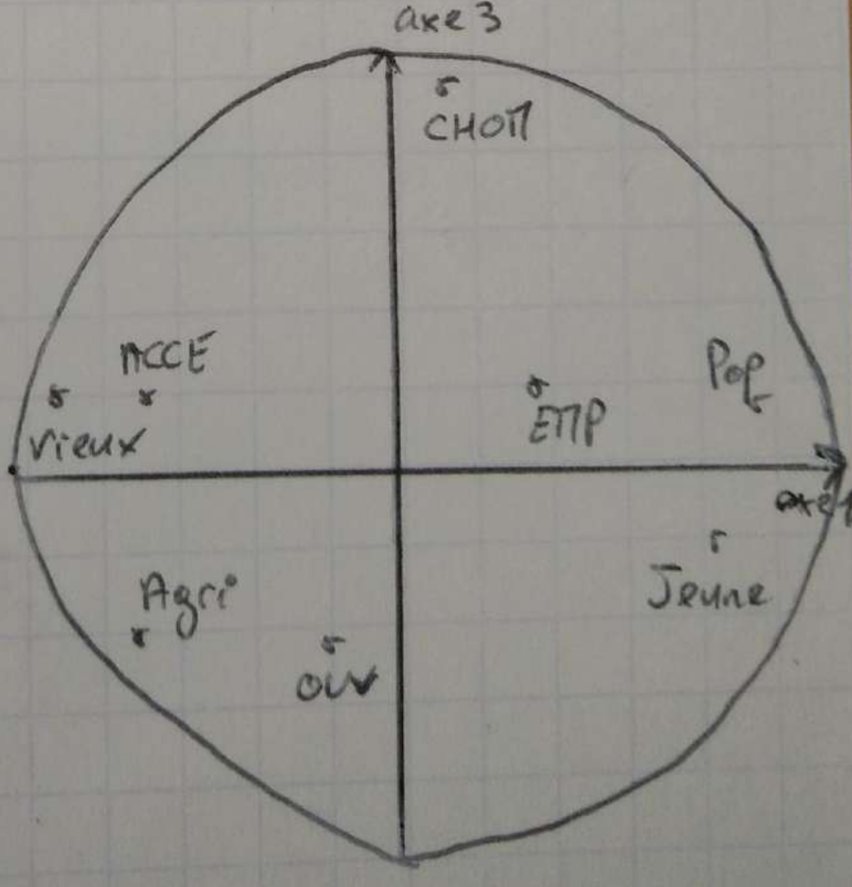


	axe 1	axe 2	axe 3
Pop	0,790	-0,279	0,213
Jeune	0,743	0,456	-0,201
Vieux	-0,945	-0,154	0,187
Agri	-0,826	-0,068	-0,312
ACCE	-0,574	-0,449	0,178
CPIS	0,724	-0,589	-0,098
PI	0,882	-0,221	-0,245
ETTP	0,539	0,315	0,238
OUV	-0,185	0,898	-0,271
CHOM	0,124	0,331	0,892

Axe 1 : Pop, PI, CPIS, jeune
- Vieux, agri

Axe 2 : OUV

Axe 3 : CHOM



⑤ Choix et analyse des individus intéressants

Relever les individus illustrants le mieux les axes, à partir du tableau des composantes principales.

Axe 1 ⊖	Axe 1 ⊕	Axe 2 ⊖	Axe 2 ⊕	Axe 3 ⊖	Axe 3 ⊕
individu 3	4	2	27		
12	5	14			
8					

⑥ Synthèse

- Conclure l'analyse

Analyse de données 6/6

AFC

① Test d'indépendance

- Comparer χ^2_{obs} et χ^2_{th}
- $\chi^2_{obs} = NR$ avec N l'effectif total et R l'inertie totale (somme des λ)

- $\chi^2_{obs} > \chi^2_{th} \Rightarrow$ dépendance

② Analyse des PFL et PFC

Pour chaque variable, on note sa fréquence totale (par exemple 25%). Puis, on détermine un pourcentage (par exemple 10%) de cette fréquence afin d'avoir un intervalle (22,5 - 27,5). Pour chaque variable, on relève alors les classes sous-représentées (sous l'intervalle) et sur-représentées (après l'intervalle) sur le tableau des PFL. Faire de même pour les PFC.

③ Choix des axes

Idem que pour l'ACP avec les λ et le QGE

④ Représentations graphiques

Idem que pour l'ACP mais sans le cercle
Plus on est proche du centre, plus on est proche du profil moyen

⑤ Synthèse

- Conclure l'analyse

ACM

- Généralisation de l'AFC

- On remplace les données qualitatives par des données quantitatives via un codage (que l'on choisit sur le coup au plus pratique).

- $R = \frac{m}{q} - 1$ m : nombre de modalité (vieux, jeune, homme, femme, etc...)
q : nombre de variables (âge, sexe, etc...)