Partie VII. Aspects pratiques

Formules barycentriques

Les coordonnées des individus Soit $\mathbf{c}_k$ le vecteur à n composantes des coordonnées des n individus sur l'axe factoriel associé à la valeur propre μ_k. D'après les résultats sur l'AFC, on a

$$\mathbf{c}_k = \frac{1}{\sqrt{\mu_k}}\frac{1}{p}\mathbf{X}\mathbf{a}_k \quad \text{et donc} \quad c_{ik} = \frac{1}{\sqrt{\mu_k}}\frac{1}{p}\sum_{j \text{ catégorie de } i} a_{jk}$$

Les seuls termes non nuls dans le calcul de $\mathbf{X}\mathbf{a}_k$ sont les coordonnées de la catégorie de chaque variable possédée par l'individu. Comme on est est dans le cadre de l'AFC, la variance de $\mathbf{c}_k$ est toujours $\operatorname{var}\mathbf{c}_k = \frac{1}{n}\mathbf{c}_k'\mathbf{c}_k = \mu_k$.

Barycentre des catégories *À $1/\sqrt{\mu_k}$ près, la coordonnée d'un individu est égale à la moyenne arithmétique simple des coordonnées des catégories auxquelles il appartient.*

Les coordonnées des catégories On a de même la seconde formule

$$\mathbf{a}_k = \frac{1}{\sqrt{\mu_k}}\mathbf{D}^{-1}\mathbf{X}'\mathbf{c}_k \quad \text{c-à-d} \quad a_{jk} = \frac{1}{\sqrt{\mu_k}}\frac{1}{n_j}\sum_{i \text{ de catégorie } j} c_{ik}$$

Les seuls termes non nuls de $\mathbf{X}'\mathbf{c}_k$ sont les coordonnées des individus ayant une catégorie donnée. Là encore, $\operatorname{var}\mathbf{a}_k = \mu_k$.

Barycentre des individus *À $1/\sqrt{\mu_k}$ près, la coordonnée d'une catégorie est égale à la moyenne arithmétique des coordonnées des n_j individus de cette catégorie.*

Barycentres et représentation

Représentation commune Les points représentatifs des catégories sont barycentres des groupes d'individus. On peut donc représenter individus et catégories dans un même plan factoriel.

Moyennes Comme $\mathbf{c}_k$ est une variable de moyenne nulle, la formule de barycentre indique que pour chaque variable $\mathcal{X}_i$, les coordonnées de ses catégories (pondérés par les effectifs) sont de moyenne nulle. Aucun centrage n'est donc nécessaire

Échelle pour que les catégories se trouvent visuellement au barycentre des individus qui les représentent on peut remplacer $\mathbf{a}_k$ par

$$\boldsymbol{\alpha}_k = \mathbf{D}^{-1}\mathbf{X}'\mathbf{c}_k = \sqrt{\mu_k}\mathbf{a}_k$$

Sélection de variables et axes

Sélection des variables on décide souvent de ne garder qu'un nombre réduit de variables actives et de garder les autres comme variables supplémentaires.

Sélection des axes

- règle courante : garder les axes tels que $\mu_k > 1/p$ (la moyenne des valeurs propres est $1/p$).
- les axes intéressants sont ceux que l'on peut interpréter, en regardant les contributions des variables actives et les valeurs-tests associées aux variables supplémentaires (définies plus tard).
- En pratique on se contente souvent d'interpréter le premier plan principal.

Inertie expliquée elle est moins intéressante qu'en ACP.

Catégories et axes factoriels

Catégorie Comme $\operatorname{var}\mathbf{a}_k = \sum_j \frac{n_j}{np}(a_{jk})^2 = \mu_k$, la contribution de la catégorie j à l'axe k est

$$\frac{n_j}{np}\frac{(a_{jk})^2}{\mu_k},$$

intéressante si elle est supérieure au poids n_j/np (à un facteur près comme en ACP et AFC).

Variable la contribution totale de la variable $\mathcal{X}_v$ à l'axe factoriel est

$$\frac{1}{\mu_k}\frac{1}{np}\sum_{j \text{ modalité de } \mathcal{X}_v} n_j(a_{jk})^2$$

Qualité de la représentation pour le sous-espace formé par les k^* premier axes, la qualité de la représentation de la catégorie j est le cosinus carré habituel

$$\frac{\sum_{k=1}^{k^*}(a_{jk})^2}{\sum_{k=1}^{q}(a_{jk})^2}.$$

Individus et axes factoriels

La normalisation de $\mathbf{c}_k$ est $\sum_{i=1}^n (c_{ik})^2 = n\mu_k$, où c_{ik} est la coordonnée de l'individu i sur l'axe factoriel k associé à la valeur propre μ_k.

Contribution d'un individu pour l'individu i, c'est

$$\frac{1}{n}\frac{(c_{ik})^2}{\mu_k}$$

Cette contribution est jugée en la comparant au poids $1/n$ comme en ACP et AFC.

Qualité de la représentation pour le sous-espace formé par les k^* premier axes, la qualité de la représentation de l'individu i est

$$\frac{\sum_{k=1}^{k^*}(c_{ik})^2}{\sum_{k=1}^{q}(c_{ik})^2}.$$

Contribution à l'inertie totale

Soit $\mathbf{x}^j = (x_i^j)$ le vecteur colonne de $\mathbf{X}$ correspondant à une catégorie j. On rappelle que l'inertie totale vaut

$$I_{\mathbf{g}} = \sum_{j \in \text{catégories}} \frac{n_j}{np} d^2(\mathbf{z}^j, \mathbf{g}) = \frac{1}{p}\sum_{v=1}^{p} m_v - 1,$$

où la distance du profil-colonne j au centre de gravité des profils-colonnes $\mathbf{g} = \mathbf{1}/n$ est

$$d^2(\mathbf{z}^j, \mathbf{g}) = \sum_{i=1}^{n} \frac{np}{p} \left(\frac{x_i^j}{n_j} - \frac{1}{n} \right)^2 = n \sum_{i=1}^{n} \left(\frac{x_i^j}{n_j^2} + \frac{1}{n^2} - \frac{2x_i^j}{nn_j} \right)$$
$$= n \left(\frac{n_j}{n_j^2} + \frac{n}{n^2} - \frac{2n_j}{nn_j} \right) = \frac{n}{n_j} - 1$$

Contribution d'une catégorie La contribution absolue de la catégorie j à l'inertie est

$$\frac{n_j}{np} d^2(\mathbf{z}^j, \mathbf{g}) = \frac{1}{p} \left(1 - \frac{n_j}{n} \right),$$

qui est une fonction décroissante de l'effectif. Il faut donc éviter les catégories d'effectif trop faible, qui d'ailleurs se retrouveront dans les premiers axes

Contribution d'une variable La contribution de la variable $\mathcal{X}_v$ est

$$\sum_{j \text{ modalité de } \mathcal{X}_v} \frac{1}{p} \left(1 - \frac{n_j}{n} \right) = \frac{m_v - 1}{p}$$

Elle est d'autant plus grande que le nombre de modalités de $\mathcal{X}_i$ est élevé. Il faut donc éviter les disparités trop grandes entre les nombre de modalités (quand on a le choix du découpage...)

Correspondances multiples et ACP non linéaire

Problème l'ACP vise à trouver une combinaison *linéaire* $u_1\mathbf{x}^1 + \cdots + u_p\mathbf{x}^p$ des variables qui soit de variance maximale. Si les relations entre variables ne sont pas linéaires, l'ACP échoue à extraire des données intéressantes.

Extension non-linéaire on cherche des fonctions $\phi^1, \ldots, \phi^p$ qui maximisent la variance

$$\text{var}\left(\phi^1(\mathbf{x}^1) + \cdots + \phi^p(\mathbf{x}^p) \right)$$

Fonctions en escalier On peut prendre des fonctions en escalier : on découpe l'intervalle de variation de $\mathbf{x}^j$ en m_j classes et on se donne un vecteur $\mathbf{a}_j = (a_{j1}, \ldots, a_{jm_j})$ de poids. Alors $\phi^j(x) = a_{j\ell}$ si x est dans la ℓ-ième classe.

Discrétisation des variables on définit le tableau disjonctif $\mathbf{X}_j$ indiquant quelle modalité (classe) de la variable j est prise par chaque individu. Alors

$$\phi^j(\mathbf{x}^j) = \mathbf{X}_j \mathbf{a}_j.$$

Reformulation de l'ACP non-linéaire on cherche le vecteur $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_p)$ qui maximise la variance

$$\text{var}\left(\mathbf{X}_1\mathbf{a}_1 + \cdots + \mathbf{X}_p\mathbf{a}_p \right) = \text{var}(\mathbf{X}\mathbf{a})$$

La solution est la première composante de l'ACM du tableau disjonctif joint $\mathbf{X}$.

Conclusion le découpage en classes des variables numériques permet d'obtenir une analyse non linéaire des données. Elle n'est possible que si on a suffisamment d'observations par classe.

Partie VIII. Interprétation externe

Les variables supplémentaires

Leur usage est très courant en analyse des correspondances multiples.

Variables quantitatives on calcule « à la main » leur corrélation avec les axes factoriels et on les place sur un cercle de corrélations. Si $\hat{\mathbf{z}}$ est une version centrée-réduite de la variable, alors

$$\text{cor}(\hat{\mathbf{z}}, \mathbf{c}_k) = \frac{1}{\sqrt{\mu_k}} \frac{1}{n} \sum_{i=1}^{n} \hat{z}_i c_{ik}$$

On peut aussi les découper en classes et les traiter comme des variables qualitatives.

Variables qualitatives on calcule directement les coordonnées de leurs modalités en utilisant la formule de barycentre des individus : la coordonnée la catégorie supplémentaire $\hat{\jmath}$ sur l'axe principal k est

$$a_{\hat{\jmath}k} = \frac{1}{\sqrt{\mu_k}} \frac{1}{n_{\hat{\jmath}}} \sum_{i \text{ de catégorie } \hat{\jmath}} c_{ik}$$

Valeurs-test pour les variables supplémentaires qualitatives

But on cherche à savoir si une catégorie $\hat{\jmath}$ d'effectif $n_{\hat{\jmath}}$ et de coordonnée $a_{\hat{\jmath}k}$ sur cet axe est liée à cet axe.

Idée du calcul si les $n_{\hat{\jmath}}$ individus d'une catégorie étaient pris au hasard, la moyenne de leurs coordonnées serait une variable aléatoire centrée (les $\mathbf{c}$ sont de moyenne nulle) et de variance $\frac{\mu_k}{n_{\hat{\jmath}}} \frac{n - n_{\hat{\jmath}}}{n-1}$. De plus, la moyenne des coordonnées est égale à $\sqrt{\mu_k} a_{\hat{\jmath}k}$.

Valeur-test c'est la version centrée et réduite de la moyenne des coordonnées

$$a_{\hat{\jmath}k} \sqrt{n_{\hat{\jmath}}} \sqrt{\frac{n-1}{n - n_{\hat{\jmath}}}}.$$

Quand $n_{\hat{\jmath}}$ et $n - n_{\hat{\jmath}} =$ sont assez grand (en général > 30), elle est significative si elle est supérieure à 2 ou 3 en valeur absolue. On ne doit pas l'utiliser sur les variables actives.

Partie IX. Récapitulatif

Notations AFC

Notation	taille	description
m_1 et m_2	entiers	nombre de modalités des variables 1 et 2
$\mathbf{N} = (n_{ij})$	$m_1 \times m_2$	table de contingence
$\mathbf{D}_1 = \mathrm{diag}(n_{i\cdot})$	$m_1 \times m_1$	effectifs (marges) de lignes
$\mathbf{D}_2 = \mathrm{diag}(n_{\cdot j})$	$m_2 \times m_2$	effectifs (marges) de colonnes
${}^{n_{ij}}/_{n_{i\cdot}}$ et ${}^{n_{ij}}/_{n_{\cdot j}}$	$m_1 \times m_2$	profils lignes et colonnes
$d^2 = n\varphi^2$	réel > 0	χ^2 d'écart à l'indépendance
$\mathbf{a}_k$ et $\mathbf{b}_k$	m_1 et m_2	coordonnées des lignes et colonnes sur l'axe k
λ_k	réel > 0	Valeur propre associée à l'axe k

Notations ACM

Notation	taille	description
$m_1, \dots, m_p$	entiers	nombres de modalités des variables
$\mathbf{X} = (x_i^j)$	$n \times$ (nb. cat.)	tableau disjonctif
$\mathbf{N}_{k\ell}$	$m_k \times m_\ell$	table de contingence des variables k et ℓ
$\mathbf{D} = \mathrm{diag}(n_i)$	nb. cat.	effectifs (marges) des catégories
$\mathbf{B}$	(nb. cat.)^2	matrice de Burt
μ_k	réel > 0	Valeur propre associée à l'axe k
$\mathbf{a}_k$	nb. cat.	coordonnées des catégories sur l'axe k
$\mathbf{c}_k$	n	coordonnées des individus sur l'axe k

nb. cat. $= m_1 + \cdots + m_p$.

Points communs entre AFC et ACM

But	décrire les liaisons entre plusieurs variables qualitatives
Cas $p = 2$	les coordonnées des modalités sont les mêmes pour les deux analyses
Représentation	toutes les modalités peuvent être représentées sur le même diagramme
Contribution d'une modalité à un axe	$\text{poids} \times \dfrac{(\text{coordonnée})^2}{\text{valeur propre}}$
Qualité de la représentation d'une modalité par un sous espace	$\cos^2\theta = \dfrac{\sum_{\text{axes du sous esp.}} (\text{coord sur l'axe})^2}{\sum_{\text{tous les axes}} (\text{coord sur l'axe})^2}$

Différences entre AFC et ACM

	AFC	ACM
Individus	non	oui
Données	tableau de contingence profils lignes/colonnes	tableau disjonctif tableau de Burt
Poids d'une modalité	$\frac{n_{i\cdot}}{n}$ (profil-ligne) $\frac{n_{\cdot j}}{n}$ (profil-colonne)	$\frac{n_j}{np}$
Nb de val. propres	$\min(m_1 - 1, m_2 - 1)$	$\sum_{v=1}^{p} m_v - p$
Axes à conserver	pas de règle Kaiser ; peut-être part d'inertie.	$\mu > \frac{1}{p}$
Variables supplémentaires	pas vraiment de sens	qualitatives et quantitatives