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 $\mathbf{a} = (\mathbf{a}_1, \dots, \mathbf{a}_p)$ qui maximise la variance

$$\operatorname{var}(\mathbf{X}_1\mathbf{a}_1 + \cdots + \mathbf{X}_p\mathbf{a}_p) = \operatorname{var}(\mathbf{X}\mathbf{a})$$

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

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

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