Décomposition de l'inertie

Nombre de valeurs propres Comme $\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'$ est carrée de dimension m_1 et que $\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}$ est de dimension m_2, le nombre de valeurs propres non nulles est $\min(m_1, m_2)$. Mais comme une des valeurs propres est 1 (associée à $\mathbf{g}$) et n'est pas intéressante :

Il y a au plus $\min(m_1 - 1, m_2 - 1)$ *valeurs propres non nulles*

φ^2 et valeurs propres l'inertie totale (et donc la somme des valeurs propres) est égale à φ^2. Donc si si $m_1 < m_2$, on obtient $\varphi^2 = \sum_{k=1}^{m_1-1} \lambda_k$.

Choix du nombre de valeurs propres On se contente souvent de regarder le premier plan principal car

- la règle de Kaiser $\lambda_k > \varphi^2/(m_1 - 1)$ s'applique mal ;
- la règle du coude reste valide, mais est subjective ;
- il existe un test sur de la part d'inertie non expliquée, mais il est un peu compliqué.

Partie V. Analyse des correspondances multiples

Analyse des correspondances multiples

But on veut étendre l'AFC au cas de $p \geq 2$ variables $\mathcal{X}_1, \mathcal{X}_2 \ldots, \mathcal{X}_p$ à $m_1, m_2, \ldots, m_p$ modalités. Ceci est particulièrement utile pour l'exploration d'enquêtes où les questions sont à réponses multiples.

Problème l'analyse des correspondances utilise une table de contingence qui nécessite $p = 2$.

Méthode on cherche un moyen différent d'analyser $p > 2$ variables et on vérifie que les résultats sont comparables à l'AFC pour $p = 2$.

Les données

Données brutes chaque individu est décrit par les numéros des modalités qu'il possède pour chacune des p variables. Il n'est pas possible de faire des calculs sur ce tableau, où les valeurs sont arbitraires.

Tableau disjonctif on remplace la v-ième colonne par m_v colonnes d'indicatrices : on met un zéro dans chaque colonne, sauf celle correspondant à la modalité de l'individu i qui reçoit 1.

Exemple On interroge 6 personnes sur la couleur de leurs cheveux (`CB`, `CC` et `CR` pour blond, châtain et roux), la couleur de leurs yeux (`YB`, `YV` et `YM` pour bleu, vert et marron) et leur sexe (`H`/`F`). On a donc trois variables (avec respectivement 3, 3 et 2 modalités) mesurées sur 6 individus. Les tableaux brut (ci-dessous à gauche) sont équivalents aux tableaux disjonctifs (à droite).

$$\begin{pmatrix} \texttt{CB} \\ \texttt{CB} \\ \texttt{CC} \\ \texttt{CC} \\ \texttt{CR} \\ \texttt{CB} \end{pmatrix} \begin{pmatrix} \texttt{YB} \\ \texttt{YV} \\ \texttt{YB} \\ \texttt{YM} \\ \texttt{YV} \\ \texttt{YB} \end{pmatrix} \begin{pmatrix} \texttt{H} \\ \texttt{H} \\ \texttt{F} \\ \texttt{H} \\ \texttt{F} \\ \texttt{F} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

Tableau disjonctif et tableau de contingence

Tableau disjonctif à la variable $\mathcal{X}_v$ on associe le tableau disjonctif $\mathbf{X}_v$ à n lignes et m_v colonnes.

Tableau de contingence on vérifie facilement que le tableau de contingence des variables $\mathcal{X}_v$ et $\mathcal{X}_w$ est donné par

$$\mathbf{N}_{vw} = \mathbf{X}_v'\mathbf{X}_w.$$

Effectifs marginaux la matrice diagonale des effectifs marginaux de la variable $\mathcal{X}_v$ est donnée par

$$\mathbf{D}_v = \mathbf{X}_v'\mathbf{X}_v.$$

Exemple (suite) Table de contingence Cheveux/Yeux et matrice d'effectif marginaux de la couleur de cheveux

$$\mathbf{N}_{12} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \qquad \mathbf{D}_1 = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Tableau disjonctif joint

Définition c'est la matrice $\mathbf{X} = (\mathbf{X}_1|\mathbf{X}_2|\cdots|\mathbf{X}_p)$, qui possède n lignes et $m_1 + \cdots + m_p$ colonnes. Chaque colonne représente une *catégorie*, c'est-à-dire une modalité d'une variable.

Exemple pour l'exemple de variables précédentes, on a le tableau disjonctif joint suivant

$$\mathbf{X} = \left(\begin{array}{ccc|ccc|cc} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right)$$

Chaque ligne somme à 3. Les sommes de colonnes sont

$$\left(\begin{array}{ccc|ccc|cc} 3 & 2 & 1 & 3 & 2 & 1 & 3 & 3 \end{array}\right)$$

Le tableau de Burt

Définition c'est un super-tableau de contingence des variables $\mathcal{X}_1, \ldots, \mathcal{X}_p$, formé de tableaux de contingence et de matrices d'effectifs marginaux. :

$$\mathbf{B} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 & \cdots & \mathbf{X}_1'\mathbf{X}_p \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 & & \\ \vdots & & \ddots & \vdots \\ \mathbf{X}_p'\mathbf{X}_1 & \cdots & & \mathbf{X}_p'\mathbf{X}_p \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{D}_1 & \mathbf{N}_{12} & \cdots & \mathbf{N}_{1p} \\ \mathbf{N}_{21} & \mathbf{D}_2 & & \\ \vdots & & \ddots & \vdots \\ \mathbf{N}_{p1} & \cdots & & \mathbf{D}_p \end{bmatrix}$$

Exemple Toujours pour les mêmes variables

$$\mathbf{B} = \left(\begin{array}{ccc|ccc|cc} 3 & 0 & 0 & 2 & 1 & 0 & 2 & 1 \\ 0 & 2 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ \hline 2 & 1 & 0 & 3 & 0 & 0 & 1 & 2 \\ 1 & 0 & 1 & 0 & 2 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ \hline 2 & 1 & 0 & 1 & 1 & 1 & 3 & 0 \\ 1 & 1 & 1 & 2 & 1 & 0 & 0 & 3 \end{array}\right)$$

Partie VI. L'ACM : une AFC sur tableau disjonctif

Comment utiliser l'AFC pour analyser p variables

But on cherche à faire une représentation des $m_1+\cdots+m_p$ catégories comme points d'un espace de faible dimension.

Méthode on fait une AFC sur le tableau disjonctif joint $\mathbf{X} = (\mathbf{X}_1|\mathbf{X}_2|\cdots|\mathbf{X}_p)$.

Les lignes la somme des éléments de chaque ligne de $\mathbf{X}$ est égale à p. Le tableau des profils-lignes est donc $\frac{1}{p}\mathbf{X}$.

Les colonnes la somme des éléments de chaque colonne de $\mathbf{X}$ est égale à l'effectif marginal de la catégorie correspondante. Le tableau des profils colonnes est donc $\mathbf{X}\mathbf{D}^{-1}$, où $\mathbf{D}$ est la matrice diagonale par blocs

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{D}_p \end{pmatrix}$$

Les coordonnées factorielles des catégories

Notation On note $\mathbf{a}_k = (\mathbf{a}_{1k}, \ldots, \mathbf{a}_{pk})'$ le vecteur à $m_1 + \cdots + m_p$ composantes des coordonnées factorielles des catégories sur l'axe k.

Calcul de l'AFC sur $\mathbf{X}$ comme la matrice des profils lignes est $\frac{1}{p}\mathbf{X}$ et celle des profils colonnes $\mathbf{X}\mathbf{D}^{-1}$, $\mathbf{a}_k$ est vecteur propre de

$$(\mathbf{X}\mathbf{D}^{-1})'\frac{1}{p}\mathbf{X} = \frac{1}{p}\mathbf{D}^{-1}\mathbf{X}'\mathbf{X} = \frac{1}{p}\mathbf{D}^{-1}\mathbf{B}$$

et donc l'équation des coordonnées des catégories est

$$\frac{1}{p}\mathbf{D}^{-1}\mathbf{B}\mathbf{a}_k = \mu_k\mathbf{a}_k$$

avec la convention de normalisation $\frac{1}{np}\mathbf{a}_k'\mathbf{D}\mathbf{a}_k = \mu_k$.

Propriétés des valeurs propres

Valeur propres triviales La valeur propre 1 est associée (comme en AFC) à la composante $\mathbf{z}^0 = (1, \ldots, 1)$ dans l'espace des individus. Les autres vecteurs propres lui sont orthogonaux, et donc de moyenne nulle.

Autres valeurs propres Si $n > \sum_{v=1}^p m_v$, le rang de $\mathbf{X}$ est $\sum_{v=1}^p m_v - p + 1$ et le nombre de valeurs propres non trivialement égales à 0 ou 1 est

$$q = \sum_{v=1}^p m_v - p.$$

Somme La somme des valeurs propres non triviales est

$$\sum_{k=1}^q \mu_k = \mathrm{Tr}\left(\frac{1}{p}\mathbf{D}^{-1}\mathbf{B}\right) - 1 = \frac{1}{p}\sum_{v=1}^p m_v - 1 = \frac{q}{p}$$

et leur moyenne vaut donc $1/p$.

Résolution dans le cas $p = 2$

On note $\mathbf{a}_k$ (resp. $\mathbf{b}_k$) les m_1 premières (resp. m_2 dernières) coordonnées de la composante principale k et μ_k la valeur propre correspondante :

$$\frac{1}{2}\mathbf{D}^{-1}\mathbf{B}\begin{bmatrix}\mathbf{a}_k\\ \mathbf{b}_k\end{bmatrix} = \frac{1}{2}\begin{bmatrix}\mathbf{I}_{m_1} & \mathbf{D}_1^{-1}\mathbf{N}\\ \mathbf{D}_2^{-1}\mathbf{N}' & \mathbf{I}_{m_2}\end{bmatrix}\begin{bmatrix}\mathbf{a}_k\\ \mathbf{b}_k\end{bmatrix} = \mu_k\begin{bmatrix}\mathbf{a}_k\\ \mathbf{b}_k\end{bmatrix}.$$

On obtient les équations

$$\begin{cases} \mathbf{D}_1^{-1}\mathbf{N}\mathbf{b}_k = (2\mu_k - 1)\mathbf{a}_k \\ \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a}_k = (2\mu_k - 1)\mathbf{b}_k \end{cases},$$

et donc on retrouve les coordonnées des modalités de lignes et de colonnes dans l'AFC classique (avec $\lambda_k = (2\mu_k - 1)^2$) :

$$\begin{cases} \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}\mathbf{b}_k = (2\mu_k - 1)^2\mathbf{b}_k \\ \mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a}_k = (2\mu_k - 1)^2\mathbf{a}_k \end{cases}.$$

Différences ACM/AFC pour $p = 2$

Nombre de valeurs propres on a *a priori* $m_1 + m_2 - 2$ valeurs propres non nulles, ce qui est plus important que dans le cas classique. En particulier pour chaque λ_k, on a deux μ_k possibles

$$\begin{cases} \mu_k = \frac{1+\sqrt{\lambda_k}}{2} & \text{associée à } \begin{bmatrix}\mathbf{a}_k\\ \mathbf{b}_k\end{bmatrix} \\ \mu_k' = \frac{1-\sqrt{\lambda_k}}{2} & \text{associée à } \begin{bmatrix}\mathbf{a}_k\\ -\mathbf{b}_k\end{bmatrix} \end{cases}$$

On ne garde donc que les valeurs $\mu_k > \frac{1}{2}$. On peut montrer qu'il y en a $\min(m_1 - 1, m_2 - 1)$.

Inertie L'interprétation de la part d'inertie expliquée par les valeurs propres est maintenant très différente. En particulier les valeurs propres qui étaient très séparées dans l'AFC de $\mathbf{N}$ le sont beaucoup moins dans celle de $\mathbf{X}$.