Analyse des Correspondances Multiples

MSPES-ENSAI

Introduction

Introduction

Le but de l'analyse des correspondances multiples (ACM) est de généraliser l'AFC en étudiant les liaisons entre plusieurs variables qualitatives.

C'est l'équivalent de l'ACP pour des variables qualitatives.

Les données décrivent n individus par p variables qualitatives ayant respectivement $m_1, \ldots, m_p$ modalités. On note m le nombre total de modalités de sorte que

$$m = \sum_{j=1}^{p} m_j.$$

Codages

Les variables qualitatives peuvent être codées avec le codage condensé (numérotation arbitraire de modalités) ou le codage disjonctif complet (composé uniquement de 0 et 1).

On notera J le tableau de données obtenu avec le codage condensé de sorte que

$$J = \left[J_{ij} \right],$$

ainsi J est une matrice de dimension $n \times p$ et J_{ij} est le numéro de la modalité prise par l'individu i pour la variable j.

On notera Z le tableau obtenu avec le code disjonctif complet de sorte que

$$Z = \left[Z_1 \dots Z_p \right] \text{ où } Z_j = \left[Z_{ji\ell} \right],$$

où Z est une matrice de dimension $n \times m$, Z_j est une matrice de dimension $n \times m_j$ avec $Z_{ji\ell} = 1$ si l'individu i prend la modalité ℓ pour la variable j et $Z_{ji\ell} = 0$ sinon.

Codage disjonctif complet

$$Z = \left[Z_1 \dots Z_p \right] \text{ où } Z_j = \left[Z_{ji\ell} \right],$$

La somme de chaque ligne de Z vaut p ($\forall i \ \sum_{j=1}^{p} \sum_{\ell=1}^{m_j} Z_{ji\ell} = p$).

La somme de chaque colonne de Z donne l'effectif marginal de chaque modalité.

Tableau de Burt

On définit le tableau de Burt, noté B, par

$$B = Z^\top Z = \begin{bmatrix} B_1 & B_{12} & \dots & \dots & \dots & B_{1p} \\ B_{21} & B_2 & \dots & & \dots & \vdots \\ \vdots & & \ddots & & B_{j\ell} & \\ \vdots & & & B_j & & \\ \vdots & & B_{\ell j} & & \ddots & \\ B_{p1} & & & & & B_p \end{bmatrix}.$$

Les blocs diagonaux sont $B_1, \dots, B_p$ avec

$$B_j = Z_j^\top Z_j.$$

Ainsi B_j est une matrice diagonale de taille $m_j \times m_j$ qui indique l'effectif de chaque modalité.

Les blocs non diagonaux sont

$$B_{j\ell} = Z_j^\top Z_\ell.$$

De plus, $B_{j\ell}$ est une matrice de taille $m_j \times m_p$ représentant le nombre d'individus croisant les modalités des variables j et ℓ.

AFC du tableau de Burt

AFC du tableau de Burt

On effectue l'AFC du tableau de Burt (on remplace donc K par B dans le cours d'AFC). On obtient alors le triplet (X, D, M) suivant:

- $X = \frac{1}{p}\Delta^{-1}B$: on considère les données issues des profils-lignes;
- $D_{f_{i\bullet}} = \frac{1}{np}\Delta$: la matrice des poids considère les effectifs de chaque profils-lignes;
- $M = D_{1/f_{\bullet j}} = np\Delta^{-1}$: on utilise la métrique du $\mathcal{X}^2$;

où Δ la matrice diagonale de taille $m \times m$ telle que sa diagonale est égale à la diagonale de B, ainsi

$$\text{diag}(\Delta) = \text{diag}(B).$$

AFC du tableau de Burt

En remarquant que la métrique est l'inverse de la matrice des poids:

$$M = D_{f_{i\bullet}}^{-1},$$

on obtient

$$\begin{aligned} VM &= X^\top D_{f_{i\bullet}} XM \\ &= \left(\frac{1}{p}\Delta^{-1}B\right)^\top \left(\frac{1}{np}\Delta\right)\left(\frac{1}{p}\Delta^{-1}B\right) np\Delta^{-1} \\ &= \left(\frac{1}{p}B\Delta^{-1}\right)^2, \end{aligned}$$

car B est une matrice symétrique.

AFC du tableau de Burt

La propriété suivante montre le lien entre l'inertie obtenue en faisant l'AFC du tableau de Burt et la dépendance entre les couples de variables.

L'inertie du nuage des profils-lignes est

$$I = \frac{m-p}{p^2} + \frac{n}{p^2} \sum_{j_1=1}^{p} \sum_{j_2 \neq j_1} \mathcal{X}^2_{j_1,j_2},$$

où

$$\mathcal{X}^2_{j_1,j_2}$$

est le coefficient de liaison entre les variables j_1 et j_2 (statistique du chi-2).

AFC du tableau de Burt

L'inertie du nuage des profils-lignes est

$$I = \frac{m-p}{p^2} + \frac{n}{p^2} \sum_{j_1=1}^{p} \sum_{j_2 \neq j_1} \mathcal{X}^2_{j_1,j_2}.$$

Ainsi l'inertie I obtenue par l'AFC de B mesure la liaison globale entre tous les couples de variables.

C'est ce qui justifie de faire l'AFC sur le tableau B.

En cas d'indépendance, on a $I = \frac{m-p}{p^2} > 0$.

L'AFC de B ne permet pas de représenter les individus.

On introduit donc une seconde approche permettant de visualiser les individus et on compare ses sorties à celles de l'AFC de B.

AFC du tableau disjonctif complet

AFC du tableau disjonctif complet

On se propose d'effectuer l'AFC du tableau Z car celui-ci contient encore toutes les informations relatives à chaque individu. On définit alors les élèments suivants:

- $X = \frac{1}{p} Z$: on considère les données issues des profils-lignes;
- $D_{f_{i\bullet}} = \frac{1}{n} I_n$: la matrice des poids considère les effectifs de chaque profils-lignes;
- $M = D_{1/f_{\bullet j}} = np\Delta^{-1}$: on utilise la métrique du $\mathcal{X}^2$.

AFC du tableau disjonctif complet

On a maintenant

$$\begin{aligned} VM &= X^\top D_{f_{i\bullet}} XM \\ &= \left(\frac{1}{p} Z\right)^\top \left(\frac{1}{n} I_n\right) \left(\frac{1}{p} Z\right) \left(np\Delta^{-1}\right) \\ &= \frac{1}{p} Z^\top Z \Delta^{-1} \\ &= \frac{1}{p} B \Delta^{-1} \end{aligned}$$

Soit $\nu_1, \ldots, \nu_q$ les vecteurs propres associés aux q valeurs propres non triviales de VM (*i.e.,* différentes de 0 et de 1) associés aux valeurs propres $\lambda_1, \ldots, \lambda_q$. On remarquera qu'on a

$$q \leq m - p.$$

AFC du tableau disjonctif complet

En supposant $q = m - p$, on a

$$\begin{aligned} I &= \text{tr}(VM) - 1 \\ &= \text{tr}(\frac{1}{p} B\Delta^{-1}) - 1 \\ &= \frac{1}{p}\text{tr}(\Delta\Delta^{-1}) - 1 \\ &= \frac{m}{p} - 1 \\ &= \frac{m-p}{p} \\ &= \frac{q}{p} \end{aligned}$$

AFC du tableau disjonctif complet

$$I = \frac{q}{p}.$$

Ainsi, l'inertie I obtenue par l'AFC de Z ne dépend pas de la liaison entre les variables.

C'est donc une statistique non intéressante.

Regardons maintenant le lien entre les AFC de B et de Z.

Relation entre les AFC de B et de Z

Relation entre les AFC de B et de Z

- $\rho_1, \ldots, \rho_q$ les valeurs propres non triviales de $VM = (\frac{B\Delta^{-1}}{p})^2$ issues de l'AFC de B;
- $\lambda_1, \ldots, \lambda_q$ les valeurs propres non triviales de $VM = (\frac{B\Delta^{-1}}{p})$ issues de l'AFC de Z.

On a $\rho_k = \lambda_k^2$ et les mêmes valeurs propres.

Cela justifie donc l'AFC de Z qui permet la représentation du nuage des individus dans le plan factoriel.

Relation entre les AFC de B et de Z

De plus, on peut interpréter les valeurs propres λ_k. En effet, lorsque les variables sont indépendantes, on a

$$\sum_k \rho_k = \frac{m-p}{p^2} = \frac{q}{p^2} \Leftrightarrow \frac{1}{q}\sum_k \lambda_k^2 = \frac{1}{p^2}.$$

Par ailleurs on a

$$\frac{1}{q}\sum_k \lambda_k = \frac{1}{p},$$

car $I = \sum_k \lambda_k = \frac{q}{p}$.

Ainsi, lorsque les variables sont indépendantes on a $\lambda_k = \frac{1}{p}$. On peut donc en déduire un critère pour choisir le nombre d'axes factoriels. On conservera les axes dont l'inertie est supérieure à $1/p$.