Analyse des correspondances multiples

Que faire avec des données catégorielles ?

L'analyse en composantes principales (ACP) est utilisée pour des données **continues**. Mais que faire lorsque les variables sont catégorielles, comme c'est le cas dans les tableaux de contingences, dans les questionnaires à choix multiples ou pour des variables qualitatives comme la couleur des yeux, la profession, ou la région d'origine ?

Il est possible de faire la même chose que pour l'ACP pour des variables catégorielles, cette méthode s'appelle l'analyse des correspondances (AC). Elle permet une représentation géométrique des relations entre les modalités d'une ou plusieurs variables qualitatives. Lorsque l'on s'intéresse à la relation entre deux variables qualitatives, on parlera d'analyse des correspondances binaires (ACB) ou alors d'analyse factorielle des correspondances (AFC). Lorsque l'on étudie simultanément plus de deux variables qualitatives, on parlera d'analyse des correspondances multiples (ACM).

Exemples

- La boussole électorale de Radio-Canada (questionnaire politique).
- Étude de segmentation de clientèle pour un opérateur télécom.
- Relation entre la couleur des yeux et des cheveux d'individus.

Notation

On considère un tableau de contingence $K = (k_{ij})$, où k_{ij} est le nombre d'individus appartenant à la classe $i \in \{1, \dots, n\}$ et à la catégorie $j \in \{1, \dots, p\}$. On travaille ensuite avec le tableau des fréquences relatives. Comme les fréquences sont proportionnelles à la taille d'échantillon n, le tableau des fréquences relatives contient plus d'information. Notons $F = (f_{ij})$, dans lequel

$$f_{ij} = \frac{k_{ij}}{k_{\bullet\bullet}} = \frac{k_{ij}}{\sum_{l=1}^{n}\sum_{m=1}^{p} k_{lm}}.$$

Les marges ligne (resp. colonne) du tableau correspondent à la somme des colonnes pour chaque ligne (resp. à la somme des lignes pour chaque colonne) :

$$f_{i\bullet} = \sum_{j=1}^{p} f_{ij} = \frac{k_{i\bullet}}{k_{\bullet\bullet}}, \quad 1 \leq i \leq n; \tag{1}$$

$$f_{\bullet j} = \sum_{i=1}^{n} f_{ij} = \frac{k_{\bullet j}}{k_{\bullet\bullet}}, \quad 1 \leq j \leq p. \tag{2}$$

Indépendance statistique

Les fréquences relatives estiment des probabilités. Dans le cas d'un tableau de fréquences croisant deux variables, sous l'hypothèse d'indépendance, les fréquences relatives devraient être telles qu'on ne s'éloigne pas trop de la relation

$$f_{ij} = f_{i\bullet} f_{\bullet j}, \quad i \in \{1, \dots, n\},\ j \in \{1, \dots, p\}.$$

On peut tester si les différences entre f_{ij} et $f_{i\bullet} f_{\bullet j}$ sont significatives à l'aide du test du χ^2 :

$$T = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{\left(k_{ij} - \mathbb{E}(k_{ij})\right)^2}{\mathbb{E}(k_{ij})} = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{\left(k_{ij} - \frac{k_{i\bullet} k_{\bullet j}}{k_{\bullet\bullet}}\right)^2}{\left(\frac{k_{i\bullet} k_{\bullet j}}{k_{\bullet\bullet}}\right)}.$$

Si les variables sont indépendantes, la statistique T doit être proche de 0.

Profils-lignes et profils-colonnes

Chaque ligne du tableau peut être vue comme un profil-ligne

$$L_i = \left(\frac{k_{i1}}{k_{i\bullet}}, \dots, \frac{k_{ip}}{k_{i\bullet}}\right) = \left(\frac{f_{i1}}{f_{i\bullet}}, \dots, \frac{f_{ip}}{f_{i\bullet}}\right)$$

et chaque colonne du tableau peut être vue comme un profil-colonne

$$C_j = \left(\frac{k_{1j}}{k_{\bullet j}}, \dots, \frac{k_{nj}}{k_{\bullet j}}\right) = \left(\frac{f_{1j}}{f_{\bullet j}}, \dots, \frac{f_{nj}}{f_{\bullet j}}\right).$$

Le profil-ligne moyen est donné par

$$\left(\sum_{i=1}^{n} f_{i\bullet} \frac{f_{i1}}{f_{i\bullet}}, \dots, \sum_{i=1}^{n} f_{i\bullet} \frac{f_{ip}}{f_{i\bullet}}\right) = \left(f_{\bullet 1}, \dots, f_{\bullet p}\right),$$

et le profil-colonne moyen est donné par

$$\left(\sum_{j=1}^{p} f_{\bullet j}\frac{f_{1j}}{f_{\bullet j}}, \dots, \sum_{j=1}^{p} f_{\bullet j}\frac{f_{nj}}{f_{\bullet j}}\right) = (f_{1\bullet}, \dots, f_{n\bullet})\,.$$

En cas d'indépendance entre les variables, tous les profils sont égaux à leurs moyennes respectives, i.e pour tout $i \in \{1, \dots, n\}$ et $j \in \{1, \dots, p\}$,

$$\left(\frac{f_{i1}}{f_{i\bullet}}, \dots, \frac{f_{ip}}{f_{i\bullet}}\right) = (f_{\bullet 1}, \dots, f_{\bullet p}) \quad \text{et} \quad \left(\frac{f_{1j}}{f_{\bullet j}}, \dots, \frac{f_{nj}}{f_{\bullet j}}\right) = (f_{1\bullet}, \dots, f_{n\bullet})\,.$$

La dépendance entre les variables est fonction de la similarité entre les profils-lignes et les profils-colonnes. On peut mesurer la distance entre deux profils-lignes avec la distance du χ^2, en tenant compte de l'importance de chaque colonne :

$$d^2(i, i') = \sum_{j=1}^{p} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'j}}\right)^2.$$

On peut écrire cela sous forme matricielle en notant $D_n = \mathrm{diag}(f_{i\bullet})$ et $D_p = \mathrm{diag}(f_{\bullet j})$. La matrice $D_n^{-1}F$ a pour lignes les profils-lignes et la matrice $D_p^{-1}F^\top$ a pour lignes les profils-colonnes. En utilisant ces matrices, la distance du χ^2 au carré est de la forme $x^\top D_p^{-1} x$ pour un point-ligne $x \in \mathbb{R}^p$ et de la forme $x^\top D_n^{-1} x$ pour un point-colonne $x \in \mathbb{R}^n$.

Analyse factorielle des correspondances

L'analyse factorielle des correspondances (AFC) est une approche graphique permettant de représenter simultanément les profils-lignes appartenant à $\mathbb{R}^p$ et les profils-colonnes appartenant à $\mathbb{R}^n$ d'un tableau de fréquences. On cherche donc un espace en deux dimensions où projeter les profils lignes et colonnes de sorte que les points dans cet espace soient le plus près possible des points originaux au sens de la distance du χ^2. L'analyse des correspondances est très similaire à une double ACP.

Analyse directe (sur les lignes) : les lignes de la matrice $D_n^{-1}F \in \mathbb{R}^p$. On cherche à les représenter dans cet espace muni de la distance $x^\top D_p^{-1} x$.

Analyse duale (sur les colonnes) : les colonnes de la matrice $D_p^{-1}F^\top \in R^n$. On cherche à les représenter dans cet espace muni de la distance $x^\top D_n^{-1} x$.

Premier axe factoriel de l'analyse directe : on cherche le vecteur $u \in \mathbb{R}^p$ tel que

$$\left(u^\top D_p^{-1} F^\top D_n^{-1}\right) D_n \left(D_n^{-1} F D_p^{-1} u\right).$$

soit maximal, sachant que $u^\top D_p^{-1} u = 1$. La solution est donnée par le vecteur propre principal de

$$D_p \left(D_p^{-1} F^\top D_n^{-1} F D_p^{-1} \right) = F^\top D_n^{-1} F D_p^{-1} \equiv S.$$

Les formules des coordonnées dans le système d'axes sont plutôt complexes et pas très parlantes… mais à remarquer :

1. S et T ont les mêmes p premières valeurs propres.
2. En centrant les profils lignes et colonnes, on peut illustrer le résultat des deux graphiques sur les mêmes axes.

Centre de gravité et inertie

Les logiciels produisent généralement un graphique centré en $(0, 0)$. Il s'agit d'une analyse relative aux centres de gravité des lignes et des colonnes. Cette pratique est à la fois commune et commode. En fait, la masse de la ie ligne est $f_{i\bullet}$, soit la proportion des observations qui tombent sur cette ligne. De façon similaire, la masse de la je colonne est $f_{\bullet j}$. Le centre de gravité des lignes est la moyenne des profils-lignes, mais pondérée par la masse de chaque ligne, et similairement pour les profils-colonnes.

Le centre de gravité des lignes est défini par $G_L = \left(g_1, \ldots, g_p\right)^t op$, où

$$g_j = \sum_{i=1}^{n} f_{i\bullet} \frac{f_{ij}}{f_{i\bullet}} = \sum_{i=1}^{n} f_{ij} = f_{\bullet j}, \quad 1 \leq j \leq p.$$

De même, le centre de gravité des colonnes est défini par

$$G_C = (f_{1\bullet}, \ldots, f_{n\bullet})^t \, op.$$

Centrage des lignes :

$$\frac{f_{ij}}{f_{i\bullet}} - g_j = \frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} = \frac{f_{ij} - f_{i\bullet} f_{\bullet j}}{f_{i\bullet}}.$$

de sorte que, pour tout $i \in \{1, \ldots, n\}$,

$$\sum_{j=1}^{p} \frac{f_{ij} - f_{i\bullet} f_{\bullet j}}{f_{i\bullet}} = 0.$$

L'analyse ne se fait plus sur

$$S = F^\top D_n^{-1} F D_p^{-1},$$

mais plutôt sur $S^\star = (s^\star_{jj'})$, où

$$s^\star_{jj'} = \sum_{i=1}^{n} \frac{\left(f_{ij} - f_{i\bullet} f_{\bullet j}\right)\left(f_{ij'} - f_{i\bullet} f_{\bullet j'}\right)}{f_{i\bullet} f_{\bullet j'}}.$$

Par définition,

$$\mathrm{tr}(S^\star) = \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{\left(f_{ij} - f_{i\bullet} f_{\bullet j}\right)^2}{f_{i\bullet} f_{\bullet j}}.$$

On retrouve l'expression de la statistique du χ^2 servant à tester l'indépendance entre deux variables.

On peut prouver que $s^\star_{jj'} = s_{jj'} - f_{\bullet j}$, où

$$s_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{f_{i\bullet} f_{\bullet j'}}.$$

Ceci entraîne que ces deux matrices ont les mêmes p premiers vecteurs propres normalisés.

Coordonnées des points-lignes. La projection du ie point-ligne sur l'axe j est donnée par

$$\left(D_n^{-1} F \phi_j\right)_i = \frac{1}{f_{i\bullet}} \sum_{j'=1}^{p} f_{ij'} \phi_{jj'} = \sqrt{\lambda_j} \Psi_{ji} \equiv \widehat{\Psi}_{ji}.$$

De plus,

$$\sum_{i=1}^{n} f_{i\bullet} \widehat{\Psi}^2_{ji} = \sum_{i=1}^{n} f_{i\bullet} \left(\sqrt{\lambda_j} \Psi_{ji}\right)^2 = \lambda_j.$$

Coordonnées des point-colonne. La projection du ke point-ligne sur l'axe j est donnée par

$$\left(D_p^{-1} F^\top \psi_j\right)_k = \frac{1}{f_{\bullet k}} \sum_{i=1}^{n} f_{ik} \psi_{ji} = \sqrt{\lambda_j} \Phi_{jk} \equiv \widehat{\Phi}_{jk}.$$

De plus,

$$\sum_{k=1}^{p} f_{\bullet k} \widehat{\Phi}^2_{jk} = \sum_{k=1}^{p} f_{\bullet k} \left(\sqrt{\lambda_j} \Phi_{jk}\right)^2 = \lambda_j.$$

L'inertie absolue du ie point-ligne sur l'axe j est $f_{i\bullet} \widehat{\Psi}^2_{ji}$. L'inertie relative du ie point-ligne sur l'axe j est

$$\frac{f_{i\bullet} \widehat{\Psi}^2_{ji}}{\lambda_j}.$$

L'inertie absolue du ke point-colonne sur l'axe j est $f_{\bullet k}\widehat{\Phi}^2_{jk}$. L'inertie relative du ke point-colonne sur l'axe j est

$$\frac{f_{\bullet k}\widehat{\Phi}^2_{jk}}{\lambda_j}.$$

L'inertie totale est $I = T/k_{\bullet\bullet}$.

la qualité de la représentation du ke point-colonne dans l'axe j est donnée par

$$\frac{d^2_j(k, G_C)}{d^2(k, G_c)} = \cos^2(\theta_{kj}),$$

où θ_{kj} est l'angle entre le point k et sa projection sur l'axe j.

Interprétation :

1. Plus les $\cos^2(\theta_{kj})$ sont élevés, mieux les points sont représentés sur l'axe j.
2. Ceci ne signifie pas pour autant que les points sont près du centre du graphique.
3. Les points éloignés du centre de gravité se distinguent du centre de gravité.
4. Une interprétation semblable existe pour les points-lignes.

Analyse des correspondances multiples

L'analyse des correspondances multiples est une généralisation de l'analyse des correspondances binaires. Elle permet la representation graphique de tableaux de fréquences contnant plus de deux variables. Un exemple classique d'un tableau de fréquences avec plus de deux variables est le tableau présentant les réponses d'individus à un questionnaire contenant Q questions à choix multiples.

Très utile pour visualiser les résultats d'une étude par questionnaire.

L'ACM peut aussi être vue comme une version de l'ACP quand les variables sont catégorielles :

— l'analyse duale permet de voir les individus ayant des profils de réponses similaires
— on peut obtenir des scores continus pour les individus qui capturent une grande partie de l'information
— donc aussi utile pour scorer les résultats d'une étude par questionnaire dans un but éventuel de partitionnement, par exemple

En général, pour un questionnaire contenant Q questions, on a un tableau de la forme suivante :

$$Z = \left[Z_1 \mid \cdots \mid Z_Q\right].$$

Notation :

— Q : nombre de questions
— n : nombre d'individus répondant au questionnaire
— p_q : nombre de modalités (choix de réponses) de la question q.
— $p = p_1 + p_Q$

Potentiel problème : plus le nombre de questions est grand, plus il y aura de cellules vides. C'est aussi le cas si le nombre de réponses aux questions est important. La proportion de cellules non vides est

$$\frac{nQ}{np} = \frac{Q}{p}.$$

.

Si toutes les questions ont le même nombre de choix de réponses, alors $p_1 = \cdots = p_Q = \frac{p}{Q}$, de sorte que

$$\frac{Q}{p} = \frac{1}{p_1} \longrightarrow 0, \quad \text{quand } p_1 \to \infty.$$

Le tableau résumé est un tableau de taille $n \times Q$. Il contient le numéro de la modalité associée à la réponse de chaque individu pour chacune des questions.

La tableau de Burt est une autre façon de présenter un tableau de fréquences contenant plus de deux variables. Étant donné un tableau logique $Z = \big[Z_1 \mid \cdots \mid Z_Q\big]$, le tableau de Burt associé est la matrice carrée $p \times p$ définie comme étant

$$B = \begin{pmatrix} Z_1^\top Z_1 & \cdots & Z_1^\top Z_Q \\ \vdots & \ddots & \vdots \\ Z_Q^\top Z_1 & \cdots & Z_Q^\top Z_Q \end{pmatrix}.$$

Propriétés de $Z_q^\top Z_q$

1. $Z_q^\top Z_q$ est une matrice diagonale $p_q \times p_q$ présentant les réponses à la qe question.
2. L'élément (j, j) de la matrice $Z_q^\top Z_q$ est égal au nombre d'individus d_{jj} qui appartiennent à la je catégorie de la qe question.
3. $Z_q^\top Z_r$ est le tableau de fréquences présentant les répones au x qe et re questions.
4. L'élément (j, k) de la matrice $Z_q^\top Z_r$ est égal au nombre d'individus d_{jk} qui appartiennent à la je catégorie de la qe question et à la ke catégorie de la re question.

D'un point de vue mathématique, l'ACM est une AFC effectuée sur la matrice logique Z ou sur le tableau de Burt B. On peut démontrer que l'on obtient les mêmes facteurs, et ce, peu importe la matrice utilisé pour l'analyse.

Note

On peut créé un graphique comme l'AFC. Cependant, en ACM, la distance entre les points de même couleur et la géométrie globale du graphique ne peuvent pas s'interpréter comme en AFC. En fait, on s'intéresse aux points qui sont dans une même direction par rapport à l'origine.