Variables qualitatives : analyse des correspondances

Jean-Marc Lasgouttes

`http://ana-donnees.lasgouttes.net/`

L'analyse factorielle des correspondances

But On cherche à décrire la liaison entre *deux* variables qualitatives.

Exemple on peut regarder la répartition de la couleur des yeux en fonction de la couleur des cheveux.

Différence avec l'ACP l'ACP se fait dans un cadre différent ; les variables sont quantitatives et donc

- il est possible de faire des opérations mathématiques sur les valeurs des variables ;
- par contre, il n'est en général pas possible de compter les individus qui ont une caractéristique donnée (taille=1,83m)

Pourquoi deux variables ? le cas de plus de deux variables est l'analyse de correspondance multiples, traité plus tard dans le cours.

Partie I. Les données qualitatives

Variables qualitatives

Soit $\mathcal{X}$ une variable qualitative. On dispose d'un échantillon de n individus sur lesquels la variable est mesurée.

Modalités (ou catégories) les valeurs que peut prendre une variable qualitative ; si la variable a m modalités (valeurs possibles), on note x_i, $1 \leq i \leq m$, ces modalités, ou plus simplement i.

Effectif le nombre d'occurrence de la modalité i dans l'échantillon ; on le note n_i, et on a $\sum_{i=1}^{m} n_i = n$.

Profil c'est l'ensemble des valeurs n_i/n ; la somme du profil sur les modalités est 1.

Tableau de contingence

Soient $\mathcal{X}_1$ et $\mathcal{X}_2$ deux variables qualitatives à m_1 et m_2 modalités respectivement décrivant un ensemble de n individus.

Définition le tableau de contingence est une matrice à m_1 lignes et m_2 colonnes renfermant les effectifs n_{ij} d'individus tels que $\mathcal{X}_1 = i$ et $\mathcal{X}_2 = j$.

$$\mathbf{N} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1m_2} \\ n_{21} & n_{22} & & \\ & \ddots & & \\ \vdots & & n_{ij} & \vdots \\ & & & \ddots \\ n_{m_1 1} & & & n_{m_1 m_2} \end{pmatrix}$$

La constitution de ce tableau est aussi appel un « tri croisé ».

Marges et profils

Marge en ligne c'est la somme $n_{i\cdot} = \sum_{j=1}^{m_2} n_{ij}$, c'est-à-dire l'effectif total de la modalité i de $\mathcal{X}_1$.

On définit aussi le profil marginal des lignes $n_{i\cdot}/n$.

Marge en colonne c'est la somme $n_{\cdot j} = \sum_{i=1}^{m_1} n_{ij}$, c'est-à-dire l'effectif total de la modalité j de $\mathcal{X}_2$.

On définit aussi le profil marginal des colonnes $n_{\cdot j}/n$.

Deux lectures possibles selon la variable que l'on privilégie, on peut définir

- le tableau des *profils-lignes* $n_{ij}/n_{i\cdot}$, qui représente la fréquence de la modalité j conditionnellement à $\mathcal{X}_1 = i$; la somme de chaque ligne est ramenée à 100%.
- le tableau des *profils-colonnes* $n_{ij}/n_{\cdot j}$, qui représente la fréquence de la modalité i conditionnellement à $\mathcal{X} = j$; la somme de chaque colonne est ramenée à 100%.

Propriétés des profils

Moyenne la moyenne des profils-lignes (avec poids correspondant aux profils marginaux des lignes) est le profil marginal des colonnes :

$$\sum_{i=1}^{m_1} \frac{n_{i\cdot}}{n} \times \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n},$$

et de même pour les colonnes $\sum_{j=1}^{m_2} \frac{n_{\cdot j}}{n} \times \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{n}$.

Indépendance empirique lorsque tous les profils lignes sont identiques, il y a indépendance entre $\mathcal{X}_1$ et $\mathcal{X}_2$, puisque la connaissance de $\mathcal{X}_1$ ne change pas la répartition de $\mathcal{X}_2$. On a pour tout j

$$\frac{n_{1j}}{n_{1\cdot}} = \frac{n_{2j}}{n_{2\cdot}} = \cdots = \frac{n_{m_1 j}}{n_{m_1\cdot}} = \frac{n_{1j} + \cdots + n_{m_1 j}}{n_{1\cdot} + \cdots + n_{m_1\cdot}} = \frac{n_{\cdot j}}{n}$$

et donc $n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$.

Partie II. Géométrie de nuages de profils

Analyse des correspondances de deux variables : les données

Effectifs on a un tableau de contingence $\mathbf{N}$ à m_1 lignes et m_2 colonnes résultant du croisement de deux variables qualitatives $\mathcal{X}_1$ et $\mathcal{X}_2$ à m_1 et m_2 modalités respectivement. On note $\mathbf{D}_1$ et $\mathbf{D}_2$ les matrices diagonales des effectifs marginaux

$$\mathbf{D}_1 = \begin{pmatrix} n_{1\cdot} & & & 0 \\ & n_{2\cdot} & & \\ & & \ddots & \\ 0 & & & n_{m_1\cdot} \end{pmatrix}$$

$$\mathbf{D}_2 = \begin{pmatrix} n_{\cdot 1} & & & 0 \\ & n_{\cdot 2} & & \\ & & \ddots & \\ 0 & & & n_{\cdot m_2} \end{pmatrix}$$

Profils le tableau des profils des lignes $n_{ij}/n_{i\cdot}$ est donné par $\mathbf{D}_1^{-1}\mathbf{N}$ et celui des profils des colonnes $n_{ij}/n_{\cdot j}$ par $\mathbf{N}\mathbf{D}_2^{-1}$.

Représentation géométrique des profils

Nuage de points les profils-lignes forment un nuage de m_1 points de $\mathbb{R}^{m_2}$. Chaque point est affecté d'un poids égal à sa fréquence marginale $n_{i\cdot}/n$, et la matrice des poids est donc $\frac{1}{n}\mathbf{D}_1$.

Centre de gravité c'est le profil marginal des colonnes car

$$\mathbf{g}_\ell = \frac{1}{n}(\mathbf{D}_1^{-1}\mathbf{N})'\mathbf{D}_1\mathbf{1}_{m_1} = \left(\frac{n_{\cdot 1}}{n}, \cdots, \frac{n_{\cdot m_2}}{n}\right)'$$

Profils-colonnes les lignes du tableau $\mathbf{D}_2^{-1}\mathbf{N}'$ forment un nuage de m_2 points de $\mathbb{R}^{m_1}$, avec matrice de poids $\frac{1}{n}\mathbf{D}_2$ et centre de gravité

$$\mathbf{g}_c = \left(\frac{n_{1\cdot}}{n}, \cdots, \frac{n_{m_1\cdot}}{n}\right)'$$

Comment étudier ces données

Cas indépendant en cas d'indépendance empirique, on a

$$\frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n} \text{ et } \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{n}.$$

Les deux nuages sont donc réduits à leurs centres de gravité.

Dimension des nuages comme les profils somment à 1, les m_1 profils-lignes sont situés dans le sous-espace W_1 de dimension $m_2 - 1$ défini par $\sum_{j=1}^{m_2} x_j = 1$ et $x_j \geq 0$.

ACP l'étude de la forme des nuages au moyen de l'analyse en composantes principales permettra de rendre compte de la structure des écarts à l'indépendance.

Partie III. L'AFC : une ACP sur un nuage de profils

La métrique du χ^2

Profils-lignes la norme du χ^2 de la différence de profils-lignes $\mathbf{e}_i - \mathbf{e}_{i'}$ est définie par

$$\|\mathbf{e}_i - \mathbf{e}_{i'}\|_{\chi^2_\ell}^2 = \sum_{j=1}^{m_2} \frac{n}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{i'j}}{n_{i'\cdot}}\right)^2,$$

ce qui revient à utiliser la métrique diagonale $n\mathbf{D}_2^{-1}$.

Inertie l'inertie totale du nuage des profils-lignes par rapport à $\mathbf{g}_\ell$ est

$$\begin{aligned} I_{\mathbf{g}_\ell} &= \sum_{i=1}^{m_1} \frac{n_{i\cdot}}{n} \|\mathbf{e}_i - \mathbf{g}_\ell\|_{\chi^2_\ell}^2 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{n_{i\cdot}}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{\cdot j}}{n}\right)^2 \\ &= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{1}{n_{i\cdot}n_{\cdot j}} \left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2 \end{aligned}$$

Cette inertie mesure l'écart à l'indépendance.

Pourquoi la métrique du χ^2 ?

Pondération la pondération $n/n_{\cdot j}$ permet de donner des importances comparables aux différentes « variables ».

Équivalence distributionnelle si deux colonnes j et j' de $\mathbf{N}$ ont le même profil, il est logique de les regrouper en une seule d'effectif $n_{ij}+n_{ij'}$; on a alors quand $n_{ij}/n_{\cdot j} = n_{ij'}/n_{\cdot j'}$

$$\begin{aligned} &\frac{n}{n_{\cdot j}}\left[\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{\cdot j}}{n}\right]^2 + \frac{n}{n_{\cdot j'}}\left[\frac{n_{ij'}}{n_{i\cdot}} - \frac{n_{\cdot j'}}{n}\right]^2 \\ &\qquad = \frac{n}{n_{\cdot j}+n_{\cdot j'}}\left[\frac{n_{ij}+n_{ij'}}{n_{i\cdot}} - \frac{n_{\cdot j}+n_{\cdot j'}}{n}\right]^2 \end{aligned}$$

La distance entre les profils-ligne est donc inchangée.

Propriétés diverses

Profils-colonnes on définit la distance entre deux profils-colonnes $\mathbf{e}_j$ et $\mathbf{e}_{j'}$ comme

$$\|\mathbf{e}_j - \mathbf{e}_{j'}\|_{\chi^2_c}^2 = \sum_{i=1}^{m_1} \frac{n}{n_{i\cdot}} \left(\frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij'}}{n_{\cdot j'}}\right)^2,$$

soit une métrique de matrice $n\mathbf{D}_1^{-1}$.

Dualité On remarque que $I_{\mathbf{g}_\ell} = I_{\mathbf{g}_c}$, notée donc $I_{\mathbf{g}}$.

Norme du vecteur centre de gravité $\mathbf{g}_\ell$

$$\|\mathbf{g}_\ell\|_{\chi^2_\ell}^2 = \sum_{j=1}^{m} \frac{n}{n_{\cdot j}} \left(\frac{n_{\cdot j}}{n}\right)^2 = 1.$$