Séance 2: Analyse Factorielle des Correspondances

Révisions

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Deuxième partie II

Analyse Factorielle des Correspondances

Données Qualitatives

Notations

- On suppose donnés 2 variables X et Y qualitatives.
- On suppose donnés n individus décrits par ces chacune de ces 2 variables.
- Recherche de la dépendance entre les différentes modalités de X et Y.
- X possède m_1 modalités, Y en possède m_2.
- ♣ Comment résumer les données ?

Tableau de contingence, nuage associés

La table de contingence associée à ces observations, de dimension $m_1 \times m_2$, est souvent notée $\mathbf{T}$ ou N ; son élément générique est $n_{\ell h}$, effectif conjoint. Elle se présente sous la forme suivante :

	y_1	$\cdots$	y_h	$\cdots$	y_{m_2}	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_{m_1}	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	$n_{m_1 +}$
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+m_2}	n

Effectifs Marginaux

On note par D_1 et D_2 les matrices diagonales des effectifs marginaux des variables X et Y :

$$D_1 = \begin{pmatrix} n_{1+} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \dots & \dots & \vdots \\ 0 & \dots & n_{i+} & \ddots & 0 \\ \vdots & \dots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & n_{m_1+} \end{pmatrix} \qquad D_2 = \begin{pmatrix} n_{+1} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \dots & \dots & \vdots \\ 0 & \dots & n_{+j} & \ddots & 0 \\ \vdots & \dots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & n_{+m_2} \end{pmatrix}$$

Profils Lignes

- Tableau des profils lignes d'éléments $\frac{n_{i,j}}{n_{i+}}$ donné par :
- On considère les profils lignes comme m_1 points dans $\mathbb{R}^{m_2}$.
- Chacun de ces points est affecté d'un poids proportionnel à sa fréquence marginale :
- Centre de gravité du nuage de points :

$$g_l =$$

- Les m_1 profils lignes appartiennent à un sous-espace W_2 de dimension $m_2 - 1$ défini par :

Profils Colonnes

- Tableau des profils colonnes d'éléments $\frac{n_{i,j}}{n_{+j}}$ donné par :
- On considère les profils lignes comme m_2 points dans $\mathbb{R}^{m_1}$.
- Chacun de ces points est affecté d'un poids proportionnel à sa fréquence marginale :
- Centre de gravité du nuage de points :

$$g_c =$$

- Les m_2 profils colonnes appartiennent à un sous-espace W_1 de dimension $m_1 - 1$ défini par :

Métrique du χ^2, Indépendance

- ♣ Dans le cas de l'indépendance statistique, on a la relation :
- Pour calculer la distance entre deux profils lignes i et i', on utilise la formule :

$$d^2_{\chi^2}(i,i') =$$

- Il s'agit de la distance basée sur la métrique M_l donnée par

$$M_l =$$

- Cette métrique revient là-encore à donner autant d'importance à chacune des variables.

Métrique du χ^2, Indépendance

- Pour calculer la distance entre deux profils colonnes j et j', on utilise la formule :

$$d^2_{\chi^2}(j,j') =$$

- Il s'agit de la distance basée sur la métrique M_c donnée par

$$M_c =$$

- La quantité φ^2 mesure l'écart à l'indépendance :

$$\varphi^2 = \dots = \dots = \dots$$

Propriétés de la distance du χ^2

Proposition : Étant données deux colonnes de N, j et j' ayant le même profil, si l'on regroupe ces 2 colonnes en une seule d'effectif $n_{ij} + n_{ij'}$ pour chacune des lignes i, alors les distances entre profils lignes est inchangée.
Preuve : ...
♣ Cette propriété est-elle vraie pour la métrique euclidienne ?
Proposition : φ^2 correspond à la fois à l'inertie des profils lignes par rapport à g_l, mais également à l'inertie des profils colonnes par rapport à g_c.

Analyse en composantes principales des deux nuages de profils

ACP profils lignes
Données $X = D_1^{-1}N$
Métrique $M = nD_2^{-1}$
Poids $D = \frac{D_1}{n}$

ACP profils colonnes
Données $X = D_2^{-1}N'$
Métrique $M = nD_1^{-1}$
Poids $D = \frac{D_2}{n}$

Nous verrons que ces deux ACP amènent à des résultats parfaitement duaux l'un de l'autre.

ACP non centrées et facteur trivial

Remarques et propriétés

- $0g_l$ est orthogonal à W_1 pour la métrique du χ^2.
- ♣ $\|g_l\|_{\chi^2} =$
- Proposition : g (g_l ou g_c) est vecteur propre associé à la valeur propre 1.
- Il est donc à chaque fois inutile de préciser ce résultat dans les AFC.

ACP non centrées et facteur trivial

On peut montrer que les facteurs principaux sont :

ACP profils lignes

Facteurs Principaux

$$VP \quad \text{de} \quad D_2^{-1}N'D_1^{-1}N$$

Composantes principales

$$VP \quad \text{de} \quad D_1^{-1}ND_2^{-1}N'$$

Normalisés par

$$a'\frac{D_1}{n}a = \lambda$$

ACP profils colonnes

Facteurs Principaux

$$VP \quad \text{de} \quad D_1^{-1}ND_2^{-1}N'$$

Composantes principales

$$VP \quad \text{de} \quad D_2^{-1}N'D_1^{-1}N$$

Normalisés par

$$b'\frac{D_2}{n}b = \lambda$$

ACP non centrées et facteur trivial

- Les 2 analyses conduisent aux mêmes valeurs propres.
- Les facteurs principaux de l'une sont les composantes principales de l'autre.
- Les coordonnées des points-lignes et points-colonnes s'obtiennent en cherchant les vecteurs propres des produits des deux tableaux de profils

Contributions

- Cercle de corrélation : aucun intérêt dans le contexte de variables qualitatives
- On utilise les contributions des profils lignes ou profils colonnes :

$$\lambda = \sum_{i=1}^{m_1} n_{i+} a_i^2$$

- $$CTR(i) = \frac{\frac{n_{i+}}{n} a_i^2}{\lambda} \qquad CTR(j) = \frac{\frac{n_{+j}}{n} b_j^2}{\lambda}$$

Contributions

Formules de transition :

$$b = \frac{1}{\sqrt{\lambda}} D_2^{-1} N' a \qquad a = \frac{1}{\sqrt{\lambda}} D_1^{-1} N b$$

Autrement dit :

$$b_j = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{m_1} \frac{n_{ij}}{n_{j+}} a_i \qquad a_i = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{m_2} \frac{n_{ij}}{n_{+i}} a_j$$

Reconstitution des données

Si $m_1 < m_2$, en éliminant la valeur propre 1, on a :

$$\varphi^2 = \sum_{k=1}^{m_1 - 1} \lambda_k$$

Les pourcentages de variance sont égaux à :

$$\% Var_k =$$

La formule de reconstitution est :

$$n_{ij} =$$

Données AGR concernent les exploitations agricoles de la région Midi-Pyrénées.

Elles proviennent des ”Tableaux Economiques de Midi-Pyrénées”, publiés par la Direction Régionale de Toulouse de l'INSEE, en 1996 (données relatives à l'année 1993 ; chiffres arrondis à la dizaine près).
Les 73 000 exploitations ont été ventilées dans une table de contingence selon le département (en lignes, 8 modalités) et la SAU (Surface Agricole Utilisée, en colonnes, 6 classes).
Départements : ARIE = Ariège ; AVER = Aveyron ; H.G. = Haute-Garonne ; GERS = Gers ; LOT = Lot ; H.P. = Hautes-Pyrénées ; TARN = Tarn ; T.G. = Tarn-et-Garonne.
SAU : inf05 = moins de 5 hectares ; s0510 = entre 5 et 10 hectares... ; sup50 = plus de 50 hectares.

Représentations graphiques

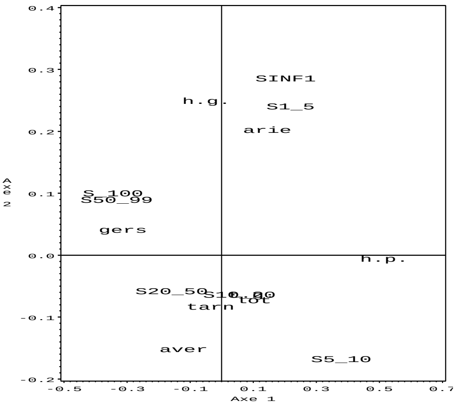

FIG.: *Biplot isométrique des données AGR.*

Interprétation

- ♣ Quelles sont les variables qui sont croisées entre elles ?
- ♣ Que met en évidence le premier axe ?
- ♣ Que met en évidence le second axe ?