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.
- Les réalisations des n individus sont notées $x_1, \ldots x_n$ et $y_1, \ldots y_n$.
- X possède m_1 modalités, Y en possède m_2.
- Les modalités de X sont notées $i_1, \ldots i_{m_1}$
- Les modalités de Y sont notées $j_1, \ldots j_{m_2}$

Données Qualitatives

Objectifs

- Recherche de la dépendance entre les différentes modalités de X et Y.
- Y a-t-il des modalités corrélées entre X et Y ?
- Pourquoi comparer les modalités de X pose problème ? Idem pour Y ?
- Comment résumer les données ?

Tableau de contingence, nuage associés

Définition

- On construit une table de contingence associée à ces observations
- La dimension de la table est $m_1 \times m_2$
- La table est souvent notée $\mathbf{T}$ ou N
- Son élément générique est $n_{\ell h}$, effectif conjoint

$$n_{\ell h} = Card \left\{ i \quad | \quad x_i = i_\ell \quad \text{et} \quad y_i = j_h \right\}$$

Tableau de contingence, nuage associés

Elle se présente sous la forme suivante :

	j_1	$\cdots$	j_h	$\cdots$	j_{m_2}	sommes
i_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i_{m_1}	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	$n_{m_1 +}$
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+m_2}	n

Les effectifs $n_{\ell +}$ et n_{+h} sont définis par

$$n_{\ell +} = \sum_{h=1}^{m_2} n_{\ell h} \quad \text{et} \quad n_{+h} = \sum_{\ell=1}^{m_1} n_{\ell h}$$

Effectifs Marginaux

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

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

La taille de D_1 est $m_1 \times m_1$ alors que D_2 est de taille $m_2 \times m_2$.

Profils Lignes

- On construit à partir de T un tableau de fréquences marginales pour la variable X.
- Tableau des Profils Lignes est composé des éléments

$$\frac{n_{i,j}}{n_{i+}}$$

- C'est la fraction des individus ayant la modalité $X = i$ qui ont la également modalité $Y = j$
- Proposition Ce tableau des profils lignes est donné par la multiplication matricielle

$$PL = D_1^{-1} \times T$$

Profils Colonnes

- On construit à partir de T un tableau de fréquences marginales pour la variable Y.
- Tableau des Profils Colonnes est composé des éléments

$$\frac{n_{i,j}}{n_{+j}}$$

- C'est la fraction des individus ayant la modalité $Y = j$ qui ont la également modalité $X = i$
- Proposition Ce tableau des profils colonnes est donné par la multiplication matricielle

$$PC = D_2^{-1} \times T'$$

Positionnement dimensionnel Profils Lignes

- On considère les Profils Lignes comme m_1 points dans $\mathbb{R}^{m_2}$.
- On le note

$$PL_i = \begin{pmatrix} \frac{n_{i,1}}{n_{i+}} \\ \vdots \\ \frac{n_{i,m_2}}{n_{i+}} \end{pmatrix}$$

- Chacun de ces points est affecté d'un poids proportionnel à sa fréquence marginale $\frac{n_{i+}}{n}$
- Centre de gravité du nuage de points :

$$g_l = \frac{1}{n}(D_1^{-1}T)'D_1\mathbf{1} = \begin{pmatrix} n_{+1}/n \\ \vdots \\ n_{+m_2}/n \end{pmatrix}$$

Positionnement dimensionnel Profils Colonnes

- On considère les Profils Colonnes comme m_1 points dans $\mathbb{R}^{m_1}$.
- On le note

$$PC_j = \begin{pmatrix} \frac{n_{1j}}{n_{+j}} \\ \vdots \\ \frac{n_{m_1 j}}{n_{+j}} \end{pmatrix}$$

- Chacun de ces points est affecté d'un poids proportionnel à sa fréquence marginale $\frac{n_{+j}}{n}$
- Centre de gravité du nuage de points :

$$g_c = \begin{pmatrix} n_{1+}/n \\ \vdots \\ n_{m_1+}/n \end{pmatrix}$$

Positionnement dimensionnel

- Les m_1 profils lignes appartiennent à un sous-espace affine W_2 de $\mathbb{R}^{m_2}$.
- W_2 est de dimension $m_2 - 1$ défini par :

$$\forall i \in \{1, \dots m_1\} \qquad \sum_{j=1}^{m_2} PL_i(j) = 1$$

- Les m_2 profils colonnes appartiennent à un sous-espace affine W_1 de $\mathbb{R}^{m_1}$.
- W_1 est de dimension $m_1 - 1$ défini par :

$$\forall j \in \{1, \dots m_2\} \qquad \sum_{i=1}^{m_1} PC_j(i) = 1$$

Métrique du χ^2, Indépendance

- Dans le cas de l'indépendance statistique entre la modalité i de X et la modalité j de Y, on a

$$P(X = i, Y = j) = P(X = i)P(Y = j)$$

- Proposition : Le pendant empirique de cette relation est :

$$n_{ij} = \frac{n_{i+}n_{+j}}{n}$$

- Pour calculer la distance entre deux profils lignes i et i', on utilise la formule :

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{m_2} \frac{n}{n_{+j}} \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{i'j}}{n_{i'+}} \right)^2 = D_{M_l}(PL_i, PL_{i'})$$

- Proposition La métrique M_l est donnée par $M_l = nD_2^{-1}$
- Cette métrique revient là-encore à donner autant d'importance à chacune des modalités de Y.

Métrique du χ^2, Indépendance

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

$$d^2_{\chi^2}(j,j') = \sum_{i=1}^{m_1} \frac{n}{n_{i+}} \left(\frac{n_{ij}}{n_{+j}} - \frac{n_{ij'}}{n_{+j'}} \right)^2 = D_{M_c}(PC_j, PC_{j'})$$

- Proposition La métrique M_c est donnée par $M_c = nD_1^{-1}$
- Cette métrique revient là-encore à donner autant d'importance à chacune des modalités de X.

Métrique du χ^2, Indépendance

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

$$\varphi^2 = \frac{1}{n} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}$$

- Proposition : L'inertie des Profils Lignes et l'inertie des Profils Colonnes coïncident et valent le φ^2.

Propriétés de la distance du χ^2

- Proposition : Étant données deux colonnes de T, 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.
- Proposition : Étant données deux lignes de T, i et i' ayant le même profil, si l'on regroupe ces 2 lignes en une seule d'effectif $n_{i'j} + n_{ij}$ pour chacune des colonnes j, alors les distances entre profils colonnes est inchangée.
- Cette propriété est-elle vraie pour la métrique euclidienne ?

Analyse en composantes principales des deux nuages de profils

ACP profils lignes
Données $X = D_1^{-1} T$
Métrique $M = n D_2^{-1}$
Poids $D = \frac{D_1}{n}$

ACP profils colonnes
Données $X = D_2^{-1} T'$
Métrique $M = n D_1^{-1}$
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

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

ACP non centrées et facteur trivial

Théorème :

ACP profils lignes

Facteurs Principaux

$$VP \quad \text{de} \quad D_2^{-1}T'D_1^{-1}T$$

Composantes principales

$$VP \quad \text{de} \quad D_1^{-1}TD_2^{-1}T'$$

Normalisés par

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

ACP profils colonnes

Facteurs Principaux

$$VP \quad \text{de} \quad D_1^{-1}TD_2^{-1}T'$$

Composantes principales

$$VP \quad \text{de} \quad D_2^{-1}T'D_1^{-1}T$$

Normalisés par

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

ACP non centrées et facteur trivial

Théorème :

- 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 a la relation entre les valeurs propres λ et les vecteurs propres :

$$\lambda = \frac{1}{n}\sum_{i=1}^{m_1} n_{i+}a_i^2 = \frac{1}{n}\sum_{j=1}^{m_2} n_{+j}b_j^2$$

- On définit la contribution des profils lignes et colonnes par :

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

Formules de transition

Théorème :

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

C'est-à-dire :

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

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 :

$$\% Var_k = \frac{\lambda_k}{\varphi^2}$$

La formule de reconstitution est :

$$n_{ij} = \frac{n_{i+}n_{+j}}{n} \left(1 + \sum_k \frac{a_i^k b_j^k}{\sqrt{\lambda_k}} \right)$$

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

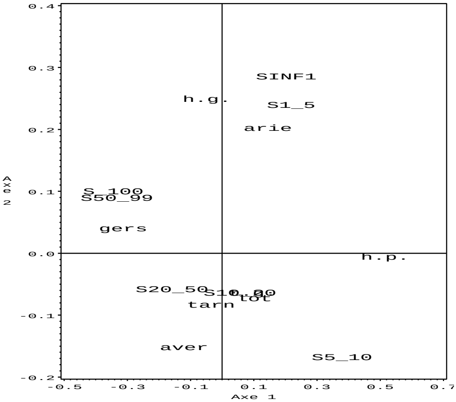

FIG.: *Biplot isométrique des données AGR.*

Interprétation

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