Séance 4: Analyse Factorielle des Correspondances Multiples

Révisions

Sébastien Gadat

Laboratoire de Statistique et Probabilités
UMR 5583 CNRS-UPS

www.lsp.ups-tlse.fr/gadat

Quatrième partie IV

Analyse Factorielle des Correspondances Multiples

Tableau disjonctif complet

- Généralisation de l'AFC pour $p > 2$ observées sur n individus
- Parfois utilisée pour la construction de "scores" afin d'effectuer une méthode de classification
- X variable à c modalités, on définit la variable indicatrice comme

$$\forall k \in \{1 \dots c\} \qquad X_{(k)}(i) = 1 \qquad \text{si} \qquad X(i) = \mathcal{X}_k \qquad \text{et 0 sinon}$$

- On note n_k l'effectif de $\mathcal{X}_k$
- La matrice des indicatrices de X est donnée par son terme général

$$x_i^k = X_{(k)}(i)$$

-

$$\clubsuit \sum_{i=1}^{n} x_i^k = \dots \qquad \clubsuit \sum_{k=1}^{c} x_i^k = \dots$$

Tableau disjonctif complet

- On considère p variables notées $X^1, \dots X^p$
- c_j est le nombre de modalités de X^j
- Le nombre de modalités total c est donné par

$$c = \dots$$

- Le tableau disjonctif complet $\mathbf{X}$ de taille $n \times c$ est obtenue par concaténation

$$\mathbf{X} = |X_1| \dots |X_p|$$

- Chaque sous-matrice X_j est obtenue comme précédemment
- $\mathbf{X}$ vérifie

$$\clubsuit \sum_{i=1}^{n} \sum_{k=1}^{p} x_i^k = \dots \qquad \clubsuit \sum_{k=1}^{p} x_i^k = \dots$$

Tableau de Burt

- On construit à partir de $\mathbf{X}$ le tableau de Burt :

$$\mathcal{B} = \mathbf{X}'\mathbf{X}$$

- $\mathcal{B}$ a pour taille $c \times c$
- On peut écrire $\mathcal{B} = (\mathcal{B}_{j,l}), j, l = 1 \dots p$
- La taille de $\mathcal{B}_{j,l}$ est $c_j \times c_l$

$$\mathcal{B}_{j,l} = \mathbf{X}_\mathbf{j}'\mathbf{X}_\mathbf{l}$$

- Si $j \neq l$, $\mathcal{B}_{j,l}$ est la table de contingence croisant X^j avec X^l
- Si $j = l$, $\mathcal{B}_{j,j}$ est diagonale vérifiant

$$\mathcal{B}_{j,j} = diag(n_1^j, \dots n_{c_j}^j)$$

- $\mathcal{B}$ est symétrique, d'effectifs marginaux $n_j^l p$ et d'effectif total np^2

Démarche

- On s'intéresse aux résultats fournis par l'AFC réalisée sur $|X_1|X_2|$ (table de contingence relative à 2 variables qualitatives)
- On généralise les propriétés obtenues dans ce cas à un nombre plus important de variables (p)
- On définit ainsi l'AFCM

Démarche

- $$T = \mathbf{X} = |X_1|X_2|$$
- r valeurs pour X_1 et c pour X_2
- Matrice des poids $\bar{P} = 1/nI_n$ (cas équipondéré)
- La matrice de la métrique $\bar{D}$ est donnée par

$$\bar{D} = \frac{1}{2}\begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} = \frac{1}{2}\Delta$$

- Le tableau des profils lignes est donné par

$$\bar{PL} = \frac{1}{2}\mathbf{X}'$$

- Le tableau des profils colonnes est donné par

$$\bar{PC} = \frac{1}{n}\mathbf{X}\Delta^{-1}$$

ACP du Profil Ligne

- L'ACP du profil ligne issue de l'AFC réalisée sur $\mathbf{T}$ conduit à l'analyse spectrale de $\bar{PL} \times \bar{PC}$ avec $\bar{PL} \times \bar{PC} = \begin{pmatrix} I_r & B \\ A & I_c \end{pmatrix}$
- Les $r + c$ valeurs propres sont

$$\mu_k = \frac{1 + \sqrt{\lambda_k}}{2} \qquad \text{où} \qquad Sp(\bar{Pl} \times \bar{Pc}) = (\lambda_k)_{k \geq 0} \qquad M = diag(\mu_1, \ldots \mu_{r+c})$$

- Les vecteurs $\bar{D}$ normés se mettent sous la forme

$$\frac{1}{2} \begin{bmatrix} U \\ V \end{bmatrix}$$

où U et V vecteurs propres obtenus en diagonalisant AB et BA.

- Dans la pratique, on ne garde que $\inf(r-1, c-1)$ axes
- La matrice des composantes principales vaut

$$\bar{C}_r = \frac{1}{2}(X_1 C_r + X_2 C_c)\Delta^{-1/2}$$

C_r et C_c composantes principales de l'AFC classique

ACP du Profil Colonne

- On obtient des résultats similaires en opérant une AFC sur les profils colonnes
- On diagonalise la matrice $\bar{PC} \times \bar{PL}$
- Les $r + c$ valeurs propres non nulles sont les μ_k
- Les vecteurs propres associés se mettent sous la forme

$$U = \frac{1}{n} \bar{C}_r M^{-1/2}$$

- Les composantes principales s'écrit

$$\bar{C}_c = \frac{1}{2} \begin{bmatrix} C_r \\ C_c \end{bmatrix} \Delta^{-1/2} M^{1/2}$$

- On obtient ainsi la représentation des modalités des variables

AFC du tableau de Burt

- L'ACP issue de l'AFC du tableau de Burt conduit à l'analyse spectrale de
$$\tilde{PL} \times \tilde{PC} = [\bar{PL} \times \bar{PC}]^2$$
- Les valeurs propres associées vérifient $\nu_k = \mu_k^2$
- Les composantes principales s'écrit
$$\begin{bmatrix} C_r \\ C_c \end{bmatrix} \Delta^{-1/2} M$$

AFC du tableau de Burt

- $\mathbf{X} = |\mathbf{X_1}| \dots |\mathbf{X_p}|$ tableau disjonctif complet
- $\mathcal{B} = \mathbf{X}'\mathbf{X}$ Tableau de Burt
- L'AFCM est l'AFC effectuée sur le tableau de Burt
- On définit $D_j = diag(n_1^j, \dots n_{c_j}^j)/n$ et $\Delta = diag(D_1, \dots D_p)$
- On reprend les notations

$$\mathbf{T} = \mathbf{X} \qquad \bar{P} = I_n/n \qquad \mathcal{D} = \Delta/p \qquad \bar{PL} = \mathbf{X}'/p \qquad \bar{PC} = \mathbf{X}\Delta^{-1}/n$$

- On effectue l'ACP des Profils Lignes via l'analyse spectrale de

$$\bar{PL} \times \bar{PC} = \frac{1}{n}\mathcal{B}\Delta^{-1}$$

- On effectue l'ACP des Profils Colonnes via l'analyse spectrale de

$$\bar{PC} \times \bar{PL} = \frac{1}{np}\mathbf{X}\Delta^{-1}\mathbf{X}'$$

- On effectue l'ACP du tableau de Burt via l'analyse spectrale de

$$[\bar{PL} \times \bar{PC}]^2$$

Interprétation

- L'interprétation se fait de façon comparable aux AFC
- On interprète les proximités et les oppositions entre les modalités des différentes variables
- On privilégie les interprétations sur les modalités suffisamment éloignées du centre du graphique
- Les rapports de valeurs propres ne sont pas interprétables mais on regarde la décroissance des valeurs propres pour choisir la dimension
- Seules les contributions des modalités à l'inertie selon les axes sont interprétables

Interprétation

Données : trois centres hospitaliers (Boston, Glamorgan, Tokyo) sur des patientes atteintes d'un cancer du sein. Étudier la survie de ces patientes, trois ans après le diagnostic. En plus de cette information, quatre autres variables sont connues pour chacune des patientes :

- le centre de diagnostic,
- la tranche d'âge,
- le degré d'inflammation chronique,
- l'apparence relative (bénigne ou maligne).

L'objectif de cette étude est une analyse descriptive de cette table en recherchant à mettre en évidence les facteurs de décès.

Données

TAB.: *Données sous la forme d'une table de contingence complète*

			Histologie			
			Inflammation minime		Grande inflammation	
Centre	*Âge*	*Survie*	Maligne	Bénigne	Maligne	Bénigne
Tokyo	< 50	non	9	7	4	3
		oui	26	68	25	9
	$50 - 69$	non	9	9	11	2
		oui	20	46	18	5
	> 70	non	2	3	1	0
		oui	1	6	5	1
Boston	< 50	non	6	7	6	0
		oui	11	24	4	0
	$50 - 69$	non	8	20	3	2
		oui	18	58	10	3
	> 70	non	9	18	3	0
		oui	15	26	1	1
Glamorgan	< 50	non	16	7	3	0
		oui	16	20	8	1
	$50 - 69$	non	14	12	3	0
		oui	27	39	10	4
	> 70	non	3	7	3	0
		oui	12	11	4	1

Résultats

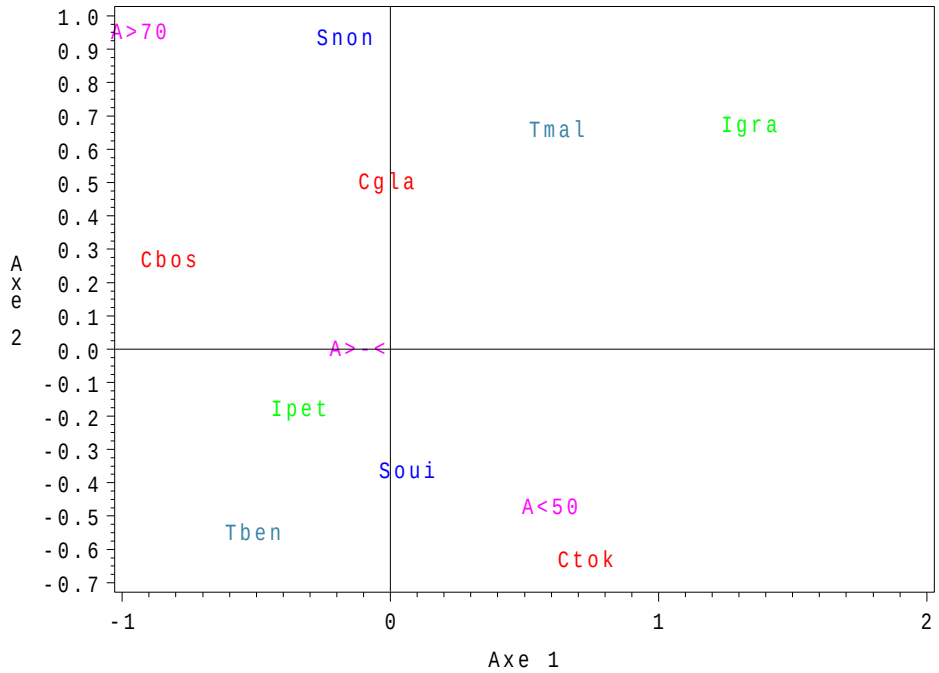

FIG.: *Cancer du sein : analyse des données brutes.*

La variable survie, qui joue en quelques sortes le rôle de variable à expliquer, est très proche de l'axe 2 et semble liée à chacune des autres variables.

Résultats

- Les variables ”centre” et ”âge” sont croisées, pour construire une variable ”c x âge”, à 9 modalités.
- Les variables ”inflam” et ”appar” sont également croisées pour définir la variable ”histol”, à 4 modalités.

Une nouvelle analyse est alors réalisée en considérant comme actives les deux variables nouvellement créées, ainsi que la variable ”survie”, et comme illustratives les variables initiales : ”centre, âge, inflam, appar”. Les résultats sont donnés dans la figure suivante.

Résultats

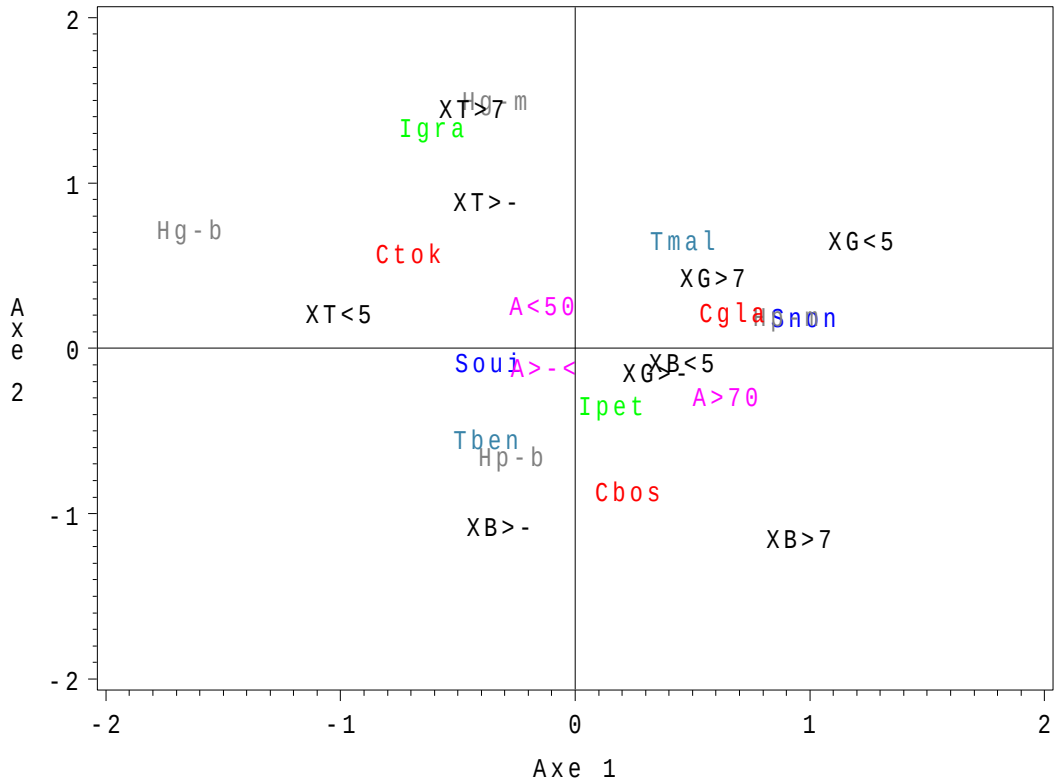

FIG.: *Cancer du sein : analyse des interactions.*