L'Analyse Factorielle des Correspondences (AFC)

Marie Chavent

http://www.math.u-bordeaux.fr/ machaven/

2014-2015

Le but est l'analyse des relations entre deux variables qualitatives. L'AFC s'applique au tableau de contingence $\mathbf{K}$ obtenu à partir du croisement de deux variables qualitatives X_1 et X_2 sur un échantillon de taille n :

$$\mathbf{K} = \begin{array}{c|ccc|c} & 1\dots & s & \dots m & \\ \hline 1 & & & & \\ \vdots & & \vdots & & \\ i & \dots & n_{is} & \dots & n_{i.} \\ \vdots & & \vdots & & \\ q & & & & \\ \hline & & n_{.s} & & \end{array}$$

1 Rappels et notations

Matrice des fréquences $\mathbf{F}$:

$$\mathbf{F} = \begin{array}{c|ccc|c} & 1\dots & s & \dots m & \\ \hline 1 & & & & \\ \vdots & & \vdots & & \\ i & \dots & f_{is} = \frac{n_{is}}{n} & \dots & f_{i.} \\ \vdots & & \vdots & & \\ q & & & & \\ \hline & & f_{.s} & & \end{array}$$

On note :

— $\mathbf{r} = (f_{1.}, \dots, f_{i.}, \dots, f_{q.})^t \in \mathbb{R}^q$

— $\mathbf{c} = (f_{1.}, \dots, f_{.s}, \dots, f_{.m})^t \in \mathbb{R}^m$

— $\mathbf{D}_r$=diag($\mathbf{r}$)

— $\mathbf{D}_c$=diag($\mathbf{c}$)

Matrice des profil-lignes **L** :

$$\mathbf{L} = \begin{array}{c|ccc|}
 & 1\ldots & s & \ldots m \\
\hline
1 & & & \\
\vdots & & & \\
i & \ldots & f_{is}/f_{i.} & \ldots \\
\vdots & & & \\
\hline\hline
\mathbf{c}' & & f_{.s} & \\
\end{array}$$

On a :

- $\mathbf{L} = \mathbf{D}_r^{-1}\mathbf{F}$
- Profil ligne moyen : $\mathbf{c}$

Matrice des profil-lignes centrés **L** :

$$\mathbf{L} = \begin{array}{c|ccc|}
 & 1\ldots & s & \ldots m \\
\hline
1 & & & \\
\vdots & & & \\
i & \ldots & \frac{f_{is}-f_{i.}f_{.s}}{f_{i.}} & \ldots \\
\vdots & & & \\
q & & & \\
\hline
 & & & \\
\end{array}$$

On a :

- $\mathbf{L} = \mathbf{D}_r^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)$
- Profil-ligne moyen : origine de $\mathbb{R}^m$

Matrice des profil-colonnes **C** :

$$\mathbf{C} = \begin{array}{c|ccc||c}
 & 1\ldots & s & \ldots m & \mathbf{r} \\
\hline
1 & & & & \\
\vdots & & & & \\
i & \ldots & f_{is}/f_{.s} & \ldots & f_{i.} \\
\vdots & & & & \\
q & & & & \\
\hline
 & & & & \\
\end{array}$$

On a :

- $\mathbf{C} = \mathbf{F}\mathbf{D}_c^{-1}$
- Profil colonne moyen : $\mathbf{r}$

Matrice des profil-colonnes centrés **C** :

$$\mathbf{C} = \begin{array}{c|ccc|}
 & 1 \ldots & s & \ldots m \\
\hline
1 & & & \\
\vdots & & & \\
i & \ldots & \frac{f_{is} - f_{i.} f_{.s}}{f_{.s}} & \ldots \\
\vdots & & & \\
q & & & \\
\hline
 & & &
\end{array}$$

On a :

— $\mathbf{C} = (\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D}_c^{-1}$

— Profil-colonne moyen : origine de $\mathbb{R}^q$

Deux nuages de points pondérés :

Le nuage des q profil-lignes centrés de $\mathbb{R}^m$ avec :

— $\mathbf{r}$ comme pondération,

— $\mathbf{D}_r$ comme métrique sur $\mathbb{R}^q$ (métrique des poids),

— $\mathbf{D}_c^{-1}$ comme métrique sur $\mathbb{R}^m$ (distance du χ^2).

Le nuage des m profil-colonnes centrés de $\mathbb{R}^q$ avec :

— $\mathbf{c}$ comme pondération,

— $\mathbf{D}_r^{-1}$ comme métrique sur $\mathbb{R}^q$ (distance du χ^2),

— $\mathbf{D}_c$ comme métrique sur $\mathbb{R}^m$ (métrique des poids).

Les inerties de ces deux nuages de points avec ces métriques et pondérations vérifient la propriétée suivante :

$$I(\mathbf{L}) = I(\mathbf{C}) = \chi^2/n$$

Objectif de l'AFC et plan du cours :

Il s'agira d'analyser les deux nuages de points pondérés c'est à dire le nuage des profil-lignes (les lignes de la matrice **L**) et le nuage des profil-colonnes (les colonnes de la matrice **C**). On va donc analyser les lignes et les colonnes de deux matrices différentes (alors qu'en ACP, on analyse les lignes et les colonnes de la même matrice de données quantitatives **Z**).

Pour cela, on va projeter "au mieux" :

— les profil-lignes (les modalités de X_1) dans une sous-espace vectoriel (s.e.v.) de $\mathbb{R}^m$,

— les profil-colonnes (les modalités de X_2) dans une sous-espace vectoriel (s.e.v.) de $\mathbb{R}^q$.

Dans ce cours nous allons présenter l'AFC comme une double ACP (ACP de **L** et ACP de **C**). Puis nous montrerons que les résultats de cette double ACP peuvent être obtenus à partir de l'ACP d'une seule et même matrice c'est à dire à partir de la décomposition en valeurs singulières (DVSG) de la matrice dîte des écarts à l'indépendance. Nous utiliserons ce résultat pour démontrer les propriétés barycentriques qui sont fondamentale pour l'interpétation des résultats.

2 ACP de la matrice des profil-lignes centrés

Les q modalités de la variable X_1 sont décrites par les lignes de la matrice des profil-lignes centrés $\mathbf{L} = \mathbf{D}_r^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)$. Les profil-lignes centrés sont des points de $\mathbb{R}^m$:

— Ces points sont pondérés par les poids des lignes (vecteur $\mathbf{r}$).

— On utilise comme métrique pour comparer deux profil-ligne la distance du χ^2 définie par $\mathbf{D}_c^{-1}$.

On veut projeter "au mieux" les q modalités de X_1 sur un s.e.v. de $\mathbb{R}^m$ de dimension k (pour k=2 on projette sur un plan par exemple) : on veut que les distances entre les modalités projetées soient "aussi proche que possible" des distances entre les modalités dans leur espace d'origine. Ce s.e.v est définit par k axes $\Delta_1, \ldots, \Delta_k$, tels que pour chaque axe, la variance des $\mathbf{D}_c^{-1}$-projections (projections $\mathbf{D}_c^{-1}$ orthogonales) des profil-lignes soit maximale ($\mathbf{D}_r$ norme maximale). Ces axes sont engendrés par des vecteurs $\mathbf{v}_1, ..., \mathbf{v}_k$ de $\mathbb{R}^m$. Ces vecteurs doivent être $\mathbf{D}_c^{-1}$-normés à 1 ($\mathbf{v}_\alpha^t \mathbf{D}_c^{-1} \mathbf{v}_\alpha = 1$, $\alpha = 1, \ldots, k$) et $\mathbf{D}_c^{-1}$-orthogonaux ($\mathbf{v}_\alpha^t \mathbf{D}_c^{-1} \mathbf{v}_{\alpha'} = 0$, $\forall \alpha \neq \alpha'$).

On note $\mathbf{x}^\alpha = \mathbf{L}\mathbf{D}_c^{-1}\mathbf{v}_\alpha$ le vecteur de $\mathbb{R}^q$ des projections des q modalités sur l'axe Δ_α avec $\mathbf{v}_\alpha$ qui maximise $\text{Var}(\mathbf{x}^\alpha)$. Les vecteurs $\mathbf{v}_\alpha$ sont les colonnes d'une matrice notée $\mathbf{V}_k$ de dimension $m \times k$. On note enfin $\mathbf{X} = \mathbf{L}\mathbf{D}_c^{-1}\mathbf{V}_k$ la matrice de dimension $q \times k$ des coordonnées des profil-lignes projetés sur $\Delta_1, \ldots, \Delta_k$:

— $\mathbf{X}$ est la matrice des coordonnées factorielles des profil-lignes.

— $\mathbf{x}^\alpha$ est la αème composante principale des profil-lignes.

L'ACP du triplet $(\mathbf{L}, \mathbf{D}_r, \mathbf{D}_c^{-1})$ donne les résultats suivant :

a) $\mathbf{X} = \mathbf{L}\mathbf{D}_c^{-1}\mathbf{V}_k$ où $\mathbf{V}_k$ est la matrice dont les colonnes sont les k vecteurs propres associés aux k plus grandes valeurs propres $\lambda_1, \ldots, \lambda_k$ de $\mathbf{L}^t\mathbf{D}_r\mathbf{L}\mathbf{D}_c^{-1}$.
$\mathbf{V}_k$ est $\mathbf{D}_c^{-1}$-orthonormée : $\mathbf{V}_k^t\mathbf{D}_c^{-1}\mathbf{V}_k = \mathbb{I}_k$.

b) $\text{Var}(\mathbf{x}^\alpha) = \lambda_\alpha$ et $\bar{\mathbf{x}}^\alpha = 0$.

c) Si $k = \text{rang}(\mathbf{L})$ alors $I(\mathbf{X}) = \lambda_1 + \ldots + \lambda_k = I(\mathbf{L}) = \chi^2/n$.

Remarque notation : En ACP, on notait Ψ la matrice des coordonnées factorielles des lignes d'une matrice quantitative $\mathbf{Z}$ et Φ la matrice des coordonnées factorielles des colonnes de cette même matrice. Ici $\mathbf{X}$ correspond à la matrice Ψ des coordonnées factorielles de $\mathbf{L}$. On ne s'interrèsse pas à la matrice Φ des coordonnées factorielles des colonnes de $\mathbf{L}$.

Exercice 1 : Démontrer a) b) et c) en vous aidant du poly "Rappels sur l'ACP avec métrique".

Exemple

3 ACP de la matrice des profil-colonnes centrés

Les m modalités de la variable X_2 sont décrites par les colonnes de la matrice des profil-colonnes centrés $\mathbf{C} = (\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D}_c^{-1}$. Les profil-colonnes centrés sont de points de $\mathbb{R}^q$:

— Ces points sont pondérés par les poids des colonnes (vecteur $\mathbf{c}$).

— On utilise comme métrique pour comparer deux profil-colonnes la distance du χ^2 définie par $\mathbf{D}_r^{-1}$.

On veut maintenant projeter "au mieux" les m modalités de X_2 sur un s.e.v. de $\mathbb{R}^q$ de dimension k. Ce s.e.v est définit par k axes $G_1, \ldots, G_k$, tels que pour chaque axe, la variance des $\mathbf{D}_r^{-1}$-projections des profil-colonnes soit maximale ($\mathbf{D}_c$ norme maximale). Ces axes sont engendrés par des vecteurs $\mathbf{u}_1, ..., \mathbf{u}_k$ de $\mathbb{R}^q$. Ces vecteurs doivent être $\mathbf{D}_r^{-1}$-normés à 1 et $\mathbf{D}_r^{-1}$-orthogonaux.

On note $\mathbf{y}^\alpha = \mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{u}_\alpha$ le vecteur de $\mathbb{R}^m$ des projections des m modalités sur l'axe G_α avec $\mathbf{u}_\alpha$ qui maximise $\mathrm{Var}(\mathbf{y}^\alpha)$. Les vecteurs $\mathbf{u}_\alpha$ sont les colonnes de la matrice $\mathbf{U}_k$ de dimension $q \times k$. On note enfin $\mathbf{Y} = \mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{U}_k$ la matrice de dimension $m \times k$ des coordonnées des profil-lignes projetés sur $G_1, \ldots, G_k$:

— $\mathbf{Y}$ est la matrice des coordonnées factorielles des profil-colonnes.

— $\mathbf{y}^\alpha$ est la αème composante principale des profil-colonnes.

L'ACP du triplet $(\mathbf{C}, \mathbf{D}_r^{-1}, \mathbf{D}_c))$ donne les résultats suivant :

a) $\mathbf{Y} = \mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{U}_k$ où les colonnes de la matrice $\mathbf{U}_k$ sont les k vecteurs propres associés aux k plus grandes valeurs propres $\lambda_1, \ldots, \lambda_k$ de la matrice $\mathbf{C}\mathbf{D}_c\mathbf{C}^t\mathbf{D}_r^{-1}$.
$\mathbf{U}_k$ est $\mathbf{D}_c^{-1}$-orthonormée : $\mathbf{U}_k^t\mathbf{D}_r^{-1}\mathbf{U}_k = \mathbb{I}_k$.

b) $\mathrm{Var}(\mathbf{y}^\alpha) = \lambda_\alpha$ et $\bar{\mathbf{y}}^\alpha = 0$

c) Si $k = \mathrm{rang}(\mathbf{C})$ alors $I(\mathbf{X}) = \lambda_1 + \ldots + \lambda_k = I(\mathbf{C}) = I(\mathbf{L}) = \chi^2/n$.

Remarque notation : Ici $\mathbf{Y}$ correspond à la matrice Φ des coordonnées factorielles des colonnes de $\mathbf{C}$. On ne s'intéresse pas à la matrice Ψ des coordonnées factorielles des lignes de $\mathbf{C}$.

Exemple

4 AFC : la SVD généralisée d'une seule matrice

Les matrices $\mathbf{X}$ et $\mathbf{Y}$ des coordonnées factorielles des profil-lignes et des profil-colonnes obtenus dans les deux sections précédentes à partir de l'ACP de deux triplets, peuvent être obtenus à partir de l'ACP du seul triplet $(\mathbf{Z}, \mathbf{N}, \mathbf{M})$ avec :

— $\mathbf{Z} = \mathbf{D}_r^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D}_c^{-1}$ la matrice des écarts à l'indépendance

— $\mathbf{N} = \mathbf{D}_r$

— $\mathbf{M} = \mathbf{D}_c$

On effectue donc la DVS de $\mathbf{Z}$ avec les métriques $\mathbf{N}$ et $\mathbf{M}$:

$$\mathbf{Z} = \mathbf{U}\Lambda\mathbf{V}^t$$

où

— $\Lambda = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ est la matrice des valeurs singulières de $\mathbf{Z}\mathbf{N}\mathbf{Z}^t\mathbf{M}$ et $\mathbf{Z}^t\mathbf{N}\mathbf{Z}\mathbf{M}$.

— $\mathbf{U}$ est la matrice de dimension $n \times r$ dont les colonnes sont les vecteurs propres de $\mathbf{Z}\mathbf{M}\mathbf{Z}^t\mathbf{N}$ et $\mathbf{U}^t\mathbf{N}\mathbf{U} = \mathbb{I}_r$ (les vecteurs propres sont $\mathbf{N}$-orthonormés).

— $\mathbf{V}$ est la matrice de dimension $p \times r$ dont les colonnes sont les vecteurs propres de $\mathbf{Z}^t\mathbf{N}\mathbf{Z}\mathbf{M}$ et $\mathbf{V}^t\mathbf{M}\mathbf{V} = \mathbb{I}_r$ (les vecteurs propres sont $\mathbf{M}$-orthonormés).

On a alors

$$\begin{cases} \mathbf{X} = \mathbf{Z}\mathbf{M}\mathbf{V}_k \\ \mathbf{Y} = \mathbf{Z}^t\mathbf{N}\mathbf{U}_k \end{cases}$$

et on en déduit

$$\begin{cases} \mathbf{X} = \mathbf{U}_k\Lambda_k \\ \mathbf{Y} = \mathbf{V}_k\Lambda_k \end{cases}$$

En pratique, pour effectuer la SVD généralisée d'une matrice $\mathbf{Z}$ avec les métriques $\mathbf{N}$ et $\mathbf{M}$, on effectue la SVD de $\tilde{\mathbf{Z}} = \mathbf{N}^{1/2}\mathbf{Z}\mathbf{M}^{1/2}$ avec les métriques $\mathbb{I}_n$ et $\mathbb{I}_p$ (implémentée dans les logiciels comme R). On trouve $\tilde{\mathbf{Z}} = \tilde{\mathbf{U}}\tilde{\Lambda}\tilde{\mathbf{V}}^t$ et on a ensuite :

$$\begin{aligned} \mathbf{U} &= \mathbf{N}^{-1/2}\tilde{\mathbf{U}} \\ \mathbf{V} &= \mathbf{M}^{-1/2}\tilde{\mathbf{V}} \\ \Lambda &= \tilde{\Lambda} \end{aligned}$$

5 Propiétés barycentriques

Une composante principale standardisée est une composante divisée par son écart-type : $\mathbf{x}_\alpha/\sqrt{\lambda_\alpha}$ ou $\mathbf{y}_\alpha/\sqrt{\lambda_\alpha}$. La matrice Λ_k étant la matrice diagonale des racines carrés des valeurs propres, les matrices $\mathbf{X}^*$ et $\mathbf{Y}^*$ des coordonnées factorielles standardisés s'écrivent :

$$\begin{cases} \mathbf{X}^* = \mathbf{X}\Lambda_k^{-1} \\ \mathbf{Y}^* = \mathbf{Y}\Lambda_k^{-1} \end{cases} \text{donc} \begin{cases} \mathbf{X}^* = \mathbf{U}_k \\ \mathbf{Y}^* = \mathbf{V}_k \end{cases}$$

On a les relations suivantes :

$$\begin{cases} \mathbf{X} = \mathbf{D}_r^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{Y}^* \\ \mathbf{Y} = \mathbf{D}_c^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)^t\mathbf{X}^* \end{cases}$$

Exercice 2 : Retrouvez ces relations.

Ces deux relations s'interprètent aussi en terme de moyennes réciproques : la coordonée factorielle d'une modalité d'une variable est la moyenne (pondérée) des coordonnées factorielles (standardisées) des modalités de l'autre variable.
En effet on a les relations barycentriques suivantes :

$$\begin{cases} x_{i\alpha} = \displaystyle\sum_{s=1}^{m} \frac{f_{is}}{f_{i.}} y^*_{s\alpha} \\ y_{s\alpha} = \displaystyle\sum_{i=1}^{q} \frac{f_{is}}{f_{.s}} x^*_{i\alpha} \end{cases}$$

Et on en déduit les relations quasi-barycentriques suivantes :

$$\begin{cases} x_{i\alpha} = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{s=1}^{m} \frac{f_{is}}{f_{i.}} y_{s\alpha} \\ y_{s\alpha} = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{i=1}^{q} \frac{f_{is}}{f_{.s}} x_{i\alpha} \end{cases}$$

Exercice 3 : Retrouvez ces relations.

Interprétations et conséquences de ces relations :

— Au coefficient de dilatation $\frac{1}{\sqrt{\lambda_\alpha}}$ près, les coordonnées factorielles d'un nuage de points sont, sur un axe, les barycentres des coordonnées factorielles de l'autre nuage.

— Les relations quasi-barycentriques justifient la représentation simultanée des profil-lignes et des profil-colonnes sur un même graphique. Mais attention, la distance entre un profil-ligne et un profi-colonne sur ce graphique s'interpète en terme de liaison.

— La coordonnée de la modalité i est la moyenne des coordonnées des modalités s de l'autre variable, pondérée par les fréquences conditionnelles de s sachant i.

Exemple

6 Interprétation des résultats d'une AFC

Les nuages des profil-lignes et des profil-colonnes sont représentés dans les plans de projection formés par les axes factoriels pris deux à deux. La lecture de ces graphiques nécessite des règles d'interpétation.

6.1 Inertie et test d'indépendance

En ACP normée, l'inertie totale du nuage des point-individus est égale à p le nombre de variables. En AFC, on a vu que l'inertie totale du nuage des profil-lignes est égale à l'inertie totale du nuage des profil-colonnes, et est égale au χ^2 d'indépendance entre les deux variables qualitatives :

$$I(\mathbf{L}) = I(\mathbf{C}) = \chi^2(X_1, X_2)/n$$

La valeur de l'inertie est donc un indicateur de la dispertion des nuages de points et une mesure de liaison entre les deux variables qualitatives encore appellée mesure d'écart à l'indépendance.

De plus, on a vu que l'inertie des nuages de points est égale à l'inertie des matrices des coordonnées factorielles $\mathbf{X}$ et $\mathbf{Y}$ "complètes" (lorsque $k = r$). En AFC, il y a au plus $r = \min(q-1, m-1)$ valeurs propres non nulles et l'inertie totale vaut $\lambda_1 + \ldots + \lambda_r$. Chaque composante principale explique donc une partie de l'inertie mesurée par :

$$\frac{\lambda_\alpha}{\lambda_1 + \ldots + \lambda_r} * 100$$

qui s'interprète comme :

- le pourcentage de l'inertie totale expliquée par l'axe α,
- la part de la liaison entre X_1 et X_2 expliquée par cet axe.

En pratique :

- On peut d'abord réaliser un test du χ^2 pour conclure ou non à l'indépendance entre X_1 et X_2. On ne réalisera à priori une AFC que si l'on conclue que X_1 et X_2 ne sont pas indépendantes.
- pour savoir combien d'axes retenir, on peut comme en ACP utiliser l'une des règles suivantes :
  - On peut utiliser le pourcentage d'inertie expliquée par les k premiers axes et choisir le nombre k d'axes tel que cette inertie expliquée dépasse un certain seuil (75% par exemple). Attention, il reste néanmoins la nécessité de ne retenir que des axes principaux utiles pour l'interprétation, c'est à dire interprétable.
  - On peut ne retenir que les valeurs propres supérieures à leur moyenne (règle empirique de Kaiser)
  - On peut utiliser la règle du coude :
    i) calculer les différence premières : $\epsilon_1 = \lambda_1 - \lambda_2$, $\epsilon_2 = \lambda_2 - \lambda_2$, ...
    ii) calculer les différence secondes : $\delta_1 = \epsilon_1 - \epsilon_2$, $\delta_2 = \epsilon_2 - \epsilon_2$, ...
    iii) retenir le nombre k tel que $\delta_1, \ldots, \delta_{k-1}$ soient toutes positives et que δ_k soit négative.

  D'autres critères peuvent être trouvés p.209 du livre de G. Saporta (2006).

Remarque : les valeurs propres sont toujours inférieures ou égales à 1.

Exemple

6.2 Contributions

La contribution d'une modalité i de X_1 et d'une modalité s de X_2 à l'inertie de l'axe α sont :

$$\begin{cases} \mathrm{Ctr}_\alpha(i) = \dfrac{f_{i.}x_{i\alpha}^2}{\lambda_\alpha} \\ \mathrm{Ctr}_\alpha(s) = \dfrac{f_{.s}y_{s\alpha}^2}{\lambda_\alpha} \end{cases}$$

La contribution $\mathrm{Ctr}_\alpha(i)$ est la part de la variance de l'axe α expliquée par la modalité i. Ce coefficient permet de connaître les modalités responsables de la construction de l'axe α, et permet de trouver une eventuelle signification aux axes.

Attention : En AFC, les points les plus excentrés sur les axes ne sont pas nécessairement ceux qui contribuent le plus (à cause des poids $f_{i.}$ et $f_{.s}$).

Exercice 4 : Dans l'exemple, retrouvez le calcul de la contribution de la modalité marron à l'axe 1.

6.3 Cosinus carrés

Le cosinus carré de l'angle entre le profil-ligne $\mathbf{l}_i$ et l'axe Δ_α mesure la qualité de la projection de ce profil sur cet axe :

$$\cos^2_\alpha(i) = \frac{x^2_{i\alpha}}{d^2(\mathbf{l}_i, \mathbf{c})}$$

où $d^2(\mathbf{l}_i, \mathbf{c}) = \sum_{s=1}^m \frac{1}{f_{.s}}(f_{is}/f_{i.} - f_{.s})^2$ est la distance du χ^2 entre le profil-ligne $\mathbf{l}_i$ et le profil-ligne moyen $\mathbf{c}$.
De même pour les modalités s de la variableX_2, on calcule le cosinus carré de l'angle entre le profil-colonne $\mathbf{c}_s$ et l'axe G_α pour mesurer la qualité de la projection de ce profil sur cet axe :

$$\cos^2_\alpha(s) = \frac{y^2_{s\alpha}}{d^2(\mathbf{c}_i, \mathbf{r})}$$

où $d^2(\mathbf{c}_i, \mathbf{r}) = \sum_{s=1}^m \frac{1}{f_{i.}}(f_{is}/f_{.s} - f_{i.})^2$ est la distance du χ^2 entre le profil-ligne $\mathbf{c}_i$ et le profil-colonne moyen $\mathbf{r}$.
Pour analyser les proximités entre les points sur les graphiques factoriels, on s'intéresse surtout aux points bien projetés (ayant un cos2 élevé) car les proximités entre ces points observée sur le graphique est "proche" de celle dans l'espace d'origine.

Exercice 5 : Dans l'exemple, retrouvez le calcul du cos2 de la modalité marron sur l'axe 1.

Attention : Pour interpéter des proximités entre deux points sur un graphique il faut prendre des précautions :

- Si deux modalités d'une même variable sont proches et bien projetées (bien représentées), cela signifie que leurs profils sont semblables.
- Par contre, la proximité entre une modalité d'une variable et une modalité de l'autre, est plus délicate à interpréter. Elle s'interprète avec les relations barycentriques.

7 Références

- "Statistique exploratoire multidimensionnelle", Lebart & al., Dunod.
- "Probabilité, analyse des données, statistique", G. Saporta, Technip.