Analyse factorielle multiple des correspondances (AFCM)

Résumé

Méthode factorielle de réduction de dimension pour l'exploration statistique de données qualitatives complexes. Cette méthode est une généralisation de l'Analyse Factorielle des Correspondances, permettant de décrire les relations entre p ($p > 2$) variables qualitatives simultanément observées sur n individus. Elle est aussi souvent utilisée pour la construction de scores *comme préalable à une méthode de classification (kmeans) nécessitant des données quantitatives.*

Travaux pratiques *de complexité croissante par l'études de données élémentaires, puis épidémiologiques avec interactions.*

Retour au plan du cours.

1 Codages de variables qualitatives

1.1 Tableau disjonctif complet

Soit X une variable qualitative à c modalités. On appelle *variable indicatrice* de la k-ième modalité de x ($k = 1, \ldots, c$), la variable $X_{(k)}$ définie par

$$X_{(k)}(i) = \begin{cases} 1 & \text{si } X(i) = \mathcal{X}_k, \\ 0 & \text{sinon,} \end{cases}$$

où i est un individu quelconque et $\mathcal{X}_k$ est la k–ième modalité de X. On notera n_k l'effectif de $\mathcal{X}_k$.

On appelle *matrice des indicatrices* des modalités de X, et l'on notera $\mathbf{X}$, la matrice $n \times c$ de terme général :

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

On vérifie :

$$\sum_{k=1}^{c} x_i^k = 1, \forall i \text{ et } \sum_{i=1}^{n} x_i^k = n_k.$$

Considérons maintenant p variables qualitatives $X^1, \ldots, X^p$. On note c_j le nombre de modalités de X^j, $c = \sum_{j=1}^{p} c_j$ et $\mathbf{X}_j$ la matrice des indicatrices de X^j.

On appelle alors *tableau disjonctif complet* la matrice $\mathbf{X}$, $n \times c$, obtenue par concaténation des matrices $\mathbf{X}_j$:

$$\mathbf{X} = [\mathbf{X}_1 | \cdots | \mathbf{X}_p].$$

$\mathbf{X}$ vérifie :

$$\sum_{k=1}^{c} x_i^k = p, \forall i \text{ et } \sum_{i=1}^{n} \sum_{k=1}^{c} x_i^k = np.$$

D'autre part, la somme des éléments d'une colonne de $\mathbf{X}$ est égale à l'effectif marginal de la modalité de la variable X^j correspondant à cette colonne.

1.2 Tableau de Burt

On observe toujours p variables qualitatives sur un ensemble de n individus. On appelle *tableau de Burt* la matrice $\boldsymbol{\mathcal{B}}$, $c \times c$, définie par :

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

On peut écrire $\boldsymbol{\mathcal{B}} = [\boldsymbol{\mathcal{B}}_{jl}]$ $(j = 1, \ldots, p\,;\, l = 1, \ldots, p)$; chaque bloc $\boldsymbol{\mathcal{B}}_{jl}$, de dimension $c_j \times c_l$, est défini par :

$$\boldsymbol{\mathcal{B}}_{jl} = \mathbf{X}_j'\mathbf{X}_l.$$

Si $j \neq l$, $\boldsymbol{\mathcal{B}}_{jl}$ est la table de contingence obtenue par croisement des variables X^j en lignes et X^l en colonnes. Si $j = l$, le bloc diagonal $\boldsymbol{\mathcal{B}}_{jj}$ est lui–même une matrice diagonale vérifiant :

$$\boldsymbol{\mathcal{B}}_{jj} = \text{ diag } (n_1^j, \ldots, n_{c_j}^j).$$

La matrice $\boldsymbol{\mathcal{B}}$ est symétrique, d'effectifs marginaux $n_l^j p$ et d'effectif total np^2.

1.3 La démarche suivie dans ce chapitre

La généralisation de l'AFC à plusieurs variables qualitatives repose sur certaines propriétés observées dans le cas élémentaire où $p = 2$. On s'intéresse

tout d'abord aux résultats fournis par l'AFC usuelle réalisée sur le tableau disjonctif complet $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$ relatif à 2 variables qualitatives X^1 et X^2 ; $\mathbf{X}$ est alors considéré comme une table de contingence (paragraphe 2). Ensuite, on suit la même démarche avec l'AFC réalisée sur le tableau de Burt $\mathcal{B}$ relatif à X^1 et X^2 (paragraphe 3). Enfin, en utilisant les propriétés obtenues dans les deux premiers cas, on généralise cette double approche à un nombre quelconque p de variables qualitatives ; on définit ainsi l'Analyse Factorielle des Correspondances Multiples (paragraphe 4).

2 AFC du tableau disjonctif complet relatif à 2 variables

2.1 Données

On note toujours X^1 et X^2 les 2 variables qualitatives considérées et r et c leurs nombres respectifs de modalités.

Les matrices intervenant dans l'AFC usuelle sont reprises ici avec les mêmes notations, mais surlignées. On obtient ainsi :

$$\begin{aligned}
\overline{\mathbf{T}} &= \mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2] \, ; \\
\overline{\mathbf{D}}_r &= \frac{1}{n}\mathbf{I}_n \, ; \\
\overline{\mathbf{D}}_c &= \frac{1}{2}\begin{bmatrix} \mathbf{D}_r & 0 \\ 0 & \mathbf{D}_c \end{bmatrix} = \frac{1}{2}\mathbf{\Delta} \, ; \\
\overline{\mathbf{A}} &= \frac{1}{2n}\overline{\mathbf{T}}'\overline{\mathbf{D}}_r^{-1} = \frac{1}{2}\mathbf{X}' \, ; \\
\overline{\mathbf{B}} &= \frac{1}{2n}\overline{\mathbf{T}}\,\overline{\mathbf{D}}_c^{-1} = \frac{1}{n}\mathbf{X}\mathbf{\Delta}^{-1}.
\end{aligned}$$

On considère ici l'AFC comme une double ACP : celle des profils–lignes $\overline{\mathbf{A}}$, puis celle des profils–colonnes $\overline{\mathbf{B}}$.

2.2 ACP des profils–lignes

Les profils–lignes, provenant de $\overline{\mathbf{T}}$, sont associés aux n individus observés. Leur ACP conduit ainsi à une représentation graphique des individus, inconnue en AFC classique.

PROPOSITION 1. — *L'ACP des profils–lignes issue de l'AFC réalisée sur le tableau disjonctif complet associé à 2 variables qualitatives conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}}_c^{-1}$–symétrique et positive :*

$$\overline{\mathbf{A}}\,\overline{\mathbf{B}} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_r & \mathbf{B} \\ \mathbf{A} & \mathbf{I}_c \end{bmatrix}.$$

Les $r + c$ valeurs propres de $\overline{\mathbf{A}}\,\overline{\mathbf{B}}$ s'écrivent

$$\mu_k = \frac{1 \pm \sqrt{\lambda_k}}{2},$$

où les λ_k sont les valeurs propres de la matrice $\mathbf{AB}$ (donc celles de l'AFC classique de X^1 et X^2).

Les vecteurs propres $\overline{\mathbf{D}}_c^{-1}$–orthonormés associés se mettent sous la forme

$$\overline{\mathbf{V}} = \frac{1}{2}\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix};$$

la matrice $\mathbf{U}$ (resp. $\mathbf{V}$) contient les vecteurs propres $\mathbf{D}_r^{-1}$–orthonormés (resp. $\mathbf{D}_c^{-1}$-orthonormés) de la matrice $\mathbf{BA}$ (resp. $\mathbf{AB}$) ; autrement dit, les matrices $\mathbf{U}$ et $\mathbf{V}$ sont les matrices de vecteurs propres obtenues en faisant l'AFC classique de la table de contingence croisant X^1 et X^2.

La matrice des composantes principales s'écrit

$$\overline{\mathbf{C}}_r = \frac{1}{2}\left[\mathbf{X}_1\mathbf{C}_r + \mathbf{X}_2\mathbf{C}_c\right]\mathbf{\Lambda}^{-1/2},$$

où $\mathbf{C}_r$ et $\mathbf{C}_c$ sont encore les matrices de composantes principales de l'AFC classique.

Dans la pratique, on ne considère que les $d = \inf(r-1, c-1)$ plus grandes valeurs propres différentes de 1, ainsi que les vecteurs propres associés. Les valeurs propres sont rangées dans la matrice

$$\mathbf{M} = \text{ diag }(\mu_1, \ldots, \mu_d) = \frac{1}{2}\left[\mathbf{I}_d + \mathbf{\Lambda}^{1/2}\right].$$

Les autres valeurs propres non nulles sont dues à l'artifice de construction de la matrice à diagonaliser ; elles n'ont donc pas de signification statistique.

On notera que la matrice $\overline{\mathbf{C}}_r$, $n \times d$, fournit les coordonnées permettant la représentation graphique des individus sur les axes factoriels.

2.3 ACP des profils–colonnes

Les profils–colonnes sont associés aux $r + c$ modalités des variables. Leur ACP conduit donc à une représentation graphique de ces modalités dont on verra qu'elle est très voisine de celle fournie par une AFC classique.

PROPOSITION 2. — *L'ACP des profils–colonnes issue de l'AFC réalisée sur le tableau disjonctif complet associé à 2 variables conduit à l'analyse spectrale de la matrice* $\overline{\mathbf{D}}_r^{-1}$*–symétrique et positive :*

$$\overline{\mathbf{B}}\,\overline{\mathbf{A}} = \frac{1}{2n}\left[\mathbf{X}_1\mathbf{D}_r^{-1}\mathbf{X}_1' + \mathbf{X}_2\mathbf{D}_c^{-1}\mathbf{X}_2'\right].$$

Les $r + c$ *valeurs propres non nulles de* $\overline{\mathbf{B}}\,\overline{\mathbf{A}}$ *sont les* μ_k.

Les vecteurs propres $\overline{\mathbf{D}}_r^{-1}$*–orthonormés associés se mettent sous la forme :*

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

La matrice des composantes principales s'écrit :

$$\overline{\mathbf{C}}_c = \begin{bmatrix} \mathbf{C}_r \\ \mathbf{C}_c \end{bmatrix} \mathbf{\Lambda}^{-1/2}\mathbf{M}^{1/2}.$$

Ainsi, l'AFC du tableau disjonctif complet permet, grâce aux coordonnées contenues dans les lignes de la matrice $\overline{\mathbf{C}}_c$, une représentation simultanée des modalités des 2 variables. Cette représentation est très voisine de celle obtenue par l'AFC classique, définie au chapitre précédent. Une simple homothétie sur chaque axe factoriel, de rapport $\sqrt{\frac{1+\sqrt{\lambda_k}}{2\lambda_k}}$, permet de passer de l'une à l'autre.

De plus, cette approche permet aussi de réaliser une représentation graphique des individus avec les coordonnées contenues dans les lignes de la matrice $\overline{\mathbf{C}}_r$. À un facteur près, chaque individu apparaît comme le barycentre des 2 modalités qu'il a présentées. Dans le cas où n est grand, le graphique des individus a néanmoins peu d'intérêt ; seule sa forme générale peut en avoir un.

Remarque. — Si, dans l'AFC classique, on choisit d'utiliser, pour la représentation simultanée des modalités de X^1 et de X^2, les lignes des matrices

$$\mathbf{C}_r^* = \mathbf{D}_r^{-1}\mathbf{U} = \mathbf{C}_r\mathbf{\Lambda}^{-1/2} \text{ et } \mathbf{C}_c^* = \mathbf{D}_c^{-1}\mathbf{V} = \mathbf{C}_c\mathbf{\Lambda}^{-1/2}$$

(voir chapitre précédent, sous–section 4.4), alors on obtient par AFC du tableau disjonctif complet la matrice

$$\overline{\mathbf{C}}_c^* = \overline{\mathbf{C}}_c\mathbf{M}^{-1/2} = \begin{bmatrix} \mathbf{C}_r^* \\ \mathbf{C}_c^* \end{bmatrix};$$

il y a invariance de la représentation des modalités lorsque l'on passe d'une méthode à l'autre. Pour les individus, on obtient

$$\overline{\mathbf{C}}_r^* = \frac{1}{2}\left[\mathbf{X}_1\mathbf{C}_r^* + \mathbf{X}_2\mathbf{C}_c^*\right]\mathbf{M}^{-1/2}$$

(le commentaire est alors le même qu'avec $\overline{\mathbf{C}}_r$).

3 AFC du tableau de Burt relatif à 2 variables

Dans cette section, on s'intéresse aux résultats fournis par l'AFC réalisée sur le tableau de Burt $\mathcal{B} = \mathbf{X}'\mathbf{X}$, $(r + c) \times (r + c)$, relatif aux 2 variables X^1 et X^2 ; $\mathcal{B}$ est encore considéré comme une table de contingence. La matrice $\mathcal{B}$ étant symétrique, les profils–lignes et les profils–colonnes sont identiques ; il suffit donc de considérer une seule ACP

Les notations des matrices usuelles de l'AFC sont maintenant réutilisées surmontées d'un tilde. On obtient ainsi :

$$\begin{aligned}
\widetilde{\mathbf{T}} &= \mathcal{B} = \begin{bmatrix} n\mathbf{D}_r & \mathbf{T} \\ \mathbf{T}' & n\mathbf{D}_c \end{bmatrix}; \\
\widetilde{\mathbf{D}_r} &= \widetilde{\mathbf{D}_c} = \frac{1}{2}\begin{bmatrix} \mathbf{D}_r & 0 \\ 0 & \mathbf{D}_c \end{bmatrix} = \frac{1}{2}\mathbf{\Delta} = \overline{\mathbf{D}}_c\,; \\
\widetilde{\mathbf{A}} &= \widetilde{\mathbf{B}} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_r & \mathbf{B} \\ \mathbf{A} & \mathbf{I}_c \end{bmatrix} = \overline{\mathbf{A}}\,\overline{\mathbf{B}}.
\end{aligned}$$

On considère encore l'AFC comme l'ACP des profils–lignes $\widetilde{\mathbf{A}}$ (ou des profils–colonnes $\widetilde{\mathbf{B}}$).

PROPOSITION 3. — *L'ACP des profils–lignes (ou des profils–colonnes) issue de l'AFC réalisée sur le tableau de Burt associé à 2 variables qualitatives conduit à l'analyse spectrale de la matrice* $\widetilde{\mathbf{D}_c}^{-1}$ *–symétrique et positive :*

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \left[\overline{\mathbf{A}}\,\overline{\mathbf{B}}\right]^2.$$

Elle admet pour matrice de vecteurs propres $\widetilde{\mathbf{D}_c}^{-1}$ *–orthonormés*

$$\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}} = \overline{\mathbf{V}} = \frac{1}{2}\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}.$$

Les valeurs propres associées vérifient : $\nu_k = \mu_k^2$.

La matrice des composantes principales s'écrit :

$$\widetilde{\mathbf{C}_r} = \widetilde{\mathbf{C}_c} = \begin{bmatrix}\mathbf{C}_r\\ \mathbf{C}_c\end{bmatrix}\mathbf{\Lambda}^{-1/2}\mathbf{M}.$$

La matrice $\widetilde{\mathbf{C}_r}$ fournit les coordonnées permettant une représentation simultanée des modalités des deux variables. À une homothétie près, cette représentation est identique à celle de l'AFC classique, réalisée sur la table de contingence $\mathbf{T}$ (mais le rapport d'homothétie, sur chaque axe, n'est plus le même qu'avec $\overline{\mathbf{C}}_c$).

Remarque. —

- En reprenant les notations de la remarque 2.3, on obtient ici :

$$\widetilde{\mathbf{C}_r^*}\ (=\widetilde{\mathbf{C}_c^*}) = \widetilde{\mathbf{C}_r}\mathbf{M}^{-1} = \overline{\mathbf{C}}_c^* = \begin{bmatrix}\mathbf{C}_r^*\\ \mathbf{C}_c^*\end{bmatrix}.$$

  Ainsi, si l'on utilise ce mode de représentation graphique, les trois approches de l'AFC que nous avons présentées conduisent à la même représentation simultanée des modalités des 2 variables : il y a donc invariance de cette représentation.

- Dans les deux cas d'AFC considérés dans ce chapitre (sur tableau disjonctif complet et sur tableau de Burt) on trouve, par construction, des valeurs propres non nulles sans signification statistique. En conséquence, les critères de qualité s'exprimant comme une "part d'inertie expliquée" n'ont plus de signification.
- L'AFC sur tableau de Burt ne prend en compte que l'information contenue dans $\mathcal{B}$ qui ne considère que les croisements de variables prises deux à deux. En conséquence, les interactions de niveau plus élevé sont ignorées par cette approche, à moins de procéder à des recodages de variables comme l'explique l'exemple présenté dans la section 5.

4 Analyse Factorielle des Correspondances Multiples

4.1 Définition

On considère maintenant p variables qualitatives ($p \geq 3$) notées $\{X^j\ ; j = 1,\ldots,p\}$, possédant respectivement c_j modalités, avec $c = \sum_{j=1}^{p} c_j$. On suppose que ces variables sont observées sur les mêmes n individus, chacun affecté du poids $1/n$.

Soit $\mathbf{X} = [\mathbf{X}_1|\cdots|\mathbf{X}_p]$ le tableau disjonctif complet des observations ($\mathbf{X}$ est $n \times c$) et $\boldsymbol{\mathcal{B}} = \mathbf{X}'\mathbf{X}$ le tableau de Burt correspondant ($\boldsymbol{\mathcal{B}}$ est carré d'ordre c, symétrique).

DÉFINITION 4. — *On appelle Analyse Factorielle des Correspondances Multiples (AFCM) des variables* $(X^1,\ldots,X^p)$ *relativement à l'échantillon considéré, l'AFC réalisée soit sur la matrice* $\mathbf{X}$ *soit sur la matrice* $\boldsymbol{\mathcal{B}}$.

On note n_k^j $(1 \leq j \leq p, 1 \leq k \leq c_j)$ l'effectif de la k–ième modalité de X^j, $\mathbf{D}_j = \dfrac{1}{n}\,\text{diag}\,(n_1^j,\ldots,n_{c_j}^j)$ et $\mathbf{\Delta} = \text{ diag }(\mathbf{D}_1\ldots\mathbf{D}_p)$ ($\mathbf{\Delta}$ est carrée d'ordre c et diagonale).

4.2 AFC du tableau disjonctif complet

Comme dans le cas $p = 2$, on reprend les notations de l'AFC classique en les surlignant. On obtient ainsi :

$$\begin{aligned}
\overline{\mathbf{T}} &= \mathbf{X}\,; \\
\overline{\mathbf{D}}_r &= \frac{1}{n}\mathbf{I}_n\,; \\
\overline{\mathbf{D}}_c &= \frac{1}{p}\boldsymbol{\Delta}\,; \\
\overline{\mathbf{A}} &= \frac{1}{p}\mathbf{X}'\,; \\
\overline{\mathbf{B}} &= \frac{1}{n}\mathbf{X}\boldsymbol{\Delta}^{-1}.
\end{aligned}$$

ACP des profils–lignes

PROPOSITION 5. — *L'ACP des profils–lignes issue de l'AFC réalisée sur le tableau disjonctif complet de p variables qualitatives conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}}_c^{-1}$–symétrique et positive :*

$$\overline{\mathbf{A}}\,\overline{\mathbf{B}} = \frac{1}{np}\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}^{-1}.$$

Il y a m ($m \leq c-p$) valeurs propres notées μ_k, $(0 < \mu_k < 1)$ rangées dans la matrice diagonale $\mathbf{M}$.

La matrice des vecteurs propres $\overline{\mathbf{D}}_c^{-1}$–orthonormés associés se décompose en blocs de la façon suivante :

$$\overline{\mathbf{V}} = \begin{bmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_p \end{bmatrix};$$

chaque bloc $\mathbf{V}_j$ est de dimension $c_j \times m$.

La matrice des composantes principales s'écrit :

$$\overline{\mathbf{C}}_r = \sum_{j=1}^{p} \mathbf{X}_j\mathbf{D}_j^{-1}\mathbf{V}_j.$$

Comme dans le cas $p = 2$, la matrice des composantes principales permet de réaliser une représentation graphique des individus dans laquelle chacun apparaît, à un facteur près, comme le barycentre des p modalités qu'il a présentées.

Remarque. — La généralisation au cas $p > 2$ restreint les propriétés. Ainsi, les vecteurs des blocs $\mathbf{V}_j$ ne sont pas les vecteurs propres $\mathbf{D}_j^{-1}$–orthonormés d'une matrice connue.

ACP des profils–colonnes

PROPOSITION 6. — *L'ACP des profils–colonnes issue de l'AFC réalisée sur le tableau disjonctif complet de p variables conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}}_r^{-1}$–symétrique et positive :*

$$\overline{\mathbf{B}}\,\overline{\mathbf{A}} = \frac{1}{np}\mathbf{X}\boldsymbol{\Delta}^{-1}\mathbf{X}' = \frac{1}{np}\sum_{j=1}^{p}\mathbf{X}_j\mathbf{D}_j^{-1}\mathbf{X}_j'.$$

La matrice des vecteurs propres $\overline{\mathbf{D}}_r^{-1}$–orthonormés vérifie :

$$\overline{\mathbf{U}} = \overline{\mathbf{B}}\overline{\mathbf{V}}\mathbf{M}^{-1/2}.$$

La matrice des composantes principales s'écrit :

$$\overline{\mathbf{C}}_c = p\boldsymbol{\Delta}^{-1}\overline{\mathbf{V}}\mathbf{M}^{1/2}\,;$$

elle se décompose en blocs sous la forme :

$$\overline{\mathbf{C}}_c = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_p \end{bmatrix}.$$

Chaque bloc $\mathbf{C}_j$, de dimension $c_j \times m$, fournit en lignes les coordonnées des modalités de la variable X^j permettant la représentation graphique simultanée.

4.3 AFC du tableau de Burt

Le tableau de Burt $\mathcal{B} = \mathbf{X}'\mathbf{X}$, carré d'ordre c, étant symétrique, les profils–lignes et les profils–colonnes sont identiques ; on ne considère donc ici qu'une seule ACP

En utilisant encore le tilde dans ce cas, les matrices usuelles de l'AFC deviennent :

$$\begin{aligned}\widetilde{\mathbf{T}} &= \mathcal{B}\,;\\ \widetilde{\mathbf{D}_r} &= \widetilde{\mathbf{D}_c} = \frac{1}{p}\mathbf{\Delta} = \overline{\mathbf{D}}_c\,;\\ \widetilde{\mathbf{A}} &= \widetilde{\mathbf{B}} = \frac{1}{np}\mathcal{B}\mathbf{\Delta}^{-1} = \overline{\mathbf{A}}\,\overline{\mathbf{B}}.\end{aligned}$$

PROPOSITION 7. — *L'ACP des profils–lignes (ou des profils–colonnes) issue de l'AFC réalisée sur le tableau de Burt associé à p variables qualitatives conduit à l'analyse spectrale de la matrice $\widetilde{\mathbf{D}_c}^{-1}$–symétrique et positive :*

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \left[\overline{\mathbf{A}}\,\overline{\mathbf{B}}\right]^2.$$

Elle admet pour matrice de vecteurs propres $\widetilde{\mathbf{D}_c}^{-1}$–orthonormés $\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}} = \overline{\mathbf{V}}$.

Les valeurs propres associées vérifient $\nu_k = \mu_k^2$.

La matrice des composantes principales s'écrit :

$$\widetilde{\mathbf{C}_r} = \widetilde{\mathbf{C}_c} = \overline{\mathbf{C}}_c\mathbf{M}^{1/2} = \begin{bmatrix}\mathbf{C}_1\\ \vdots\\ \mathbf{C}_p\end{bmatrix}\mathbf{M}^{1/2}.$$

La matrice $\widetilde{\mathbf{C}_r}$ fournit les coordonnées permettant la représentation simultanée des modalités de toutes les variables (on ne peut pas faire de représentation des individus si l'on fait l'AFC du tableau de Burt).

4.4 Variables illustratives

Soit X^0 une variable qualitative, à c_0 modalités, observée sur les mêmes n individus que les X^j et n'étant pas intervenue dans l'AFCM Soit $\mathbf{T}_{0j}$ la table de contingence $c_0 \times c_j$ croisant les variables X^0 en lignes et X^j en colonnes. L'objectif est maintenant de représenter les modalités de cette variable supplémentaire X^0 dans le graphique de l'AFCM réalisée sur $X^1, \ldots, X^p$. Pour cela, on considère les matrices :

$$\begin{aligned}\mathbf{B}_0 &= [\mathbf{T}_{01}|\ldots|\mathbf{T}_{0p}]\ ;\\ \mathbf{D}_0 &= \frac{1}{n}\,\mathrm{diag}\,(n_1^0, \ldots, n_{c_0}^0)\ ;\\ \mathbf{A}_0 &= \frac{1}{np}\mathbf{D}_0^{-1}\mathbf{B}_0.\end{aligned}$$

Les coordonnées des modalités de la variable supplémentaires X^0 sur les axes factoriels sont alors fournies dans les lignes de la matrice

$$\mathbf{C}_0 = \mathbf{A}_0\widetilde{\mathbf{D}_c}^{-1}\widetilde{\mathbf{V}} = p\mathbf{A}_0\mathbf{\Delta}^{-1}\overline{\mathbf{V}}.$$

4.5 Interprétation

Les représentations graphiques sont interprétées de manière analogue à ce qui est fait dans l'AFC de deux variables, bien que la représentation simultanée des modalités de toutes les variables ne soit pas, en toute rigueur, réellement justifiée.

Les "principes" suivants sont donc appliqués :

- on interprète globalement les proximités et les oppositions entre les modalités des différentes variables, comme en AFC, en privilégiant les modalités suffisamment éloignées du centre du graphique (attention aux modalités à faible effectif !) ;
- les rapports de valeurs propres ne sont pas interprétables comme indicateurs de qualité globale ; on peut néanmoins regarder la décroissance des premières valeurs propres pour choisir la dimension ;
- les coefficients de qualité de chaque modalité ne peuvent pas être interprétés ; seules les contributions des modalités à l'inertie selon les axes sont interprétées, selon le même principe qu'en AFC

5 Exemple

L'AFCM ne donne pas de résultats très intéressants sur les données bancaires à l'exception du graphe présenté dans le chapitre d'introduction qui est relativement plus sophistiqué car il fait préalablement appel à une classification. Il en est de même pour les données d'expression qui sont quantitatives.

TABLE 1 – *Données sous la forme d'une table de contingence complète*

			Histologie			
			Inflam. minime		Grande inflam.	
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

En revanche, l'AFCM est très indiquée et très utilisée dans des enquêtes de nature épidémiologique.

5.1 Les données

La littérature anglo-américaine présente souvent des données relatives à plusieurs variables qualitatives sous la forme d'une table de contingence *complète* (5). C'est le cas de l'exemple ci–dessous qui décrit les résultats partiels d'une enquête réalisée dans trois centres hospitaliers (Boston, Glamorgan, Tokyo) sur des patientes atteintes d'un cancer du sein. On se propose d'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 cherchant à mettre en évidence les facteurs de décès.

5.2 Analyse brute

On se reportera à la figure 5. 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.

5.3 Analyse des interactions

Pour essayer de mettre en évidence d'éventuelles interactions entre variables, les données sont reconsidérées de la façon suivante :

- 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 2.

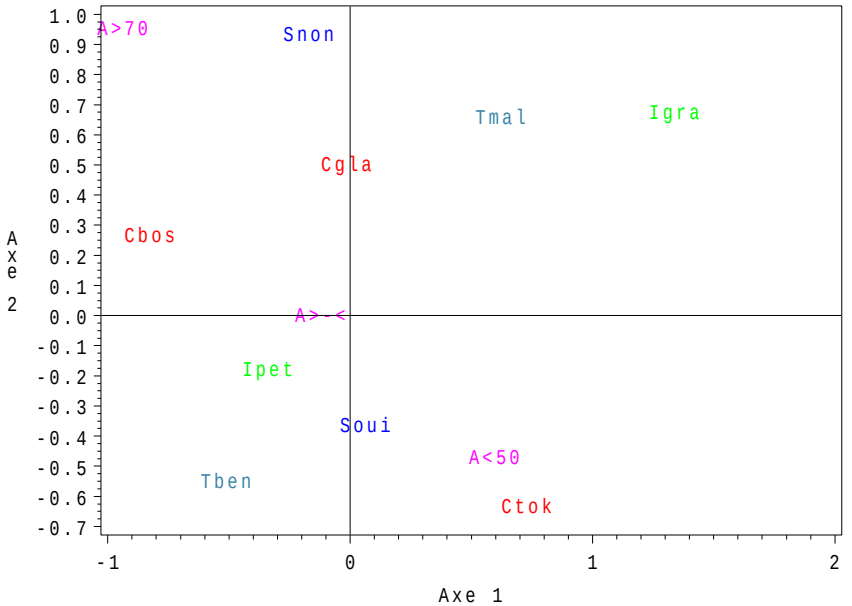

FIGURE 1 – *Cancer du sein : analyse des données brutes.*

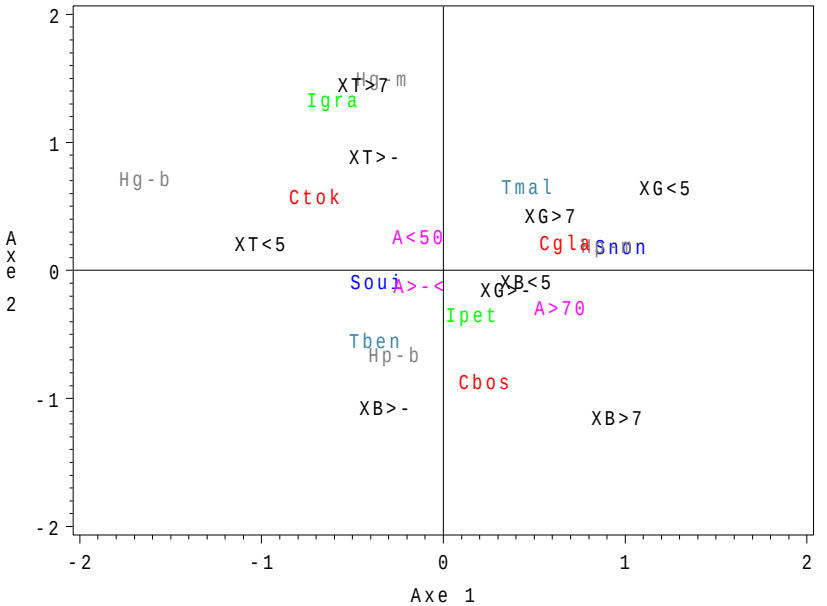

FIGURE 2 – *Cancer du sein : analyse des interactions.*