Chapitre 13

Analyse factorielle des correspondances (A.F.C.) : une introduction

L'analyse factorielle des correspondances (A.F.C.) entre dans la catégorie des statistiques bivariées, mais elle a la particularité d'être multidimensionnelle au niveau des modalités prises par les variables qualitatives qu'elle étudie. Elle fut développée dans les années 1970, notamment par J.-P. Benzécri . Elle constitue une introduction à l'analyse factorielle et à sa propre généralisation multivariée, l'analyse factorielle des correspondances multiples (A.C.M.). Toutefois, la méthode optimisant l'A.F.C. sera étudiée dans le chapitre portant sur l'analyse factorielle en composantes principales (A.C.P.). Ainsi, lors du chapitre portant sur la généralisation de l'A.F.C. en A.C.M., certains aspects seront repris et réinterprétés.

Jean-Paul Benzécri (1932-2019)

L'A.F.C. est une méthode statistique permettant de transformer un tableau de nombres en graphique, appelé *mapping*. Dit autrement, elle permet de visualiser la nature de la liaison entre deux variables. Elle est souvent utilisée en dépouillement d'enquête dans le cas où l'on ne se pose que deux questions.

13.1 Les principes généraux de l'analyse factorielle des correspondances

13.1.1 Définition d'une analyse factorielle

« L'analyse factorielle traite des tableaux de nombres. Elle remplace un tableau de nombres difficile à analyser par une série de tableaux plus simples qui sont une bonne approximation de celle-ci » [Cibois, 1991]. Un tableau est dit simple lorsqu'il est exprimable sous forme graphique.

L'analyse est dite factorielle, car l'objectif est de décomposer le tableau original en une somme de tableaux qui sont chacun le **produit de facteurs simples**. La matrice est mise en facteur.

13.1.2 Notion de correspondance

À l'instar de la corrélation pour les variables quantitatives, les variables nominales liées entraînent une correspondance.

13.1.3 Exemple pour illustrer les étapes du calcul

Soient deux variables qualitatives, la série du baccalauréat et le choix du type d'établissement post-baccalauréat. En croisant les deux variables, il est possible d'obtenir un tableau de contingence (Tab. 13.1). À partir de là, peut-on donner un sens à ce tableau ? L'objectif d'une A.F.C. est de mettre en évidence ce qui est inattendu par rapport à une répartition uniforme. Pour ce, l'A.F.C. procède en quatre étapes :

1. évaluer ce qui serait une situation d'uniformité c'est-à-dire d'indépendance ;
2. calculer de quelle manière la situation observée diffère de cette situation d'uniformité théorique ;
3. exprimer graphiquement cette différence afin de pouvoir l'analyser ;
4. interpréter et optimiser le *mapping* obtenu.

		Orientation post-BAC			**Marge**
		Université	Classe préparatoire	Autres	**des lignes**
Série	A	13	2	5	20
	BDD'	20	2	8	30
	CE	10	5	5	20
	FGH	7	1	22	30
Marge des colonnes		50	10	40	100

TABLE 13.1 – Le devenir des bacheliers en 1975 en pourcentage d'après le ministère de l'Éducation nationale sur 204 489 lycéens [Cibois, 1991]

13.2 Le processus du calcul

Mathématiquement, les tableaux de nombres sont des matrices. De fait, on note T la matrice des données d'entrées. La première étape consiste à calculer

la matrice R des écarts à l'indépendance. La deuxième étape met en facteur la matrice R afin de l'exprimer plus simplement.

$$T = \begin{bmatrix} 13 & 2 & 5 \\ 20 & 2 & 8 \\ 10 & 5 & 5 \\ 7 & 1 & 22 \end{bmatrix} \tag{13.1}$$

13.2.1 Évaluation de la situation d'indépendance théorique

On calcule les marges des lignes et des colonnes. On obtient deux vecteurs représentant la somme des colonnes $[50, 10, 40]$ et la somme des lignes $\begin{bmatrix} 20 \\ 30 \\ 20 \\ 30 \end{bmatrix}$. Grâce à ces deux vecteurs, il est possible de construire la matrice T_0 correspondant la **matrice de situation d'indépendance**. En cas d'indépendance entre les variables, il faut effectuer le produit entre la valeur marginale des colonnes et la valeur marginale des lignes, comme cela a été vu lors du test du χ^2.

$$T_0 = \begin{bmatrix} 10 & 2 & 8 \\ 15 & 3 & 12 \\ 10 & 2 & 8 \\ 15 & 3 & 12 \end{bmatrix} \tag{13.2}$$

13.2.2 Calcul de la matrice des écarts

On calcule ici la différence entre T et T_0 qui matérialise les écarts à l'indépendance.

$$R = T - T_0 = \begin{bmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ -8 & -2 & 10 \end{bmatrix} \tag{13.3}$$

R est appelée **matrice du reste**. Elle représente ce qui est surprenant. Par ailleurs, R possède une propriété important. La somme de chacune de ses lignes, ainsi que celle de chacune de ses colonnes, sont nulles. Soient respectivement $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ et $[0, 0, 0]$.

Il est possible d'envisager de proposer une expression plus simple de R en la décomposant en somme de matrices :

$$R = T_1 + T_2 \tag{13.4}$$

Chaque matrice T_1 et T_2 doit être factorisable par le produit d'un vecteur ligne L_1 ou L_2 et d'un vecteur colonne C_1 et C_2.

$$T_1 = C_1 L_1 \tag{13.5}$$

et

$$T_2 = C_2 L_2 \tag{13.6}$$

d'où

$$T - T_0 = T_1 + T_2 = C_1 L_1 + C_2 L_2 \tag{13.7}$$

Il s'agit d'appliquer une propriété des matrices. Une matrice dont la plus petite dimension est n (rang n) est décomposable au maximum en n matrices pouvant se mettre en facteurs. Pour T, la plus petite dimension est trois.

La mise en facteur des matrices est l'opération inverse du produit matriciel. Ici, il s'agit de décomposer la matrice T en un produit de matrices colonnes et de matrices lignes. Si une matrice $A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}$ a une dimension 2×2, alors la matrice colonne de sa factorisation aura une dimension 1×2 et la matrice ligne de sa factorisation aura une dimension 2×1. Trivialement, il est facile de vérifier qu'une solution est $C = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ et $L = [1, 2]$. $C \times L = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \times [1, 2] = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} = A$.

Appliquer à $R = C_1 L_1 + C_2 L_2$, l'une des solutions est :

$$C_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \\ -4 \end{bmatrix} \tag{13.8}$$

et

$$L_1 = [1, 1, -2] \tag{13.9}$$

ainsi que

$$C_2 = \begin{bmatrix} 1 \\ 2 \\ -1 \\ -2 \end{bmatrix} \tag{13.10}$$

et

$$L_2 = [2, -1, -1] \tag{13.11}$$

Comment passer de la décomposition des lignes et des colonnes en un graphique ?

13.2.3 Production et interprétation du *mapping*

Un vecteur colonne (ou ligne) correspond à une modalité des données en colonnes (ou lignes). En regroupant, C_1 et L_1 au sein d'une même, on obtient un axe unidirectionnel (ou composante), de même en regroupant C_2 et L_2. Les deux axes peuvent être projetés dans un repère.

On trace les vecteurs entre l'origine des deux axes et l'un des points projetés. Trois situations peuvent se présenter :

1. la **situation de conjonction**, qui correspond à un produit scalaire strictement positif. Les deux modalités ont une certaine affinité ;
2. la **situation d'opposition**, qui correspond à un produit scalaire strictement négatif. Les deux modalités ont une certaine répulsivité ;
3. la **situation de quadrature**, qui correspond à un produit scalaire nul. Géométriquement, cela signifie que les deux vecteurs sont orthogonaux. Les deux modalités sont dans une situation d'équilibre.

13.3 La manière d'optimiser la factorisation

Par définition, les matrices T_1 et T_2 ne sont pas uniques pour la matrice R. Quelle est la meilleure décomposition possible pour la matrice R ?

Pour toute matrice de rang n, on chercher à déterminer la meilleure matrice T_1, puis la meilleure matrice T_2 en fonction d'un critère retenu. Le premier axe doit exprimer le plus de sens pour l'analyse.

Par ailleurs, il faut utiliser une métrique. Dans le cas des variables qualitatives la métrique retenue est celle du χ^2 qui représente l'écart à l'indépendance, mesuré ici par T_0.

Le processus d'optimisation suit deux étapes :

1. calculer le carré de l'écart entre T et T_0, à proprement dit, il s'agit de la distance χ^2 ;

2. diviser la distance χ^2 obtenue par l'effectif théorique T_0 ; on parle alors de χ^2 pondérée. Grâce à cette manipulation, le poids des cellules à faible effectif est renforcée dans la matrice obtenue.

N.B. Les fréquences sont davantages utilisées que les effectifs.

La matrice de la distance du ${\chi^2}_{\text{obs}}$ vaut :

$$ {\chi^2}_{\text{obs}} = \begin{bmatrix} 0,900 & 0,000 & 1,125 \\ 1,667 & 0,333 & 1,333 \\ 0,000 & 4,500 & 1,125 \\ 4,267 & 1,333 & 8,333 \end{bmatrix} \tag{13.12} $$

Le χ^2 correspond à la somme de toutes les contributions au χ^2 de toutes les cellules de la matrice résultat ; il mesure et teste la dépendance. Ici, $\chi^2 \approx 24,916$, donc la liaison est significative avec un seuil de confiance à 5 % et à 1 %. Le pourcentage des contributions de T_1 et T_2 par rapport au χ^2 de la matrice R aboutit aux contributions relatives de T_1 et de T_2 au χ^2 de la matrice T.

$$ \chi^2 (R) = \chi^2 (T_1) + \chi^2 (T_2) \tag{13.13} $$

Ici, on obtient $2491 = 1998 + 443$, d'où le fait que T_1 contribue à 82,2 % au χ^2 et T_2 contribue à 19,8 % au χ^2. La concentration correspond au pourcentage de la variance expliquée par un axe. Sur un *mapping*, il n'est possible que de représenter deux axes à la fois, donc autant choisir de représenter les plus significatifs. En effet, lors de la représentation graphique, l'importance de la contribution des axes fixe la dilatation appliquée sur chacun d'entre eux qui est proportionnelle au χ^2 qu'ils expriment. Néanmoins, ce choix de dilatation implique la perte de la visibilité d'une quadrature. Dit autrement, il faut étudier chaque point en fonction de leur position par rapport aux axes, et non par rapport à leur produit scalaire. Les points peuvent être représentés proportionnellement à l'effectif qu'ils représentent.

Il est également possible de mesurer l'**intensité de liaison** ϕ^2, c'est-à-dire l'écart entre les probabilités observées et théoriques.

$$ \phi^2 = \frac{\chi^2}{n} \tag{13.14} $$

Ici,

$$ \phi^2 \approx \frac{24,916}{100} \approx 0,249 \tag{13.15} $$

et

$$\phi^2{}_{\text{obs}} = \begin{bmatrix} 0,009 & 0,000 & 0,011 \\ 0,017 & 0,003 & 0,013 \\ 0,000 & 0,045 & 0,011 \\ 0,043 & 0,013 & 0,083 \end{bmatrix} \tag{13.16}$$

Le ϕ^2 permet d'identifier la nature de la liaison, c'est-à-dire la nature de l'association des modalités.

À partir de là, il est possible d'opérer une **analyse factorielle**. Le critère d'optimisation est la somme des distances projetées sur l'axe recherché pondéré par rapport au nombre de points. On cherche à la maximiser. De fait, contrairement à la méthode des moindres carrés, on cherche un axe orthogonal à tous les points et à maximiser les écarts entre les points et l'axe. Pour des raisons pédagogiques, la méthode sera détaillée lors du chapitre sur l'analyse factorielle en composantes principales (A.C.P.). Après avoir expliqué la méthode d'optimisation, lors du chapitre sur l'analyse factorielle des correspondances multiples (A.C.M.), l'A.F.C. sera reprise, puis généralisée.

13.4 La généralisation de l'A.F.C. en A.C.M.

L'A.F.C. utilise des catégories mutuellement exclusives. La matrice s'y rapportant est un **tableau disjonctif**. Il n'est pas possible d'opérer deux choix simultanés.

L'A.F.C. se généralise avec plusieurs variables. On parle alors d'analyse des composantes multiples (A.C.M.). Le tableau disjonctif est remplacé par le **tableau de Burt**. Lors du chapitre portant sur l'A.C.M., la première partie commencera par un exemple d'A.F.C.

Sir Cyril Cecil Barrow Burt (1883-1971)

Bibliographie

[Cibois, 1991] CIBOIS, P. (1991). L'analyse factorielle. Analyse en composantes principales et analyse factorielle des correspondances. Que sais-je ? PUF, Paris.