

Analyse Factorielle des Correspondances AFC

Prof. Mustapha RACHDI



Université Grenoble Alpes
UFR SHS, BP. 47
38040 Grenoble cedex 09
France

Bureau : C08 au Bât. Michel Dubois
e-mail : mustapha.rachdi@univ-grenoble-alpes.fr



Introduction

- Depuis les années 60, à la suite des travaux de J. P. Benzecri et de son équipe, l'**Analyse Factorielle des Correspondances (AFC)** occupe une place de premier plan dans la littérature statistique de langue française.
- L'AFC est le cas particulier de l'**Analyse Canonique** pour lequel chacun des deux tableaux de données est constitué par les indicatrices d'une variable qualitative. Ainsi, le coefficient du Khi-deux mesure globalement l'intensité des relations entre deux variables qualitatives et l'AFC décrit plus précisément ces relations liant deux variables qualitatives.
- L'AFC est une analyse canonique particulière, mais l'AFC peut aussi être présentée comme la juxtaposition de deux ACP particulières menées à partir du tableau de contingence : des graphiques et des aides à l'interprétation très proches de ceux de l'ACP sont alors obtenus.

Exemple 1

- Si les modalités de l'une **des deux variables qualitatives** sont constituées par les **22 régions françaises**, et si les modalités de l'autre variable sont **des classes d'âge**, c'est-à-dire, si une case du tableau de contingence donne le nombre d'individus appartenant à une région donnée et une classe d'âge donnée, **l'AFC répond à des questions du type :**
 - **la structure démographique** de la Lorraine est-elle proche de la structure démographique de la Bretagne ? plus généralement, quelles sont les régions dont la structure démographique est proche, et, à l'opposé, quelles sont les régions qui ont des pyramides des âges totalement différentes ?
 - quelles sont les régions où les *jeunes* sont relativement nombreux ?
 - et la **sur-représentation des *jeunes*** dans certaines régions se fait-elle systématiquement au détriment des classes d'âge moyen ou des personnes âgées ?

Exemple 2

- Dans ce chapitre, on considérera le tableau déjà présenté au chapitre 2 qui croise **Niveau de diplôme** et **Classe d'âge**.
- L'AFC met alors en évidence :
 - les classes d'âge qui sont sur-diplômées ou, à l'opposé, celles qui sont sous-diplômées,
 - ou encore distingue les diplômes correspondant à des classes d'âge *jeunes* des diplômes plutôt possédés par des classes d'âge *moins jeunes*.

Exemple 3

- Une seconde application a trait aux relations entre le niveau de diplôme de la personne de référence et le type de ménage. Autrement dit :
 - a-t-on le même nombre d'enfants si l'on est titulaire d'un CAP ou si l'on a effectué des études universitaires ?
 - le pourcentage de personnes seules ou le pourcentage de couples avec un enfant est-il plus important chez les bacheliers que chez les titulaires d'un CAP ?
 - Quels sont les profils, en terme de diplôme des familles nombreuses ?
- C'est à ce type de questions que l'AFC apporte des réponses, en synthétisant l'information contenue dans le tableau de contingence.

- A la suite de l'AFC, l'**Analyse Factorielle des Correspondances Multiples (AFCM)**, technique qui généralise, **en quelque sorte**, l'AFC au cas de plusieurs variables qualitatives, est présentée et illustrée par la description des relations entre trois variables qualitatives :
 - le lieu d'habitation,
 - le niveau de diplôme,
 - et le type de ménage.

- Les données peuvent se présenter sous **deux formes équivalentes** :
 - soit sous la forme de **deux tableaux disjonctifs complets** X_1 et X_2 : chacun de ces deux tableaux décrit les modalités d'une variable qualitative.
 - soit sous la forme du **tableau de contingence** ${}^t X_1 \ X_2$ (ou, bien entendu, ${}^t X_2 \ X_1$).
- Le nombre de colonnes de X_1 , c'est-à-dire, le nombre de modalités de la première variable qualitative est p_1 , et le nombre de colonnes de X_2 est p_2 .
- Les tableaux X_1 et X_2 possèdent n lignes, chaque ligne correspondant à un individu. Pour ne pas alourdir les notations, nous supposerons que $p_2 > p_1$.

- Le tableau de contingence est évidemment plus maniable, **en pratique**, que les tableaux disjonctifs complets, surtout si les individus sont nombreux.
- Aussi, les données sont présentées le plus souvent sous la forme d'un tableau de contingence.
- Les tableaux disjonctifs complets servent surtout à exposer le principe de l'AFC.
- Les notations utilisées relatives au tableau de contingence (n_{ij} , f_{ij} , ...) sont celles définies dans le chapitre 2.

Exemple d'application

- On considère par exemple un ensemble de 18282 individus pour lesquels on connaît la CSP (modalités agriculteur AGRI, cadre supérieur CADR, inactif INAC, et ouvrier OUVR) et le choix de l'hébergement pour les vacances HEB (modalités camping CAMP, HOTEL, location LOCA, et résidence secondaire RESI).
- Le tableau des données brutes serait de la forme :

	CSP	HEB	effectif
AGRI	CAMP		239
AGRI	HOTEL		155
AGRI	LOCA		129
AGRI	RESI		0
CADR	CAMP		1003
CADR	HOTEL		1556
CADR	LOCA		1821
CADR	RESI		1521
INAC	CAMP		682
INAC	HOTEL		1944
INAC	LOCA		967
INAC	RESI		1333
OUVR	CAMP		2594
OUVR	HOTEL		1124
OUVR	LOCA		2176
OUVR	RESI		1038

individu	CSP	HEB
1	OUVR	CAMP
2	INAC	CAMP
3	AGRI	HOTEL
...
18281	INAC	RESI
18282	CADR	LOCA

FIGURE – Tableaux de données brutes et de données groupées en colonnes dépliés

Exemple d'application

- On considère par exemple un ensemble de 18282 individus pour lesquels on connaît la CSP (modalités agriculteur AGRI, cadre supérieur CADR, inactif INAC, et ouvrier OUVR) et le choix de l'hébergement pour les vacances HEB (modalités camping CAMP, HOTEL, location LOCA, et résidence secondaire RESI).
- Cependant, les identifications des individus ne nous importent peu dans l'étude de ce lien, on préfèrera présenter ces données sous forme de données groupées : en colonnes ("déplié") ou en tableau de contingence.
- Le tableau de contingence serait de la forme :

CSB/HEB	CAMP	HOTEL	LOCA	RESI
AGRI	239	155	129	0
CADR	1003	1556	1821	1521
INAC	682	1944	967	1333
OUVR	2594	1124	2176	1038

Exemple d'application

- Dans cet exemple, le but de l'AFC sera de représenter les éventuels liens entre la CSP et le type d'hébergement choisi HEB.
- Cette étude de lien peut se passer de l'analyse par AFC, via le calcul de la statistique du khi-deux, ou le calcul des profils-lignes et des profils-colonnes.
- D'autres méthodes d'étude de ce type de lien existent. Cependant, l'avantage qu'a l'AFC par rapport à ces méthodes est :
 - la hiérarchisation des différents types de lien,
 - la représentation, s'il y a lieu, des individus par rapport à ces liens,
 - et des représentations graphiques qui facilitent la communication de l'information.

Etude des liens entre deux variables qualitatives sans AFC

La statistique du khi-deux et le test associé

- On appelle effectifs observés (O) les effectifs du tableau de contingence. Les totaux des lignes et des colonnes s'appellent les marges.
- Pour calculer la statistique du khi-deux, il faut calculer les effectifs théoriques (T) en faisant les produits des marges divisés par l'effectif total général n.
- Les effectifs théoriques :

CSB/HEB	CAMP	HOTEL	LOCA	RESI	Total
AGRI	129.248	136.715	145.697	111.340	523
CADR	1458.304	1542.549	1643.901	1256.246	5901
INAC	1217.354	1287.679	1372.285	1048.681	4926
OUVR	1713.094	1812.057	1931.117	1475.733	6932
Total	4518	4779	5093	3892	18282

Etude des liens entre deux variables qualitatives sans AFC

La statistique du khi-deux et le test associé

- La statistique du khi-deux est alors :

$$\sum_{i,j} (O_{ij} - T_{ij})^2 / T_{ij}$$

Ce calcul donne la valeur 2067.911 pour ces données.

- Le test du khi-deux consiste à comparer cette valeur observée du khi-deux (χ^2_{obs}) à une valeur théorique (χ^2_{theo}).
- Les hypothèses :
 H_0 : Il n'y a pas de lien entre l'HEB et la CSP
versus
 H_1 : Il y a un lien significatif entre HEB et la CSP.
- Dans cet exemple, $\chi^2_{theo}(5\%) = 16.919$ avec $ddl = (4 - 1)(4 - 1) = 9$.

Etude des liens entre deux variables qualitatives sans AFC

La statistique du khi-deux et le test associé

- La règle du test est alors la suivante :

si $\chi^2_{obs} < \chi^2_{theo}(\alpha)$, alors on ne peut pas affirmer la dépendance entre les variables. Dans notre exemple, on a $\chi^2_{obs} > \chi^2_{theo}(\alpha)$, donc on peut affirmer, avec un risque $\alpha = 5\%$ de se tromper, qu'il y a dépendance entre CSP et HEB.

Autrement dit, le choix de l'hébergement pour les vacances semble dépendre de la CSP.

- Une autre façon de faire le test est de comparer la *p*-valeur (calculée par les logiciels qui font le test) à 5%. Ici, *p*-valeur < 0.0001.
- La règle de test, équivalente à la première, est alors :
si *p*-valeur > α , alors on ne peut pas affirmer la dépendance entre les variables.

Les profils lignes et colonnes

Ce sont les pourcentages (ou taux) d'individus d'une catégorie d'une des deux variables répartis selon les modalités de l'autre variable. On donne ci-dessous les profils-lignes et colonnes pour l'exemple.

Proportions / Ligne :

	CAMP	HOTEL	LOCA	RESI	Total
AGRI	0,457	0,296	0,247	0,000	1
CADR	0,170	0,264	0,309	0,258	1
INAC	0,138	0,395	0,196	0,271	1
OUVR	0,374	0,162	0,314	0,150	1
Total	0,247	0,261	0,279	0,213	1

Proportions / Colonne :

	CAMP	HOTEL	LOCA	RESI	Total
AGRI	0,053	0,032	0,025	0,000	0,029
CADR	0,222	0,326	0,358	0,391	0,323
INAC	0,151	0,407	0,190	0,342	0,269
OUVR	0,574	0,235	0,427	0,267	0,379
Total	1	1	1	1	1

FIGURE — Tableaux des profils lignes et colonnes

- Ils permettent de décrire les liens entre variables.
- Par exemple, on voit qu'il y a 24.7% d'individus qui choisissent CAMP sur l'ensemble. Dans la catégorie AGRI, il y en a 45.7%. Cela signifie que les agriculteurs choisissent plus souvent le camping que les individus des autres CSP.

Les profils lignes et colonnes

Proportions / Ligne :					
	CAMP	HOTEL	LOCA	RESI	Total
AGRI	0,457	0,296	0,247	0,000	1
CADR	0,170	0,264	0,309	0,258	1
INAC	0,138	0,395	0,196	0,271	1
OUVR	0,374	0,162	0,314	0,150	1
Total	0,247	0,261	0,279	0,213	1

Proportions / Colonne :					
	CAMP	HOTEL	LOCA	RESI	Total
AGRI	0,053	0,032	0,025	0,000	0,029
CADR	0,222	0,326	0,358	0,391	0,323
INAC	0,151	0,407	0,190	0,342	0,269
OUVR	0,574	0,235	0,427	0,267	0,379
Total	1	1	1	1	1

FIGURE – Tableaux des profils lignes et colonnes

- De même, dans la catégorie INAC, seulement 13.8% choisissent CAMP. Donc les inactifs choisissent moins souvent le camping que l'ensemble. Par une analyse de tous les profils-ligne, on décrirait ainsi les liens (préférences/non préférences) entre les CSP et les HEB choisis.
- Une analyse des profils-colonne donnerait une même conclusion, mais dite différemment.

Ces profils font souvent l'objet de représentations graphiques :

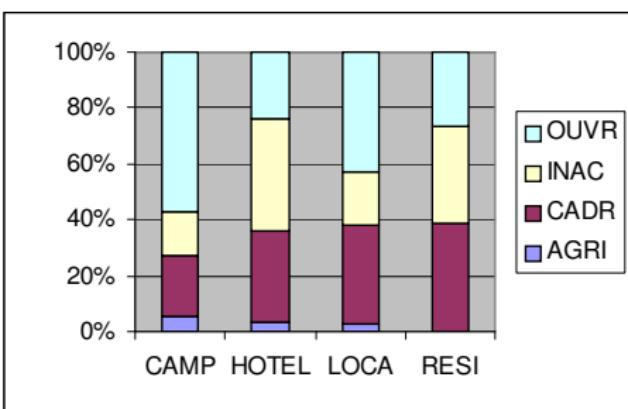
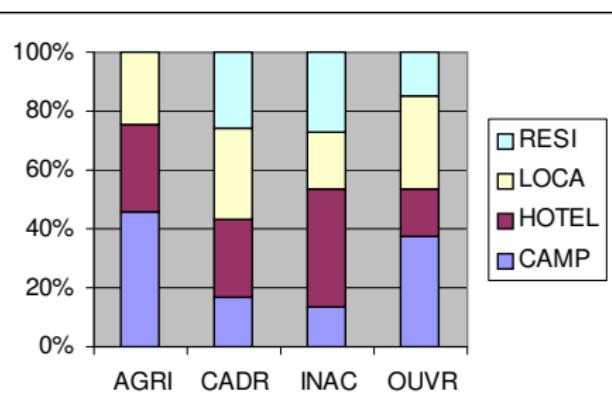


FIGURE – Représentations des profils lignes et colonnes

AFC = une ACP bien choisie

- Elle consiste à faire une ACP bien choisie du tableau des profils-ligne, ce qui est équivalent à l'ACP du tableau des profils-colonne. Le vocabulaire employé sera donc assez voisin de celui de l'ACP. Cependant, leur calcul et usage diffèrera.
- Les résultats, pour un tableau de contingence à r lignes et c colonnes est le suivant.

Valeurs propres et pourcentages d'inertie :			
	F1	F2	F3
Valeur propre	0,098	0,014	0,001
Les lignes dépendent des colonnes (%)	86,855	12,256	0,889
% cumulé	86,855	99,111	100,000

FIGURE – Valeurs propres et pourcentage d'inertie

- Elles sont au maximum au nombre de $\min(r, c) - 1$.
- Dans l'exemple, $r = c = 4$, donc nous aurons 3 axes.
- Leur valeur étant toujours inférieure à 1, la règle de Kaiser pour choisir le nombre d'axes doit toujours être adaptée, en choisissant les valeurs propres supérieures à la moyenne.

- Pour les autres méthodes de choix du nombre d'axes à interpréter, les règles sont les mêmes que pour l'ACP.
- Dans cet exemple, comme il est recommandé de choisir au moins 2 axes, et qu'il y en a 3 en tout, le choix se porte sur 2 axes quelle que soit la méthode.
- **Propriété.** La statistique du khi-deux égale la somme des valeurs propres multipliée par n .
- Cela fait pour cet exemple la relation suivante :

$$18282 * (0,098 + 0,014 + 0,001) = 2067,911.$$

Les coordonnées des modalités et leur représentation graphique

- Une des particularités de l'AFC par rapport à l'ACP est la représentation sur un même graphique des lignes et des colonnes du tableau.

Coordonnées principales (lignes) :			
	F1	F2	F3
AGRI	0,441	0,431	0,137
CADR	-0,140	-0,129	0,027
INAC	-0,379	0,109	-0,020
OUVR	0,355	-0,001	-0,019

Coordonnées principales (colonnes)			
	F1	F2	F3
CAMP	0,443	0,088	-0,022
HOTEL	-0,325	0,139	0,019
LOCA	0,130	-0,124	0,036
RESI	-0,286	-0,110	-0,045

FIGURE — Aides à l'interprétation : Contributions et cosinus carrés

- Ces coordonnées sont centrées, comme en ACP, mais en tenant compte de la pondération de chaque modalité par son effectif total.
- Exemple avec la dimension 1 des coordonnées des lignes :

$$(523 * 0,441 + 5901 * (-0,140) + 4926 * (-0,379) + 6932 * 0,355) / 18282 = 0.$$

Les coordonnées des modalités et leur représentation graphique

Coordonnées principales (lignes) :			
	F1	F2	F3
AGRI	0,441	0,431	0,137
CADR	-0,140	-0,129	0,027
INAC	-0,379	0,109	-0,020
OUVR	0,355	-0,001	-0,019

Coordonnées principales (colonnes)			
	F1	F2	F3
CAMP	0,443	0,088	-0,022
HOTEL	-0,325	0,139	0,019
LOCA	0,130	-0,124	0,036
RESI	-0,286	-0,110	-0,045

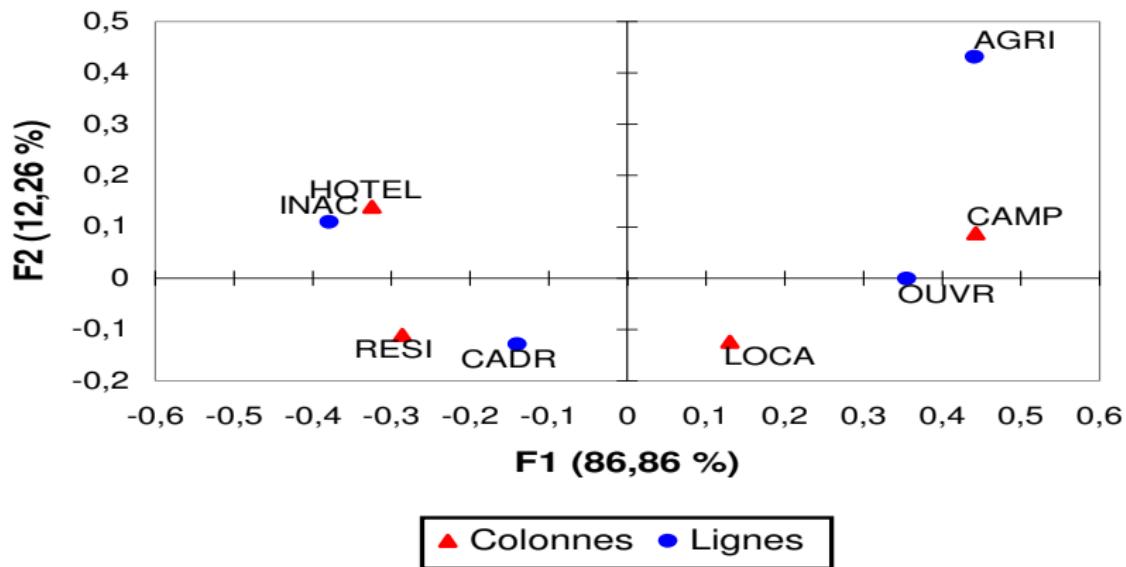
FIGURE — Aides à l'interprétation : Contributions et cosinus carrés

- Cela entraîne le fait qu'il y a toujours des modalités de part et d'autre de chaque axe (coordonnées négatives et positives).
- La variance de chaque axe est aussi, comme en ACP, égale à la valeur propre de l'axe. Cette variance est, comme pour la moyenne pondérée. Cela donne la relation, par exemple :

$$(523*0,4412+5901*(-0,140)^2+4926*(-0,379)^2+6932*0,3552)/18282 =$$

Graphique

**Graphique symétrique
(axes F1 et F2 : 99,11 %)**



Aides à l'interprétation : Contributions et cosinus carrés

Contributions (lignes) :			Cosinus carrés (lignes) :			
	F1	F2	F3	F1	F2	F3
AGRI	0,057	0,383	0,531	0,488	0,465	0,047
CADR	0,064	0,385	0,228	0,532	0,449	0,019
INAC	0,393	0,232	0,105	0,921	0,077	0,003
OUVR	0,486	0,000	0,135	0,997	0,000	0,003

Contributions (colonnes) :			Cosinus carrés (colonnes)			
	F1	F2	F3	F1	F2	F3
CAMP	0,494	0,137	0,122	0,960	0,038	0,002
HOTEL	0,281	0,366	0,092	0,842	0,155	0,003
LOCA	0,048	0,310	0,364	0,504	0,457	0,039
RESI	0,178	0,187	0,422	0,852	0,127	0,021

FIGURE – Aides à l'interprétation : Contributions et cosinus carrés

- Comme en ACP, pour chaque tableau, la somme des contributions d'une colonne égale 1 (ou 100%). Ce qui entraîne une moyenne des contributions égale à $1/r$ pour les modalités ligne, et $1/c$ pour les modalités colonne.
- Dans notre exemple, $r = c = 4$, donc les moyennes des contributions sont égales à $1/4 = 0,25$.
- De même, les sommes des cosinus carrés pour une ligne égalent 1. Donc la moyenne des cosinus carrés égale $1/\text{nombre d'axes total} = 1/3 = 0,333$ dans l'exemple.

Aides à l'interprétation : Contributions et cosinus carrés

- Donc l'axe 1 est basé sur l'opposition entre d'une part le fait que les inactifs choisissent plus souvent l'hôtel et moins souvent le camping, alors que c'est l'inverse chez les ouvriers.
- L'axe 2 est construit sur la base du fait que pour un sous-groupe d'individus, les cadres vont plus souvent en location et moins à l'hôtel, et en cela ils s'opposent aux agriculteurs, qui ont, pour un sous-groupe d'entre eux, un comportement inverse.
- Sur l'axe 1, les cadres sup ont un comportement voisin des inactifs, choisissant plus volontiers l'hôtel, ou la résidence secondaire. A l'opposé, les agriculteurs ont un comportement voisin des ouvriers, choisissant plus le camping ou la location.
- Sur l'axe 2, la modalité HOTEL disparaît par rapport à ce que nous avons dit avec les contributions. Cela signifie que le sous- groupe concerné par l'axe 2 comporte moins de personnes choisissant l'hôtel que les autres modes d'hébergement cités.

Quelques plus de l'AFC

- Reprenons le graphique des modalités. Il permet de voir que les modalités sont disposées en arc de cercle. Ce phénomène est connu sous le nom d'effet Guttman. Il apparaît quand un ordre sous-tend les modalités.
- Dans notre exemple, l'ordre est le suivant : hotel, inac, resi, cadr, loca, ouvr, camp, agri. On peut sans trop s'avancer soupçonner un ordre dû au coût de ces hébergements, et au moyen financier consacré par chaque type de CSP.
- D'autre part, un regard plus attentif permet de constater que, si l'axe 1 est celui des moyens financiers consacrés aux vacances, l'axe 2 est plutôt celui du type de vacances choisi. Les modes d'hébergement côté négatif (RESI, LOCA) sont plutôt de type sédentaire, ceux côté positif sont plutôt destinés à des vacances itinérantes.
- Ces constatations, dont l'explication est grandement aidée par la représentation graphique, sont particulières à l'AFC, et auraient été plus difficilement décelables par une autre analyse de ce tableau.
- On peut facilement généraliser la méthode AFC par l'AFCM, qui consiste à étudier plusieurs variables qualitatives, et à en faire une représentation graphique comme en AFC.

comme une AC particulière

L'AFC est une analyse canonique (AC) particulière

Analyse canonique entre 2 variables qualitatives

Définition

L'AFC est l'AC des 2 tableaux X_1 et X_2 .

- A l'étape k , le problème de l'AC est la détermination de z_1^k et de z_2^k de telle manière que $R^2(z_1^k, z_2^k)$ ait une valeur maximale (cf. Chapitre 4).

Proposition

z_1^k et z_2^k sont alors les vecteurs propres d'ordre k de $P_1 P_2$ et $P_2 P_1$ respectivement.

L'AFC est une analyse canonique (AC) particulière

Analyse canonique entre 2 variables qualitatives

- En reprenant les notations de l'AC, (cf. Chapitre 4), c'est-à-dire, en notant V_{ij} la matrice $\frac{1}{n}({}^t(X_i)X_j)$, les facteurs vérifient les équations suivantes :

$$(V_{11})^{-1} V_{12} a_2^k = R(z_1^k, z_2^k) a_1^k$$

$$(V_{22})^{-1} V_{21} a_1^k = R(z_1^k, z_2^k) a_2^k$$

ou encore

$$(V_{11})^{-1} V_{12} (V_{22})^{-1} V_{21} a_1^k = R^2(z_1^k, z_2^k) a_1^k$$

$$(V_{22})^{-1} V_{21} (V_{11})^{-1} V_{12} a_2^k = R^2(z_1^k, z_2^k) a_2^k$$

où

- V_{21} est le tableau de dimension $p_1 \times p_2$ des fréquences relatives des deux variables qualitatives
- et $(V_{11})^{-1}$ est la matrice diagonale des inverses des fréquences relatives des modalités de la variable qualitative numéro 1.

L'AFC est une analyse canonique (AC) particulière

- Par conséquent, l'élément à l'intersection de la ligne i et de la colonne j du tableau $(V_{11})^{-1} V_{12}$ est égal à $\frac{f_{ij}}{f_{i\bullet}}$, où $(V_{11})^{-1} V_{12}$ est le tableau des profils des lignes.
- De façon symétrique, $(V_{22})^{-1} V_{21}$ est le tableau de dimension p_2 et p_1 dont l'élément à l'intersection de la ligne j et de la colonne i est $\frac{f_{ij}}{f_{\bullet j}}$, où $(V_{22})^{-1} V_{21}$ est le tableau transposé des profils des colonnes.

L'AFC est une analyse canonique (AC) particulière

- Les facteurs et les variables canoniques possèdent la propriété suivante :

Propriété

Les variables canoniques obtenues à partir des indicatrices non centrées sont centrées et les facteurs canoniques sont centrés.

Remarque

- 1 *Centrer une variable consiste à projeter cette variable orthogonalement au vecteur $\mathbf{1}\mathbf{l}$; il n'est donc pas nécessaire de centrer les colonnes des tableaux X_1 et X_2 avant d'effectuer l'AC.*
- 2 *L'AC effectuée à partir des variables non centrées fournit les mêmes résultats aux valeurs triviales près.*

Corrélations canoniques et le Khi-deux

Le lien entre les corrélations canoniques de l'AFC des deux variables qualitatives et la distance du Khi-deux qui mesure la liaison entre ces variables qualitatives est le suivant.

Proposition

A la valeur propre trivial près, la somme des carrés des corrélations canoniques de l'AFC de deux variables qualitatives est égale au χ^2 divisé par n.

$$\sum_{k=2}^{p_1} R^2(z_1^k z_2^k) = \frac{\chi^2}{n}.$$

Par conséquent

Proposition

Le Khi-deux est une mesure globale de la liaison entre deux variables qualitatives. L'AFC décrit plus précisément cette liaison, en la décomposant en plusieurs relations entre des couples de variables canoniques. L'intensité de chacune de ces relations est mesurée par le coefficient de corrélation canonique.

Remarque

Où l'AFC peut se présenter :

comme la juxtaposition de deux
ACP

ACP des deux tableaux de profils

- L'AFC constitue **un cas particulier de l'AC**, mais, et ceci est pour beaucoup dans l'intérêt que présente cette technique
- l'AFC peut aussi être présentée comme **la juxtaposition de deux ACP** :
 - l'une portant sur **les profils des lignes** du tableau de contingence
 - et l'autre sur **les profils des colonnes** du tableau de contingence

ACP des deux tableaux de profils

ACP des profils de la première variable qualitative

V_{12} est le tableau à p_1 lignes et p_2 colonnes des fréquences relatives des deux variables qualitatives.

Le tableau des profils des lignes est le tableau $(V_{11})^{-1} V_{12}$: l'élément situé à l'intersection de la ligne i et de la colonne j de ce dernier tableau est $\frac{f_{ij}}{f_{i\bullet}}$.

Chaque ligne du tableau $(V_{11})^{-1} V_{12}$ peut être représentée par un point dans \mathbb{R}^{p_2} . La distance retenue ici pour calculer la proximité entre les p_1 points est la distance du Khi-deux.

Cette distance du Khi-deux entre 2 individus i et i' , $d(i, i')$, est la racine carrée positive de

$$d^2(i, i') = \sum_{j=1}^{p_2} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2.$$

Ainsi, deux modalités de la première variable qualitative sont proches si les profils des lignes correspondantes sont proches. La différence entre cette distance et la distance usuelle est que chaque modalité de la seconde variable qualitative a pour poids l'inverse de sa fréquence relative, ceci afin de ne pas donner une importance trop grande dans le calcul de la distance aux modalités.

ACP des deux tableaux de profils

La distance du Khi-deux possède une propriété intéressante (appelée *principe d'équivalence distributionnelle*) que ne possède pas la métrique euclidienne usuelle.

Proposition

Le calcul des distances entre les profils des lignes n'est pas modifié (et donc le résultat de l'ACP n'est pas modifié) si on agrège deux colonnes j et j' possédant le même profil.

L'ACP permet de décrire ce nuage de p_1 points (les profils des lignes) :

Proposition

L'AFC est l'ACP du tableau des profils de la première variable qualitative, où les lignes sont les individus et les colonnes les variables. Cette ACP est effectuée en utilisant une distance particulière, la distance du Khi-deux.

Les composantes principales successives sont les vecteurs a_1^k .

ACP des deux tableaux de profils

Notons qu'ici l'ACP a été menée à partir du tableau des données non centrées.
En effet :

Proposition

Effectuer l'ACP du tableau des profils centrés est équivalent, à la valeur propre triviale près, à effectuer l'ACP du tableau des profils non centrés.

ACP des deux tableaux de profils

ACP des profils de la seconde variable qualitative

L'ACP des profils de la seconde variable qualitative repose sur les mêmes principes que l'ACP des profils de la première variable qualitative. Il suffit pour ceci de permute les rôles des indices i et j .

Les individus sont alors les profils de la seconde variable qualitative, et les lignes du tableau constituent les variables. La distance $d(j, j')$ entre 2 individus, est alors la racine carrée positive de

$$d^2(j, j') = \sum_{i=1}^{p_1} \left(\frac{f_{ij}}{f_{\bullet j} \sqrt{f_{i\bullet}}} - \frac{f_{ij'}}{f_{\bullet j'} \sqrt{f_{i\bullet}}} \right)^2.$$

Enfin, chaque individu a un poids $f_{\bullet j}$.

ACP des deux tableaux de profils

ACP des profils de la seconde variable qualitative

Proposition

L'AFC est l'ACP du tableau des profils de la seconde variable qualitative, où les lignes étant les variables et les colonnes les individus. Cette ACP est effectuée en utilisant une distance particulière, la distance du Khi-deux.

Les composantes principales sont les vecteurs propres de :

$$C_2^t (C_2) V_{11} = (V_{22})^{-1} V_{21} (V_{11})^{-1} V_{12}.$$

Ainsi

Proposition

Les composantes principales successives sont les vecteurs a_2^k .

Représentation des modalités

Deux approches sont possibles, selon que l'on considère :

- l'AFC comme deux ACP particulières
- ou comme un cas particulier de l'AC.

Ces deux approches conduisent aux mêmes types de représentations graphiques des individus.

Représentation des modalités : l'approche ACP

L'AFC consiste à effectuer deux ACP séparées, à partir des profils des modalités de la première variable qualitative et à partir des profils des modalités de la seconde variables qualitative.

Pour chacune des deux ACP, on obtient une représentation des individus-modalités qui font l'objet de l'analyse. Ainsi, pour les résultats des étapes r et s , on utilise :

- la représentation graphique croisant a'_1 et a^s_1 pour décrire les proximités entre les modalités de la première variable qualitative.
- la représentation graphique croisant a'_2 et a^s_2 pour décrire les proximités entre les modalités de la seconde variables qualitative.

Représentations graphiques et aides à l'interprétation

Représentation des modalités : l'approche ACP

Ces deux représentations graphiques ne sont pas indépendantes. En effet, il existe entre les composantes principales d'une même étape, par exemple entre a'_1 et a'_2 , les relations suivantes :

$$(V_{11})^{-1} V_{12} a'_2 = R(z'_1, z'_2) a'_1$$

$$(V_{22})^{-1} V_{21} a'_1 = R(z'_1, z'_2) a'_2$$

Si on note a'_{1i} la i ème coordonnée de a'_1 , c'est-à-dire le codage de la modalité i de la première variable qualitative à l'étape r , et a'_{2j} le codage de la modalité j de la seconde variable qualitative à cette étape r , on obtient :

Proposition

$$\sum_{j=1}^{p_2} \frac{f_{ij}}{f_{i\bullet}} a'_{2j} = R(z'_1, z'_2) a'_{1i} \text{ pour } i = 1, \dots, p_1$$

$$\sum_{i=1}^{p_1} \frac{f_{ij}}{f_{\bullet j}} a'_{1i} = R(z'_1, z'_2) a'_{2j} \text{ pour } j = 1, \dots, p_2$$

Représentation des modalités : l'approche ACP

Proposition

Sur les graphiques, la modalité i de la première variable qualitative est le barycentre, à un coefficient multiplicatif près, des modalités de la seconde variable qualitative.

La modalité j de la seconde variable qualitative est le barycentre, à un coefficient multiplicatif près, des modalités de la première variable qualitative.

Représentations graphiques et aides à l'interprétation

Représentation des modalités : l'approche ACP

- Pour utiliser pleinement cette double relation barycentrique, on superpose les graphiques des deux ACP. Ainsi pour les étapes r et s , on obtient un graphique comportant $(p_1 + p_2)$ points. Les coordonnées sont données par a_1^r et a_1^s pour les p_1 modalités de la première variable qualitative et par a_2^r et a_2^s pour les p_2 modalités de la seconde variable qualitative.
- La proximité sur le graphique entre une modalité de la première variable qualitative et une modalité de la seconde variable qualitative ne doit cependant pas être interprétée hâtivement : elle n'a pas de signification, car aucune distance n'est définie entre une ligne et entre une colonne du tableau de contingence.
- La double relation barycentrique indique simplement que la modalité d'une variable est *attirée* par les modalités de l'autre variable qualitative pour lesquelles elle possède des fréquences relatives élevées.

Représentations graphiques et aides à l'interprétation

Représentation des modalités : l'approche AC

- En AC, pour décrire les résultats de l'étape r , on représente sur le même axe les composantes principales des deux tableaux, soit z_1^r et z_2^r . La relation qui lie composante canonique et facteur est ici particulièrement simple. Considérons par exemple la composante d'ordre r du premier tableau :

$$z_1^r = X_1 a_1^r$$

z_1^r prend n valeurs, une par individu, mais seules p_1 de ces valeurs sont différentes et correspondent aux p_1 valeurs prises par le facteur a_1^r (chapitre 2).

- Les graphiques seront alors constitués par $(p_1 + p_2)$ points, qui représentent les modalités des deux variables qualitatives. On retrouve les représentations graphiques décrites dans l'approche ACP de l'AFC.
- Notons cependant une différence de normalisation des composantes entre l'approche AC et l'approche ACP : les composantes canoniques ont une variance égale à 1, alors que les composantes principales ont une variance égale à la valeur propre de l'étape dont elles sont issues. C'est cette dernière convention qui est le plus souvent admise et que nous utiliserons ici pour les représentations graphiques.



Représentations graphiques et aides à l'interprétation

Aides à l'interprétation

Comme en ACP, deux types d'aides à l'interprétation sont calculés pour permettre de mieux comprendre les graphiques de l'AFC :

- les contributions des modalités à la variance
- et la qualité de représentation des modalités.

Représentations graphiques et aides à l'interprétation

Les contributions des modalités à la variance

La variance d'une composante principale donnée, a'_1 par exemple, est égale à la valeur propre d'ordre r .

Proposition

La contribution de la modalité i de la première variable qualitative à la variance est égale à

$$\frac{f_{i\bullet}(a'_{1i})^2}{R^2(z'_1, z'_2)}.$$

Pour la modalité j de la seconde variable, la contribution est égale à

$$\frac{f_{\bullet j}(a'_{2j})^2}{R^2(z'_1, z'_2)}.$$

Remarque

Pour chacune des deux variables qualitatives, la somme des contributions des modalités est égale à 100%.

Les modalités qui contribuent fortement à la variance d'une composante principale expliquent cette composante principale.

Représentations graphiques et aides à l'interprétation

Qualité de représentation des modalités

La modalité i est représentée par un point dans \mathbb{R}^{p_2} . Comme en ACP (chapitre 3), on peut mesurer la qualité de représentation de ce point sur l'axe r .

Le centre du nuage a pour j ème coordonnée, dans le cas de la première analyse :

$$\sum_{i=1}^{p_1} f_{i\bullet} \frac{f_{ij}}{f_{i\bullet}} = \sum_{i=1}^{p_1} f_{ij} = f_{\bullet j}.$$

et le carré de la distance de i au centre du nuage est égal à

$$d^2(i, O) = \sum_{j=1}^{p_2} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2.$$

Donc, le cosinus carré, qui mesure la qualité de représentation de i , est égal à

$$\frac{(a_{1i}^r)^2}{\sum_{j=1}^{p_2} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2}$$

où a_{1i}^r désigne la i ème coordonnée de la composante principale d'ordre r du premier tableau.

Les modalités bien représentées sont expliquées par la composante principale.

Application 1

L'AFC consiste à faire une ACP bien choisie du tableau des profils-ligne, ce qui est équivalent à l'ACP du tableau des profils-colonne. Le vocabulaire employé sera donc assez voisin de celui de l'ACP. Cependant, leur calcul et usage diffèrera. Les résultats, pour un tableau de contingence à r lignes et c colonnes est le suivant.

- Les valeurs propres :
 - Elles sont au maximum au nombre de $\min(r, c) - 1$. Dans l'exemple, $r = c = 4$, donc nous aurons 3 axes.
 - Leur valeur étant toujours inférieure à 1, la règle de Kaiser pour choisir le nombre d'axes doit toujours être adaptée, en choisissant les valeurs propres supérieures à la moyenne.
 - METTRE LE TABLEAU
- Pour les autres méthodes de choix du nombre d'axes à interpréter, les règles sont les mêmes que pour l'ACP. Dans cet exemple, comme il est recommandé de choisir au moins 2 axes, et qu'il y en a 3 en tout, le choix se porte sur 2 axes quelle que soit la méthode.

Application pratique 2

Une application de l'AFC est exposée ci-dessous. Elle concerne un tableau de petite taille (le tableau du chapitre 2, croisant classe d'âge et diplôme) : la petite taille du tableau analysé permet de détailler les calculs que nécessite la mise en oeuvre d'une AFC, et donc de mieux comprendre les règles d'interprétation des résultats de cette technique.

Reprendons donc, le tableau de contingence du chapitre 2, croisant la catégorie d'âge et le diplôme élevé détenu.

	BEPC	BAC	Licence	Total
Plus de 50 ans	15	12	3	30
Entre 30 et 50 ans	10	18	4	32
Moins de 30 ans	15	5	8	28
Total	40	35	15	90

Tableau de contingence des effectifs observés

où *Total* désigne les marges du tableau : il y a 30 individus de > 50 ans dans la population et parmi lesquels 15 individus titulaires d'une licence ou plus.

Application pratique 2

A ce tableau sont associées les matrices suivantes, en utilisant la notation des paragraphes précédents : V_{11} est la matrice diagonale des fréquences relatives des lignes, V_{12} est le tableau des fréquences relatives et V_{22} la matrice diagonale des fréquences relatives des colonnes.

$$V_{11} = \begin{pmatrix} \frac{30}{90} & 0 & 0 \\ 0 & \frac{32}{90} & 0 \\ 0 & 0 & \frac{28}{90} \end{pmatrix}, \quad V_{12} = \begin{pmatrix} \frac{15}{90} & \frac{12}{90} & \frac{3}{90} \\ \frac{10}{90} & \frac{18}{90} & \frac{4}{90} \\ \frac{15}{90} & \frac{5}{90} & \frac{8}{90} \end{pmatrix} \text{ et } V_{22} = \begin{pmatrix} \frac{40}{90} & 0 & 0 \\ 0 & \frac{35}{90} & 0 \\ 0 & 0 & \frac{15}{90} \end{pmatrix}$$

Le tableau des profils des modalités de la première variable qualitative est alors :

$$(V_{11})^{-1} V_{12} = \begin{pmatrix} \frac{15}{30} & \frac{12}{30} & \frac{3}{30} \\ \frac{10}{32} & \frac{18}{32} & \frac{4}{32} \\ \frac{15}{28} & \frac{5}{28} & \frac{8}{28} \end{pmatrix}$$

Et le tableau des profils des modalités de la seconde variable qualitative est :

$$V_{12}(V_{22})^{-1} = \begin{pmatrix} \frac{15}{40} & \frac{12}{35} & \frac{3}{15} \\ \frac{18}{40} & \frac{18}{35} & \frac{4}{15} \\ \frac{15}{40} & \frac{5}{35} & \frac{8}{15} \end{pmatrix}$$

Application pratique

Pour obtenir les décompositions principales de la première ACP, il faut diagonaliser la matrice suivante, qui est le profil de la matrice des profils des lignes avec la transposée de la matrice des profils des colonnes :

$$(V_{11})^{-1} V_{12} (V_{22})^{-1} V_{21} = \begin{pmatrix} .345 & .357 & .298 \\ .335 & .400 & .365 \\ .320 & .301 & .379 \end{pmatrix}$$

Les valeurs propres de cette matrice sont : 1 (valeur propre triviale), 0.1092 et 0.0149.

On vérifie bien que la somme des valeurs propres permet de retrouver la valeur du χ^2 égale, aux arrondis près, à 11.18 : $\chi^2 = 90(0.1092 + 0.0149)$, ou encore à partir des éléments diagonaux de la matrice précédente :

$$\chi^2 = 90(0.345 + 0.400 + 0.379 - 1).$$

Les pourcentages de variance expliquée sont donc de :

$$88\% = \frac{0.1092}{0.109 + 0.0149} \text{ (pour le 1er axe)}$$

et

$$12\% = \frac{0.0149}{0.109 + 0.0149} \text{ pour le 2nd axe.}$$

Application pratique

A ces valeurs propres correspondent les vecteurs propres $\begin{pmatrix} 0.068 \\ 0.340 \\ -0.462 \end{pmatrix}$ pour la valeur propre 0.1092 et $\begin{pmatrix} -0.171 \\ 0.106 \\ 0.062 \end{pmatrix}$ pour la valeur propre 0.0149.

Ces vecteurs propres sont centrés et la variance d'un vecteur propre est égale à la valeur propre correspondante, ce que l'on vérifie bien ici (aux arrondis de calcul près) :

pour la 1ère étape :

$$\frac{30}{90}(0.068) + \frac{32}{90}(0.34) + \frac{28}{90}(-0.462) = 0$$

et

$$\frac{30}{90}(0.068)^2 + \frac{32}{90}(0.34)^2 + \frac{28}{90}(-0.462)^2 = 0.1092$$

Application pratique

pour la seconde étape :

$$\frac{30}{90}(-0.171) + \frac{32}{90}(0.106) + \frac{28}{90}(0.062) = 0$$

et

$$\frac{30}{90}(-0.171)^2 + \frac{32}{90}(0.106)^2 + \frac{28}{90}(0.062)^2 = 0.0149$$

Ces vecteurs propres constituent la 1ère et la 2nde composante principale de la 1ère ACP (l'ACP des profils des lignes), c'est-à-dire donnent les coordonnées des points *Plus de 50 ans*, *entre 30 et 50 ans* et *moins de 30 ans* sur les deux premiers axes.

Application pratique

Pour déterminer les coordonnées des modalités de la seconde variable qualitative, il est possible de procéder de deux manières différentes :

- soit on diagonalise :

$$(V_{22})^{-1} V_{21} (V_{11})^{-1} V_{12} = \begin{pmatrix} 0.466 & 0.358 & 0.176 \\ 0.409 & 0.452 & 0.139 \\ 0.469 & 0.325 & 0.206 \end{pmatrix}$$

Les valeurs propres de cette matrice sont : 1 (valeur propre triviale), 0.1092 et 0.0149, comme pour la 1ère analyse.

On vérifie bien que la somme des valeurs propres permet de retrouver la valeur du χ^2 égale, aux arrondis près, à 11.18 :

$$\chi^2 = 90(0.466 + 0.452 + 0.206 - 1).$$

A ces valeurs propres correspondent les vecteurs propres $\begin{pmatrix} -0.189 \\ 0.401 \\ -0.430 \end{pmatrix}$

pour la valeur propre 0.1092 et $\begin{pmatrix} -0.117 \\ 0.039 \\ 0.222 \end{pmatrix}$ pour la valeur propre 0.0149.

Application pratique

- Si on utilise la relation barycentrique : par exemple, il est possible d'obtenir la coordonnée de *Bac* pour la 1ère composante principale, à partir des coordonnées de *Plus de 50 ans*, *Entre 30 ans et 50 ans* et *Moins de 30 ans*, en calculant :

$$\frac{1}{\sqrt{0.1092}} \left(\frac{12}{35}(0.068) + \frac{18}{35}(0.34) + \frac{5}{35}(-0.462) \right)$$

et on obtient bien la valeur de la 1ère composante principale pour *Bac*, soit 0.401.

Application pratique

- Une fois déterminées les composantes principales, il reste à calculer les valeurs des aides à l'interprétation.
- Ainsi, la contribution de la modalité *Licence* à la variance de 1er axe est :

$$\frac{1}{0.1092} \left(\frac{15}{90} (-0.430)^2 \right) = 28.2\%$$

- Le carré de la distance du profil de la modalité *Licence* à l'origine étant égal à :

$$\frac{90}{30} \left(\frac{3}{15} - \frac{30}{90} \right)^2 + \frac{90}{32} \left(\frac{4}{15} - \frac{32}{90} \right)^2 + \frac{90}{28} \left(\frac{8}{15} - \frac{28}{90} \right)^2 = 0.234$$

- Pour *Licence* et pour l'axe 1, la qualité de représentation est donc égale à

$$\frac{(-0.430)^2}{0.234}$$

Application pratique

En procédant de la même manière avec les six modalités et les deux axes, on obtient les résultats suivants :

	Axe 1 Coord.	Axe 1 Q.R.	Axe 1 C. V.	Axe 2 Coord.	Axe 2 Q. R.	Axe 2 C. V.
Plus de 50 ans	0.068	0.138	1.4	-0.171	0.862	65.2
Entre 30 et 50 ans	0.340	0.912	37.7	0.106	0.088	26.7
Moins de 30 ans	-0.462	0.982	60.9	0.062	0.018	8.1
BEPC	-0.189	0.723	14.6	-0.117	0.277	40.9
BAC	0.401	0.991	57.2	0.039	0.009	3.9
Licence	-0.430	0.789	28.2	0.222	0.211	55.2

où *Coord.* : coordonnées, *Q.R.* : qualité de représentation et *C.R* : contribution à la variance.

Application pratique

Les coordonnées des deux 1ères composantes principales permettent d'obtenir le graphique croisant les deux 1ers axes.

Application pratique : interprétation

- Les catégories *Moins de 30 ans* et *Bac*, et dans une moindre mesure les catégories *Licence* et *Entre 30 et 50 ans* contribuent fortement à la variance du 1er axe.
- Sur cet axe *Licence* et *Moins de 30 ans* s'opposent à *Bac* et entre *30 et 50 ans*. Ceci traduit la sur-représentation des titulaires d'une Licence chez les moins de 30 ans (l'effectif observé est de 8 alors que l'effectif théorique est de 4.66 (chapitre 1)) et la sous-représentation des bacheliers dans cette catégorie d'âge (effectif théorique de 10.88 pour un effectif observé de 5). A l'inverse, les *30 – 50 ans* sont nombreux dans la catégorie des bacheliers (l'effectif théorique est 12.44 pour un effectif observé de 18) et relativement peu nombreux dans la catégorie des licenciés (4 pour un effectif théorique de 5.33).

Application pratique : interprétation

- La modalité *BEPC* ne contribue pas de façon importante à la variance de l'axe, mais est bien représentée sur cet axe : sa position traduit une sur-représentation des moins de 30 ans (15, alors que les effectifs théoriques sont de 12.44) et une sous-représentation des 30 – 50 ans pour les classes d'âge, *BEPC* et *Licence* pour les diplômes sont les modalités qui contribuent le plus à la variance.
- *Plus de 50 ans* s'oppose à *Licence*(effectivement, l'effectif théorique des licenciés de plus de 50 ans est de 5, alors que dans la réalité, ils ne sont que 3), alors que *BEPC* est situé du même côté de l'axe que *Plus de 50 ans* (13.3 en théorie, 15 en réalité).
- Cet exemple illustre la manière dont l'AFC décrit les liaisons entre les modalités de deux variables qualitatives. Une application portant sur un tableau de taille plus importante est décrite dans le TP.