Chapitre 5

Analyse des Correspondances

5.1 Introduction

L'analyse factorielle des correspondances est un cas particulier de l'analyse canonique. Elle a été développée essentiellement par J.-P. Benzecri durant la période 1970-1990. L'analyse des correspondances est une technique d'analyse fatorielle destinée à mettre en évidence et décrire des associations entre deux variables qualitatives. On considère dans cette section deux variables qualitatives observées simultanément sur n individus de poids identiques $1/n$. En pratique, on va travailler avec une table de contingence qui est un tableau croisé contenant les effectifs des occurences simultannées de deux modalités.

Prenons des exemples,

1. Ponctuation dans l'oeuvre de Zola *(exemple emprunté M. Tenenhaus)* - L'étude de la ponctuation ou de la présence de certains mots dans des textes est utilisée pour reconnaitre l'auteur d'un document (article, roman, nouvelle, etc.). Les données se présentent selon le tableau Tab. **??**.

   Et une analyse factorielle des correspondances permet de faire le graphique suivant sur lequel on projette simultannément les modalités des deux variables (`Titre du roman` et `Ponctuation`) comme représenté dans la figure **??**.

2. Origine sociale des étudiants de première année et choix d'un secteur disciplinaire *(exemple emprunté à F.-G. Carpentier)*

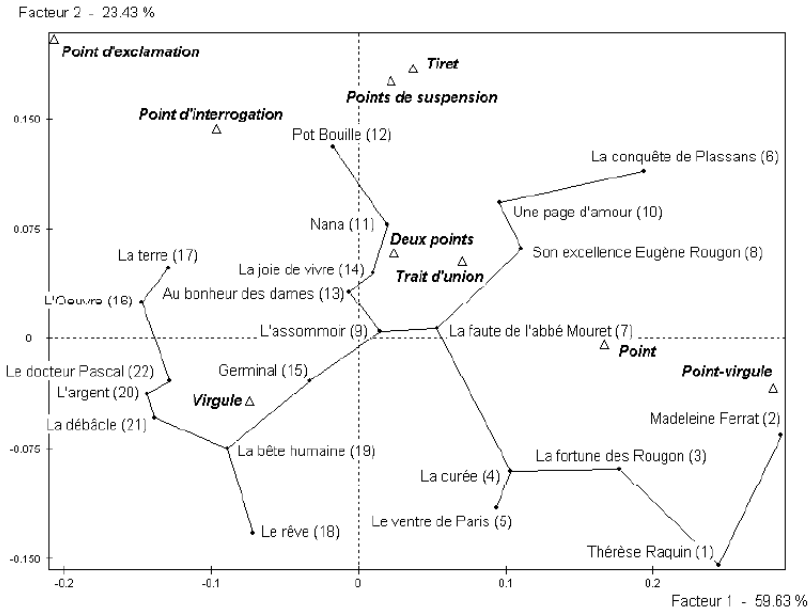
	Droit	Science	Médecine	IUT	**Total**
Exp. agri.	80	99	65	58	**302**
Patron	168	137	208	62	**575**
Cadre sup.	470	400	876	79	**1825**
Employé	145	133	135	54	**467**
Ouvrier	166	193	127	129	**615**
Total	**1029**	**962**	**1411**	**382**	**3784**

Soient [1] deux variables nominales X et Y, comportant respectivement p et q modalités. On a

1. Certaines parties de ce chapitre et notamment ce paragraphe sont fortement inspirées du cours de F.G. Carpentier

Roman	.	...	!	?	,	;	:	—	-
1. Thérèse Raquin	3468	236	138	76	6195	691	168	285	543
2. Madeleine Ferrat	5131	362	236	245	8012	922	291	518	1115
3. La fortune des Rougon	6157	238	534	229	11346	936	362	711	1301
4. La curée	4958	443	357	232	11164	738	364	679	1200
5. Le ventre de Paris	5538	534	426	232	13234	1015	318	734	1201
6. La conquête de Plassans	6292	943	756	512	11585	1285	402	1432	1916
7. La faute de l'abbé Mouret	6364	679	859	462	13948	634	377	1067	1564
8. Son excellence Eugène Rougon	7258	728	1002	496	14295	889	543	1469	1907
9. L'assommoir	7820	769	1929	443	19244	1399	436	995	2272
10 Une page d'amour	6206	843	918	492	11953	647	347	1235	1409
11. Nana	7821	1007	1796	611	17881	1087	509	1523	1797
12. Pot Bouille	6875	1045	1873	651	17044	912	675	1669	1935
13. Au bonheur des dames	6916	808	1313	651	18402	972	642	1531	2114
14. La joie de vivre	5803	710	972	623	13917	602	420	1142	1590
15. Germinal	7944	606	1463	729	21388	908	621	1362	2083
16. L'Œuvre	5000	774	1692	668	18292	811	566	1107	1489
17. La terre	6979	957	2307	796	23417	947	657	1681	2113
18. Le rêve	3052	292	385	237	9551	345	230	416	650
19. La bête humaine	5484	601	929	557	18264	673	467	957	1721
20. L'argent	5022	850	1235	569	19267	684	399	1049	1677
21. La débâcle	7440	860	1833	690	26482	832	564	1398	2197
22. Le docteur Pascal	4586	621	1072	464	15598	462	315	955	1218

Ponctuation dans les romans de Zola



Premier plan factoriel de l'ACM de la ponctuation dans la romans de Zola.

observé les valeurs de ces variables sur une population et on dispose d'un tableau de contingence à p lignes et q colonnes donnant les effectifs conjoints c'est-à-dire les effectifs observés pour chaque combinaison d'une modalité i de X et d'une modalité j de Y. Les valeurs de ce tableau seront notées n_{ij} , l'effectif total sera noté N.

L'AFC vise à analyser ce type de tableaux en apportant des réponses à des questions telles que :
- Y a-t-il des lignes du tableau (modalités de X) qui se "ressemblent", c'est-à-dire telles que les distributions des modalités de Y soient analogues ?
- Y a-t-il des lignes du tableau (modalités de X) qui s'opposent, c'est-à-dire telles que les distributions des modalités de Y soient très différentes ?
- Mêmes questions pour les colonnes du tableau.
- Y a-t-il des associations modalité de X - modalité de Y qui s'attirent (effectif conjoint particulièrement élevé) ou qui se repoussent (effectif conjoint particulièrement faible) ?

La méthode se fixe également comme but de construire des représentations graphiques mettant en évidence ces propriétés des données.

Notations

Soit $\mathbf{N} = (n_{ij})_{i=1,\cdots,p,j=1,\cdots,q}$ un tableau de contingence. On définit les marges du tableau par

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij} \ , \ n_{\bullet j} = \sum_{i=1}^{p} n_{ij}, \ n = n_{\bullet\bullet} \sum_{i,j} n_{ij}$$

Ceci correspond aux totaux en lignes et en colonne. Selon le mÃƉme principe, on peut définir les marges en fréquence avec $f_{ij} = n_{ij}/n$

$$f_{i\bullet} = \sum_{j=1}^{q} f_{ij}, \ f_{\bullet j} = \sum_{i=1}^{p} f_{ij}, \ f_{\bullet\bullet} \sum_{i,j} f_{ij} = 1$$

5.2 Modèle d'indépendance

5.2.1 Test du chi 2

Comme en ACP, on s'intéresse alors aux directions de "plus grande dispersion" de chacun de ces nuages de points, mais on utilise la distance du χ^2 entre ces deux variables (à la place de la distance euclidienne). Cette distance permet de comparer l'effectif de chacune des cellules du tableau de contingence à la valeur qu'elle aurait si les deux variables étaient indépendantes. Notons E_{ij} l'effectif attendu sous l'hypothèse d'indépendance ; par définition

$$E_{ij} = \frac{\text{Total ligne } i \times \text{Total ligne } j}{\text{Total général}} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}}$$

ce qui correspondant bien au produit des probabilités marginales. Et la distance du χ^2 est définie par

$$d^2_{\chi^2}(\mathbf{N}, \mathbf{E}) = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

On appelle *résidus standardisés*, les variables (centrées et de variance 1) :

$$c_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

Plus la distance $d^2_{\chi^2}(\mathbf{N}, \mathbf{E})$ est grande, plus le tableau observé est éloigné du tableau attendu sous l'hypothèse d'indépendance.

Pourquoi utiliser cette métrique plutôt que la métrique euclidienne ? Deux raisons fortes peuvent être avancées :

- Avec la métrique du χ^2, la distance entre deux lignes ne dépend pas des poids respectifs des colonnes. Ceci a pour conséquence, dans l'exemple, des étudiants de première année que les catégories socio-professionnelles sur-représentées ne prennent pas plus de poids que les autres dans le calcul de la distance.
- La métrique du χ^2 possède la propriété d'équivalence distributionnelle : si on regroupe deux modalités lignes, les distances entre les profils-colonne, ou entre les autres profils-lignes restent inchangées.

Notons qu'en revanche, il n'existe pas d'outil mesurant une "distance" entre une ligne et une colonne.

Sous l'hypothèse d'indépendance des deux variables, la statistique $d^2_{\chi^2}$ suit une loi du χ^2 à $(p-1)(q-1)$ degrés de liberté. Cette loi sert, par exemple, à définir une règle de décision du type : *On conclut que les variables sont indépendantes avec un risque α de se tromper si $d_{\chi^2}(\mathbf{N}, \mathbf{E}) < F^{-1}_{(p-1)(q-1)}(1-alpha)$ vec F la fonction de répartition de la loi du χ^2 à $(p-1)(q-1)$ degrés de liberté*

Dans l'exemple des étudiants de première année, la distance du χ^2 observée est

$$d^2_{\chi^2,\text{obs}}(\mathbf{N}, \mathbf{E}) = 320.2$$

et on la compare à $F^{-1}_{12}(.95) = 21.0$. La valeur de la statistique observée $d^2_{\chi^2,\text{obs}}(\mathbf{N}, \mathbf{E})$ étant supérieure au seuil, on conclut ici que le tableau observé est significativement éloigné du tableau attendu sous l'hypothèse d'indépendance et donc que les deux variables sont liées.

5.2.2 AFC et indépendance

L'analyse d'un tableau de contingence doit donc se faire en référence à la situation de d'indépendance. C'est ce que fait l'AFC en écrivant le modèle d'indépendance sous la forme suivante :

$$\forall i = 1, \cdots, p, \forall j = 1, \cdots, q, \quad \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}$$

La quantité $f_{ij}/f_{i\bullet}$ est la probabilité conditionnelle de posséder la modalité j de la variable X_2 sachant que l'on possède la modalité i de la variable X_1. De faÃĄon symétrique, on peut écrire

$$\forall i = 1, \cdots, p, \forall j = 1, \cdots, q, \quad \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}$$

Définition 7 – *L'ensemble de probabilités* $\{f_{ij}/f_{i\bullet}; j = 1, \cdots, q\}$ *est appelée profil ligne.*
– *L'ensemble de probabilités* $\{f_{ij}/f_{\bullet j}; i = 1, \cdots, p\}$ *est appelée profil colonne.*
– $\{f_{i\bullet}; j = 1, \cdots, q\}$ *(resp.* $\{f_{\bullet j}; i = 1, \cdots, p\}$ *est le profil moyen correspondant au profil ligne (resp. colonne).*

Remarque - Si on a indépendance, le profil ligne d'une part et colonne d'autre part est égal au profil moyen correspondant.

5.3 Analyse factorielle des correspondances

On va voir que l'AFC est une double ACP : ACP des profils ligne et ACP des profils colonne.

5.3.1 Nuages de points

Intéressons nous aux profils ligne, l'analyse des profils colonne étant symétrique. On peut définir la notion de nuage d'individus (ou de modalité) partir du tableau de contingence en fréquence. En pratique, on construit un nuage de points dans l'espace $\mathbb{R}^q$ en définissant pour chaque ligne i, un point dont la coordonnées dans la dimension j est $f_{ij}/f_{i\bullet}$. Ce nuage est complété par le point moyen G_I dont la $j$ème coordonnée vaut $f_{\bullet j}$. Chaque point i est affecté du poids $f_{i\bullet}$.

On remarque que la distance entre les points i et i' (c'est à dire deux modalités de X_1) est

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

On utilise donc ici la métrique du χ^2 dans laquelle les inverses des fréquences marginales des modalités de Y sont introduites comme pondérations des écarts entre éléments de deux profils relatifs à X. Cette métrique attribue donc plus de poids aux écarts correspondants à des modalités de faible effectif (rares) pour Y . L'inertie du point i par rapport à G_I s'écrit

$$\begin{aligned} \text{Inertie}(i/G_I) &= f_{i\bullet} d^2_{\chi^2}(i, G_I) \\ &= f_{i\bullet} \sum_{j=1}^{q} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2 \\ &= \sum_{j=1}^{q} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}} \end{aligned}$$

5.3.2 l'AFC proprement dite

Pour étudier les lignes, on peut réaliser une ACP de la matrice A (telle que $a_{ij} = n_{ij}/n_{i\bullet}$)) puis de représenter les modalités de la première variable. En raison du changement de métrique, on introduit la matrice $M = D_C^{-1}$ avec $D_C = diag(n_{\bullet 1}, \cdots, n_{\bullet q})$ et on considère la matrice de poids $D = D_L^{-1}$ avec $D_C L = diag(n_{1\bullet}, \cdots, n_{q\bullet})$ (pour favoriser les gros effectifs, ce qui est discutable mais permet de faire facilement les calculs). On remarque $A = D_L^{-1}\mathbf{N}$. De façon symétrique, on peut définir $B = \mathbf{N} D_C^{-1}$.

Proposition 4 *Les éléments de l'ACP de* (A, D_C^{-1}, D_L) [2] *sont fournis pas l'analyse spectrale de la matrice carrée,* D_L^{-1}*-symétrique et semi-definie positive* AB*.*

Preuve - Elle se construit en remarque successivement que
- le barycentre du nuage des profils ?colonnes est le vecteur g_C des fréquences marginales de X_2,
- la matrice $A'D_LA - g_CD_Lg_C'$ joue le rôle de la matrice des variances ?covariances,
- la solution de l'ACP est fournie par la D.V.S. de $(A-1g_L', D_C^{-1}, D_L)$ qui conduit à rechercher les valeurs et vecteurs propres de la matrice (SM)

$$A'D_LAD_C^{-1} - G_CD_LG_C' = AB - G_CG_C'D_R^{-1}(\text{car } D_C^{-1}A' = BD_L^{-1})$$

- les matrices $AB - G_CG_C'D_R^{-1}$ et AB ont les mêmes vecteurs propres associées aux mêmes valeurs propres, à l'exception du vecteur g_L associé à la valeur propre $\lambda_0 = 0$ de $AB - G_CG_C'D_R^{-1}$ et à la valeur propre $\lambda_0 = 1$ de AB.

◇

On note U la matrice contenant les vecteurs propres D_C^{-1}-othonormés de AB. La représentation des ?individus ? de l'ACP réalisée fournit une représentation des modalités de la variable X_1 . Elle se fait au moyen des lignes de la matrice des composantes principales (XMV) :

$$C_L = AD_C^{-1}U.$$

Les composantes principales permettent de représenter les modalités des variables sur les axes 2 et 3 (le premier est constant égal à 1). Une proximité de deux points i et i' indique que la distribution de la seconde variable sachant que la première vaut i est similaire à celle sachant i'.

Pour les colonnes, on fait les mêmes calculs en inversant les lignes et les colonnes. Il s'agit donc de l'ACP des 'individus' modalités de X_2 ou profils colonne (la matrice des données est B), pondérés par les fréquences marginales des lignes de N (la matrice diagonale des poids est D_C) et utilisant la métrique du χ^2. Il s'agit donc de l'ACP de (B, D_C^{-1}, D_L).

Proposition 5 *Les éléments de l'ACP de* (B, D_L^{-1}, D_C) *sont fournis par l'analyse spectrale de la matrice carrée,* D_L^{-1} *?symétrique et semi ?définie positive* BA*.*

En notant V la matrice des vecteurs propres de la matrice BA ; les coordonnées permettant la représentation des modalités de la variable X_2 sont fournies par la matrice :

$$C_C = BD_L^{-1}V.$$

Sachant que V contient les vecteurs propres de BA et U ceux de AB, montre qu'il suffit de réaliser une seule analyse, car les résultats de l'autre s'en déduisent simplement :

$$U = A'V\Lambda^{-1/2},$$

$$V = B'U\Lambda^{-1/2};$$

Λ est la matrice diagonale des valeurs propres (exceptée $\lambda_0 = 0$) commune aux deux ACP.

$$C_C = BD_L^{-1}V = BD_L^{-1}B'V\Lambda^{-1/2} = D_C^{-1}A'B'U\Lambda^{-1/2} = D_C^{-1}U\Lambda^{1/2},$$

2. Matrice, Métrique, Pondération

$$C_L = AD_C^{-1}U = D_L^{-1}V\Lambda^{1/2}.$$

On en déduit les formules de transition

$$C_C = BC_L\Lambda^{-1/2}$$

$$C_L = AC_C\Lambda^{-1/2}$$

.

On est alors tenté de mettre toutes les modalités sur un même graphique (option par défaut dans SAS). La proximité de modalités de variables différentes reste néanmoins difficile à interpréter.

5.4 Représentation graphique

5.4.1 Biplot

La décomposition de la matrice $\frac{1}{n}\mathbf{N}$ se transforme encore en :

$$\frac{f_{ij} - f_{i\bullet}f_{\bullet j}}{f_{i\bullet}f_{\bullet j}} = \sum_{k=0}^{\min(p-1,q-1)} \sqrt{\lambda_k}\frac{v_{ik}}{f_{i\bullet}}\frac{u_{jk}}{f_{\bullet j}}$$

En se limitant au rang r, on obtient donc, pour chaque cellule (i, j) de la table $\mathbf{N}$, une approximtion de son écart relatif à l'indépendance comme produit scalaire des deux vecteurs

$$\frac{v_{ik}}{f_{i\bullet}}\lambda^{1/4} \text{ et } \frac{u_{jk}}{f_{\bullet j}}\lambda^{1/4}$$

termes génériques respectifs des matrices

$$D_L^{-1}V\Lambda^{1/4} \text{ et } D_C^{-1}U\Lambda^{1/4}$$

Leur représentation (par exemple avec $r = 2$) illustre alors la correspondance entre les deux modalités x_{1i} et x_{2j} : lorsque deux modalités, éloignées de l'origine, sont voisines (resp. opposées), leur produit scalaire est de valeur absolue importante ; leur cellule conjointe contribue alors fortement et de manière positive (resp. négative) à la dépendance entre les deux variables.

L'AFC apparaît ainsi comme la meilleure reconstitution des fréquences f_{ij}, ou encore la meilleure représentation des écarts relatifs à l'indépendance.

5.4.2 Représentation barycentrique

La représentation graphique usuelle dite *représentation quasi-barycentrique*, place les points $(c_L(1, i), c_L(2, i))$ et $(c_C(1, i), c_C(2, i))$.

$$C_L = D_L^{-1}V\Lambda^{1/2} \text{ et } C_C = D_C^{-1}\Lambda^{1/2}$$

Même si la représentation simultanée n'a plus alors de justification, elle reste couramment employée. En fait, les graphiques obtenus diffèrent très peu de ceux du biplot ; ce dernier sert donc de ?caution ? puisque les interprétations des graphiques sont identiques. On notera que cette représentation issue de la double ACP est celle réalisée par la plupart des logiciels statistiques (c'est en particulier le cas de SAS).

C'est cette représentation s'étend plus facilement au cas de plusieurs variables.

La *représentation barycentrique* est une autre représentation proposée par les logiciels. Elle utilise les matrices

$$D_L^{-1} V \Lambda^{1/2} \text{ et } D_C^{-1} U \Lambda$$

ou

$$D_L^{-1} V \Lambda \text{ et } D_C^{-1} U \Lambda^{1/2}.$$

Si l'on considère alors la formule de transition

$$C_L = A C_C \Lambda^{1/2} \Leftrightarrow C_L \Lambda^{1/2} = A C_C \Leftrightarrow D_L^{-1} V \Lambda = A D_C^{-1} U \Lambda^{1/2}$$

Dans cette représentation, chaque modalité j de la deuxième variable est représentée comme barycentre des modalités i de la première variable avec un poids qui est la probabilite de i sachant j.

La formule suivante

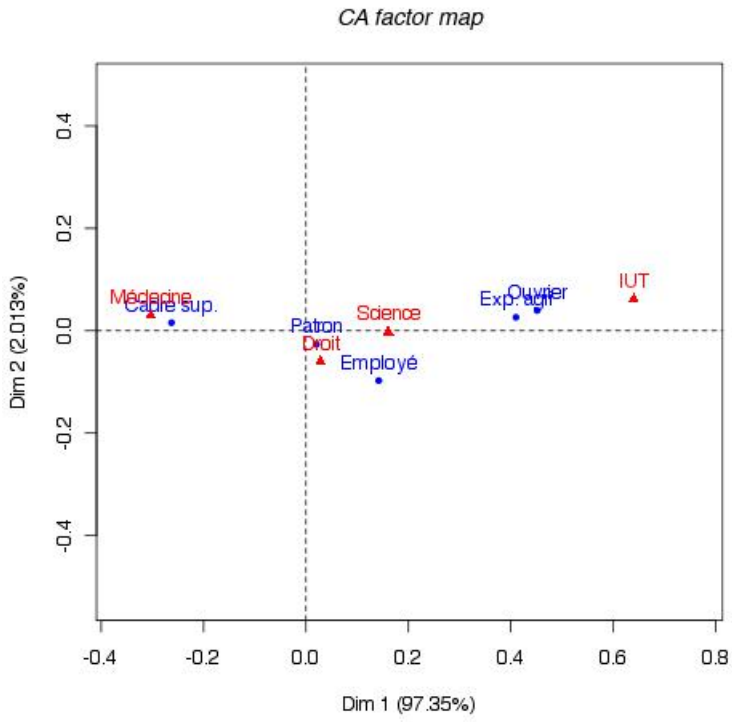
$$n_{ij} \simeq \frac{n_{i\bullet} n_{\bullet j}}{n} \left(1 + \frac{1}{\lambda_1} C_L(1,i) C_C(1,j) + \frac{1}{\lambda_2} C_L(2,i) C_C(2,j) \right)$$

indique que deux modalités formant un angle aigu (resp. obtus) s'attirent (resp. se repoussent) et ceci est d'autant plus marqué que les points sont éloignés du centre de gravité.

5.4.3 Exemples

Étudiants en première année

Dans l'exemple des étudiants en première année, on obtient le graphique suivant. On observe que toutes les modalités sont concentrées autour du premier axe. Ceci signifie qu'on a essentiellement une seule variable latente (ou facteur) structurante.

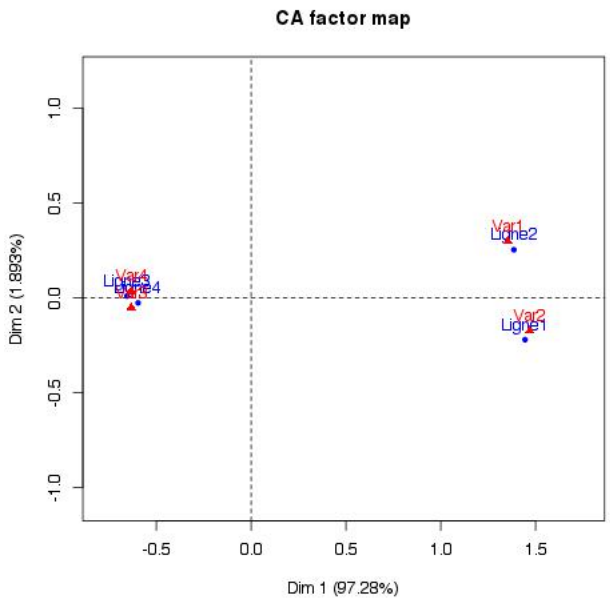


Sous groupes dans les données

Quand il existe des sous groupes dans les données, on obtient des résultats typiques. Par exemple, si on fait l'AFC du tableau suivant (tableau 1.)

	Var 1	Var 2	Var 3	Var 4
Ligne 1	20	45	2	0
Ligne 2	25	32	0	3
Ligne 3	1	0	78	112
Ligne 4	2	1	45	44

on obtient la projection ci-dessous sur le premier plan factoriel.

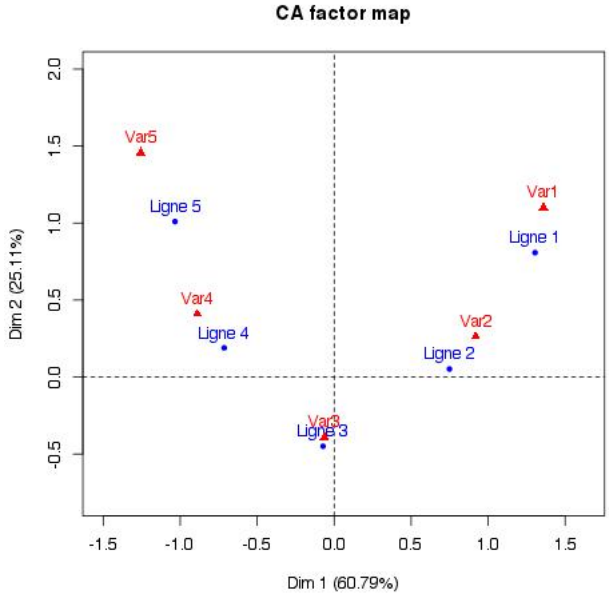


Effet Guttman

Un nuage de points de forme parabolique indique une redondance entre les deux variables étudiées : la connaissance de la ligne i donne pratiquement celle de la colonne j. Dans un tel cas, pratiquement toute l'information est contenue dans le premier facteur. Cette configuration se rencontre notamment lorsque les deux variables sont ordinales, et classent les sujets de la même façon. Dans ce cas, le premier axe oppose les valeurs extrêmes et classe les valeurs, tandis que le deuxième axe oppose les intermédiaires aux extrêmes.

	Var1	Var2	Var3	Var4	Var5
Ligne 1	10	30	7	0	0
Ligne 2	3	100	70	4	0
Ligne 3	2	32	200	35	1
Ligne 4	1	6	80	100	2
Ligne 5	0	3	5	25	5

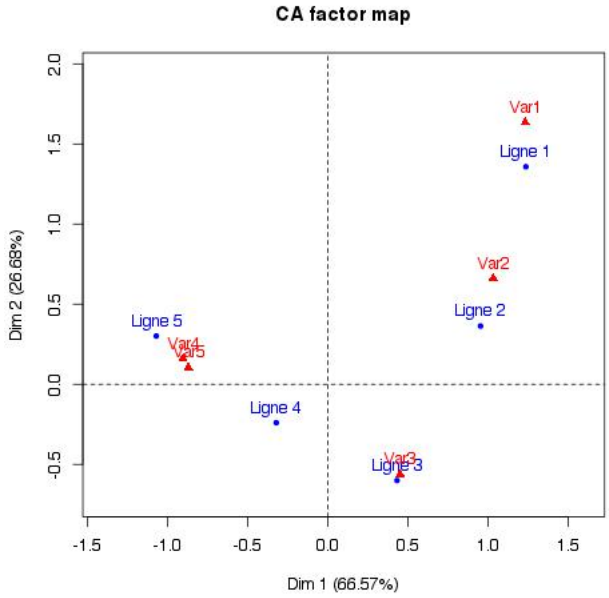
On obtient la projection ci-dessous sur le premier plan factoriel.



Exercice : Expliquer le graphique suivant associé au tableau ci-dessous en regard des résultats du jeux de données précédent.

	Var1	Var2	Var3	Var4	Var5
Ligne 1	10	30	7	0	0
Ligne 2	3	100	70	4	0
Ligne 3	2	32	200	35	1
Ligne 4	1	6	80	100	2
Ligne 5	0	3	5	250	5

On obtient la projection ci-dessous sur le premier plan factoriel.



5.5 Interprétation des résultats de l'AFC

5.5.1 Valeurs propres

On note tout d'abord que la la première valeur propre est une valeur propre triviale égale à 1. En général les logiciels l'ignorent.

On rappelle qu'on note

$$c_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}}.$$

On remarque alors que

$$d^2_{\chi^2}(\mathbf{x}, \mathbf{E}) = \sum_{i=1}^{p}\sum_{j=1}^{q} c_{ij}^2 = \mathrm{tr}(CC^T) = \sum_{k=1}^{\min(p-1,q-1)} \lambda_k$$

ce qui montre que la décomposition en valeurs singulières de C décompose le χ^2 total de même qu'en ACP on décompose l'inertie totale. La somme des valeurs propres non triviales multipliée par l'effectif total peut se comparer à un quantile de la loi du χ^2 à $(p-1)(q-1)$ degrés de liberté. La somme de toutes les valeurs propres est égale à l'inertie totale, c'est à dire à la distance $d^2(x, E)$. Elle donne donc une information sur l'écart à l'indépendance et on peut la comparer aux quantiles de la loi du χ^2.

Interprétation des valeurs propres -

- Si une valeur propre est proche de un, ça traduit le fait qu'il existe deux sous groupes de modalités dans les données. Il est alors intéressant de reconstruire la matrice N pour mettre en évidence ces deux sous groupes et de réaliser des AFC indépendamment sur les deux sous groupes.
  Par exemple l'analse factorielle des correspondances du tableau 1, renvoie les valeurs propres suivantes : 0.90, 0.01, 7e-3.
- De même, l'existence de deux valeurs propres proches de 1 indique une partition des observations en 3 groupes. Si toutes les valeurs propres sont proches de 1, cela indique une correspondance entre chaque modalité ligne et une modalité colonne "associée". Avec une réorganisation convenable des modalités, les effectifs importants se trouvent alors le long de la diagonale.

Choix de la dimension - Comme en ACP, les valeurs propres peuvent être interprétées comme la proportion d'inertie expliquée par le facteur correspondant. On peut s'en servir pour aider au choix de la dimension $r < min(1-p, 1-q)$ de l'espace de projection. En pratique, on utilise le fait que

$$K_r = \sum_{i=1}^{p}\sum_{j=1}^{q}(\frac{n_{ij} - \widehat{n_{ij}^r})^2}{\widehat{n_{ij}^r}} \simeq \sum_{k=r+1}^{\min(1-p,1-q)} \lambda_k$$

suit approximatiement une loi du χ^2 à $(p-r-1)(q-r-1)$ degrés de liberté. On peut donc retenir pour valeur de r la plus petite dimension pour laquelle K_r est inférieure à la valeur limite de cette loi. Le choix $r = 0$ correspond à la situation où les variables sont proches de l'indépendance en

probabilités ; les fréquences conjointes sont alors bien approchées par les produits des fréquences marginales.

Dans l'exemple des étudiants en première année, on obtient le tableau de valeurs propres suivant :

Valeurs propres	8.24e-02	1.70e-03	5.40e-04	1.52e-34
Proportions	0.973	0.02	0.00	0.00
Prop. cumulées	0.973	0.994	1.00	1.00

On en déduit que le premier plan fatoriel explique presque toute l'inertie de la table de contingence. C'est souvent le cas en AFC.

5.5.2 Contribution des modalités

Pour chaque modalité de X_1 (resp. de X_2), la qualité de sa représentation en dimension r se mesure par le cosinus carré de l'angle entre le vecteur représentant cette modalité dans $\mathbb{R}^p$ (resp. dans $\mathbb{R}^q$) et sa projection D_C^{-1} ?orthogonale (resp. D_L^{-1} ?orthogonale) dans le sous ?espace principal de dimension r. Ces cosinus carrés s'obtiennent en faisant le rapport des sommes appropriées des carrés des coordonnées extraites des lignes de C_L (resp. de C_C).

AUtrement dit, la "qualité" de la représentation d'une modalité contribution de la modalité i de la variable X sur l'axe k est donnée par le cosinus carré de l'angle formé avec l'axe.

$$\cos_k^2(i) = \frac{d_k^2(i,G)}{d^2(i,G)}$$

avec G le centre de gravité et $d^2(i,G) = \sum_k d_k^2(i,G)$

5.5.3 Interprétation en terme de reconstruction des effectifs

La décomposition de la matrice $\mathbf{N}$ est (formule $X = CU^T M^{-1}$))

$$\mathbf{N} = \frac{1}{n} D_L \left(1 + \sum_{k=2}^{r} c_k u_k^T\right) D_C$$

où 1 est la matrice de 1. La terme de première approximation, $n^{-1} D_L 1 D_C$, correspond à deux variables indépendantes. Si on approche par les trois premiers axes :

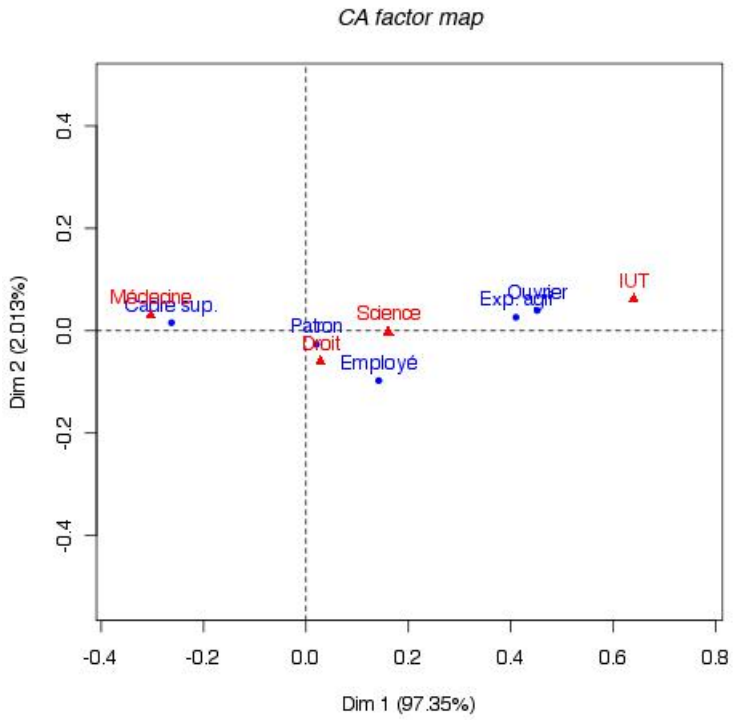
$$n_{ij} \simeq \frac{n_{i\bullet} n_{\bullet j}}{n} \left(1 + \frac{1}{\lambda_1} c_1(i) d_1(j) + \frac{1}{\lambda_2} c_2(i) d_2(j)\right) \tag{5.1}$$

5.6 Exemple

Dans l'exemple des étudiants de première année, on obtient le graphique ci-dessous. D'autre part, on obtient les contributions suivantes des modalités aux axes factoriels

En ligne		
	Dim 2	Dim 3
Exp. agri	16.29	3.22
Patron	0.07	6.30
Cadre sup.	40.40	6.89
Employé	3.02	68.63
Ouvrier	40.21	14.95

En colonne		
	Dim 2	Dim 3
Droit	0.26	58.76
Science	7.94	0.11
Médecine	41.58	19.26
IUT	50.21	21.87



Récapitulatif

Dans $\mathbb{R}^p$ (Lignes)		Dans $\mathbb{R}^q$ Colonnes)
$S = N^T D_L^{-1} N D_C^{-1}$ $Su_k = \lambda_k u_k$	Matrice à diagonaliser Axe facoriel	$T = N D_C^{-1} N^T D_L^{-1}$ $Tv_k = \lambda_k v_k$
$\psi_k = D_L^{-1} N D_C^{-1} u_k$ $\psi_{ki} = \sum_{j=1}^{p} \frac{n_{ij}}{n_{i\bullet} n_{\bullet j}} u_{ki}$	Coordonnées	$\phi_k = D_C^{-1} N^T D_L^{-1} v_k$ $\phi_{ki} = \sum_{i=1}^{q} \frac{n_{ij}}{n_{i\bullet} n_{\bullet j}} v_{ki}$