ANALYSE FACTORIELLE DES CORRESPONDANCES (AFC)

L'Analyse Factorielle des Correspondances (AFC)

A. Introduction & données
B. Objectifs de l'AFC
C. Tableaux de profils
D. Interprétation de l'AFC
E. Graphiques

A. Introduction

	none	light	medium	heavy
SM	4	2	3	2
JM	4	3	7	4
SE	25	10	12	4
JE	18	24	33	13
SC	10	6	7	2

Smoke dataset

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

Les signes de ponctuation chez Zola

Introduction

- Introduite par Guttman, 1941 & Benzécri, 1973, permet une visualisation en 2 dimensions des tableaux de contingence

- Généralisation de l'ACP appliquée aux données qualitatives

Données

Tableau de contingence

= Croisement de deux variables qualitatives X (à n modalités) et Y (à p modalités)

		Y				
		1	…	j	…	J
X	1	n_{11}		n_{1j}		n_{1J}
	:			:		
	i	n_{i1}	…	n_{ij}	…	n_{iJ}	$n_{i.}$
	:			:		
	I	n_{I1}		n_{Ij}		n_{IJ}
		$n_{.j}$				n

n_{ij} = Nombre d'observations ayant la modalité x_i de X et y_j de Y.
$n_{i.}$ = effectif marginal : Nombre d'observations ayant la modalité x_i de Y
$n.j$ = effectif marginal :Nombre d'observations ayant la modalité y_j de Y.

Exemple

CSP	Hotel	Location	Res.Second	Parents	Amis	Camping	Sej.org	Autres	Total
Agriculteurs	195	62	1	499	44	141	49	65	1056
Patrons	700	354	229	959	185	292	119	140	2978
Cadres.sup	961	471	633	1580	305	360	162	148	4620
Cadre.moy	572	537	279	1689	206	748	155	112	4298
Employes	441	404	166	1079	178	434	178	92	2972
Ouvriers	783	1114	387	4052	497	1464	525	387	9209
Autres.actifs	142	103	210	1133	132	181	46	59	2006
Inactifs	741	332	327	1789	311	236	102	102	3940
Total	4535	3377	2232	12780	1858	3856	1336	1105	31079

- X=CSP, I=8
- Y=hébergement, J=8

B. Objectif de l'AFC?

- Étudier les correspondances entre les modalités des deux variables qualitatives;
- Mettre en évidence des liaisons contenues dans un tableau de contingence;
- Résumer et représenter les principales liaisons pouvant exister entre les modalités de deux variables qualitatives.

écarts à l'indépendance, proximité des profils

➔➔ Réduction de la dimension en effectuant la décomposition factorielle des nuages de points associés aux profils lignes et aux profils colonnes du tableau de contingence croisant les modalités des deux variables

(L'AFC est une double ACP sur les deux tableaux de profils).

Le test du khi-deux d'indépendance

Test :

H_0 : Les variables X et Y sont indépendantes (pas de correspondance)

vs

H_1 : Les variables X et Y sont liées entre elles (liaison significative)

Sous H_0 , $\chi 2 \approx \chi 2 ((I-1)(J-1)$

Statistique utilisée :

n_{ij} = Effectif observé

$\frac{ni.n.j}{n}$ = Effectif attendu sous l'hypothèse d'indépendance

$\frac{nij - \frac{ni.n.j}{n}}{\sqrt{\frac{ni.n.j}{n}}}$ = Résidu standardisé (moyenne 0, écart-type 1)

$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(nij - \frac{ni.n.j}{n})^2}{\frac{ni.n.j}{n}}$

Khi-deux

- **L’indice du chi2 :** une mesure classique de la liaison entre deux variables qualitatives d’un tableau de contingence ;

- **Coefficient Beta** $(\beta) = \frac{\chi 2 - (I-1)(J-1)}{\sqrt{(I-1)(J-1)}}$
  - Si $\beta > 3$ ➔ liaison significative

Idée?

✓ *Mesure de la liaison entre X et Y*

- En probabilité, si il y a indépendance entre X et Y, on a:

(1) $\forall(i,j)\ P(X=i \text{ et } Y=j) = P(X=i)P(Y=j)$

(2) $\forall(i,j) P(X=i\ /Y=j) = P(X=i)$

(3) $\forall(i,j) P(Y=j\ /X=i) = P(Y=j)$

- En statistiques, ces relations équivalent à

(1) $\forall(i,j)\ f_{ij} = f_{i.} \times f_{.j} \quad (\Leftrightarrow f = p, \quad f = (f_{ij}); p = (f_{i.} \times f_{.j}))$

(2) $\forall(i,j)\ f_j^i = f_{i.} \quad (\Leftrightarrow Y^{(j)} = p_X, \quad p_X = (f_{i.}))$

(3) $\forall(i,j) f_i^j = f_{.j} \quad (\Leftrightarrow X^{(i)} = p_Y, \quad p_Y = (f_{.j}))$

- Conclusion : lorsque X et Y sont indépendants

  - Toutes les distributions conditionnelles de X/Y (profils colonnes) sont égales et égales à la loi marginale de X (profil moyen)

  - Toutes les distributions conditionnelles de Y/X (profils lignes) sont égales et égales à la loi marginale de Y (profil moyen)

  - La distribution jointe (fréquence) est égale au produit des marginales (fréquences marginales)

La mesure de la liaison entre X et Y se fait en évaluant un de ces écarts entre distributions

Tableaux utilisés

- Tableau de fréquence
- Profils-lignes
- Profils-colonnes

Fréquences

Fréquence de (i,j) : $f_{ij} = \frac{n_{ij}}{n_{i.}}$

Fréquence marginale de la modalité i : $f_{i.} = \frac{n_{i.}}{n}$

Fréquence marginale de la modalité j : $f_{.j} = \frac{n_{.j}}{n}$

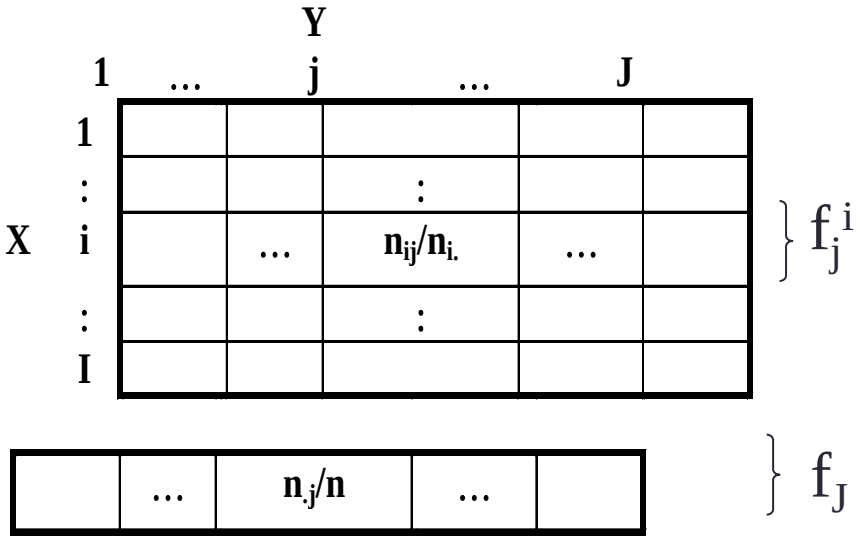
Fréquence conditionnelle j sachant i : $f_i^j = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}$

Fréquence conditionnelle i sachant j : $f_j^i = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}}$

$$\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j} = n \quad ; \quad \sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij} = \sum_{i=1}^{I} f_{i.} = \sum_{j=1}^{J} f_{.j} = \sum_{j=1}^{J} f_i^j = \sum_{i=1}^{I} f_j^i = 1$$

$$\sum_{j=1}^{J} n_{ij} = n_{i.} \; ; \; \sum_{i=1}^{I} n_{ij} = n_{.j} \; ; \; \sum_{j=1}^{J} f_{ij} = f_{i.} \; ; \; \sum_{i=1}^{I} f_{ij} = f_{.j}$$

Tableau profil-lignes

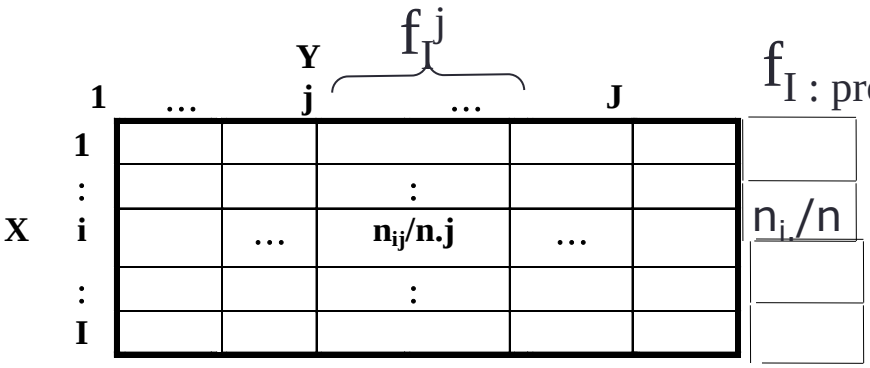


- On appelle i^o profil ligne, le vecteur de dimension J des fréquences de la variable Y conditionnellement à la valeur x_i de X;
- ➔ La répartition des valeurs de Y dans les différentes modalités de X.

CSP	Hotel	Location	Res.Second	Parents	Amis	Camping	Sej.org	Autres	Total
Agriculteurs	0,184659091	0,058712121	0,00094697	0,472537879	0,041666667	0,133522727	0,046401515	0,06155303	1
Patrons	0,235057085	0,118871726	0,076897246	0,322028207	0,06212223	0,098052384	0,039959704	0,047011417	1
Cadres.sup	0,208008658	0,101948052	0,137012987	0,341991342	0,066017316	0,077922078	0,035064935	0,032034632	1
Cadre.moy	0,133085156	0,124941833	0,064913913	0,392973476	0,047929269	0,174034435	0,036063285	0,026058632	1
Employes	0,148384926	0,135935397	0,055854643	0,363055182	0,059892328	0,14602961	0,059892328	0,030955585	1
Ouvriers	0,085025519	0,120968618	0,042024107	0,440004344	0,053968943	0,158974916	0,057009447	0,042024107	1
Autres.actifs	0,070787637	0,051345962	0,104685942	0,564805583	0,065802592	0,090229312	0,022931206	0,029411765	1
Inactifs	0,188071066	0,084263959	0,082994924	0,454060914	0,07893401	0,059898477	0,025888325	0,025888325	1
Total	1,253079138	0,796987669	0,565330733	3,351456926	0,476333356	0,938663939	0,323210747	0,294937493	8

- 23.5% des patrons vont à l'hôtel. Les patrons vont plus fréquemment à l'hôtel que dans les autres profession.. Dans toutes les professions on va majoritairement chez les parents.

Tableau profil-colonnes



- On appelle j° profil colonne, le vecteur de dimension I des fréquences de la variable X conditionnellement à la valeur j de Y;
- ➔ La répartition des valeurs de X dans les différentes modalités de Y.

CSP	Hotel	Location	Res.Second	Parents	Amis	Camping	Sej.org	Autres	Total
Agriculteurs	0,042998897	0,018359491	0,000448029	0,039045383	0,023681378	0,03656639	0,036676647	0,058823529	0,256599744
Patrons	0,154355017	0,104826769	0,102598566	0,075039124	0,099569429	0,075726141	0,089071856	0,126696833	0,827883735
Cadres.sup	0,211907387	0,139472905	0,283602151	0,123630673	0,164155005	0,093360996	0,121257485	0,133936652	1,271323253
Cadre.moy	0,126130099	0,159016879	0,125	0,132159624	0,110871905	0,193983402	0,116017964	0,101357466	1,06453734
Employes	0,09724366	0,11963281	0,07437276	0,084428795	0,095801938	0,112551867	0,133233533	0,083257919	0,800523282
Ouvriers	0,172657111	0,32987859	0,173387097	0,317057903	0,267491927	0,37966805	0,392964072	0,350226244	2,383330994
Autres.actifs	0,031312018	0,030500444	0,094086022	0,088654147	0,071044133	0,046939834	0,034431138	0,053393665	0,450361401
Inactifs	0,16339581	0,098312111	0,146505376	0,139984351	0,167384284	0,06120332	0,076347305	0,092307692	0,94544025
Total	1	1	1	1	1	1	1	1	8

- 15.4% des personnes allant à l'hôtel sont des patrons. Parmi les personnes allant à l'hôtel on trouve une majorité de cadres sup, bien que ces derniers aillent préférentiellement chez leurs parents (cf profils lignes).
- La part des cadres sup est plus importante parmi ceux-qui vont en résidence secondaires qu'ailleurs.

Interprétation

A. Indice d'attraction / répulsion:

$$dij = \frac{fij}{fi.f.j}$$

$$dij \in \begin{cases} > 1 : \ les\ modalités\ i\ et\ j\ s'attirent \\ = 1 : indépendance\ parfaite \\ < 1 : les\ modalités\ i\ et\ j\ se\ repoussent \end{cases}$$

Interprétation = Nuage de points

- Chaque profil ligne (point--ligne) représente un point dans l'espace de dimension J des profils--colonnes
- Chaque profil colonne (point--colonne) représente un point dans l'espace de dimension I des profils--colonnes
- Les tableaux des profils lignes et colonnes définissent chacun un nuage de points:
  - Le nuage de points--lignes **N(I)** est constitué des I points--lignes dans l'espace de dimension J des points--colonnes.
  - le nuage de points--colonnes **N(J)** est constitué des J points--colonnes dans l'espace de dimension I des points--lignes.

Notions

- Poids des points-ligne et points-colonne

✓ Chaque point-ligne $X^{(i)}$ est doté d'un poids relatant l'importance de la modalité i de X:

$$f_{i.} = \frac{n_{i.}}{n}$$

✓ Chaque point colonne est doté d'un poids relatant l'importance de la modalité j de Y:

$$f_{.j} = \frac{n_{.j}}{n}$$

- Centre de gravité

✓ Centre de gravité du nuage N(I) = distribution marginale de Y= p_Y

$$G_X = (g_{X1}, \ldots, g_{XJ})$$
$$g_{Xj} = \sum_{i=1}^{I} f_{i.} f_i^j = \sum_{i=1}^{k} \frac{n_{i.}}{n} \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n} = f_{.j}$$

✓ Centre de gravité du nuage N(J) : distribution marginale de X= p_X

$$G_Y = (g_{Y1}, \ldots, g_{YI})$$
$$g_{Yi} = \sum_{j=1}^{J} f_{.j} f_j^i = \sum_{j=1}^{J} \frac{n_{.j}}{n} \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n} = f_{i.}$$

- Inertie
  - ✓ Distance entre points-lignes : les points lignes étant des distributions, on utilise la métrique du chi2 centrée sur la distribution moyenne GX (=pY)

$$\chi^2_{G_x}(X^{(i)}, X^{(i')}) = \sum_{j=1}^{J} \frac{\left(f_i^j - f_{i'}^j\right)^2}{f_{.j}}$$

La distance est d'autant plus grande que i et i' sont réparties de façon différente dans les modalités de Y

Ex : La distance entre deux CSP est d'autant plus grande que ces deux CSP sont réparties de façon différentes dans les lieux de vacances (que les structures diffèrent).

- ✓ Inertie des nuage N(I) et N(J) :

$$I_X = \sum_{i=1}^{I} f_{i.} \chi^2_{G_X}(X^{(i)}, G_X) = \sum_j \sum_i f_{i.} \frac{\left(f_i^j - f_{.j}\right)^2}{f_{.j}} = \sum_i \sum_j \frac{\left(f_{ij} - f_{i.} f_{.j}\right)^2}{f_{i.} f_{.j}} = \frac{\chi^2}{n}$$

$$I_Y = \sum_{j=1}^{J} f_{.j} \chi^2_{G_Y}(Y^{(j)}, G_Y) = \sum_j \sum_i f_{.j} \frac{\left(f_j^i - f_{i.}\right)^2}{f_{i.}} = \sum_i \sum_j \frac{\left(f_{ij} - f_{i.} f_{.j}\right)^2}{f_{i.} f_{.j}} = \frac{\chi^2}{n}$$

- L'inertie est nulle lorsque tous les profils-lignes (resp. colonnes) sont égaux au centre de gravité ⇔ lorsque toutes les distributions de Y sachant X=i (resp. X sachant Y=j) sont égales et égales à la distribution marginale de Y (resp. de X) ⇔ lorsque X et Y sont indépendantes
- Au plus la dépendance est forte, au plus l'inertie est grande

AFC : décomposition factorielle

- Les objectifs de l'analyse factorielle des correspondances (AFC) :
  - ✓ comparer les profils-lignes entre eux (les distributions de Y dans les différentes modalités de X),
  - ✓ comparer les profils-colonnes entre eux (les distributions de X dans les différentes modalités de Y),
  - ✓ Repérer les cases du tableau de contingence où les effectifs observés nij sont nettement différents des effectifs théoriques (effectifs sous l'hypothèse d'indépendance) pour mettre en évidence les modalités I de X et j de Y qui s'attirent (fij > pij) et celles qui se repoussent (fij < pij)

- L'AFC est une ACP sur le tableau de contingence constitué des deux variables X et Y, en utilisant la métrique du chi2.

De façon équivalente, elle consiste en la décomposition factorielle des nuages de points N(I) (analyse directe) et N(J) (analyse duale)

➔ Il faut chercher les axes orthogonaux passant le mieux possible par le milieu du nuage de points N(I).

- Chaque axe factoriel supporte une part de l'inertie totale. Cette part est mesurée par les valeurs propres, inférieures ou égales à 1.

- Des valeurs proches de 1 indiquent d'intéressants liens entre modalités de variables différentes.

- Nombre axe k <= min(J-1, I-1)