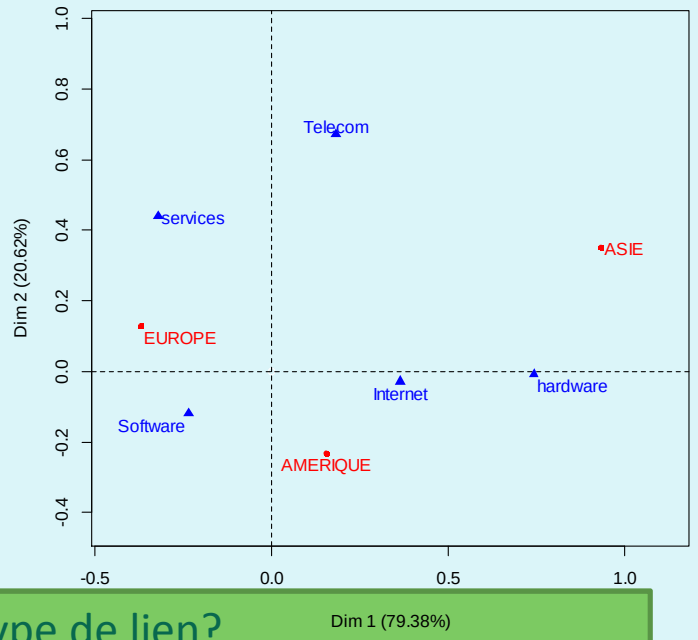
Analyse Factorielle des correspondances (A.F.C.)

- ***L' AFC est une méthode descriptive permettant de représenter graphiquement l'essentielle de l'information contenu dans le tableau des données qualitatives.***
- ***Dans ce cours, nous traitons de l'AFC simple (2 variables qualitatives).***
- ***Les données sont répertoriées dans un tableau de contingence***

Exemple : Etude des 275 plus importantes industries du numériques réparties selon le continent et le secteur d'activité

Classement effectué par Eurostat, basé sur le capital en 2005

	Hardware	Internet	Services	Software	Mobile Telecom
AMERIQUE	33	5	6	73	1
ASIE	18	2	4	5	2
EUROPE	9	3	25	85	4



- Y-a-t'il un lien entre ces deux variables? Si oui, quel type de lien?
- Y-a-t'il des modalités que se ressemblent, qui s'opposent?
- Comment quantifier l'information contenue dans ce tableau? Quelle métrique?

Comment évaluer les liaisons entre variables?

- Comment quantifier la liaison entre deux variables qualitatives?
- Comment interpréter l'écart d'une modalité à l'indépendance?

On rappelle que dans le cas où deux évènements A et B sont indépendants, alors $P(A \cap B)=P(A)\times P(B)$. *Ce qui se traduit en terme de fréquences par* $f_{ij}=f_{i.} \times f_{.j}$. *D'où les tableaux*

	Hardware	Internet	Services	Software	Mobile Telecom	Freq. Marg.
AMERIQUE	12,00%	1,82%	2,18%	26,55%	0,36%	42,91%
ASIE	6,55%	0,73%	1,45%	1,82%	0,73%	11,27%
EUROPE	3,27%	1,09%	9,09%	30,91%	1,45%	45,82%
Freq. Marg.	21,82%	3,64%	12,73%	59,27%	2,55%	100,00%

Fréquences observées

	Hardware	Internet	Services	Software	Mobile Telecom	Freq. Marg.
AMERIQUE	9,36%	1,56%	5,46%	25,43%	1,09%	42,91%
ASIE	2,46%	0,41%	1,43%	6,68%	0,29%	11,27%
EUROPE	10,00%	1,67%	5,83%	27,16%	1,17%	45,82%
Freq. Marg.	21,82%	3,64%	12,73%	59,27%	2,55%	100,00%

Fréquences théoriques

On s'intéresse à l'écart entre ces deux tableaux

Comment évaluer les liaisons entre variables?

Tableau des écarts :

$$e_{ij} = \frac{f_{ij} - f_{i\bullet} \times f_{\bullet j}}{f_{i\bullet} \times f_{\bullet j}}$$

	Hardware	Internet	Services	Software	Mobile Telecom
AMERIQUE	0,28	0,17	-0,60	0,04	-0,67
ASIE	1,66	0,77	0,01	-0,73	1,53
EUROPE	-0,67	-0,35	0,56	0,14	0,25

Cela signifie par exemple qu'il y a 60% d'entreprises américaines dans les services en moins qu'il ne devrait y en avoir dans le cas de l'indépendance. A contrario, il y a 56% d'entreprises européennes dans les services en plus que le nombre théorique.

Remarque :

- *Les écarts positifs prennent des valeurs quelconques (même supérieure à 100%).*
- *Les écarts négatifs sont entre -1 et 0 (déficit maximum de 100%)*

L'information est donc contenue dans l'écart à l'indépendance. On mesure l'inertie du tableau à l'aide du chi-deux

$$\chi^2 = n \times \sum_i \sum_j \frac{\left(f_{ij} - f_{i\bullet} \times f_{\bullet j}\right)^2}{f_{i\bullet} \times f_{\bullet j}}$$

L'AFC simple s'intéresse plus particulièrement à la structure de ces écarts.

Dans l'exemple ci-dessus on χ^2=59.6 et la p-valeur du test d'indépendance p=5.588e-10 montre une forte dépendance entre les deux variables

Comment représenter la structure des écarts ?

Comment représenter la structure des écarts à partir des profils lignes?

	Hardware	Internet	Services	Software	Mobile Telecom	
AMERIQUE	27,97%	4,24%	5,08%	61,86%	0,85%	100%
ASIE	58,06%	6,45%	12,90%	16,13%	6,45%	100%
EUROPE	7,14%	2,38%	19,84%	67,46%	3,17%	100%
FREQ. MARG.	**21,82%**	**3,64%**	**12,73%**	**59,27%**	**2,55%**	100%

Profils lignes

Pourcentages par ligne :

$$f_{j|i} = \frac{n_{ij}}{n_{i\bullet}}$$

où i est la ligne et j la colonne

On considère chaque ligne (continent) comme un individu dans un espace de dimension 5 (le secteur d'activité). Il s'agit alors d'analyser un nuage de points de dimension 5, ce qui n'est pas sans rappeler le problème de l'ACP.

La ''distance'' entre deux individus-lignes est définie par,

$$d^2(L_{i_1}, L_{i_2}) = \sum_j \frac{\left(f_{j|i_1} - f_{j|i_2}\right)^2}{f_{\bullet j}}$$

Remarque : Diviser par la fréquence marginale permet d'attribuer le même poids à chaque colonne. Sinon Software aurait un poids trop important dans le calcul comparé à Internet

On s'intéresse plus particulièrement à la ''distance'' entre un individu-ligne est le profil ligne moyen (fréquences marginales lignes).

Dans l'exemple ci-dessus, on a: $d^2(\text{AMERIQUE}, \text{ASIE}) = \frac{(0,28-0,58)^2}{0,22} + \frac{(0,042-,065)^2}{0,036} + \ldots = 0,953$

$d^2(\text{AMERIQUE}, \text{Moyenne}) = \frac{(0,28-0,22)^2}{0,22} + \frac{(0,042-,036)^2}{0,036} + \ldots = 0,077$

$d^2(\text{ASIE}, \text{Moyenne}) = \frac{(0,58-0,22)^2}{0,22} + \frac{(0,064-,036)^2}{0,036} + \ldots = 0,998$

On constate que l'Asie s'éloigne beaucoup plus du profil moyen que l'Amérique

Comment représenter la structure des écarts à partir des profils colonnes?

	Hardware	Internet	Services	Software	Mobile Telecom	Freq. Marg.
AMERIQUE	55,00%	50,00%	17,14%	44,79%	14,29%	**42,91%**
ASIE	30,00%	20,00%	11,43%	3,07%	28,57%	**11,27%**
EUROPE	15,00%	30,00%	71,43%	52,15%	57,14%	**45,82%**
	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Profils colonnes

Pourcentages par colonne :

$$f_{i|j} = \frac{n_{ij}}{n_{\bullet j}}$$

où i est la ligne et j la colonne

On peut procéder de la même façon avec les profiles colonnes. On considère chaque colonne (secteur d'activité) comme un individu dans un espace de dimension 3 (le continent).

Dans l'exemple ci-dessus, on a: $d^2(\text{Harware},\text{Internet}) = \frac{(0,55-0,50)^2}{0,43} + \frac{(0,3-0,2)^2}{0,11} + \ldots = 0,14$

$$d^2(\text{Hardware},\text{Moyenne}) = \frac{(0,55-0,43)^2}{0,43} + \frac{(0,30-0,11)^2}{0,11} + \ldots = 0,55$$

$$d^2(\text{Internet},\text{Moyenne}) = \frac{(0,50-0,43)^2}{0,43} + \frac{(0,20-0,11)^2}{0,11} + \ldots = 0,13$$

On constate que le Hardware est plus éloigné du profil colonne moyen que l'Internet

Remarque : il n'existe pas de métrique mesurant la ''distance'' entre une ligne et une colonne

Quel lien avec l'A.C.P.?

L'idée est toujours de rechercher les directions de plus grandes dispersion de ces nuages de points. La matrice dont on cherche les valeurs propres ici n'a pas de signification particulière comme en ACP. Il faut tout d'abord construire le tableau des

$$\frac{f_{ij}}{\sqrt{f_{i\bullet} \times f_{\bullet j}}}$$

La matrice en question est alors constituée des produits scalaires des lignes (ou colonnes) entre elles. On peut montrer que cela revient au même de travailler sur le nuage de points des profils lignes ou bien sur celui des profils colonnes.

	AMERIQUE	ASIE	EUROPE
AMERIQUE	0,46	0,23	0,40
ASIE	0,23	0,23	0,16
EUROPE	0,40	0,16	0,53

Valeurs propres : 1.0; 0.172; 0.045

$\chi^2/n=59.6/275=0.22=0.172 + 0.045$

	Hardware	Internet	Services	Software	Mob. Tel.
Hardware	0,34	0,11	0,13	0,30	0,08
Internet	0,11	0,04	0,06	0,13	0,03
Services	0,13	0,06	0,17	0,28	0,07
Software	0,30	0,13	0,28	0,63	0,11
Mob.Tel.	0,08	0,03	0,07	0,11	0,04

Valeurs propres : 1.0; 0.172; 0.045; $6.\ 2\times10^{-10}$; $-\ 1.\ 1\times10^{-10}$

Le nombre de valeurs propres de la méthode est le minimum entre le nombre de lignes et le nombre de colonnes. La première valeur propre est toujours égale à 1. La somme des autres est égale à χ^2/n. Comme pour l'ACP, les logiciels retournent un tableau de pourcentages et pourcentages cumulés mais il ne s'agit plus ici de variance expliquée mais de chi-deux expliqué.

```
> res$eig
      eigenvalue percentage cumulative percentage
                    of variance     of variance

dim 1 0.17203917   79.38283        79.38283
dim 2 0.04468172   20.61717        100.00000
```

Les nouveaux axes s'appellent les axes factoriels. Le choix du nombre d'axes à retenir se fait comme pour l'ACP

Interprétation graphique

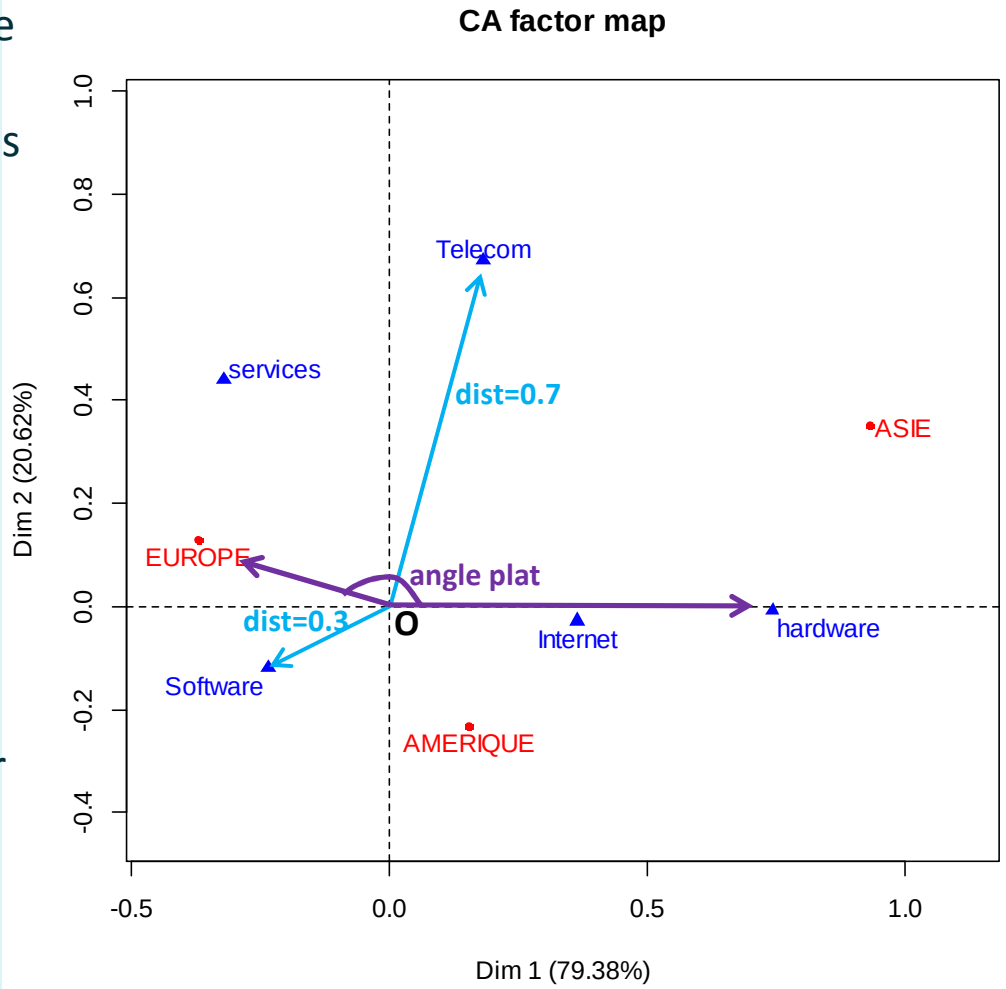
- Plus la distance euclidienne entre un individu et l'origine du graphique est grande, plus l'individu est éloigné du profil moyen (freq. marginales)

	Software	Mobile Telecom	Freq. Marg.
AMERIQUE	44,79%	14,29%	**42,91%**
ASIE	3,07%	28,57%	**11,27%**
EUROPE	52,15%	57,14%	**45,82%**

Remarque : La distance entre un individu ligne et un individu colonne ne s'interprète pas

- L'angle de sommet O de cotés passant par deux modalités s'interprète de la façon suivante :
  - ✓Angle aigu: les modalités s'attirent
  - ✓Angle plat: les modalités se repoussent
  - ✓ Angle droit: les modalités n'interagissent pas.

Remarque : Les deux modalités ne sont pas nécessairement de la même variable



Présentation d'un cas

On a interrogé 10005 étudiants à qui on a demandé le type d'études qu'ils poursuivent et la catégorie socio-professionnelle (CSP) de leur chef de famille.

On cherche à savoir

- s'il y a un lien entre le type d'étude et la CSP.
- si la réponse est oui alors quelle est la nature de ce lien

Les données du cas

	Tableau d'effectifs CSP x Type d'études poursuivies								
	Droit	Sci Eco	Lettres	Sciences	Méd. & Dent.	Pharma.	Plur. Disc.	IUT	Totaux
Exp. Agr.	80	36	134	99	65	28	11	58	511
Sal. Agr.	6	2	15	6	4	1	1	4	39
Patron	168	74	312	137	208	53	21	62	1035
Pro. Lib.	470	191	806	400	876	164	45	79	3031
Cad. Moy.	236	99	493	264	281	56	36	87	1552
Employé	145	52	281	133	135	30	20	54	850
Ouvrier	166	64	401	193	127	23	28	129	1131
Pers. Serv.	16	6	27	11	8	2	2	8	80
Autres.	305	115	624	247	301	47	47	90	1776
Totaux	1592	639	3093	1490	2005	404	211	571	10005

Tableau de contingence

Tableau d'effectifs CSP x Type d'études poursuivies									
	Droit	Sci Eco	Lettres	Sciences	Méd. & Dent.	Pharma.	Plur. Disc.	IUT	Totaux
Exp. Agr.	80	36	134	99	65	28	11	58	511
Sal. Agr.	6	2	15	6	4	1	1	4	39
Patron	168	74	312	137	208	53	21	62	1035
Pro. Lib.	470	191	806	400	876	164	45	79	3031
Cad. Moy.	236	99	493	264	281	56	36	87	1552
Employé	145	52	281	133	135	30	20	54	850
Ouvrier	166	64	401	193	127	23	28	129	1131
Pers. Serv.	16	6	27	11	8	2	2	8	80
Autres.	305	115	624	247	301	47	47	90	1776
Totaux	1592	639	3093	1490	2005	404	211	571	10005

Table de contingence									
%	Droit	Sci Eco	Lettres	Sciences	Méd# & Dent#	Pharma#	Plur# Disc#	IUT	Totaux
Exp. Agr.	0,80	0,36	1,34	0,99	0,65	0,28	0,11	0,58	5,11
Sal. Agr.	0,06	0,02	0,15	0,06	0,04	0,01	0,01	0,04	0,39
Patron	1,68	0,74	3,12	1,37	2,08	0,53	0,21	0,62	10,35
Pro. Lib.	4,70	1,91	8,06	4,00	8,76	1,64	0,45	0,79	30,30
Cad. Moy.	2,36	0,99	4,93	2,64	2,81	0,56	0,36	0,87	15,51
Employé	1,45	0,52	2,81	1,33	1,35	0,30	0,20	0,54	8,50
Ouvrier	1,66	0,64	4,01	1,93	1,27	0,23	0,28	1,29	11,30
Pers. Serv.	0,16	0,06	0,27	0,11	0,08	0,02	0,02	0,08	0,80
Autres.	3,05	1,15	6,24	2,47	3,01	0,47	0,47	0,90	17,75
Totaux	15,91	6,39	30,92	14,89	20,04	4,04	2,11	5,71	100,00

Profils lignes

Profils des lignes									
%	Droit	Sci Eco	Lettres	Sciences	Méd# & Dent#	Pharma#	Plur# Disc#	IUT	Total
Exp. Agr.	15,66	7,05	26,22	19,37	12,72	5,48	2,15	11,35	100,00
Sal. Agr.	15,38	5,13	38,46	15,38	10,26	2,56	2,56	10,26	100,00
Patron	16,23	7,15	30,14	13,24	20,10	5,12	2,03	5,99	100,00
Pro. Lib.	15,51	6,30	26,59	13,20	28,90	5,41	1,48	2,61	100,00
Cad. Moy.	15,21	6,38	31,77	17,01	18,11	3,61	2,32	5,61	100,00
Employé	17,06	6,12	33,06	15,65	15,88	3,53	2,35	6,35	100,00
Ouvrier	14,68	5,66	35,46	17,06	11,23	2,03	2,48	11,41	100,00
Pers. Serv.	20,00	7,50	33,75	13,75	10,00	2,50	2,50	10,00	100,00
Autres.	17,17	6,48	35,14	13,91	16,95	2,65	2,65	5,07	100,00
Profil moyen	15,91	6,39	30,92	14,89	20,04	4,04	2,11	5,71	100,00

Contributions à la statistique du khi-2 totale									
%	Droit	Sci Eco	Lettres	Sciences	Méd# & Dent#	Pharma#	Plur# Disc#	IUT	Somme
Exp, Agr,	0,00	0,07	0,76	1,45	2,87	0,55	0,00	5,98	11,69
Sal, Agr,	0,00	0,02	0,15	0,00	0,39	0,04	0,01	0,30	0,91
Patron	0,01	0,20	0,04	0,40	0,00	0,63	0,01	0,03	1,32
Pro, Lib,	0,07	0,01	3,84	1,23	24,92	2,97	1,18	10,71	44,92
Cad, Moy,	0,10	0,00	0,08	0,98	0,61	0,15	0,07	0,01	1,99
Employé	0,15	0,02	0,27	0,07	1,54	0,11	0,05	0,13	2,33
Ouvrier	0,23	0,20	1,58	0,75	9,19	2,36	0,15	13,50	27,97
Pers, Serv,	0,18	0,03	0,04	0,02	0,84	0,10	0,01	0,54	1,76
Autres,	0,37	0,01	2,15	0,24	1,78	1,79	0,51	0,27	7,11
Somme	1,11	0,55	8,91	5,13	42,13	8,70	1,98	31,47	100,00

Profils colonnes

Profils de colonnes									
%	Droit	Sci Eco	Lettres	Sciences	Méd# & Dent#	Pharma#	Plur# Disc#	IUT	Profil moyen
Exp, Agr,	5,03	5,63	4,33	6,64	3,24	6,93	5,21	10,16	5,11
Sal, Agr,	0,38	0,31	0,49	0,40	0,20	0,25	0,47	0,70	0,39
Patron	10,55	11,58	10,09	9,19	10,37	13,12	9,95	10,86	10,35
Pro, Lib,	29,52	29,89	26,06	26,85	43,69	40,59	21,33	13,84	30,30
Cad, Moy,	14,82	15,49	15,94	17,72	14,02	13,86	17,06	15,24	15,51
Employé	9,11	8,14	9,09	8,93	6,73	7,43	9,48	9,46	8,50
Ouvrier	10,43	10,02	12,96	12,95	6,33	5,69	13,27	22,59	11,30
Pers, Serv,	1,01	0,94	0,87	0,74	0,40	0,50	0,95	1,40	0,80
Autres,	19,16	18,00	20,17	16,58	15,01	11,63	22,27	15,76	17,75
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

Contributions à la statistique du khi-2 totale									
%	Droit	Sci Eco	Lettres	Sciences	Méd# & Dent#	Pharma#	Plur# Disc#	IUT	Somme
Exp, Agr,	0,00	0,07	0,76	1,45	2,87	0,55	0,00	5,98	11,69
Sal, Agr,	0,00	0,02	0,15	0,00	0,39	0,04	0,01	0,30	0,91
Patron	0,01	0,20	0,04	0,40	0,00	0,63	0,01	0,03	1,32
Pro, Lib,	0,07	0,01	3,84	1,23	24,92	2,97	1,18	10,71	44,92
Cad, Moy,	0,10	0,00	0,08	0,98	0,61	0,15	0,07	0,01	1,99
Employé	0,15	0,02	0,27	0,07	1,54	0,11	0,05	0,13	2,33
Ouvrier	0,23	0,20	1,58	0,75	9,19	2,36	0,15	13,50	27,97
Pers, Serv,	0,18	0,03	0,04	0,02	0,84	0,10	0,01	0,54	1,76
Autres,	0,37	0,01	2,15	0,24	1,78	1,79	0,51	0,27	7,11
Somme	1,11	0,55	8,91	5,13	42,13	8,70	1,98	31,47	100,00

Sur les axes

Décomposition de l'inertie et du Khi-2				
Inertie principale (valeurs propres)	**Khi-2**	**Pourcentage**	**Pourcentage cumulé**	17 34 51 68 85 ----+-------+-------+--------+----------+---
0,03978	398,009	83,5	83,5	**********
0,00557	55,698	11,69	95,19	***
0,00114	11,43	2,4	97,59	*
0,00096	9,566	2,01	99,59	*
0,00013	1,292	0,27	99,86	
0,00006	0,577	0,12	99,98	
0,00001	0,072	0,02	100	
0,04764	476,644	100		

Il y a 9 CSP et 8 types d'études poursuivies. Degrés de liberté = 56 = (9-1).(8-1).
Le seuil à 5% d'un chi-deux à 56 ddl est de 74,5. La valeur calculée 476,644 est bien supérieure à ce seuil. CSP et Type d'études sont assez liées

$$\sum_{i=1}^{9}\sum_{j=1}^{8}\left(\frac{\left(n_{i,j-}\frac{n_{i,.}.n_{.,j}}{n}\right)^2}{\frac{n_{i,.}.n_{.,j}}{n}}\right)=476{,}644$$

Contributions et Qualité de représentation

Contributions partielles à l'inertie des profils lignes			
	Dim1	Dim2	Dim3
Exp. Agr.	0,0691	0,4709	0,1235
Sal. Agr.	0,0102	0,0007	0,0024
Patron	0,0011	0,0167	0,2467
Pro. Lib.	0,5309	0,0316	0,0694
Cad. Moy.	0,0094	0,0036	0,0248
Employé	0,0227	0,0161	0,0429
Ouvrier	0,3176	0,0213	0,3758
Pers. Serv.	0,017	0,0005	0,065
Autres.	0,022	0,4387	0,0496
Totaux	1,00	1,00	1,00

Carré des cosinus des profils lignes avec les axes				
	Dim1	Dim2	Dim3	Totaux
Exp, Agr,	0,4935	0,4708	0,0253	0,9896
Sal, Agr,	0,9309	0,0083	0,0062	0,9454
Patron	0,0671	0,1474	0,4478	0,6623
Pro, Lib,	0,987	0,0082	0,0037	0,9989
Cad, Moy,	0,3952	0,0212	0,0298	0,4462
Employé	0,8121	0,0807	0,044	0,9368
Ouvrier	0,9483	0,0089	0,0322	0,9894
Pers, Serv,	0,8045	0,0036	0,0884	0,8965
Autres,	0,259	0,7212	0,0167	0,9969

Contributions partielles à l'inertie des profils colonnes			
	Dim1	Dim2	Dim3
Droit	0,0001	0,0296	0,2471
Sci Eco	0,0002	0,0016	0,176
Lettres	0,0635	0,2954	0,0238
Sciences	0,0339	0,0706	0,0587
Méd# & Dent#	0,4959	0,0144	0,1912
Pharma#	0,0634	0,2353	0,2451
Plur# Disc#	0,0175	0,0358	0,0177
IUT	0,3256	0,3173	0,0404
Totaux	1,00	1,00	1,00

Carré des cosinus pour des profils colonnes				
	Dim1	Dim2	Dim3	Totaux
Droit	0,00	0,31	0,53	0,85
Sci Eco	0,03	0,03	0,76	0,83
Lettres	0,59	0,39	0,01	0,99
Sciences	0,55	0,16	0,03	0,74
Méd# & Dent#	0,98	0,00	0,01	1,00
Pharma#	0,61	0,32	0,07	0,99
Plur# Disc#	0,74	0,21	0,02	0,97
IUT	0,8639	0,1178	0,0031	0,9848

Mappings des profils sur les axes