Data exploration

Analyse factorielle des correspondances

L'analyse factorielle des correspondances (AFC) fait suite au chapitre sur le croisement de 2 variables qualitatives. Si suite à cette études les deux variables sont considérées comme dépendantes, alors on est en droit de se demander comment?

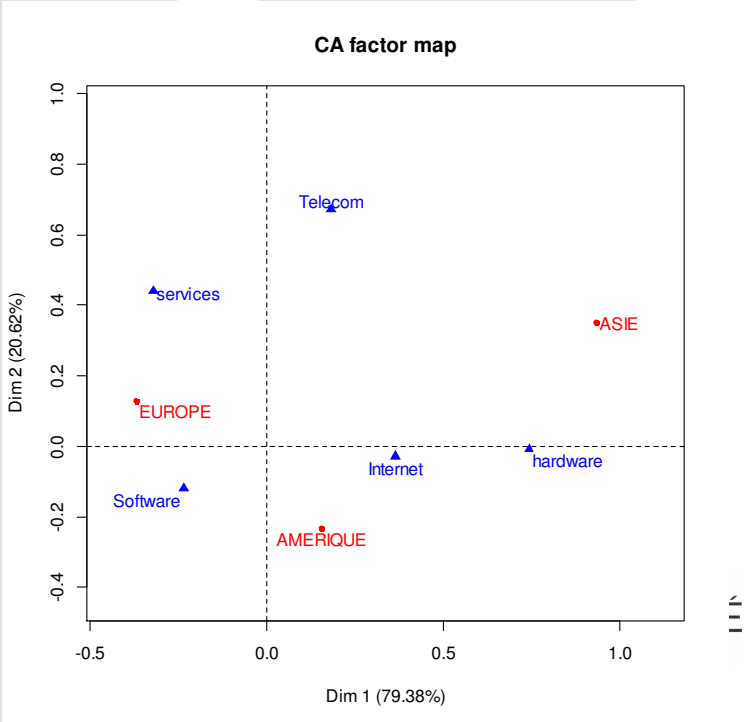
Y-a-il des modalités qui s'attirent ou se repoussent?

L' AFC est une méthode descriptive permettant de représenter graphiquement l'essentielle de l'information contenu dans le 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	26	4	15	70	3
ASIE	7	1	4	18	1
EUROPE	27	5	16	75	3



Analyse factorielle des correspondances

Distribution conjointe

Comme nous l'avons vu dans le chapitre précédent la dépendance entre deux variables qualitatives se mesure selon trois comparaisons de tableaux

- Le tableau des effectifs théoriques avec le tableau des effectifs observés
- Le tableau des profils lignes avec le profil moyen des lignes
- Le tableau des profils colonnes avec le profil moyen des colonnes

L'AFC va s'intéresser à la structure des écarts qui apparaissent entre ces tableaux

Analyse factorielle des correspondances

Tableaux des effectifs

On rappelle le calcul du tableau des effectifs théoriques,

$$f_{ij}=f_{i.} \times f_{.j}$$

	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

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

	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

Analyse factorielle des correspondances

Tableaux des effectifs

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 le seuil pour (2-1)*(5-1)=8 d.d.l. est 15,51. Il y a donc une forte dépendance entre les deux variables

Comment représenter la structure des écarts ?

Analyse factorielle des correspondances

Ecarts 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

Analyse factorielle des correspondances

Ecarts 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

Analyse factorielle des correspondances

Lien avec l'ACP

L'idée est toujours de rechercher les directions de plus grandes dispersion des ces nuages de points.

La matrice dont on cherche les valeurs propres ici n'a pas de signification particulière comme en ACP. Elle s'appelle la matrice de Burt et représente les produits scalaires entre les profils lignes et les profils colonnes. L'objectif est de choisir la combinaison avec de dimensions la moins élevée possible, afin de minimiser la perte d'information lors de la projection des données dans un espace de dimension 2.

	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; 6.2×10^{-10}; -1.1×10^{-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

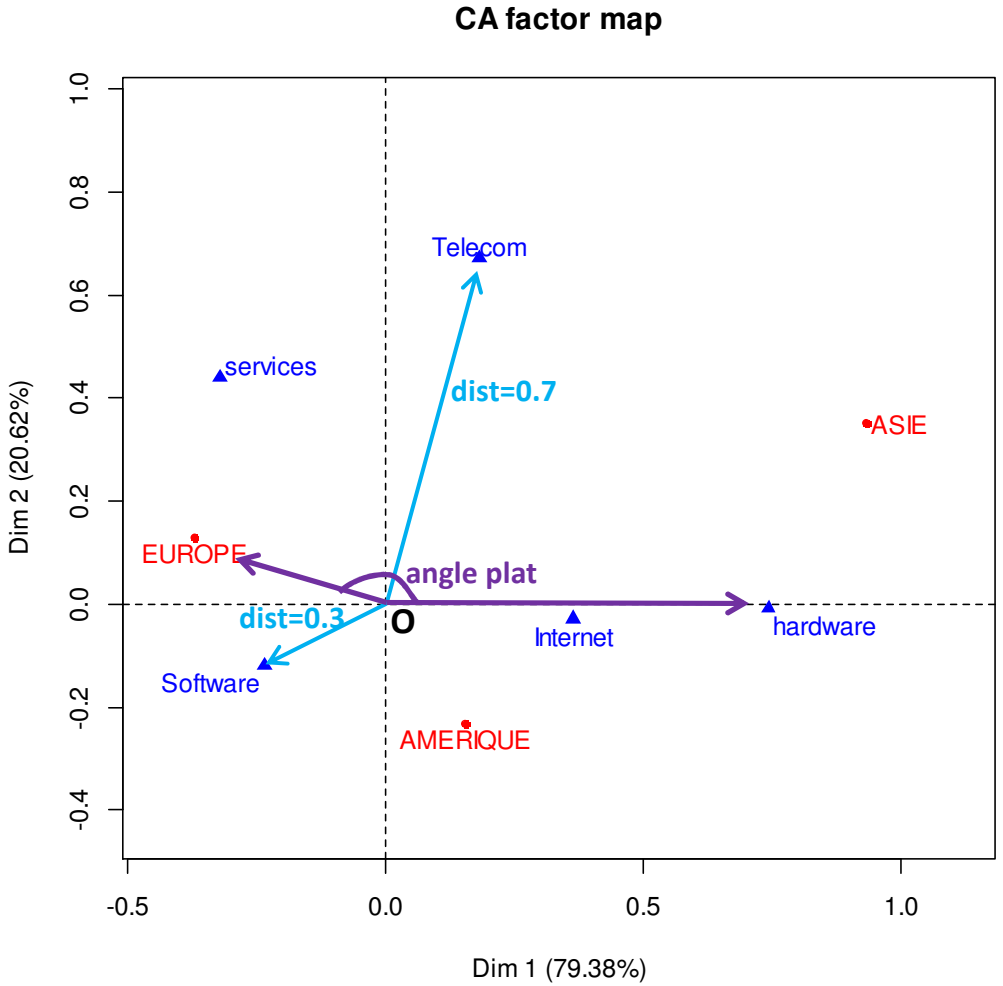
Analyse factorielle des correspondances

Lien avec l'ACP

- 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