

Analyse Factorielle des correspondances

AFC

Objectifs du cours :

- ▶ Données d'entrées et questions
- ▶ Comprendre les concepts de l'AFC
- ▶ Connaître les principes de calcul
- ▶ Savoir interpréter les résultats
- ▶ Placer l'AFC par rapport à l'ACP et aux méthodes de classification

Analyse Factorielle des Correspondances : Terminologie

- Pourquoi « des correspondances » ?
 - variables numériques \Rightarrow Corrélation
 - variables nominales \Rightarrow Correspondance
- Pourquoi « factorielle » ?
- Il s'agit de décomposer le tableau original en une somme de tableaux/matrices qui sont chacun le **produit** de facteurs simples.

Principes de l'AFC et données d'entrées

Données et questions : Exemple 1

Dans une entreprise, la répartition par sexe et catégorie socio-professionnelle (CSP) est la suivante :

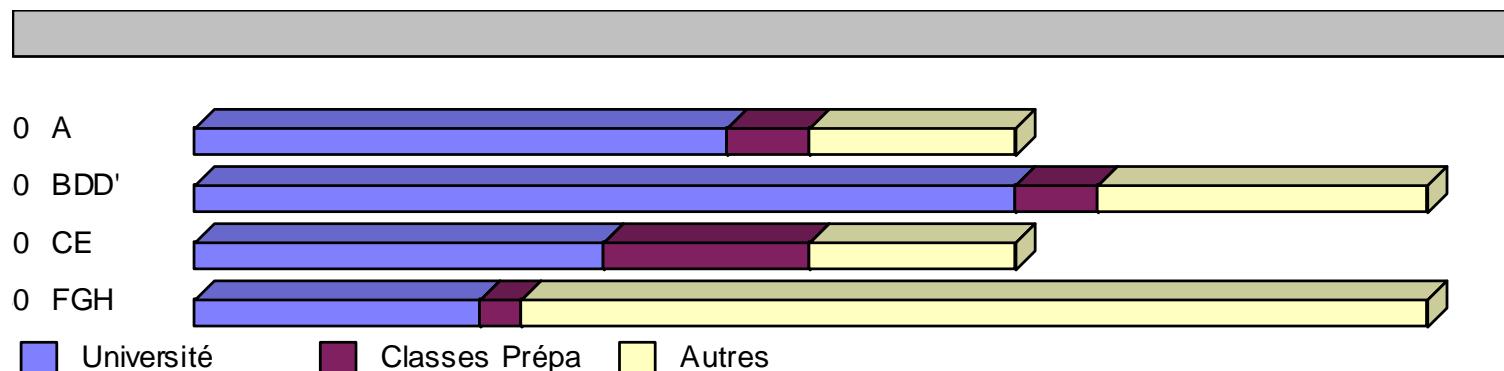
Tableau de contingence

	Ouvriers	Techniciens	Cadres	Total
Hommes	20	40	40	100
Femmes	30	60	10	100
Total	50	100	50	200

Y-a-t-il un lien entre le sexe à deux modalités et la CSP à trois modalités ?

Exemple 2 : que deviennent les bacheliers ?

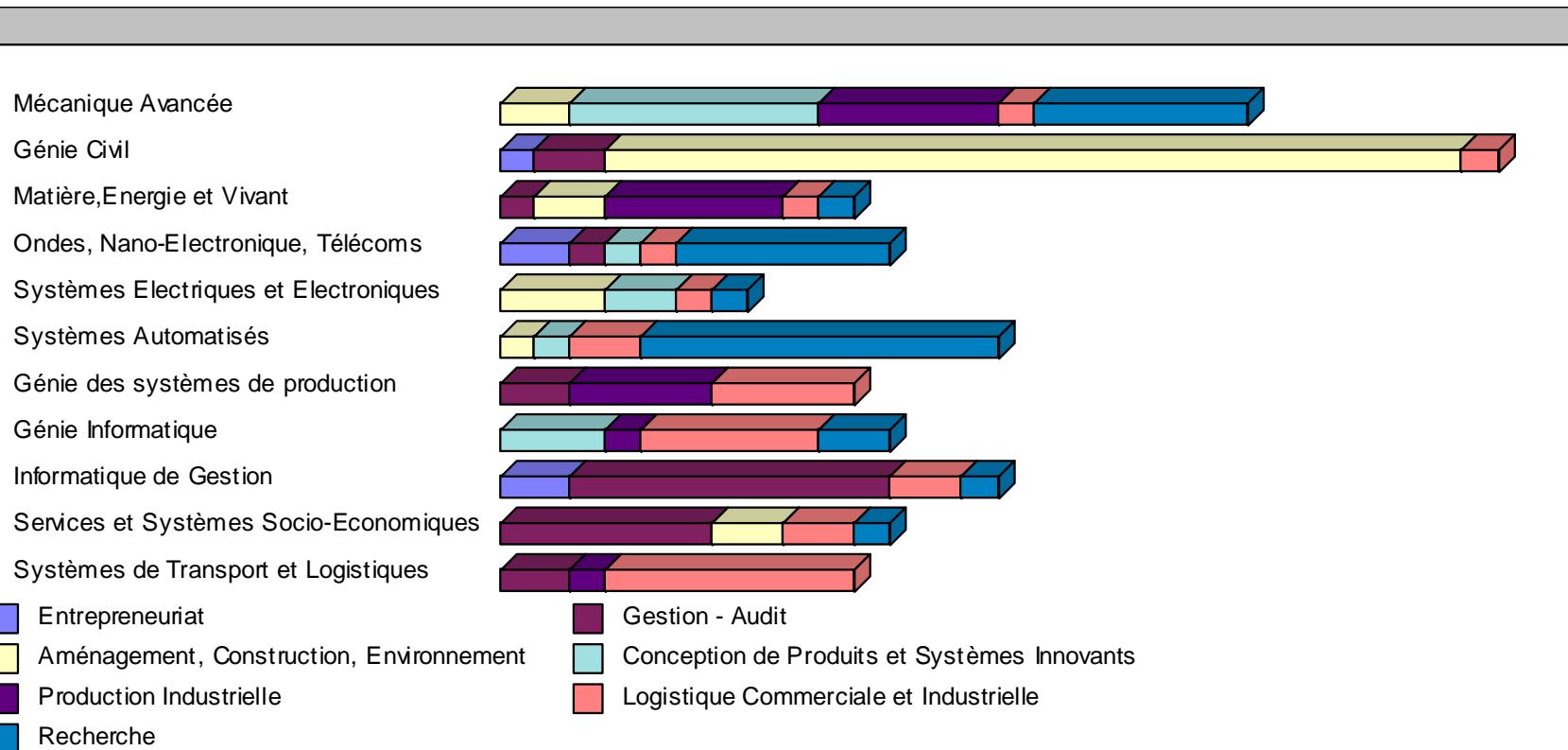
	<i>destination</i>			
	<i>université</i>	<i>classes prépa</i>	<i>autres</i>	<i>total</i>
A	13	2	5	20
BDD'	20	2	8	30
CE	10	5	5	20
FGH	7	1	22	30
total	50	10	40	100



Exemple 3 : quels souhaits d'orientation ?

Premiers vœux 2003 de Génie / filière.	Entrepreneuriat	Gestion - Audit	Aménagement, Construction, Environnement	Conception de Produits et Systèmes Innovants	Production Industrielle	Logistique Commerciale et Industrielle	Recherche
Mécanique Avancée	0	0	2	7	5	1	6
Génie Civil	1	2	24	0	0	1	0
Matière, Energie et Vivant	0	1	2	0	5	1	1
Ondes, Nano- Electronique, Télécoms	2	1	0	1	0	1	6
Systèmes Electriques et Electroniques	0	0	3	2	0	1	1
Systèmes Automatisés	0	0	1	1	0	2	10
Génie des systèmes de production	0	5	0	0	4	4	0
Génie Informatique	0	0	0	3	1	5	2
Informatique de Gestion	2	11	0	0	0	2	1
Services et Systèmes Socio-Economiques	1	6	3	0	0	2	1
Systèmes de Transport et Logistiques	0	2	0	0	1	8	0

Plus de complexité : avec plus de données



Intérêts et principes de l'AFC

- Problème : La lecture du tableau devient plus difficile quand il y a beaucoup de modalités
- Outil AFC : visualisation en 2 dimensions des tableaux de contingence
 - Transformation de variables qualitatives en variables quantitatives
- Intérêts :
 - Etude des liens entre les modalités de chaque variable
 - Etude des corrélations entre les modalités des 2 variables
- AFC = ACP avec une métrique particulière (celle du χ^2 pondéré)

Comment analyser ces données ?

Etapes

Trouver des valeurs inattendues dans les données, c'est-à-dire des valeurs qui dévient d'une situation attendue (uniforme)

1. Évaluer ce que serait une situation d'uniformité, d'indépendance
2. Exprimer cette différence graphiquement pour pouvoir l'analyser
3. Interpréter les graphiques obtenus
4. Optimiser la lisibilité des graphiques

Premier passage

1. Matrice T des données d'entrée
2. Matrice R des écarts à l'indépendance
3. Mise en facteur de R

Matrice « T » des données d'entrée

	<i>destination</i>			
	<i>université</i>	<i>classes prépa</i>	<i>autres</i>	<i>total</i>
A	13	2	5	20
BDD'	20	2	8	30
CE	10	5	5	20
FGH	7	1	22	30
total	50	10	40	100

Ce tableau est aussi une matrice, appelons-la « T »

Quelle matrice aurait-on si la répartition dans les filières post-Bac ne dépendait pas du type de Bac ?

Situation d'indépendance

$$10 = 50 * 20\% \quad (\text{produit matriciel en \%})$$
$$\begin{bmatrix} 10 & 2 & 8 \\ 15 & 3 & 12 \\ 10 & 2 & 8 \\ 15 & 3 & 12 \end{bmatrix} \begin{bmatrix} 20 \\ 30 \\ 20 \\ 30 \end{bmatrix}$$

50 10 40

Appelons cette matrice « T_0 »

On reconstitue la matrice à partir de ses marges

La matrice des écarts à l'indépendance

$$T - T_0 = R$$

$$\begin{bmatrix} 13 & 2 & 5 \\ 20 & 2 & 8 \\ 10 & 5 & 5 \\ 7 & 1 & 22 \end{bmatrix} - \begin{bmatrix} 10 & 2 & 8 \\ 15 & 3 & 12 \\ 10 & 2 & 8 \\ 15 & 3 & 12 \end{bmatrix} = \begin{bmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ -8 & -2 & 10 \end{bmatrix}$$

Comment exprimer simplement R ?

- ▶ Décomposition en une somme de matrice de la matrice des écarts à l'indépendance

$$R = T_1 + T_2$$

- ▶ Mise en facteur de T_1 et T_2
 - Produit d'un vecteur ligne et d'un vecteur colonne.

$$T_1 = C_1 L_1$$

$$R = T_1 + T_2 = C_1 L_1 + C_2 L_2$$

$$\begin{bmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ -8 & -2 & 10 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -2 \\ 1 & 1 & -2 \\ 2 & 2 & -4 \\ -4 & -4 & 8 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 2 \\ -4 \end{bmatrix} \begin{bmatrix} 2 & -1 & -1 \\ 4 & -2 & -2 \\ -2 & 1 & 1 \\ -4 & 2 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ -1 \\ -2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & -2 \end{bmatrix}$$

$$\begin{bmatrix} 2 & -1 & -1 \end{bmatrix}$$

D'une matrice à une présentation graphique

Production et interprétation du graphique

- Vecteurs colonne et vecteurs ligne
- Produit scalaire

Comment représenter graphiquement la décomposition ?

$$\begin{bmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ -8 & -2 & 10 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -2 \\ 1 & 1 & -2 \\ 2 & 2 & -4 \\ -4 & -4 & 8 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \\ -4 \end{bmatrix} + \begin{bmatrix} 2 & -1 & -1 \\ 4 & -2 & -2 \\ -2 & 1 & 1 \\ -4 & 2 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ -1 \\ -2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & -2 \end{bmatrix}$$

$$\begin{bmatrix} 2 & -1 & -1 \end{bmatrix}$$

Un vecteur colonne correspond à une modalité des données en colonnes

Un axe unidimensionnel + un axe unidimensionnel = un repère

A 1 1

BDD' 2 1

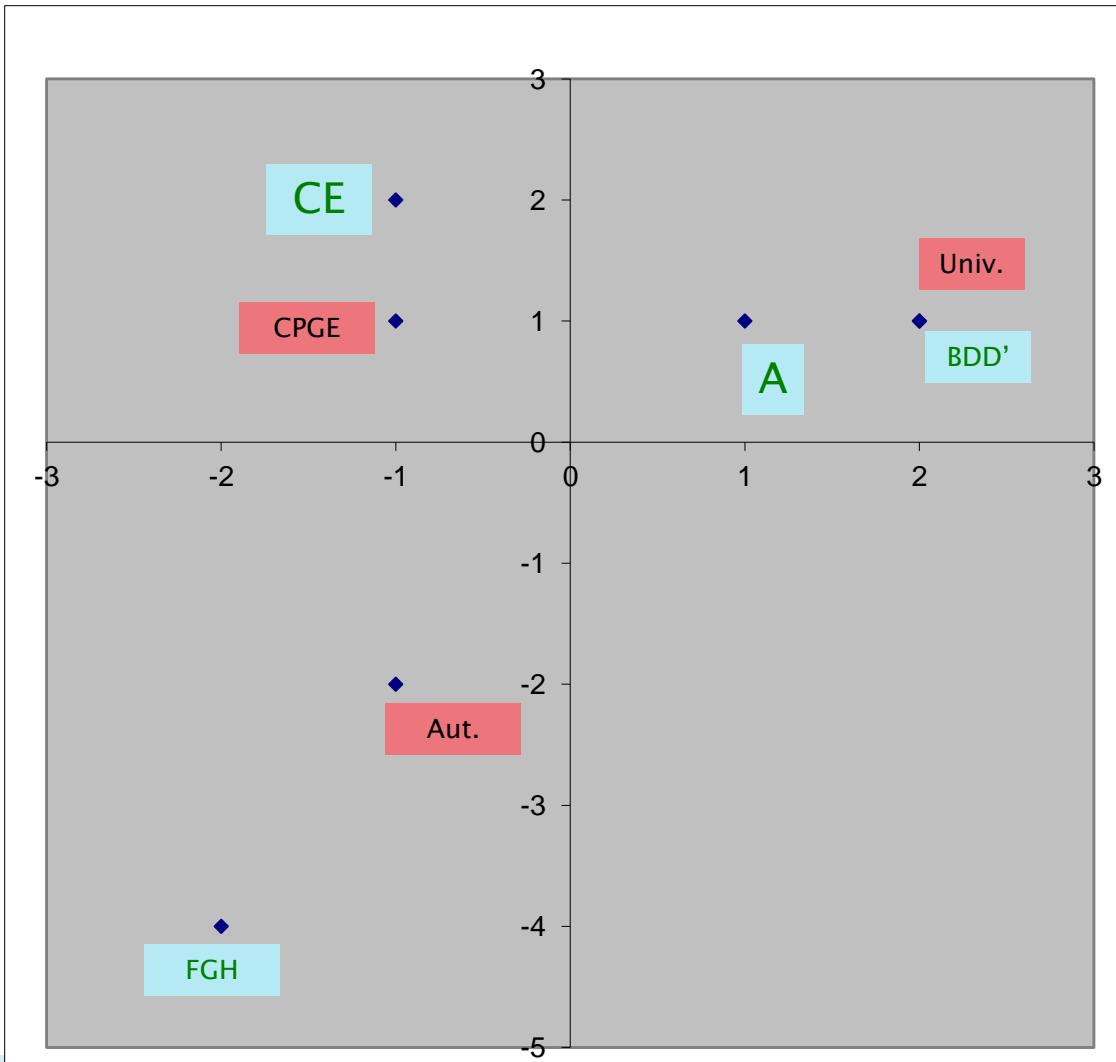
CE -1 2

FGH -2 -4

Univ 2 1

CPGE -1 1

Autres -1 -2



Que veut dire ce graphique ?

1. Produit scalaire positif :

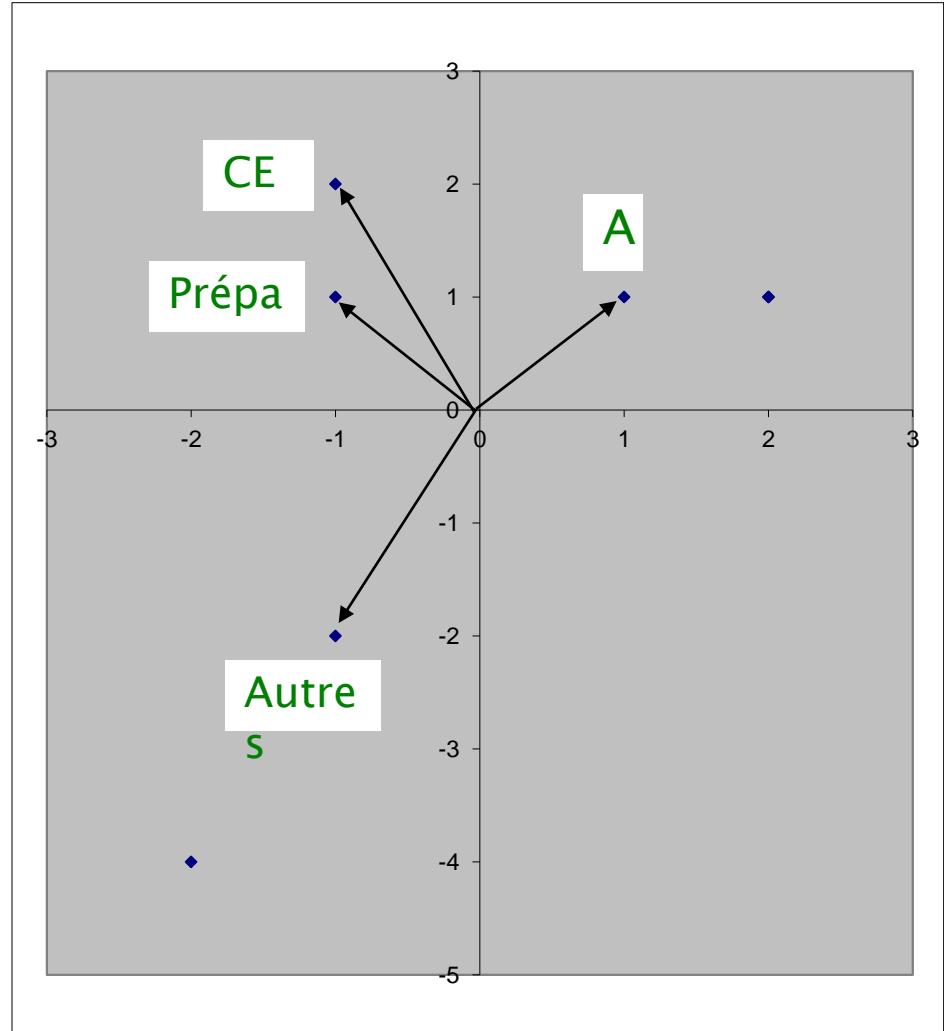
Les Bac CE ont une affinité pour la prépa

2. Produit scalaire négatif :

Les Bacs A ne vont pas vers les « autres »

3. Produit scalaire nul :

Les bacs A ne vont ni plus ni moins vers les prépas que la moyenne des bacheliers



Optimisation de la factorisation

Optimisation de la décomposition pour R ?

- ▶ R peut être écrite

$$R = T'{}_1 + T'{}_2 = T''{}_1 + T''{}_2 \dots$$

- ▶ Quel est le critère (la métrique) qui permet de définir les meilleurs T_1 et T_2 ?
- ▶ Chercher la meilleure T_1 , puis la meilleure T_2 de telle manière à ce que le premier axe soit celui qui conserve le plus d'informations possibles

La métrique que nous cherchons, c'est le Chi-2 (χ^2)

- ▶ Le χ^2 représente l'écart à l'indépendance
 - Cette indépendance, est exprimée par T_0
 - L'écart à l'indépendance est donc l'écart à T_0
- ▶ $\chi^2(R) = \chi^2(T_1) + \chi^2(T_2)$
 $2491 = 1998 + 493$

Le χ^2 en proportion de la richesse en information de la matrice = de son nombre de ddl.

Définition :

- On appelle degré de liberté par ligne (ddll) le nombre de colonnes (de modalités) diminué de 1.
- On appelle degré de liberté par colonne (ddlc) le nombre de lignes (de modalités) diminué de 1.
- Le degré de liberté du khi-deux de la matrice est le produit ddll x ddlc = ddl.
- Pour une matrice donnée, le χ^2 à prendre en compte est en fait χ^2 / ddl

Matrice T_1 maximisant le χ^2 dans notre cas

$$\chi^2(R) = \chi^2(T_1) + \chi^2(T_2)$$

$$2491 = 1998 + 493$$

$$100\% = 80.2\% + 19.8\%$$

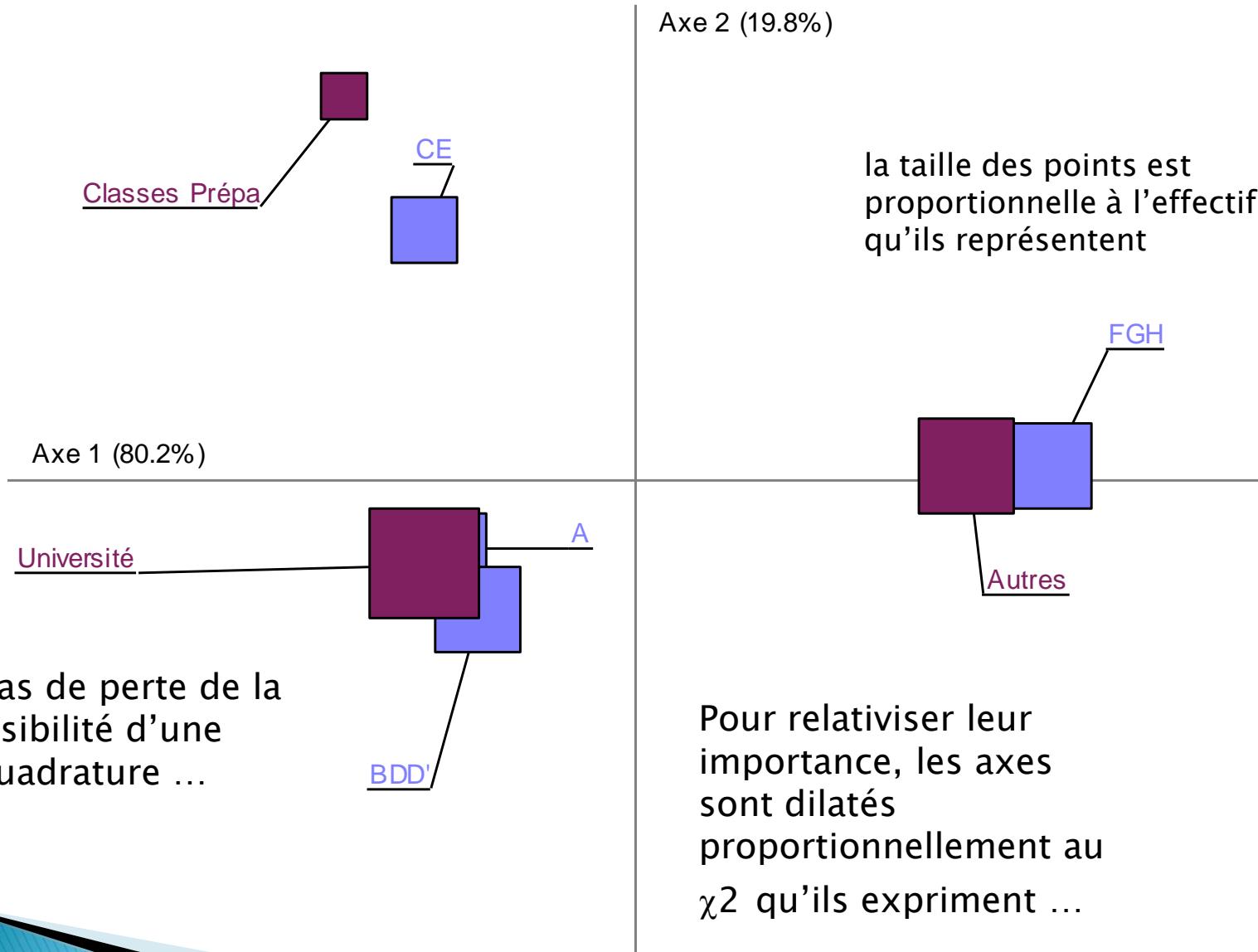
- ▶ Le pourcentage de la variance expliquée par un axe est intéressante lorsque la taille du tableau de données augmente...

$$\chi^2(R) = \chi^2(T_1) + \chi^2(T_2) + \chi^2(T_3) + \chi^2(T_4) ..$$

Pourquoi ?

- On ne peut que représenter que deux axes à la fois sur un graphique
=> représenter les plus significatifs.

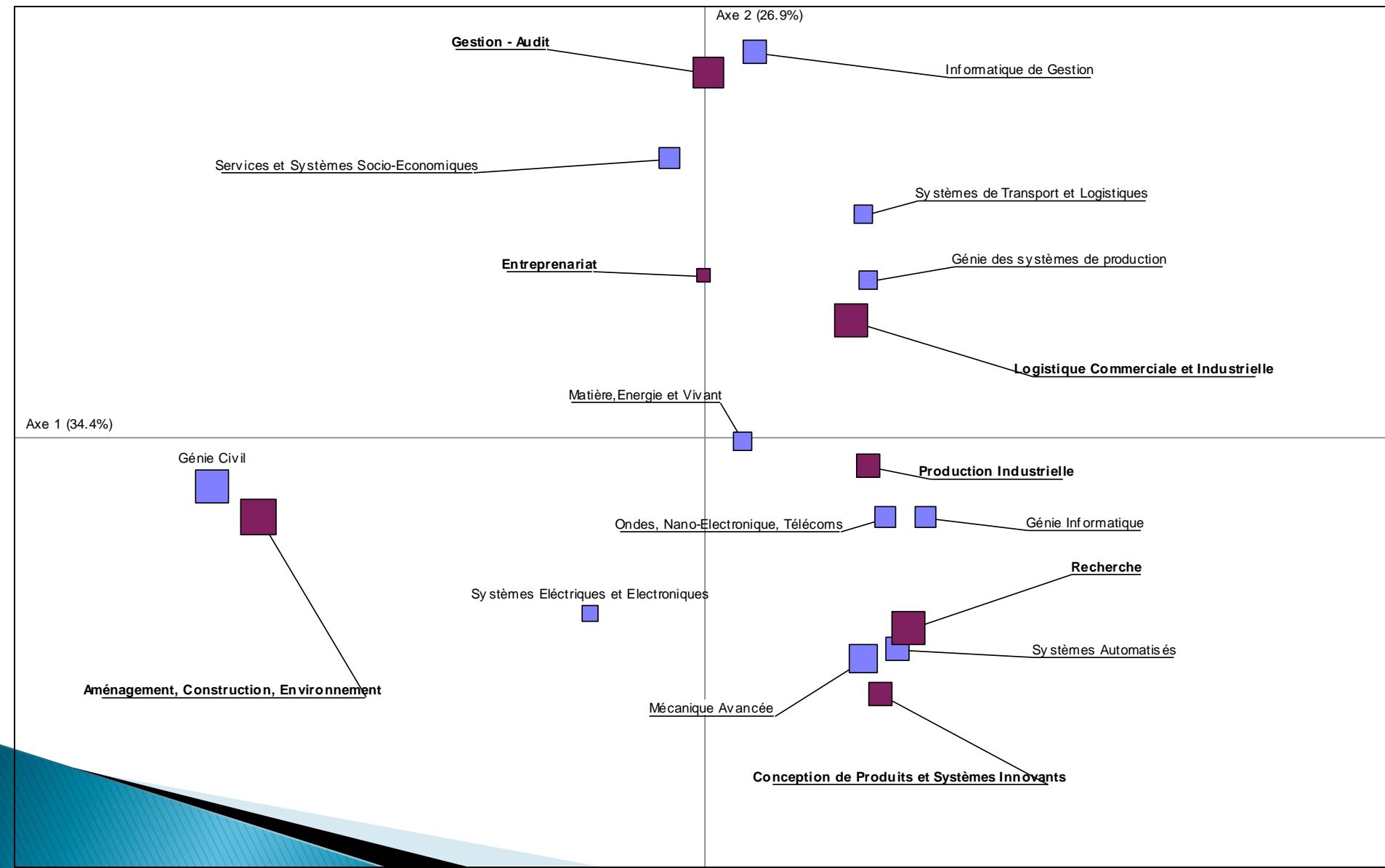
On obtient alors ce nouveau mapping



Exemple 2 : souhaits d'options?

Premiers vœux 2003 de Génie / filière.	Entrepreneuriat	Gestion - Audit	Aménagement, Construction, Environnement	Conception de Produits et Systèmes Innovants	Production Industrielle	Logistique Commerciale et Industrielle	Recherche
Mécanique Avancée	0	0	2	7	5	1	6
Génie Civil	1	2	24	0	0	1	0
Matière, Energie et Vivant	0	1	2	0	5	1	1
Ondes, Nano-Electronique, Télécoms	2	1	0	1	0	1	6
Systèmes Electriques et Electroniques	0	0	3	2	0	1	1
Systèmes Automatisés	0	0	1	1	0	2	10
Génie des systèmes de production	0	5	0	0	4	4	0
Génie Informatique	0	0	0	3	1	5	2
Informatique de Gestion	2	11	0	0	0	2	1
Services et Systèmes Socio-Economiques	1	6	3	0	0	2	1
Systèmes de Transport et Logistiques	0	2	0	0	1	8	0

Graphique des choix de filière / génie



Simulation d'accidents graves

Quantification de l'influence de certaines actions (appoints d'eau) lors d'un accident grave dans un réacteur nucléaire (fusion du cœur)

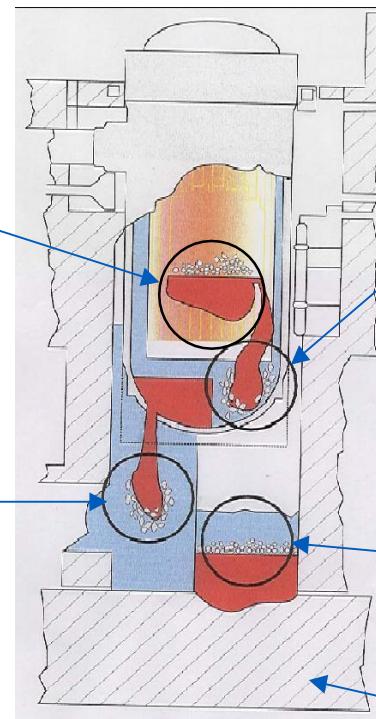
Scenario :
Dégradation cœur, transfert et interaction du corium (en cuve et hors cuve)

23 variables de sortie :
Masses de corium
Temps de percement de la cuve Temps de percement du radier

32 variables d'entrée aléatoires (lois uniformes) :
Gestion de l'eau, propriétés physique, variables scénario,
...

Dégradation et fusion du cœur

Chute du corium dans le puits de cuve



Chute du corium dans le fond de cuve

Interaction corium- béton

radier

Analyse de 2 variables catégorielles

Tableaux individus x variables

	Masse corium	Eau en cuve	Instant arrivée eau	Débit eau cuve	Percement cuve	Temps percement
1	11	oui	1000	1	oui	1500
2	15	non	NA	NA	oui	1000
3	15	oui	1000	5	non	NA
...						

tableau de contingence

On regroupe les individus

Lignes = modalités 1^{ère} variable

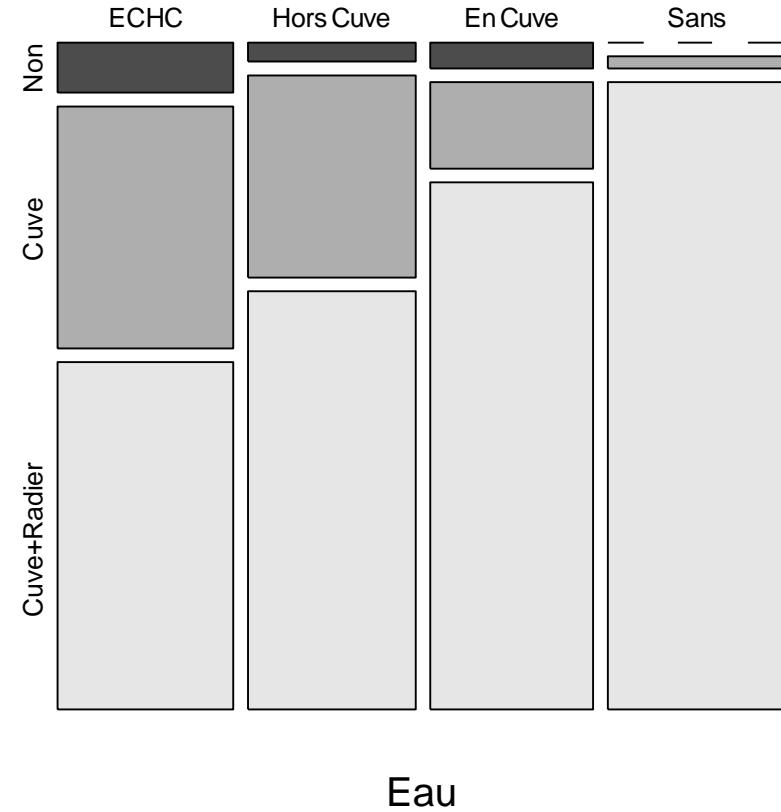
Colonnes = modalités 2^{ème} variable

Percement Eau	Non	Cuve	Cuve + Radier
Sans	0	2	100
En cuve	4	13	79
Hors cuve	3	31	64
En cuve + Hors cuve	8	39	56

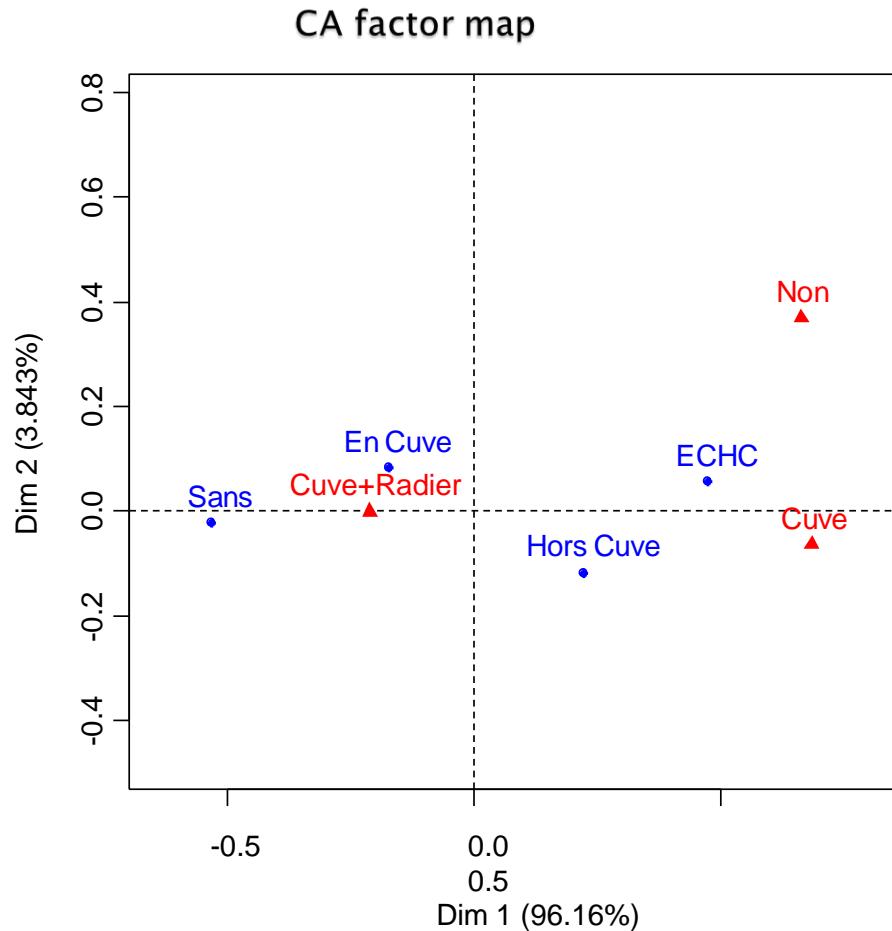
Analyse graphique

Mosaic plot

Percement Eau	Non	Cuve	Cuve + Radier
Sans	0	2	100
En cuve	4	13	79
Hors cuve	3	31	64
En cuve + Hors cuve	8	39	56



Résultat d'une analyse des correspondances



Chi-2 de l'indépendance entre les deux variables est 62.1863
La p-value associée à ce Chi-2 est 1.6167e-11.

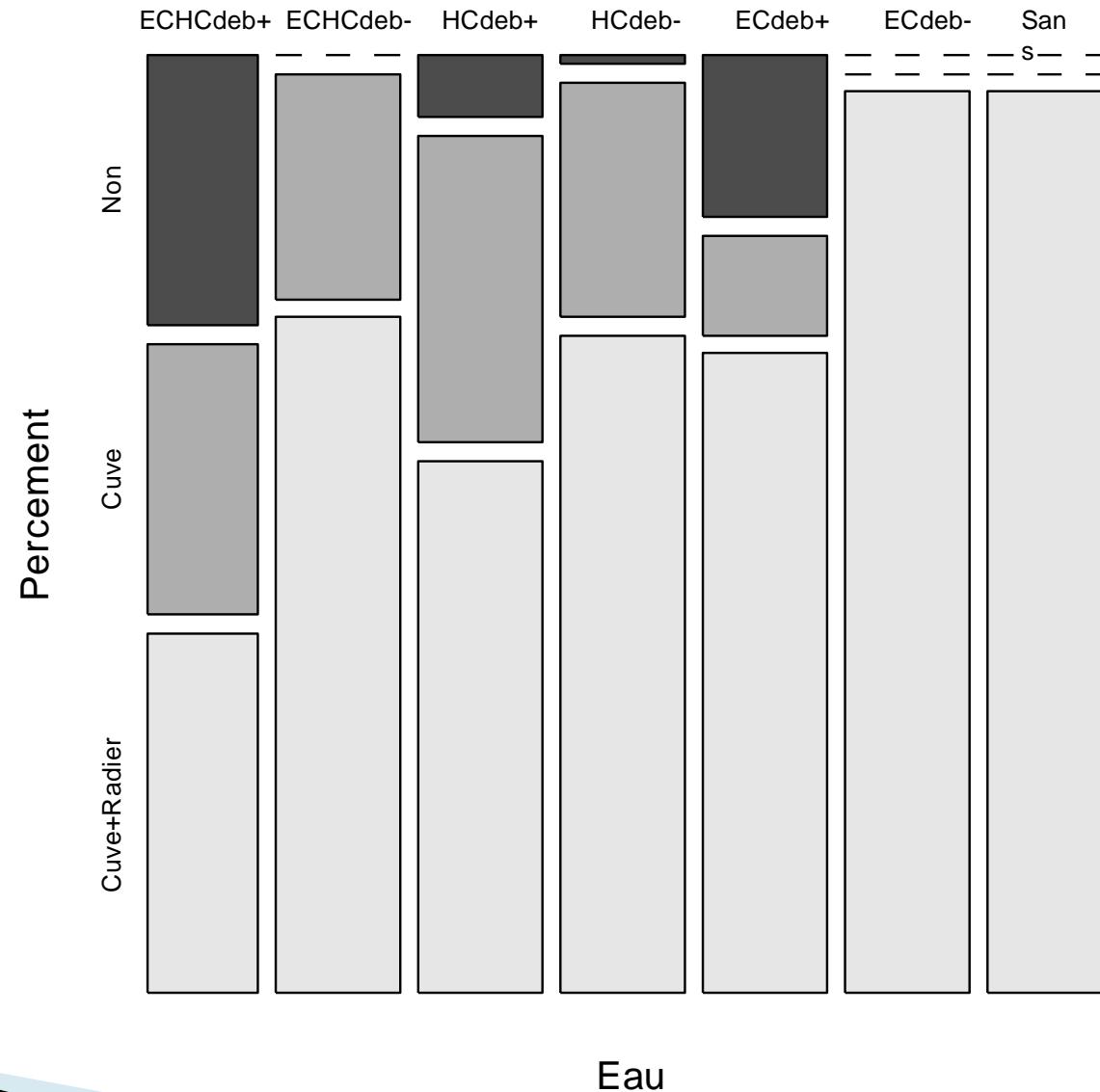
Conclusion : on rejette l'hyp. que les variables sont indépendantes avec un risque négligeable

- Analyse statistique quantitative du tableau de contingence
- Test statistique associé pour mesurer l'indépendance entre les 2 variables

Plus de modalités

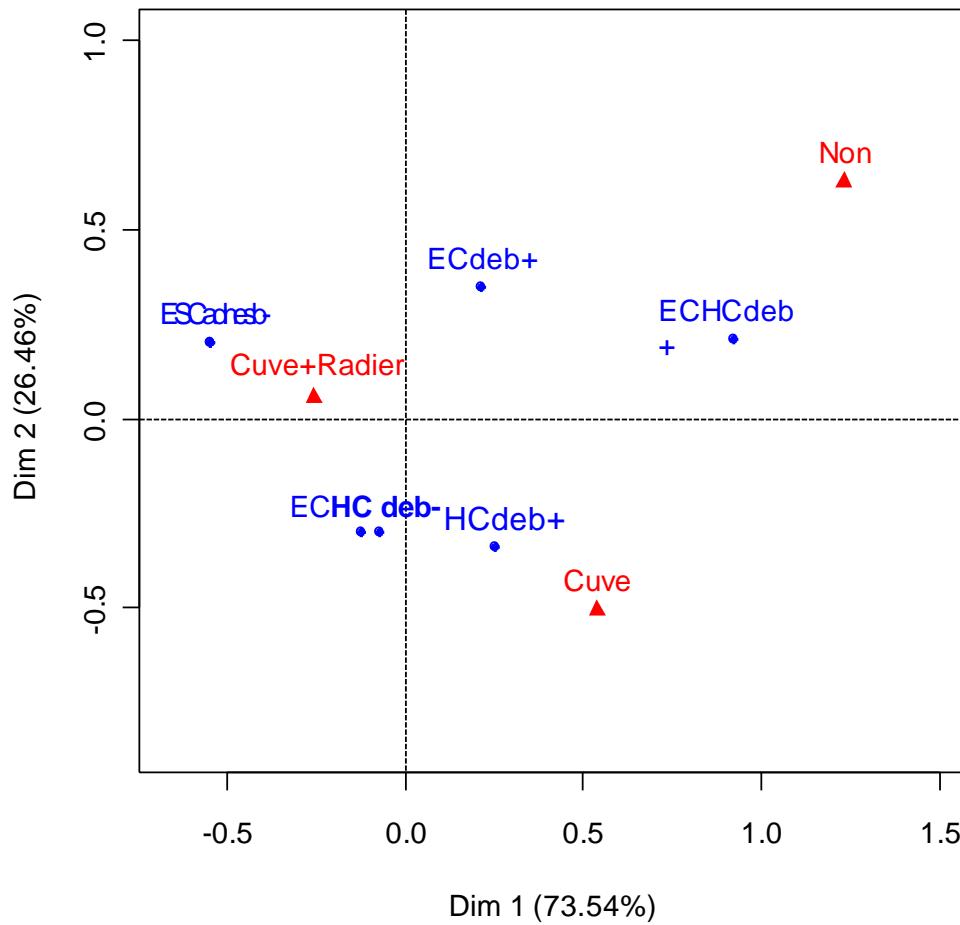
Eau	Percement	Non	Cuve	Cuve + Radier
Sans	0	0	100	
En cuve petit débit	0	0	100	
En cuve gros débit	18	11	71	
Hors cuve petit débit	1	26	73	
Hors cuve gros débit	7	34	59	
En cuve + Hors cuve petits débits	0	25	75	
En cuve + Hors cuve gros débits	27	27	36	

Plus de modalités



Plus de modalités

CA factor map



Chi-2 de
indépendance
entre les deux
variables est
203.4634
La p-value
associée à ce chi-2
est 6.284089e-37

Conclusion : on rejette
l'hyp. que les variables
sont indépendantes avec
un risque négligeable

Tableau de correspondances

Ensemble J		
	1	j
1		
i		x_{ij}
I		

x_{ij} : nombre d'individus appartenant à l'élément i de l'ensemble I et à l'élément j de l'ensemble J

⇒ Rôle symétrique des lignes et des colonnes

Personnages de Phèdre
(Racine)

Parfums

Mots

Descripteur

Nombre de fois que le personnage i a utilisé le mot j

Nombre de fois où le parfum i a été décrit par le mot j

⇒ Exemples où le test d'indépendance du χ^2 peut être appliqué

Enquête du CREDOC (N. Tabard, 1974)

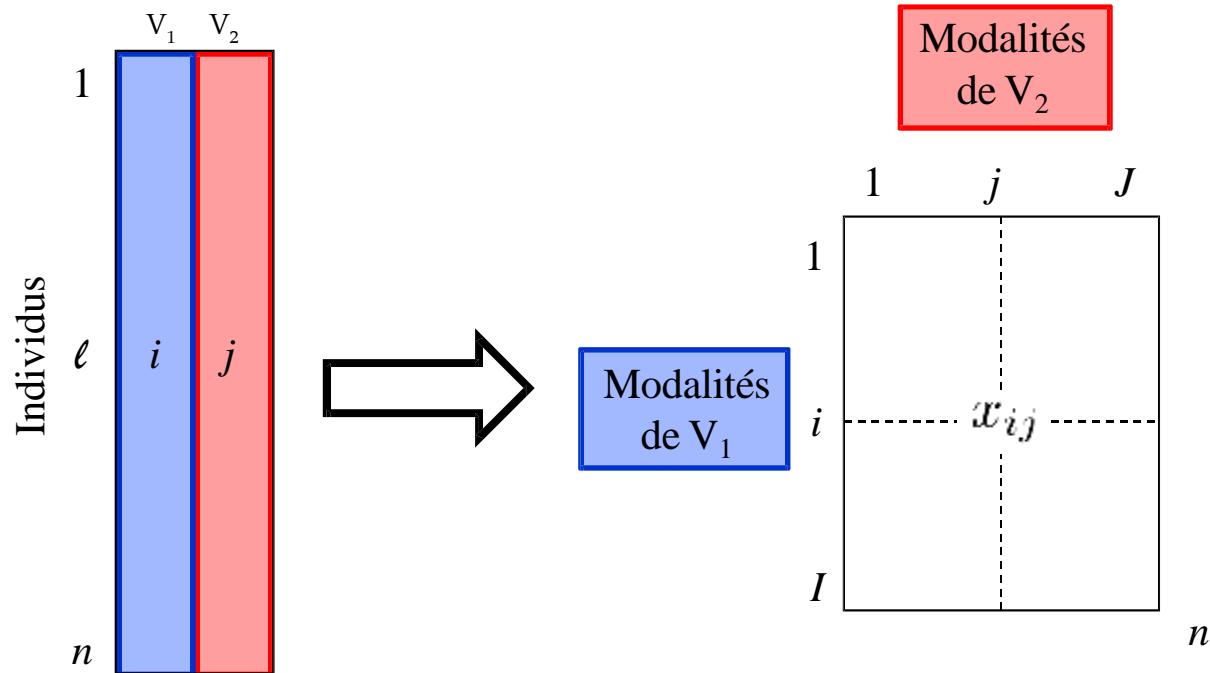
TABLEAU 37
REPONSES SIMULTANÉES A DES QUESTIONS D'OPINION

La famille idéale est celle où :	Activité convenant le mieux à une mère de famille quand les enfants vont à l'école :			
	rester au foyer	travailler à mi-temps	travailler à plein-temps	
les deux conjoints travaillent également	13	142	106	261
le mari a un métier plus absor- bant que celui de sa femme	30	408	117	555
seul le mari travaille	241	573	94	908
	284	1 123	317	1 724

⇒ Etude de la liaison entre deux variables qualitatives

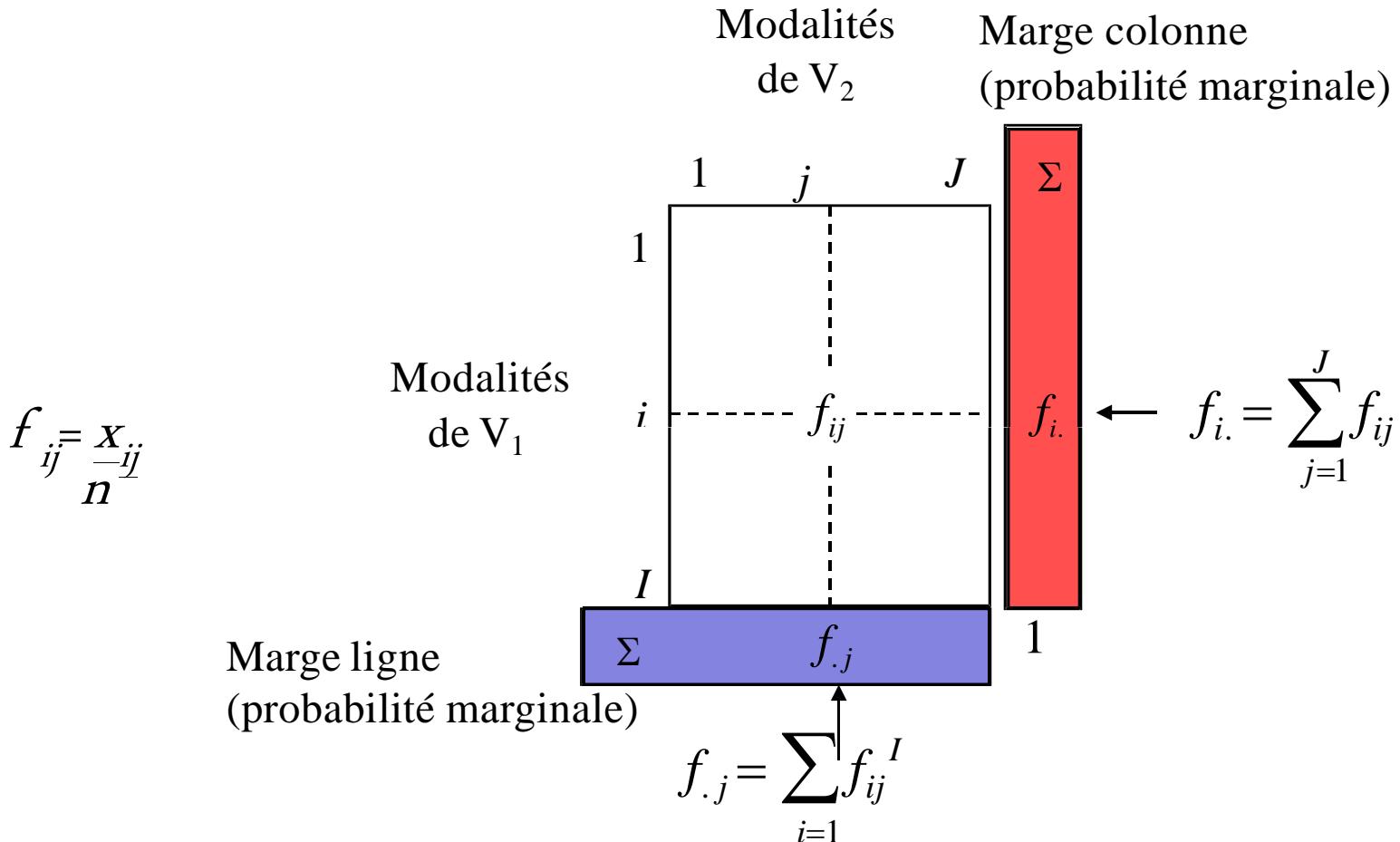
Données

n individus et 2 variables qualitatives



Distribution des n individus dans les $I \times J$ cases du tableau

Du tableau de contingences au tableau de probabilités



Liaison entre V_1 et V_2 : écart entre les données observées et le modèle d'indépendance

Liaisons et indépendance entre deux variables qualitatives

Modèle d'indépendance :

Evènements indépendants : $P(A \text{ et } B) = P(A) \times P(B)$ Variables

qualitatives indépendantes : $\forall i, \forall j, f_{ij} = f_{i.} \times f_{.j}$

\Rightarrow Probabilité conjointe = produit des probabilités marginales

Autres écritures : $\frac{f_{ij}}{f_{i.}} = f_{.j}$ $\frac{f_{ij}}{f_{.j}} = f_{i.}$

\Rightarrow Probabilité conditionnelle = probabilité marginale

Liaisons entre deux variables qualitatives

Ecart entre données obs (f_{ij}) et modèle d'indépendance (f_i, f_j)

➤ Significativité de la liaison (de l'écart) : test du χ^2

➤ H₀ : Les variables X et Y sont indépendantes

➤ H₁ : Les variables X et Y sont liées entre elles

$$\chi^2_{obs} = \sum_{i=1}^I \sum_{j=1}^J \frac{(\text{eff. observé} - \text{eff. théorique})^2}{\text{effectif théorique}} = \sum_{i=1}^I \sum_{j=1}^J \frac{(n f_{ij} - n f_{i.} f_{.j})^2}{n f_{i.} f_{.j}}$$

$$\chi^2_{obs} = \sum_{i=1}^I \sum_{j=1}^J n \frac{(\text{probabilité observée} - \text{probabilité théorique})^2}{\text{probabilité théorique}} = n \Phi^2$$

- Intensité de la liaison = Φ^2 = écart entre probabilités théoriques et observées
- Nature de la liaison = association entre modalités

Test d'indépendance

Si α est le risque associé au test, avec un niveau de confiance $1-\alpha$ (par exemple 95 %), on rejette ou on ne rejette pas l'hypothèse H_0

Remarque :

Test peu puissant et pas robuste pour de petits échantillons (< 50)

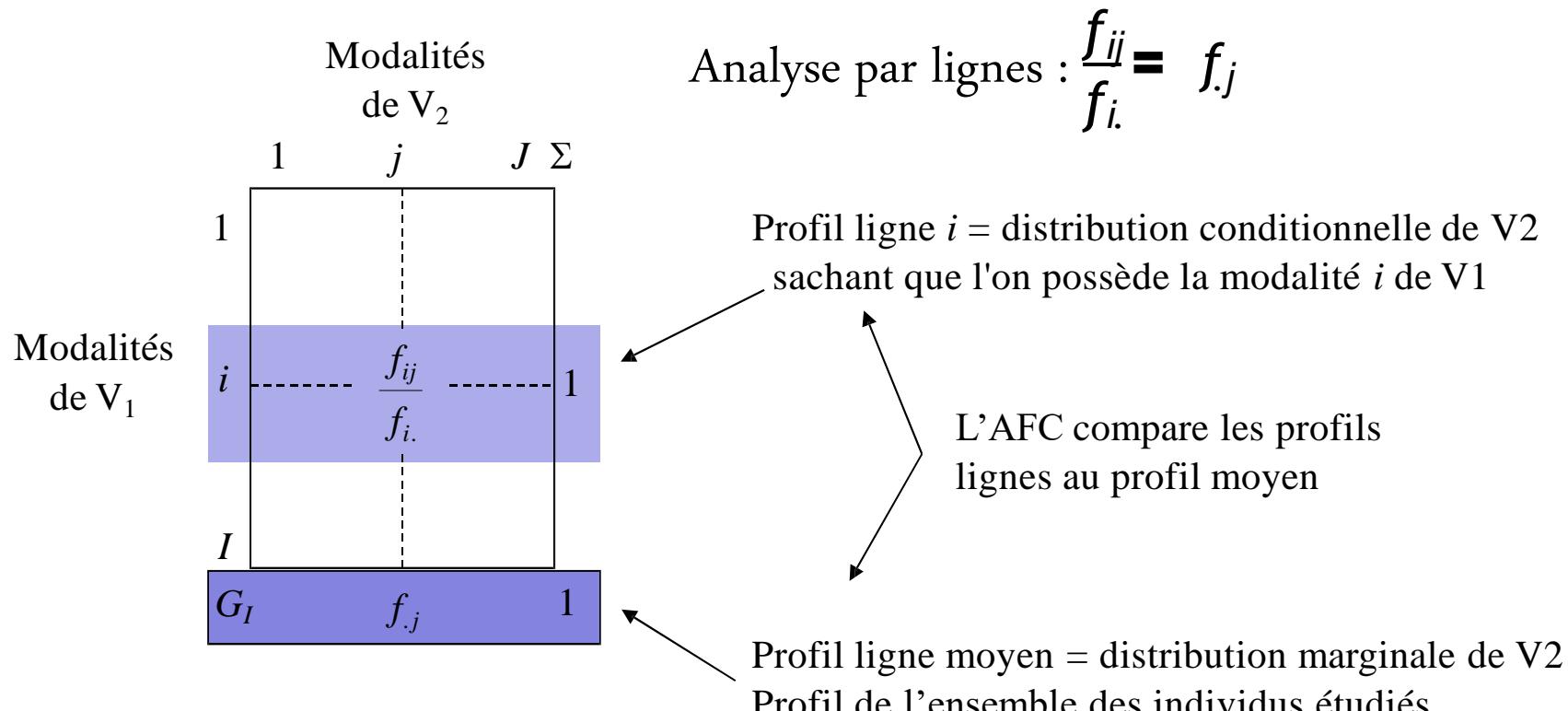
Pour les tableaux de contingence : $\chi^2_{\text{obs}} \sim \chi^2[(I - 1)(J - 1)]$

Décision :

On rejette H_0 au risque α de se tromper si $\chi^2_{\text{obs}} \geq \chi^2_{1-\alpha}[(I - 1)(J - 1)]$

Comment l'AFC appréhende l'écart à l'indépendance?

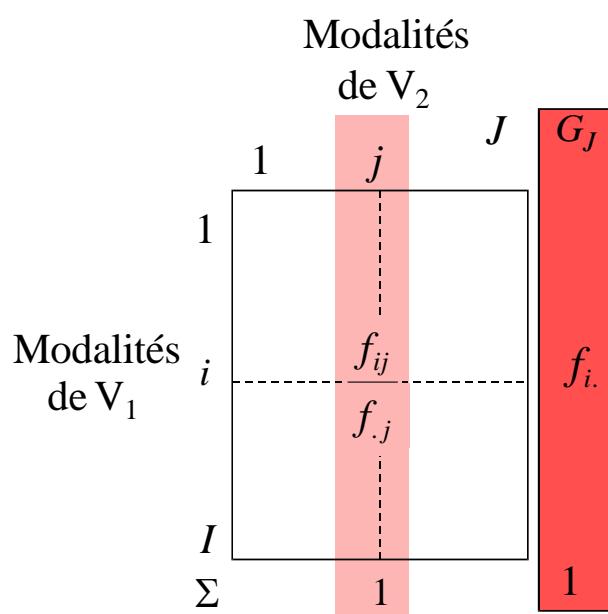
	rester au foyer	trav. à mi-temps	trav. à plein temps
2 conj. tr. également	4.98	54.41	40.61
trav. mari + absorbant	5.41	73.51	21.08
seul le mari travaille	26.54	63.11	10.35
marge ligne	16.47	65.14	18.39



Approche multidimensionnelle de l'écart à l'indépendance

Comment l'AFC appréhende l'écart à l'indépendance?

	rester au foyer	trav. à mi-temps	trav. à plein temps	marge colonne
2 conj. tr. Également	4.58	12.64	33.44	15.14
trav. mari + absorbant	10.56	36.33	36.91	32.19
seul le mari travaille	84.86	51.02	29.65	52.67



Analyse par colonnes : $\frac{f_{ij}}{f_{.j}} = f_{i.}$

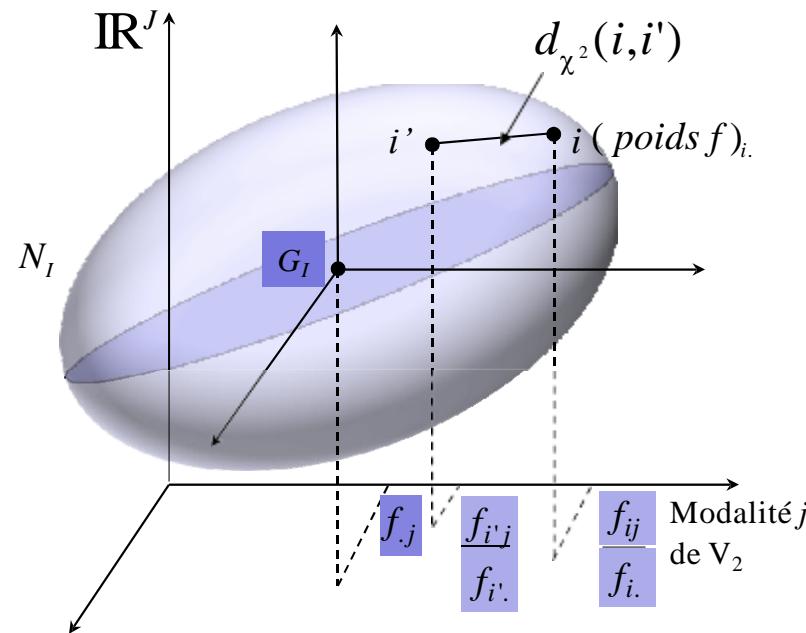
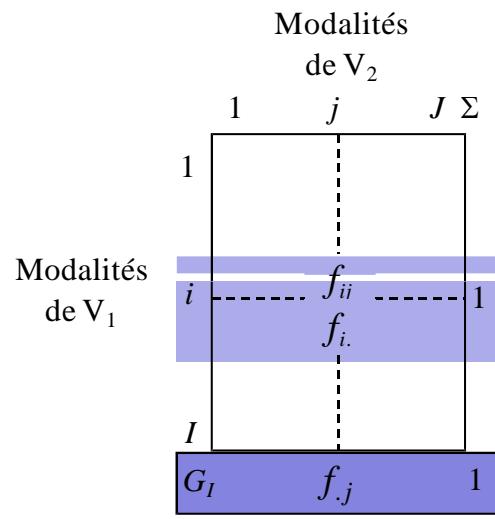
Profil colonne j
= distribution conditionnelle de V_1
sachant que l'on possède la modalité j de V_2

Profil colonne moyen = distribution marginale de V_1 Profil de l'ensemble des individus étudiés

Comparaison des profils colonnes au profil moyen

Approche multidimensionnelle de l'écart à l'indépendance

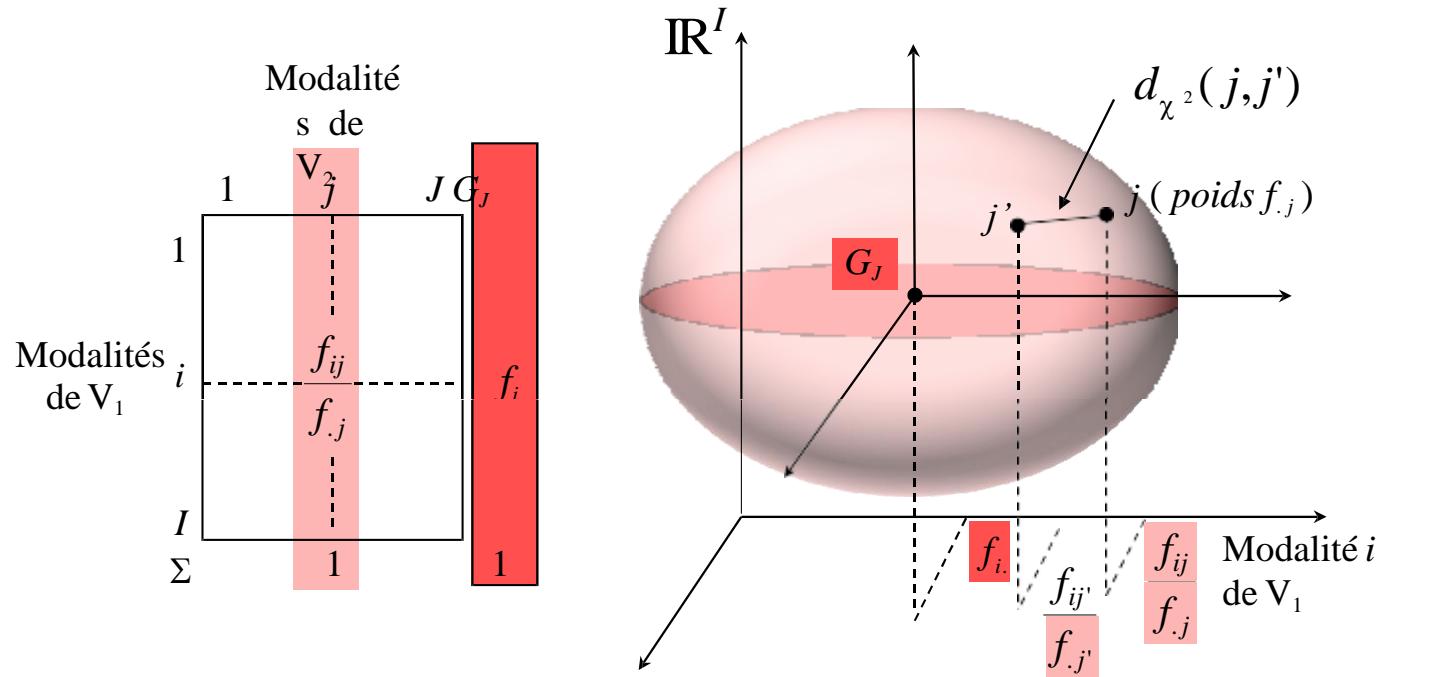
Le nuage des (profils) lignes



Distance entre deux profils : $d_{\chi^2}^2(i, i') = \sum_{j=1}^J \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$

Distance au profil moyen G : $d_{\chi^2}^2(i, G_I) = \sum_{j=1}^J \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2$

Le nuage des (profils) colonnes



Distance entre deux profils :

$$d_{\chi^2}^2(j, j') = \sum_{i=1}^I \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_j} - \frac{f_{ij'}}{f_{j'}} \right)^2$$

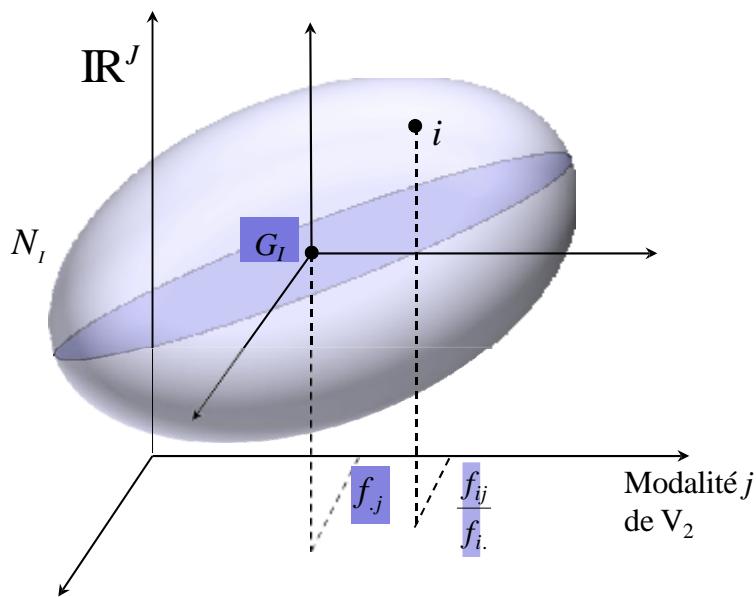
Distance au profil moyen G :

$$d_{\chi^2}^2(j, G_J) = \sum_{i=1}^I \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_j} - f_{i.} \right)^2$$

Que se passe-t-il s'il y a indépendance ?

Pour tout i , $\frac{f_{ij}}{f_{i\cdot}} = f_{\cdot j}$

- ⇒ les profils sont confondus avec le profil moyen ⇒ N_I réduit à G_I
⇒ L'inertie du nuage est nulle



Idem pour les colonnes : pour tout j , $\frac{f_{ij}}{f_{\cdot j}} = f_{i\cdot}$

Ecart à l'indépendance et inertie

Plus les données s'écartent de l'indépendance et plus les profils s'écartent de l'origine

$$\begin{aligned} Inertie(N_I/G_I) &= \sum_{i=1}^I Inertie(i/G_I) = \\ &= \frac{\chi^2}{n} = \phi^2 \end{aligned}$$

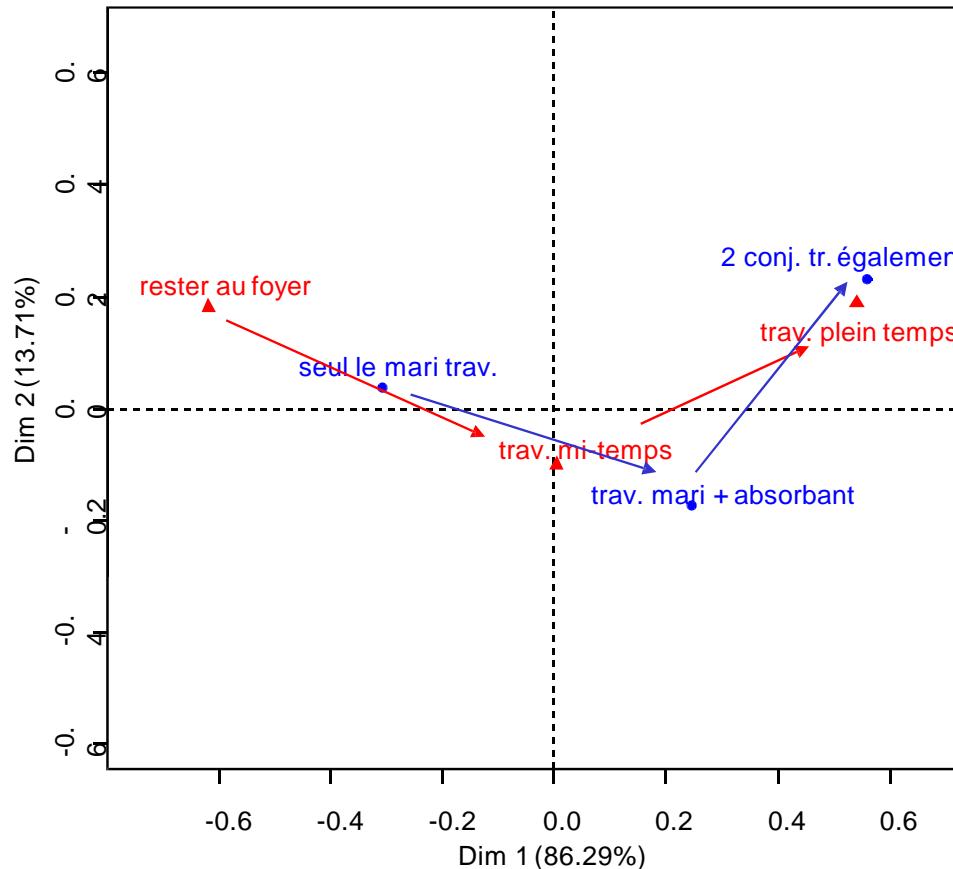
ϕ^2 mesure l'intensité de la liaison

Etudier l'inertie de N_I revient à étudier l'écart à l'indépendance Idem pour N_J :

$$Inertie(N_J/G_J) = Inertie(N_I/G_I)$$

- Avec plus de modalités, l'interprétation est difficile Il faut donc synthétiser l'information

Graphique de l'exemple



	rester au foyer	trav. à mi- temps	trav. à plein temps	$O = G_J$
2 conj. tr. également	4.58	12.64	33.44	15.14
trav. mari + absorbant	10.56	36.33	36.91	32.19
seul le mari travaille	84.86	51.02	29.65	52.67

Description d'un ensemble de profils par ACP

ACP → facteurs principaux (axes principaux + composantes principales)

- En ACP, on cherche des dimensions qui représentent au mieux la variabilité entre individus
- En AFC, on cherche des dimensions qui représentent au mieux l'écart des données à l'indépendance

Propriétés :

- ▶ Inertie = variance des facteurs
- ▶ Somme des inerties = ϕ^2
- ▶ Valeurs propres < 1
- ▶ Pourcentage de variance de chaque axe $\lambda_s / \sum \lambda_s$ où λ_s valeur propre de F_s
- ▶ Nb d'axes < min(I, J) - 1

Interprétation

▶ Poids :

chaque profil intervient d'autant plus dans l'analyse que son poids p_i ou p_j est élevé

▶ Contributions :

mesurent l'influence des profils dans le calcul des axes principaux (inertie d'un point / inertie totale λ)

▶ Qualité de la représentation : \cos^2

\cos^2 fort → point fortement expliqué et bien représenté par l'axe principal

Conclusion

Pour étudier la liaison entre deux variables qualitatives, on construit un tableau de contingence

Cette liaison réside dans l'écart entre le tableau de contingence et le modèle d'indépendance

L'analyse des correspondances :

- construit un nuage des lignes (et un nuage des colonnes) dont l'inertie totale mesure l'intensité de l'écart à l'indépendance
- décompose cette inertie totale sur une suite d'axes d'importance décroissante représentant chacun un aspect synthétique de la liaison entre les deux variables
- fournit une représentation des lignes et des colonnes dans laquelle la position d'un point reflète sa participation à l'écart à l'indépendance