Analyse factorielle des correspondances (AFC)

Angelina Roche

Executive Master Statistique et Big Data

2018–2019

Plan du cours

Plan

Exemple : attitude à l'égard du travail féminin en 1970

Réponse à deux questionnaires [1] :

- Parmi les trois modèles suivants, quel est celui qui se rapproche le plus de l'image idéale que vous vous faites d'une famille.
  - ☐ Une famille où les deux conjoints ont un métier qui les absorbent autant l'un que l'autre et où les tâches ménagères et les soins aux enfants sont partagés entre les deux.
  - ☐ Une famille où la femme a une profession moins absorbante que celle de l'homme et où elle assure une plus grande part des tâches ménagères et des soins aux enfants.
  - ☐ Une famille où l'homme seul exerce une profession et où la femme reste au foyer.

1. Source : Tabard, N. (1974). Besoins et aspirations des familles et des jeunes. CREDOC. Paris.

Exemple : attitude à l'égard du travail féminin en 1970

- En distinguant la période où les enfants sont petits et celle où tous les enfants vont à l'école, quel est selon vous le type d'activité qui convient le mieux à une mère de famille :
  - ☐ au foyer,
  - ☐ travail extérieur à mi-temps,
  - ☐ travail extérieur à plein temps.

Exemple : attitude à l'égard du travail féminin en 1970

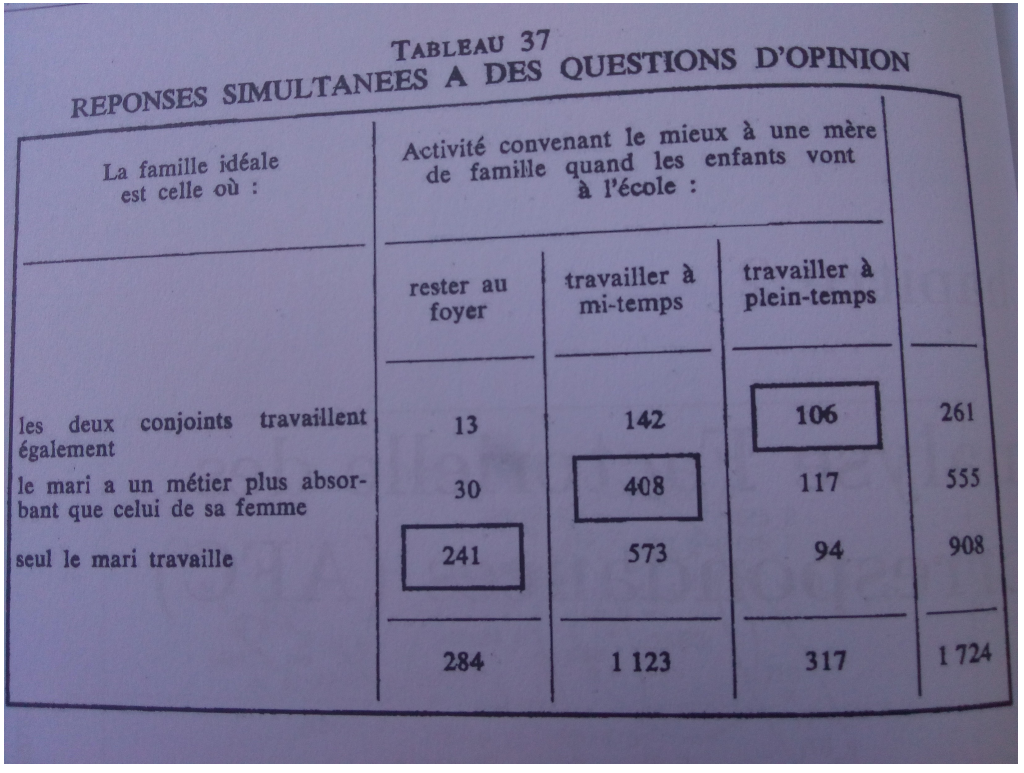


TABLEAU 37

REPONSES SIMULTANEES 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 absorbant que celui de sa femme	30	408	117	555
seul le mari travaille	241	573	94	908
	284	1 123	317	1 724

Figure – Tableau croisé des réponses aux questions reproduit dans Husson, Lê, Pages, Analyse de données avec R.

Notations et tableau de contingence

- On dispose pour n individus de leurs valeurs pour deux variables qualitatives V_1 et V_2.
- V_1 présente I modalités (= valeurs possibles) et V_2 en possède J.
- x_{ij} : nombre d'individus possédant la modalité i de V_1 et j de V_2.
- Tableau croisé (= tableau de contingence) : $(x_{ij})_{1\leq i\leq I, 1\leq j\leq J}$.
- Marges :

$$x_{i\bullet} = \sum_{j=1}^{J} x_{ij}, x_{\bullet j} = \sum_{i=1}^{I} x_{ij} \text{ et } n = x_{\bullet\bullet} = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}.$$

Tableau de probabilités et probabilités marginales

- Tableau dont les termes sont :

$$f_{ij} = \frac{x_{ij}}{n}.$$

- Probabilités marginales :

$$f_{i\bullet} = \sum_{j=1}^{J} f_{ij}, f_{\bullet j} = \sum_{i=1}^{I} f_{ij} \text{ et } 1 = f_{\bullet\bullet} = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij}.$$

Effectifs théoriques

- Si les variables V_1 et V_2 sont indépendantes alors, pour tous i et j :

$$\mathbb{P}(V_1 = i \text{ et } V_2 = j) = \mathbb{P}(V_1 = i) \times \mathbb{P}(V_2 = j).$$

- Dans ce cas : on s'attend à ce que, pour tous i et j :

$$f_{ij} \approx f_{i\bullet} f_{\bullet j} \text{ ou de même } x_{ij} (= nf_{ij}) \approx nf_{i\bullet} f_{\bullet j}$$

- L'écart entre le tableau croisé $(x_{ij})_{1 \leq i \leq I, 1 \leq j \leq J}$ et le tableau dit des *effectifs théoriques* $(nf_{i\bullet} f_{\bullet j})_{1 \leq i \leq I, 1 \leq j \leq J}$ mesure l'écart à l'indépendance.

Effectifs théorique – données sur travail des femmes

- Effectifs observés :

	rester.au.foyer	trav..à.mi.temps	trav..plein.temps
2 conj. tr. également	13.00	142.00	106.00
trav. mari + absorbant	30.00	408.00	117.00
seul le mari trav.	241.00	573.00	94.00

- Effectifs théoriques :

	rester au foyer	trav. à mi-temps	trav. plein temps
2 conj. tr. également	43.00	170.00	48.00
trav. mari + absorbant	91.40	361.50	102.10
seul le mari trav.	149.60	591.50	167.00

Test du χ^2

- $$\begin{aligned} \chi^2_{obs} &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\text{effectifs observés} - \text{effectifs théoriques})^2}{\text{effectifs théoriques}} \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(x_{ij} - nf_{i\bullet}f_{\bullet j})^2}{nf_{i\bullet}f_{\bullet j}}. \end{aligned}$$

- Sous l'hypothèse d'indépendance, la statistique χ^2_{obs} suit une loi dite du χ^2 *à* $(I-1)\times(J-1)$ *degrés de liberté.*

- p-valeur $= \mathbb{P}_{V_1 \perp V_2}\left(\chi^2 \geq \chi^2_{obs}\right)$.

- On considère que les variables V_1 et V_2 sont dépendantes si p-valeur $\leq 5\%$.

Plan

Nuages des profils lignes et colonnes

- Nuage des profils lignes

$$N_I := \{(f_{i1}/f_{i\bullet}, ..., f_{iJ}/f_{i\bullet}), i = 1, ..., I\} \subset \mathbb{R}^J.$$

On attribue à chaque ligne le poids $p_i = f_{i\bullet}$, point moyen : $G_I = (f_{\bullet 1}, ..., f_{\bullet J})$.

- Nuage des profils colonnes

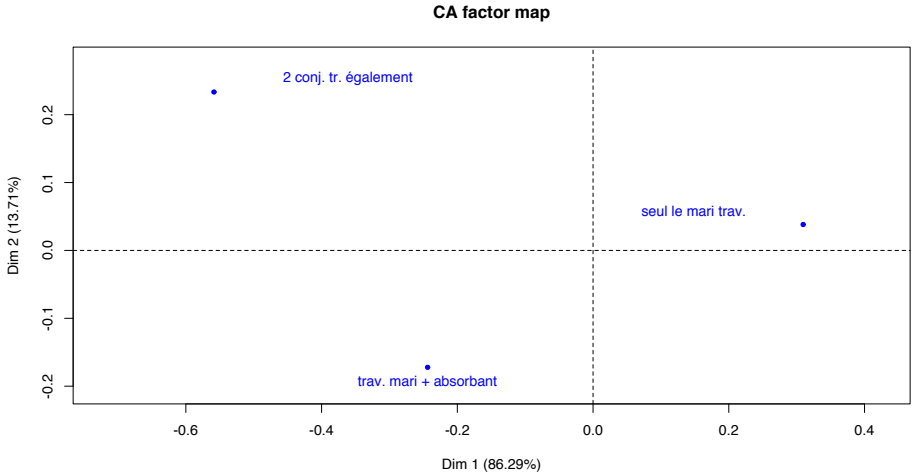
$$N_J := \{(f_{1j}/f_{\bullet j}, ..., f_{Ij}/f_{\bullet j}), j = 1, ..., J\} \subset \mathbb{R}^I.$$

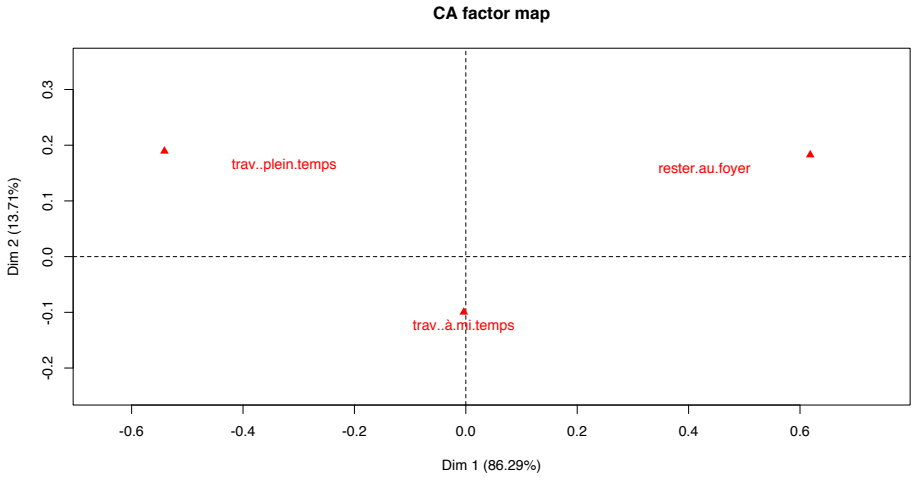
On attribue à chaque colonne le poids $p_j = f_{\bullet j}$, point moyen : $G_J = (f_{1\bullet}, ..., f_{I\bullet})$.

Axes principaux

- On procède ensuite exactement comme pour l'ACP pour la recherche des axes principaux (maximisation de l'inertie projetée).
- Le nombre d'axes maximum pour représenter parfaitement un tableau croisé de taille $I \times J$ est $\min\{I-1, J-1\}$ car :
  - la somme des coordonnées d'un profil est égale à 1 : N_i appartient donc à un sous-espace de dimension $J-1$ de $\mathbb{R}^J$,
  - N_I contient I points : il est donc possible de le représenter parfaitement avec $I-1$ dimensions.

Projection des nuages des profils lignes et colonnes





Représentation superposée des lignes et des colonnes

- Dualité des représentations de N_I et N_J : il s'agit du même tableau de données vu de 2 points de vue différents.

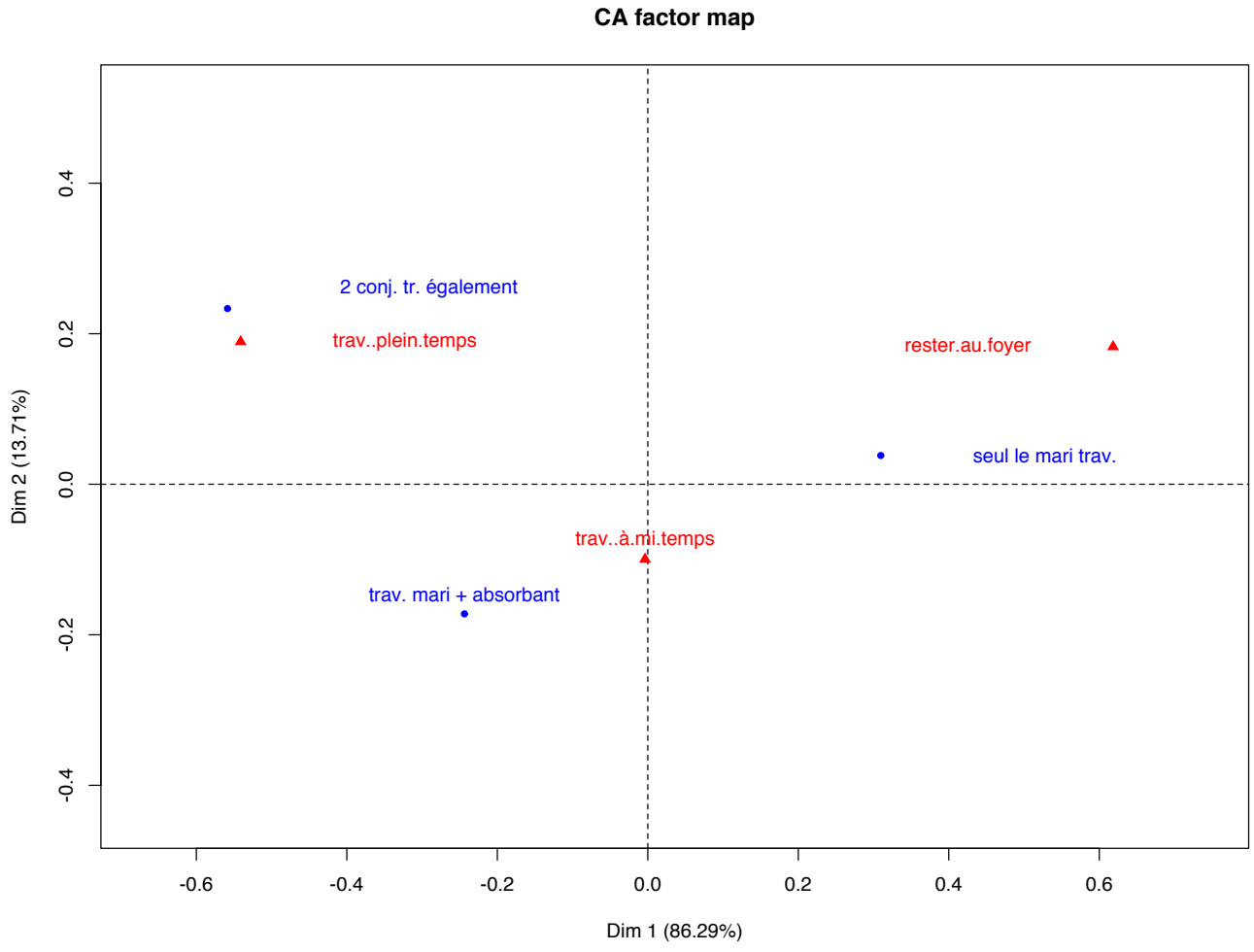
$\hookrightarrow$ même inertie totale χ^2/n,

$\hookrightarrow$ inertie projetée sur le k-ème axe factoriel de N_I = inertie projetée sur le k-ème axe factoriel de $N_J = \lambda_k$ (propriété admise),

$\hookrightarrow$ relations (admise) entre les coordonnées s_i^k (resp. t_j^k) des projections des profils lignes (resp. colonnes) sur les axes factoriels :

$$s_i^k = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{J} \frac{f_{ij}}{f_{i\bullet}} t_j^k \text{ et } t_j^k = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{\bullet j}} s_i^k.$$

Représentation superposée des lignes et des colonnes



Plan

Inertie projetée (valeurs propres)

- Particularité de l'AFC : pour tout k, $\lambda_k \leq 1$.
- $\lambda_1 = 1 \rightarrow$ liaison très forte entre les variables.
- Données sur le travail féminin :

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.12	86.29	86.29
dim 2	0.02	13.71	100.00

Contribution et qualité de représentation

- Sélectionner les points les plus contributifs ou les mieux représentés peut aider à interpréter un axe.

- Lorsqu'on s'intéresse à une modalité en particulier, on peut regarder l'axe dans lequel elle s'interprète le mieux.

	Coordonnées		Contribution		Qualité	
	Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2
2 conj. tr. également	-0.56	0.23	40.43	44.43	0.85	0.15
trav. mari + absorbant	-0.24	-0.17	16.37	51.44	0.67	0.33
seul le mari trav.	0.31	0.04	43.20	4.13	0.99	0.01

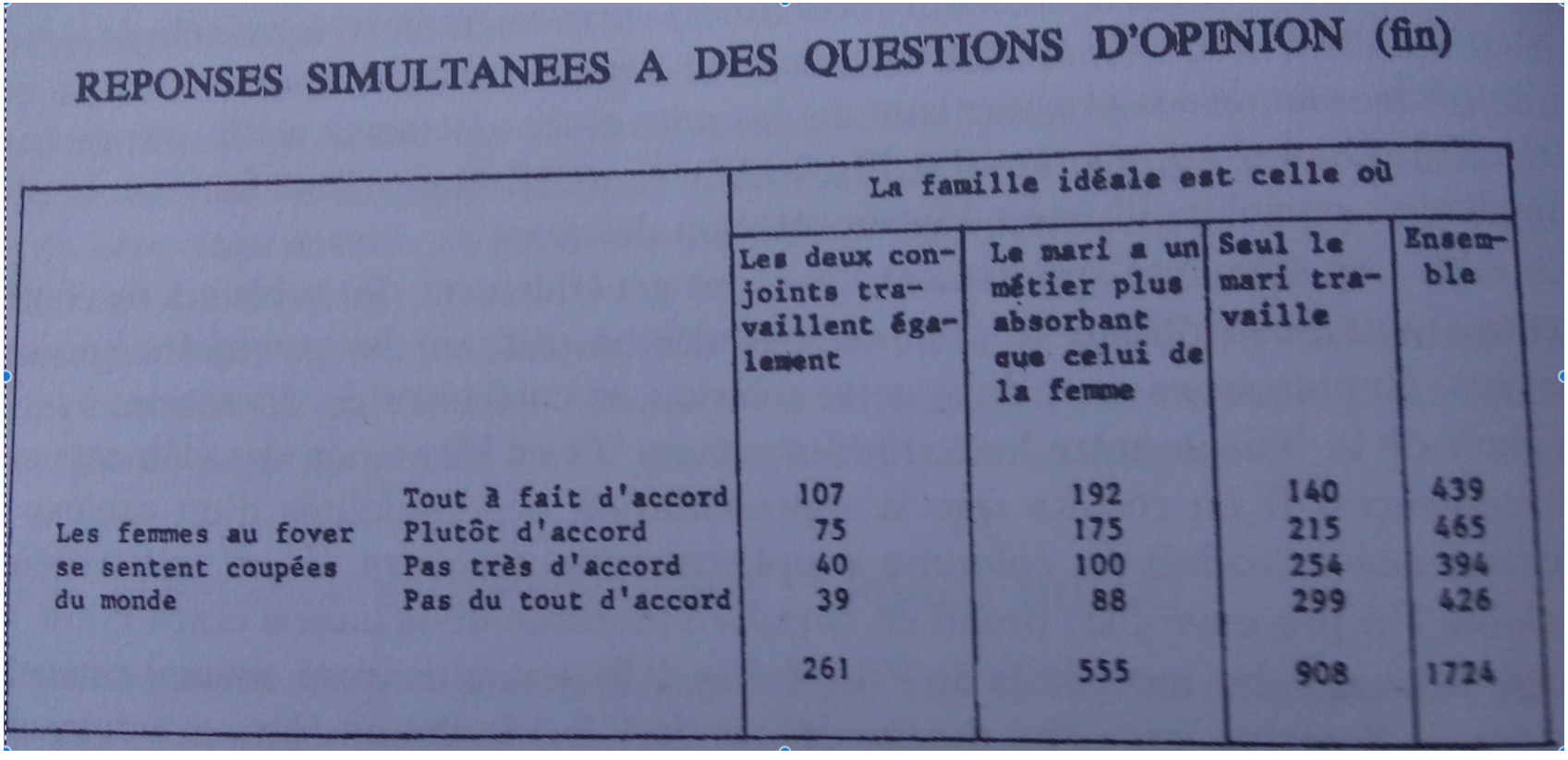
	Coordonnées		Contribution		Qualité	
	Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2
rester.au.foyer	0.62	0.18	53.91	29.61	0.92	0.08
trav..à.mi.temps	-0.00	-0.10	0.01	34.85	0.00	1.00
trav..plein.temps	-0.54	0.19	46.08	35.53	0.89	0.11

Éléments supplémentaires

- Possibilité d'ajouter dans la représentation des profils, des informations d'un autre tableau croisant les modalités d'une nouvelle variable qualitative V_3 avec V_1 ou V_2.

- Pour cela, on calcule les profils lignes ou colonnes du tableau qui est ensuite ajouté à le représentation des nuages des profils lignes ou colonnes.

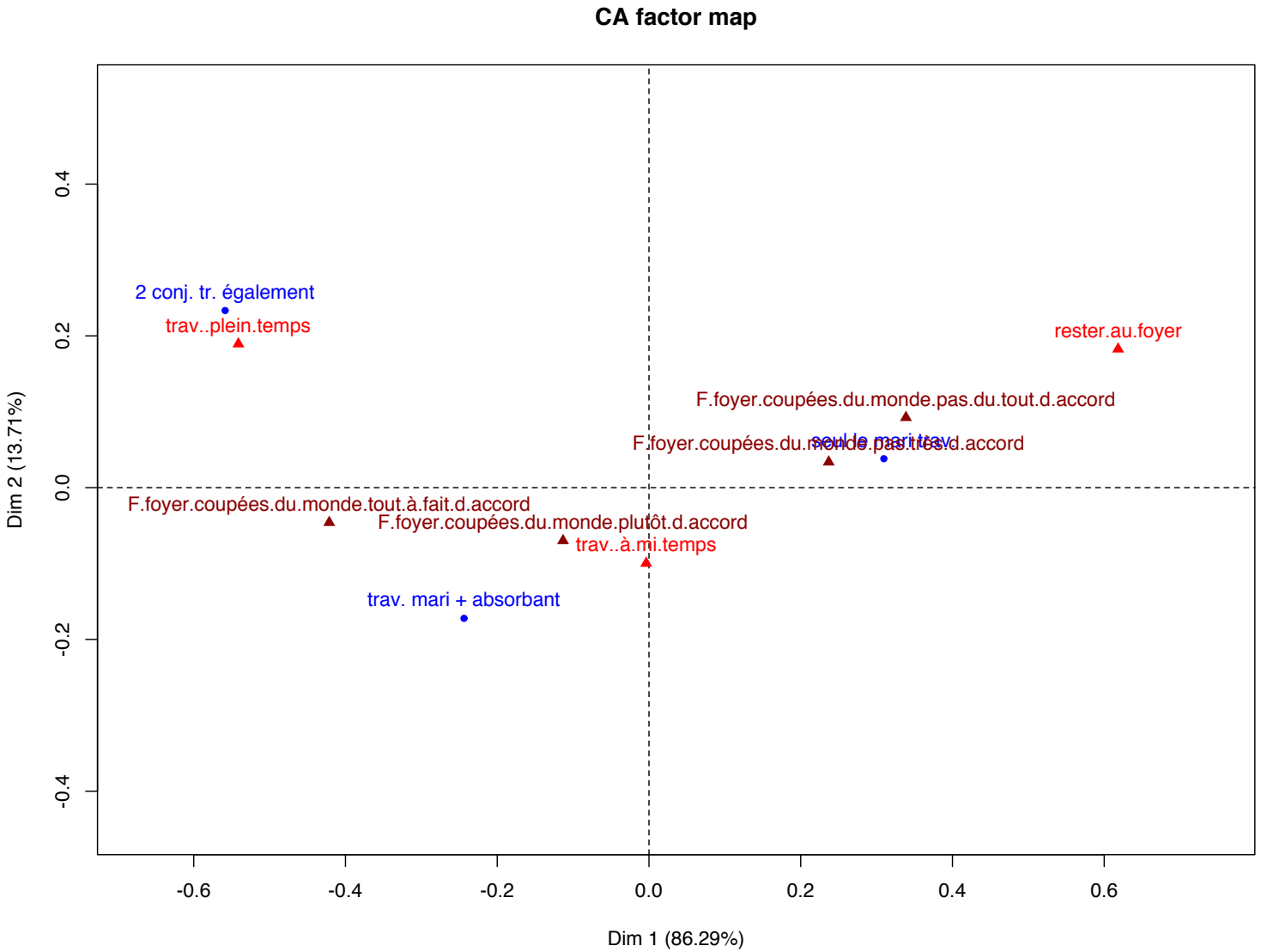
Exemple – données sur le travail féminin



REPONSES SIMULTANEES A DES QUESTIONS D'OPINION (fin)

		La famille idéale est celle où			
		Les deux conjoints travaillent également	Le mari a un métier plus absorbant que celui de la femme	Seul le mari travaille	Ensemble
Les femmes au foyer se sentent coupées du monde	Tout à fait d'accord	107	192	140	439
	Plutôt d'accord	75	175	215	465
	Pas très d'accord	40	100	254	394
	Pas du tout d'accord	39	88	299	426
		261	555	908	1724

Exemple – données sur le travail féminin



Plan

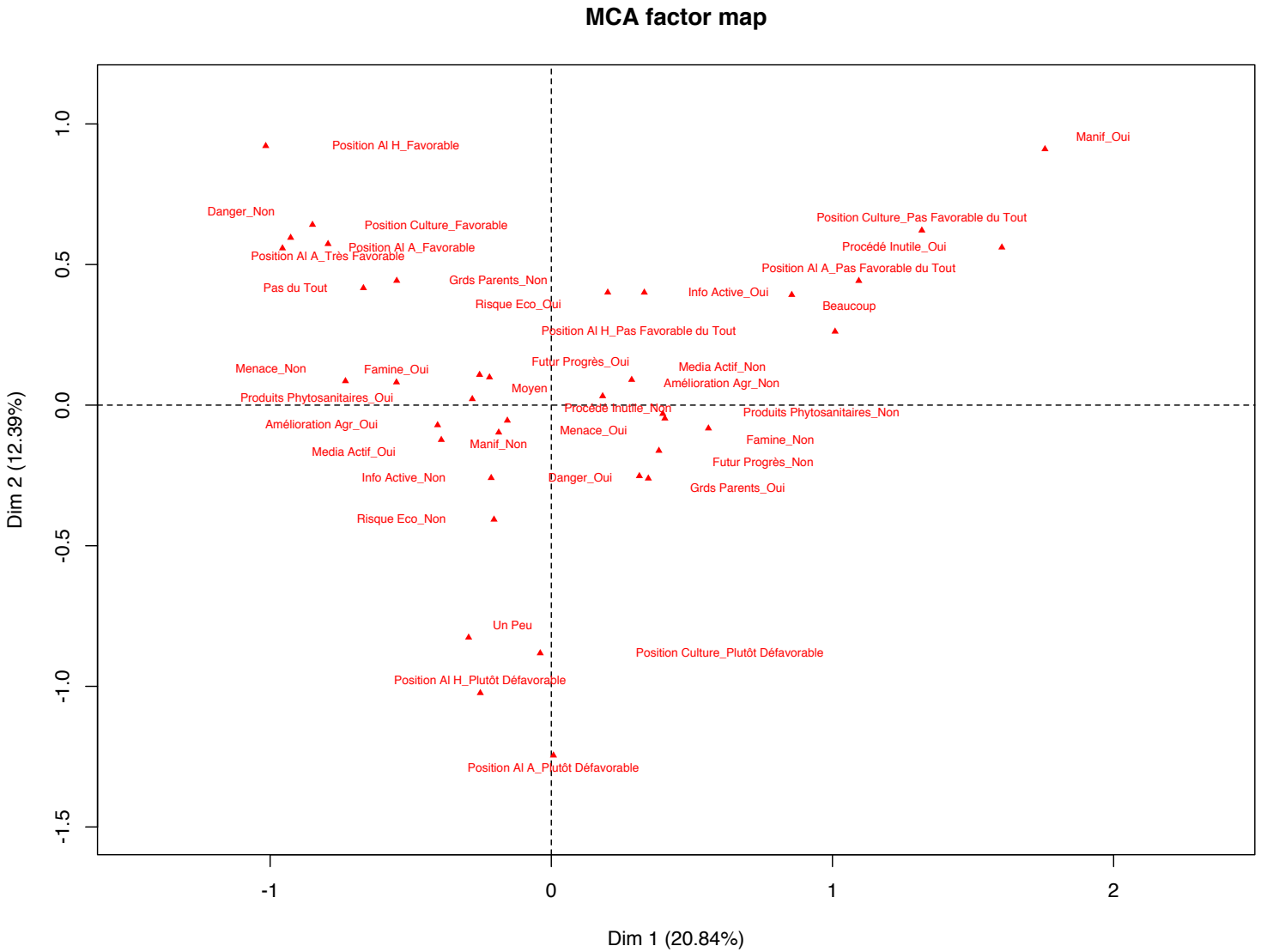
Analyse des correspondances multiples

- S'applique à des tableaux croisant n individus en ligne et p variables qualitatives en colonnes.
- On note x_i^j la modalité de l'individu i pour la variable j ayant K_j modalités.
- À partir de la donnée de $(x_i^j)_{1\leq i\leq n, 1\leq j\leq p}$, on construit le tableau disjonctif complet (TDC) :
  $T = (t_i^k)_{1\leq i\leq n, 1\leq k\leq K_1+\ldots+K_J}$, où

$$t_i^k = \begin{cases} 1 & \text{si l'individu } i \text{ possède la modalité } k \\ 0 & \text{sinon.} \end{cases}$$

- L'ACM consiste à faire une AFC sur le TDC.

Attitude à l'égard des OGM



Autres types d'analyse factorielles

- Analyse factorielle des données mixtes (AFDM) : analyse de tableaux croisant n individus et p variables quantitatives et qualitatives.
- Analyse de données textuelles : application de l'AFC. Le tableau croisé représente des textes en ligne et des mots en colonnes.
- ACP sur données fonctionnelles (voir cours de données fonctionnelles).
- ...