Enquêtes internationales
Analyse de réponses libres
en plusieurs langues

Mónica Bécue Bertaut

INED 29 mai 2022

Indice:

1. Motivation. Exemple “*Life*” (enquête Hayashi)
2. Codification. Tableau agrégé simple. Tableau agrégé multiple
3. Introduction à l’Analyse factorielle multiple pour tableaux de contingence (AFMTC) en 5 points
4. Application au corpus “*Life*” de Hayashi
5. En très bref, extension à la prise en compte de plusieurs variables, quantitatives ou qualitatives
6. Conclusion

1. Motivation: exemple

Enquête internationale
menée sous la direction du Professeur Hayashi
(Sasaki & Suzuki, fin des années 1980)

Trois échantillons sélectionnés
au Royaume Uni
en France
en Italie

ont répondu dans leur propre langue à la question :
"Qu'est-ce-qui est le plus important pour vous dans la vie?"
suivie de la relance:
"Quelles sont les autres choses très importantes pour vous?"

Données: réponses libres (non traduites) et
variables socio-économiques

2. Codification: a. un échantillon

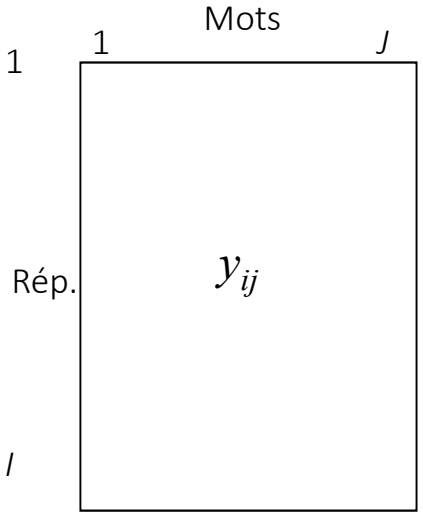


Tableau **Y**

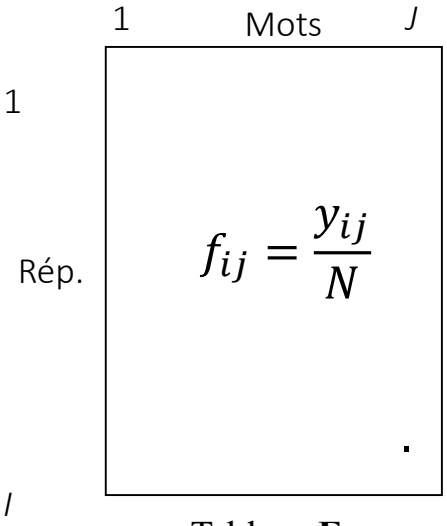


Tableau **F**

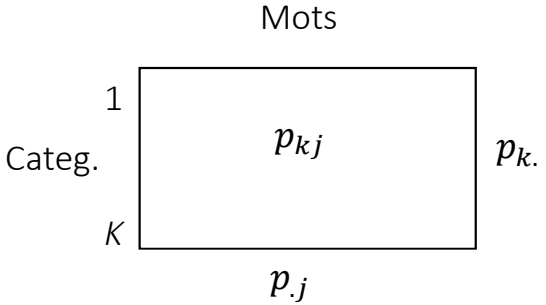


Tableau **P=(F'X)'**

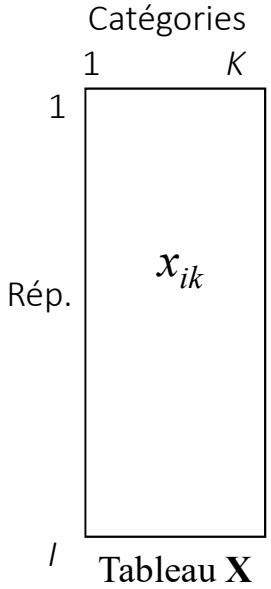
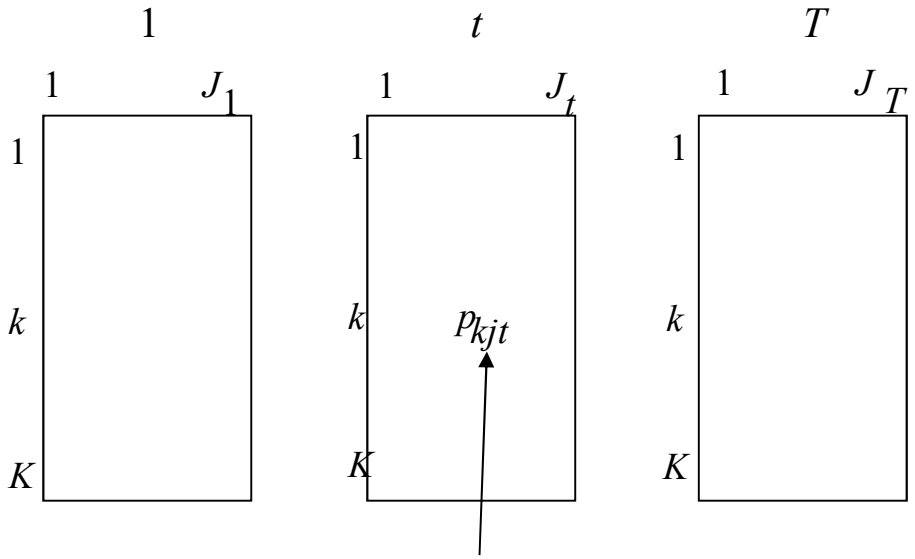


Tableau **X**

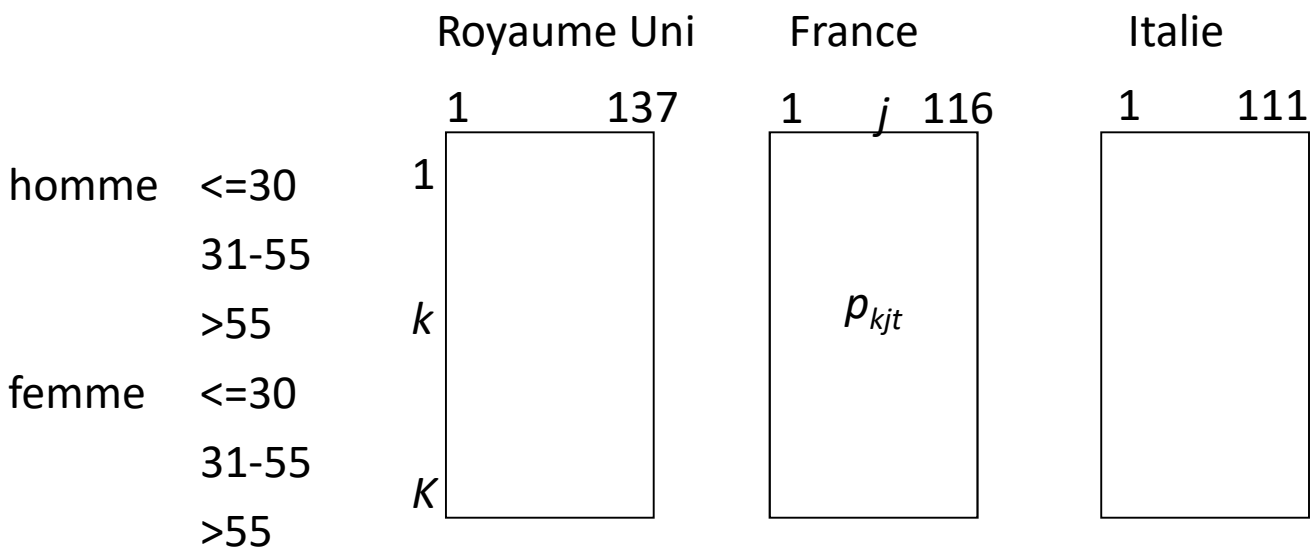
b. Plusieurs échantillons= plusieurs tableaux textuels agrégés

T tableaux de fréquence ayant les mêmes lignes (=catégories)

homme <30
31-55
>55
femme <30
31-55
>55



Proportion du mot j pour les répondants appartenant à la catégorie k dans l'échantillon t



Note: entre les questions que l'on se pose:

Quels mots caractérisent une catégorie donnée ?

Quelles catégories caractérisent un mot donné ?

Quels mots sont similaires du point de vue de leur catégorie d'utilisateur ?

Quelles catégories sont similaires du point de vue de l'utilisation du vocabulaire ?

dans tous les pays (structures communes)

dans un pays donné (structures spécifiques)

Méthodologies usuelles pour l'analyse de tableaux de contingence multiples

Analyses séparées des différents tableaux

1) les structures communes peuvent ne pas correspondre aux axes principaux
2) difficile de lire simultanément les différents tableaux

Analyse des correspondances du tableau multiple

L'influence de l'inertie entre les tableaux doit être supprimée (quotas)

Chaque tableau peut influencer différemment le premier axe global (équilibrer ?)

Manque de référence à la structure sur les lignes propre à chaque tableau

3. Introduction à l'AFMTC en 5 points

1) Équivalence entre l'Analyse des correspondances (AC) et une analyse factorielle (ou Analyse en composantes principales/ACP) particulière

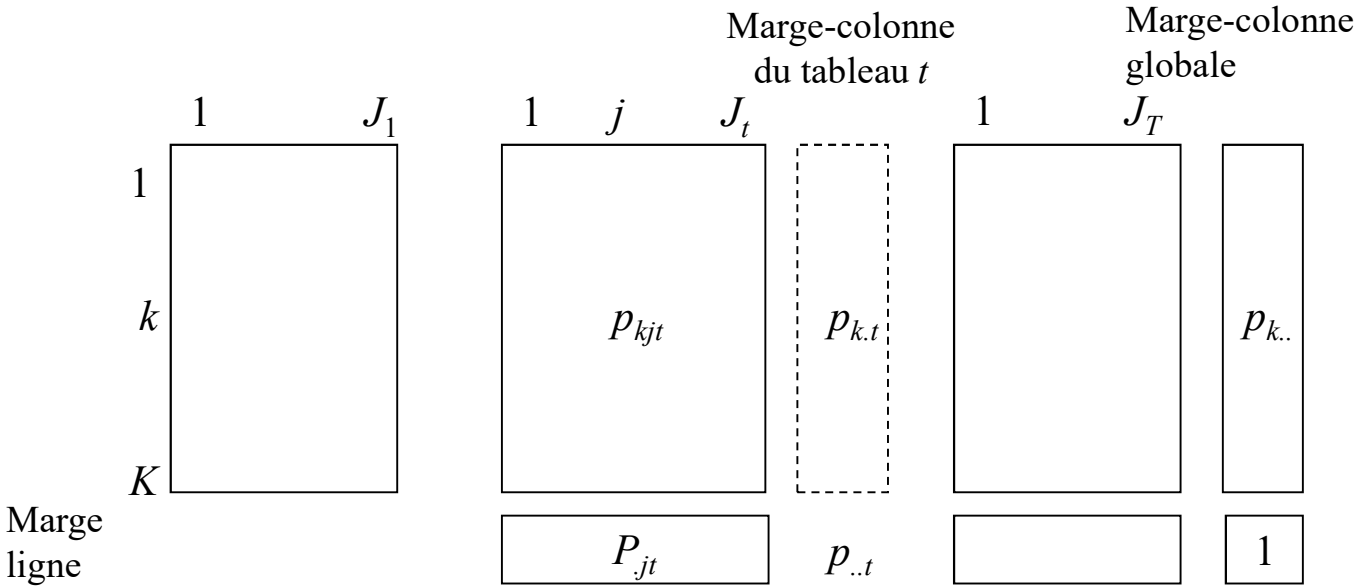
2) Généralisation de l'AC à des modèles autres que l'indépendance (Escofier, 1984)

3) Analyse des correspondances internes (ACI) (Benzécri 1983 ; Escofier & Drouet 1983 ; Cazes et al.)

4) Analyse factorielle multiple (AFM) (Escofier & Pagès 1984)

5) AFM pour tableaux de contingence (AFMTC)
   principe
   application

Structure de données: tableau multiple et marges internes et globale



Méthodologies usuelles pour l'analyse de tableaux de contingence multiples

Analyses séparées des différents tableaux

1) les structures communes peuvent ne pas correspondre aux axes principaus
2) difficile de lire simultanément les différents tableaux

Analyse des correspondances du tableau multiple

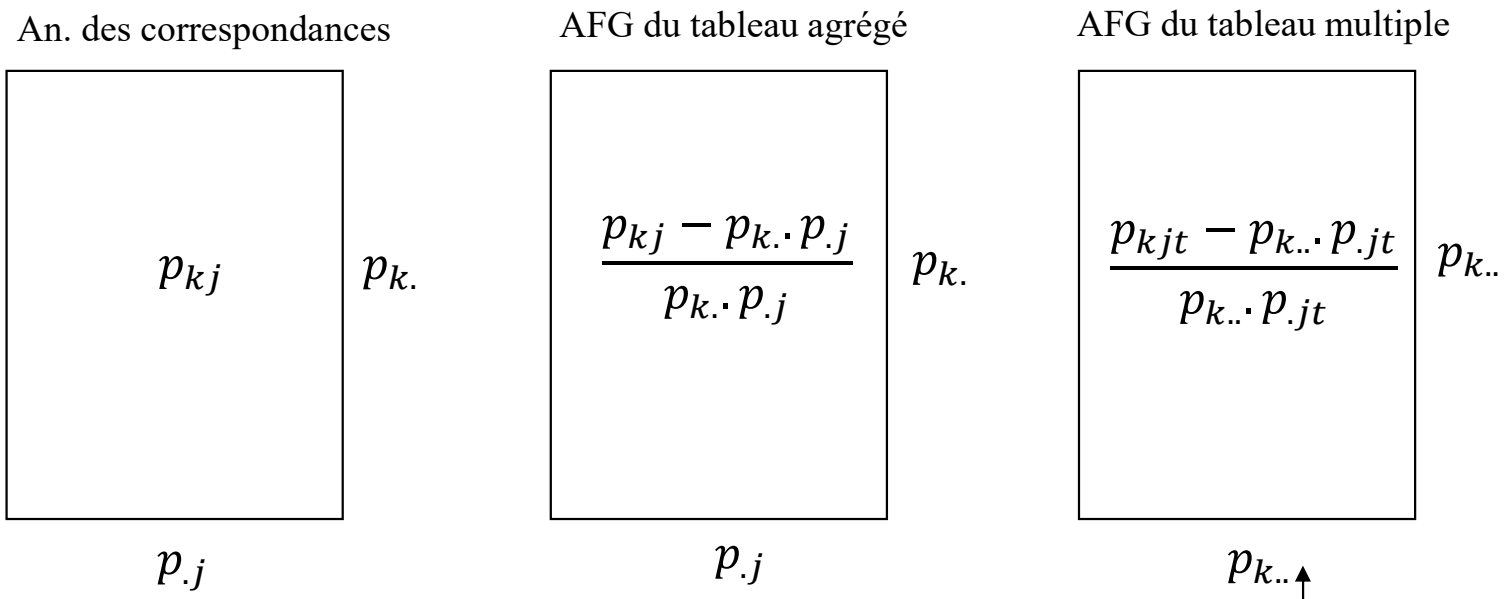
Influence de l'inertie entre-tableaux qui doit être supprimée (quotas)

Chaque tableau peut influencer différemment le premier axe global (équilibrer?)

Manque de référence à la structure sur les lignes propre à chaque tableau

3.1. AC équivalente à une ACP particulière

(AFG: analyse factorielle générale cf. Lebart = ACP en deux métriques)



$\{p_{k.}\,;\, k = 1,\ldots, K\}$ poids des lignes (métrique dans l'espace des colonnes)

$\{p_{.j}\,;\, j = 1,\ldots, J\}$ poids des colonnes (et métrique dans l'espace des lignes)

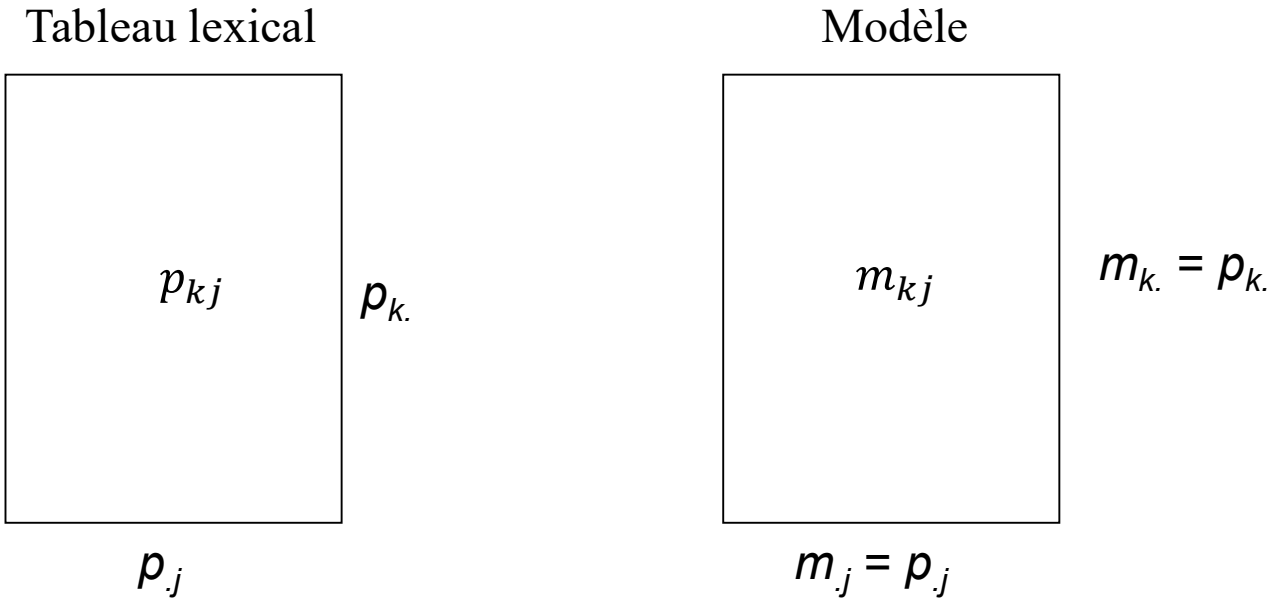
$\{p_{k..}\,;\, k = 1,\ldots, K\}$ poids des lignes (et métrique dans l'espace des colonnes)

$\{p_{.jt}\,;\, j = 1,\ldots, J_t;\ t = 1,\ldots, T\}$ poids des colonnes (et métrique dans l'espace des lignes)

3.2. Généralisation de l'AC à d'autres modèles (mêmes marges que le tableau étudié)

$\{m_{kj}\ ;\ i=1,\ldots,\ K\ ;\ j=\ 1,\ldots,\ J\}$ $\qquad$ $(m_{k.} = p_{k.}\ ;\ p_{.j} = p_{.j}\ )$

(Escofier, 1984)



AC relativement au modèle **M** = ACP du tableau

de terme général $\dfrac{\boldsymbol{p_{kj} - m_{kj}}}{\boldsymbol{p_{k.}\, p_{.j}}}$ avec $\{p_{k.}\}$ poids des lignes (métrique dans l'espace des colonnes)

$\{p_{.j}\}$ poids des colonnes (et métrique dans l'espace des lignes)

3.3. Modèle d'indépendance intra-tableaux

=Indépendance entre lignes et colonnes à l'intérieur de chaque tableau *t*

(Escofier &Drouet, Cazes, Benzécri)

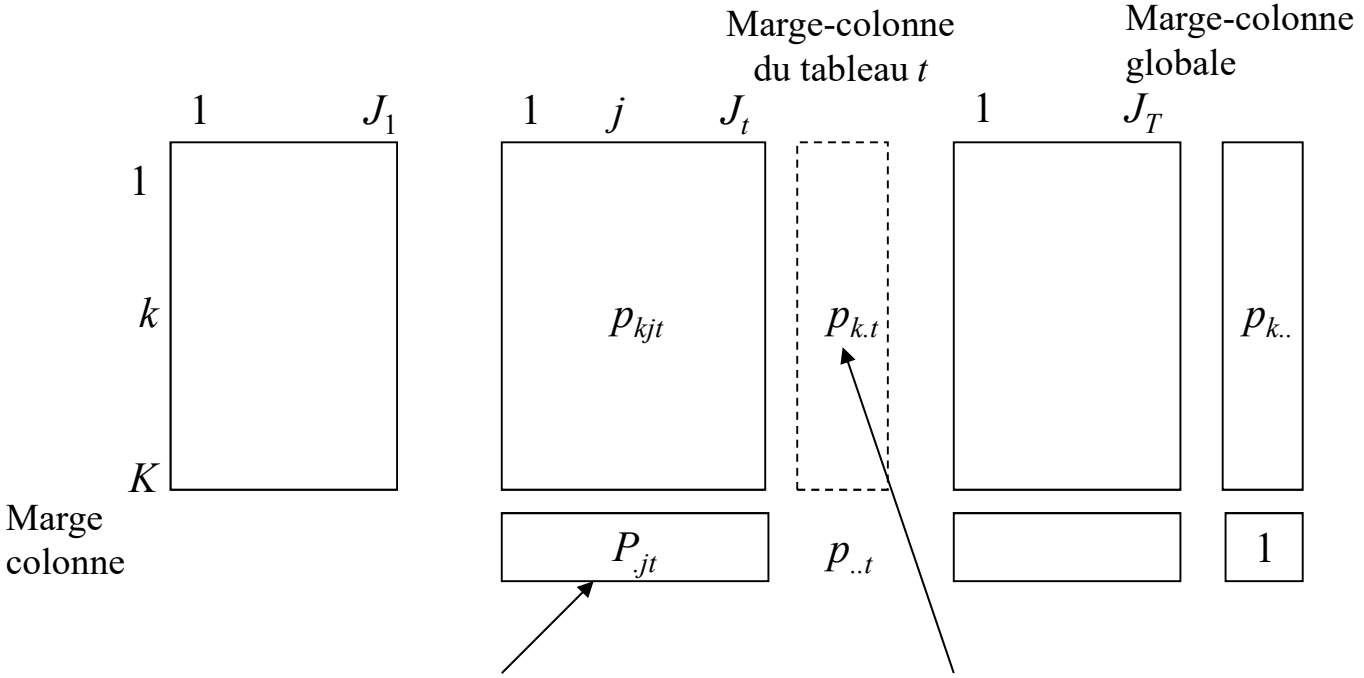
AC usuelle

$$m_{kjt} = p_{k..} \cdot p_{.jt}$$

versus

AC interne (ACI)

$$m_{kjt} = \left(\frac{p_{k.t}}{p_{..t}}\right) \cdot p_{.jt}$$



Analyse des correspondances internes (ACI)

= AC en référence au modèle d'indépendance entre lignes et colonnes

à l'intérieur de chaque tableau

$$m_{kjt} = \left(\frac{p_{k.t}}{p_{..t}}\right) \cdot p_{.jt}$$

On analysera par une ACP le tableau de terme general:

Profil de la colonne *jt*

Profil moyen des colommes du tableau *t*

$$\frac{p_{kjt} - m_{kjt}}{p_{k..}p_{.jt}} = \frac{p_{kj} - \left(\frac{p_{k.t}}{p_{..t}}\right) \cdot p_{.jt}}{p_{k..}p_{.jt}} = \frac{1}{p_{k..}}\left[\frac{p_{kjt}}{p_{.jt}} - \frac{p_{k.t}}{p_{..t}}\right]$$

Les poids et métriques étant:

$\{p_{k..}\ ;\ k = 1,\ldots, K\}$ poids des lignes (et métrique dans l'espace des colonnes)

$\{p_{.jt}\ ;\ j = 1,\ldots, J_t;\ t = 1,\ldots, T\}$ poids des colonnes (et métrique dans l'espace des lignes)

Illustration de l'effet des marges

TABLEAU 1

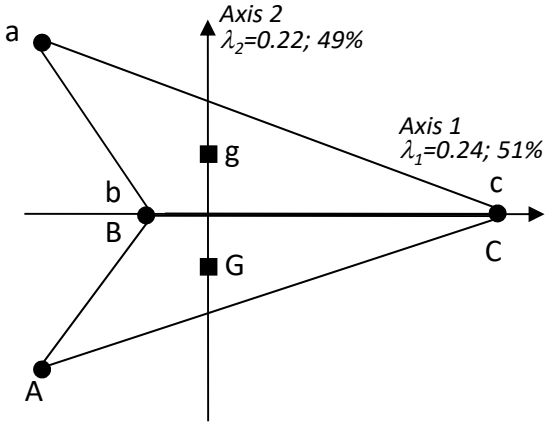
	A	B	C	Σ
C1	120	60	20	200
C2	20	60	20	100
C3	10	30	60	100
Σ	150	150	100	

TABLEAU 2

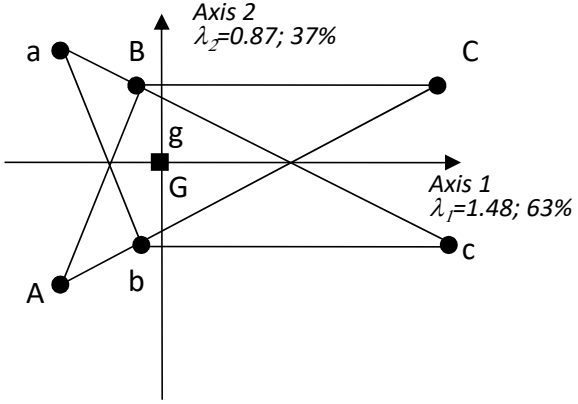
	a	b	c	Σ
C1	20	60	20	100
C2	120	60	20	200
C3	10	30	60	100
Σ	150	150	100	

Données

AC du tableau global



AC interne



Méthodologies usuelles pour l'analyse de tableaux de contingence multiples

Analyses séparées des différents tableaux

1) les structures communes peuvent ne pas correspondre aux axes principaus
2) difficile de lire simultanément les différents tableaux

Analyse des correspondances du tableau multiple

Influence de l'inertie entre-tableaux qui doit être supprimée (quotas)

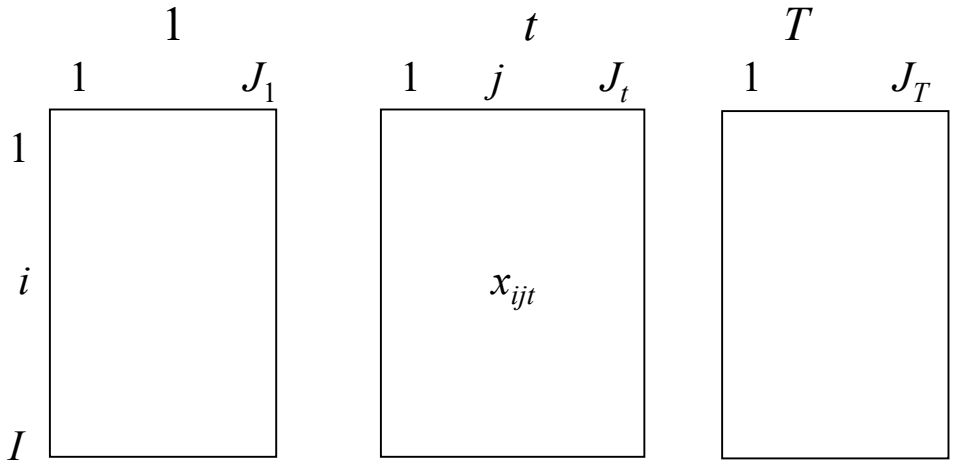
→ *Chaque tableau peut influencer différemment le premier axe global (équilibrer ?)*

Manque de référence à la structure sur les lignes propre à chaque tableau

3.4. Analyse factorielle multiple (AFM)

Structure de données :

Un ensemble d'individus est décrit par plusieurs ensembles de variables (quantitatives ou qualitatives)



Analyse factorielle multiple (AFM)

Le noyau de l'AFM

est une Analyse en Composantes Principales (ACP)
appliquée à l'ensemble des tableaux (analyse globale)

L'AFM équilibre l'inertie des sous-tableaux dans l'analyse globale

Soit λ_1^t la première valeur propre de l'ACP appliquée au groupe *t*

le poids de chaque variable appartenant au groupe *t* est divisé par λ_1^t

l'inertie axiale la plus élevée de chaque groupe est normalisée à 1

L'AFM fournit

les résultats classiques de l'ACP (les *T* groupes de variables étant équilibrés) :

coordonnées, contributions et cosinus carrés des lignes et des colonnes
...

divers outils pour comparer les structures induites par les *T* groupes de variables

une représentation superposée des *T* structures sur les lignes
(comme dans l'analyse de Procuste généralisée)

3.5. Analyse factorielle multiple pour tableaux de contingence (AFMTC)

Principe

L'AFMTC combine l'ACI (qui "gère" les différences entre les marges-colonne)
et l'AFM (qui équilibre l'influence des sous-tableaux dans l'analyse globale).

L'AFMTC applique une ACP:

au tableau analysé par l'ACI

en conservant les poids des lignes utilisés dans l'ACI ($p_{k..}$) ;

en modifiant les poids des colonnes comme dans l'AFM.

Méthodologies usuelles pour l'analyse de tableaux de contingence multiples

Analyses séparées des différents tableaux

1) les structures communes peuvent ne pas correspondre aux axes principaus
2) difficile de lire simultanément les différents tableaux

Analyse des correspondances du tableau multiple

Influence de l'inertie entre-tableaux qui doit être supprimée (quotas)

Chaque tableau peut influencer différemment le premier axe global (équilibrer ?)

Manque de référence à la structure sur les lignes propre à chaque tableau

Représentation des catégories selon les réponses de chaque pays sur les axes globaux

		Mots Anglais	Mots Français	Mots Italiens
lignes actives	hommes <=30	fréquence	fréquence	fréquence
	hommes 31-55			
	hommes >55			
	femmes <=30			
	femmes 31-55			
	femmes >55			
lignes Supp.	hommes <=30	fréquence	0	0
	hommes 31-55			
	hommes >55			
	femmes <=30			
	femmes 31-55			
	femmes >55			
	…	….	….	….
	hommes <=30	0	0	fréquence
	hommes 31-55			
	hommes >55			
	femmes <=30			
	femmes 31-55			
	femmes >55			

4. Application au corpus "*Life*"

Tailles des corpus

Reino Unido

TextData summary:

	Before	After
Doc.	1043.00	6.00
Occ.	13917.00	6415.00
Words	1334.00	137.00

France

TextData summary:

	Before	After
Doc.	1009.00	6.0
Occ.	14153.00	5517.0
Words	1246.00	116.0

Italie

TextData summary

	Before	After
Doc.	1048.00	6.0
Occ.	6156.00	3945.0
Words	785.00	111.0

Mots les plus fréquents

Family, health, happiness, money, life, job

Santé, travail, famille, enfants, argent, vie

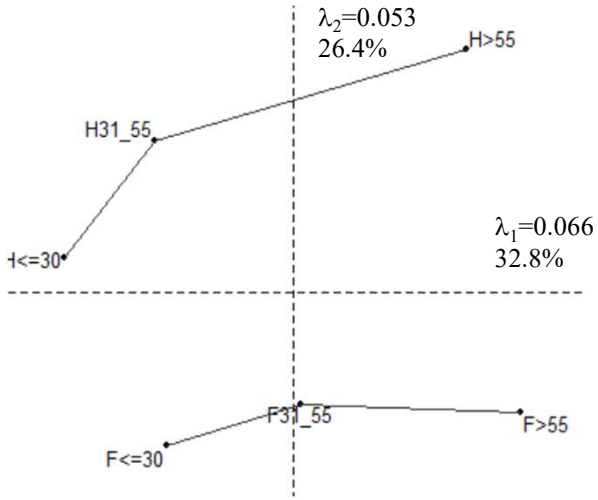
Salute, famiglia, lavoro amore, serenità, soldi

Inerties observées dans les analyses séparées

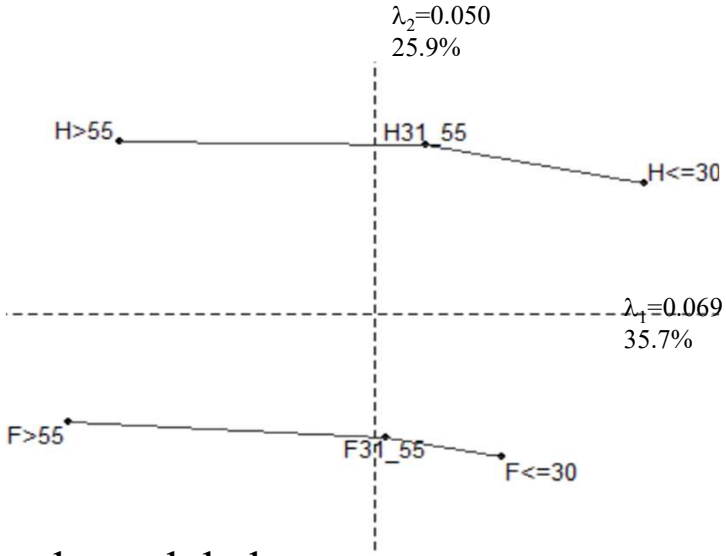
	Inertie globale (ϕ^2)	Inertie sur l'axe 1
Royaume Uni	0.202	.066
France	0.194	.069
Italie	0.198	.072

Résultats obtenus avec les packages R Xplortext et FactoMineR

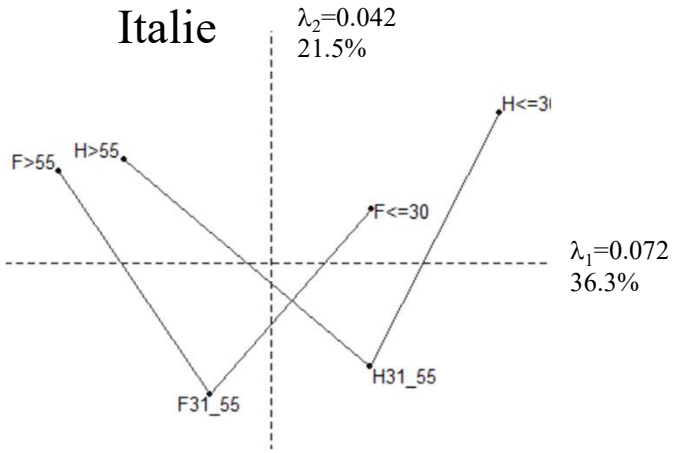
AC séparées



France



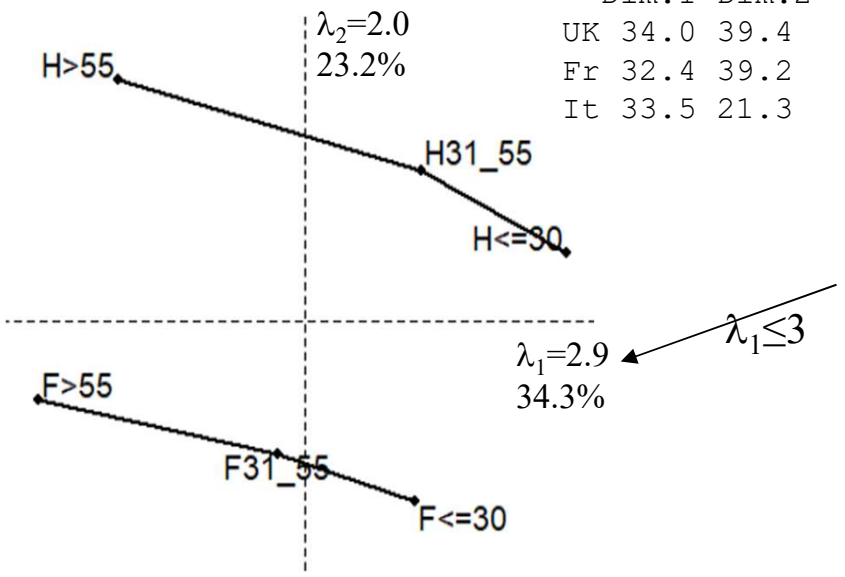
Italie

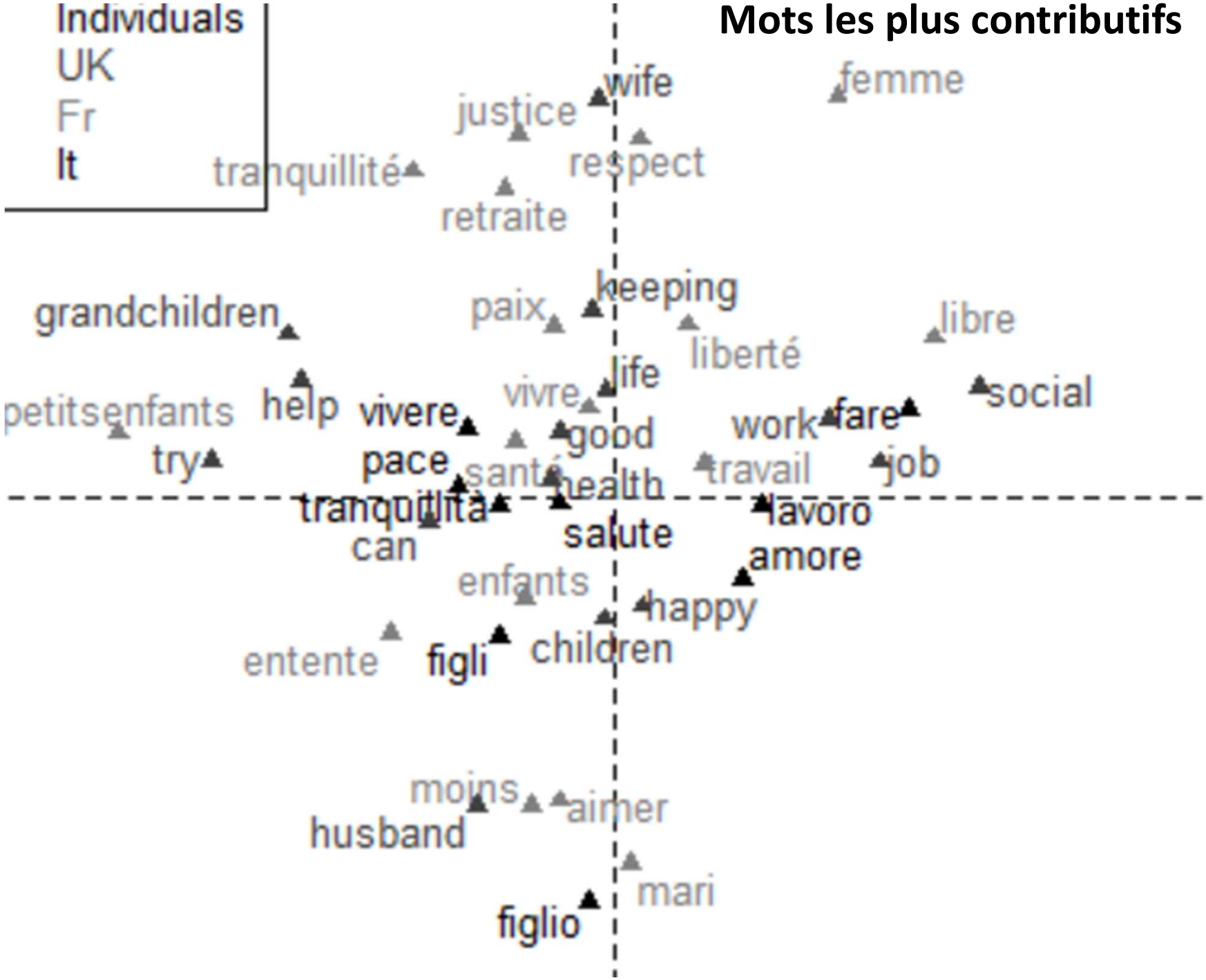


Analyse globale

Contr. Groupes

	Dim.1	Dim.2
UK	34.0	39.4
Fr	32.4	39.2
It	33.5	21.3

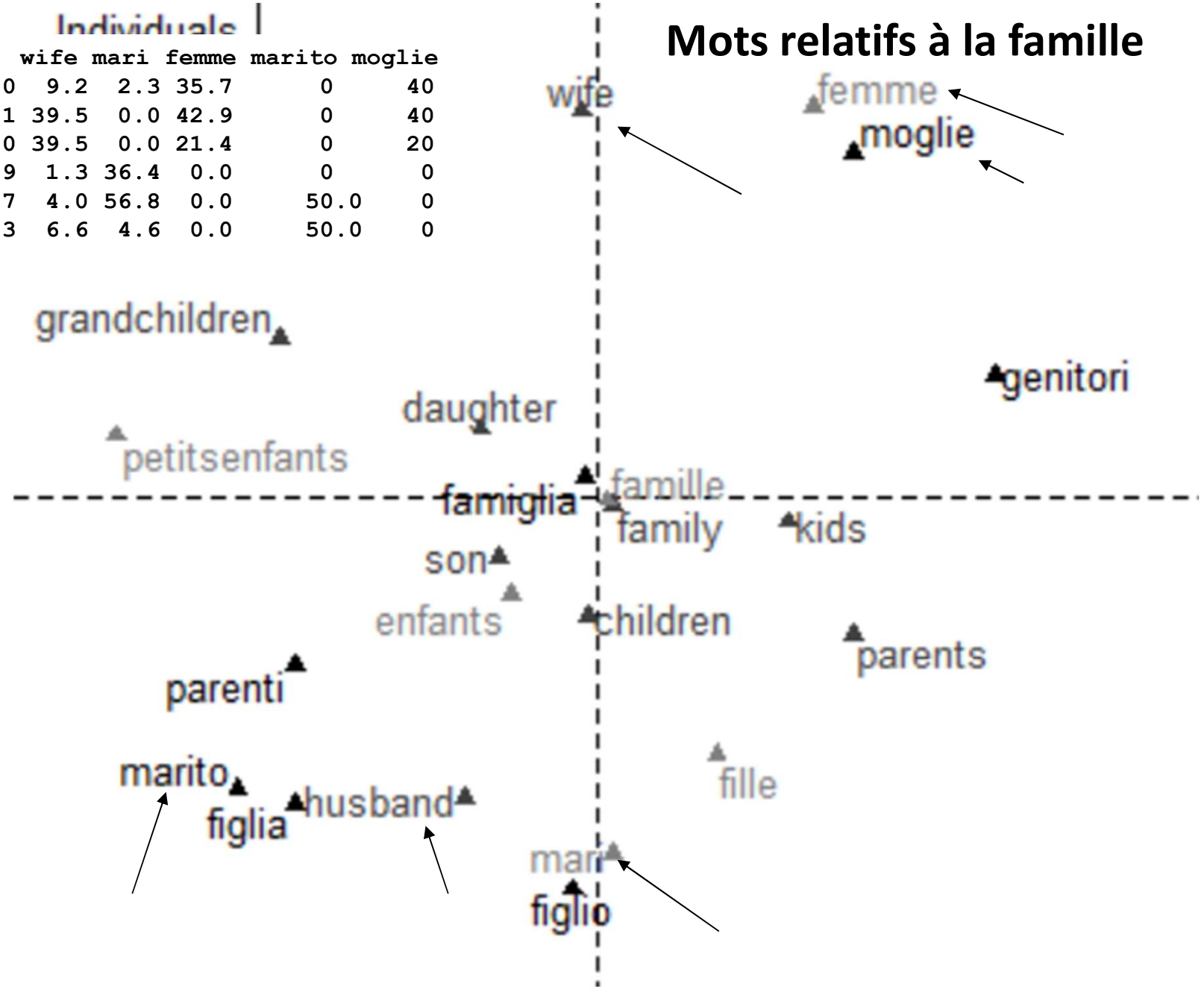



Individuals
UK
Fr
It
Mots les plus contributifs
wife
femme
justice
respect
tranquillité
retraite
keeping
paix
grandchildren
libre
liberté
life
vivre
social
petitsenfants
help
vivere
work
fare
good
try
pace
santé
travail
job
health
tranquillità
lavoro
salute
can
amore
enfants
happy
entente
figli
children
moins
aimer
husband
mari
figlio

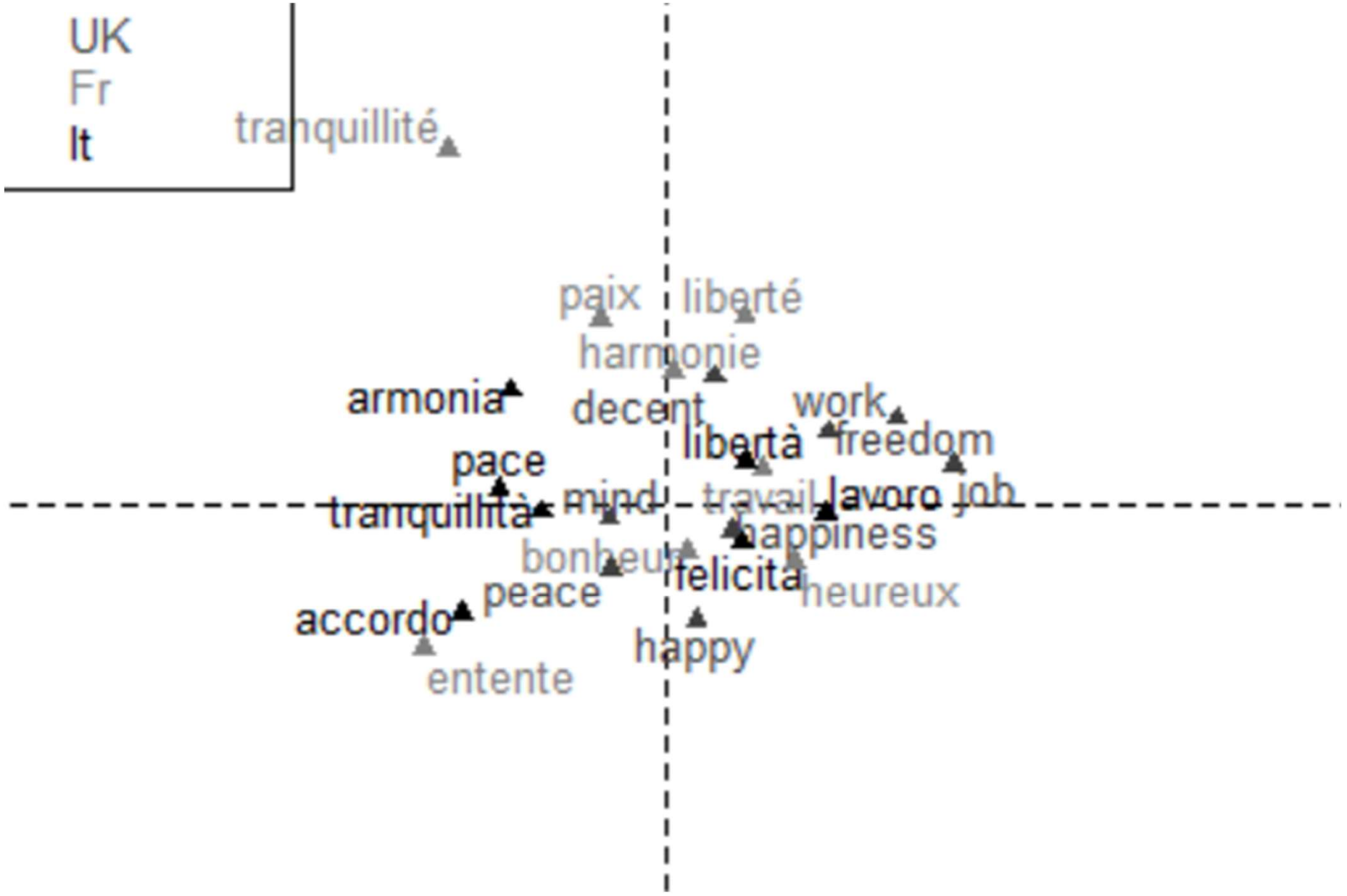
Individuals

	husband	wife	mari	femme	marito	moglie
H<=30	0.0	9.2	2.3	35.7	0	40
H31_55	3.1	39.5	0.0	42.9	0	40
H>55	1.0	39.5	0.0	21.4	0	20
F<=30	21.9	1.3	36.4	0.0	0	0
F31_55	41.7	4.0	56.8	0.0	50.0	0
F>55	32.3	6.6	4.6	0.0	50.0	0

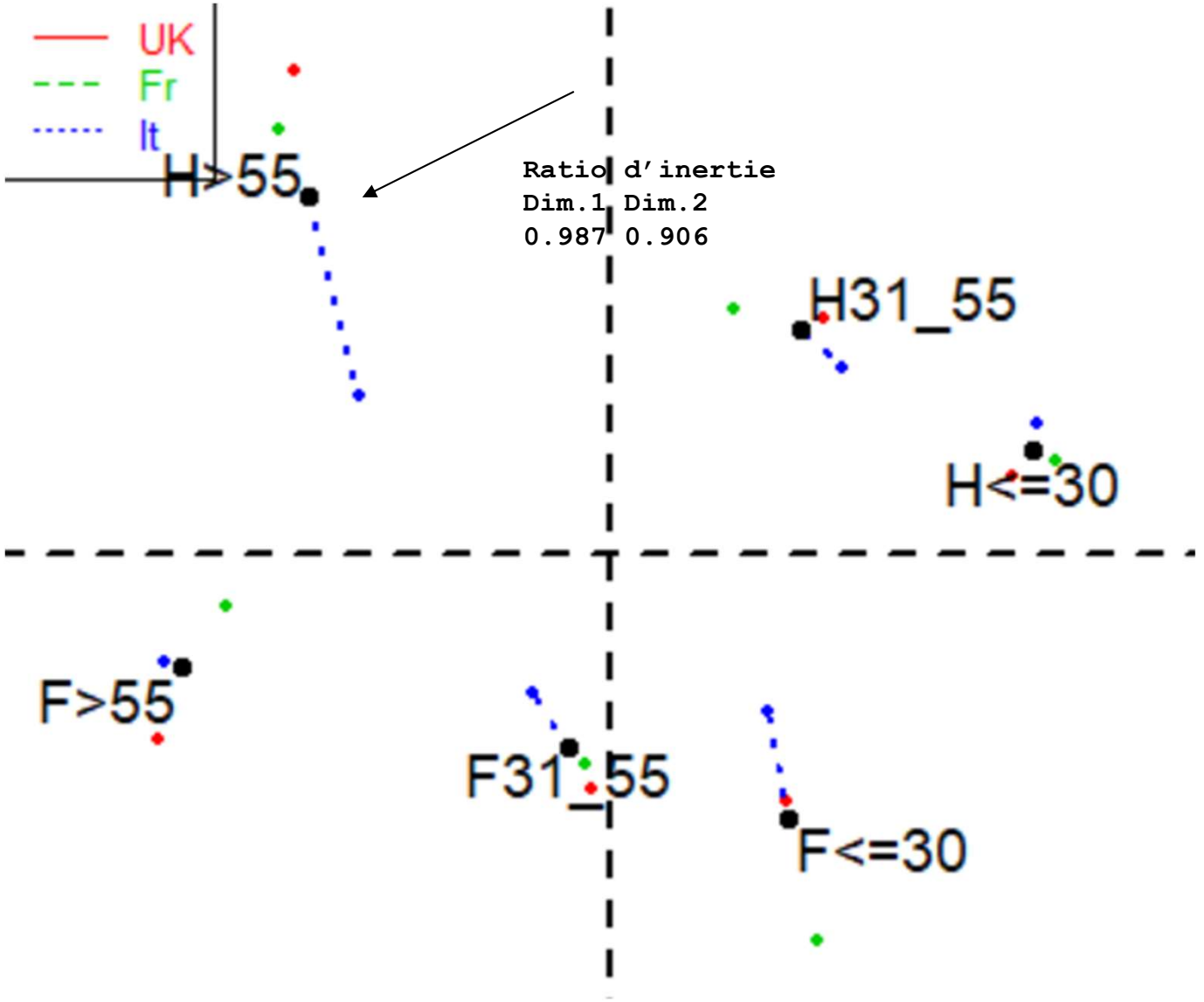
Mots relatifs à la famille

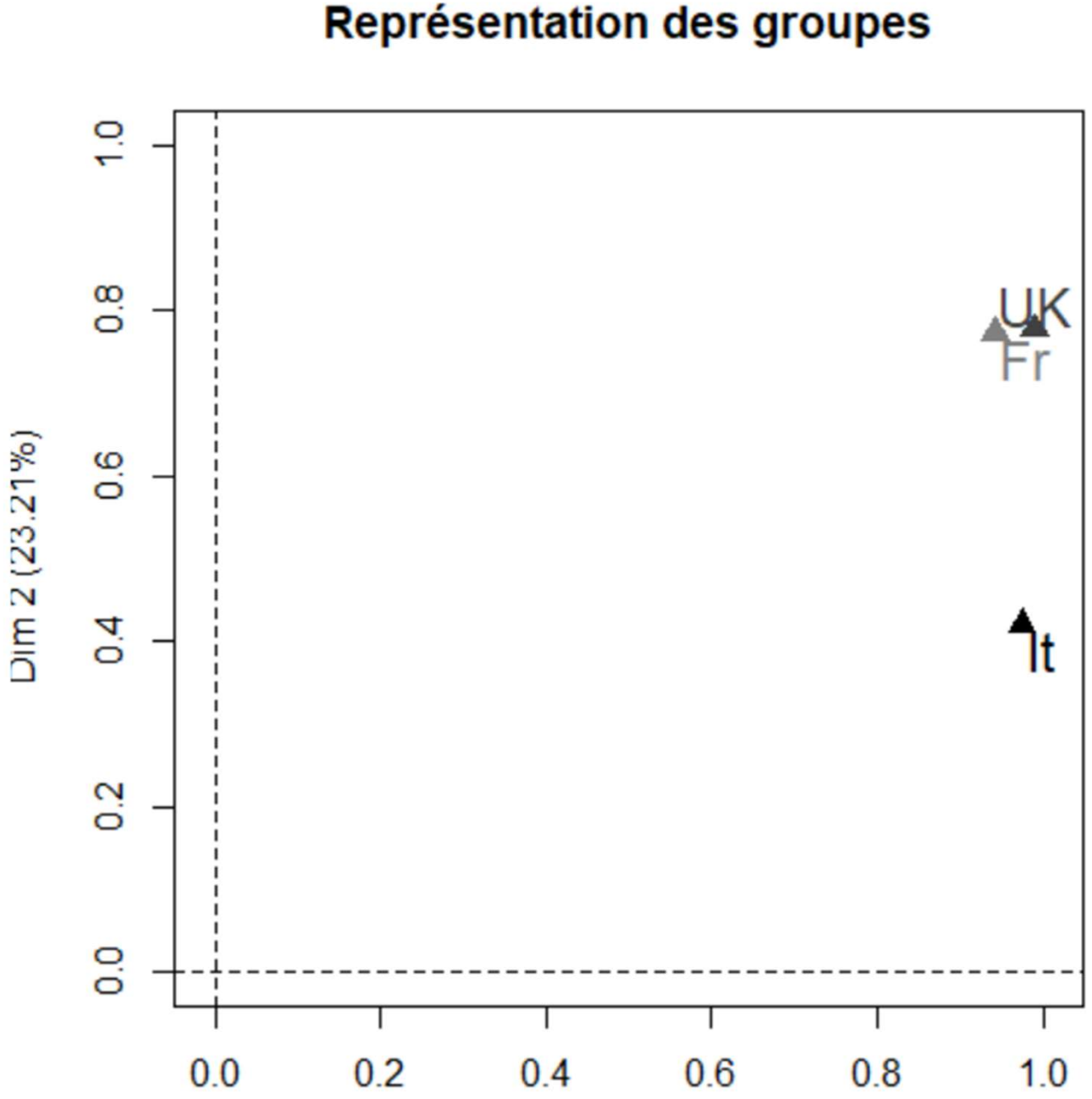


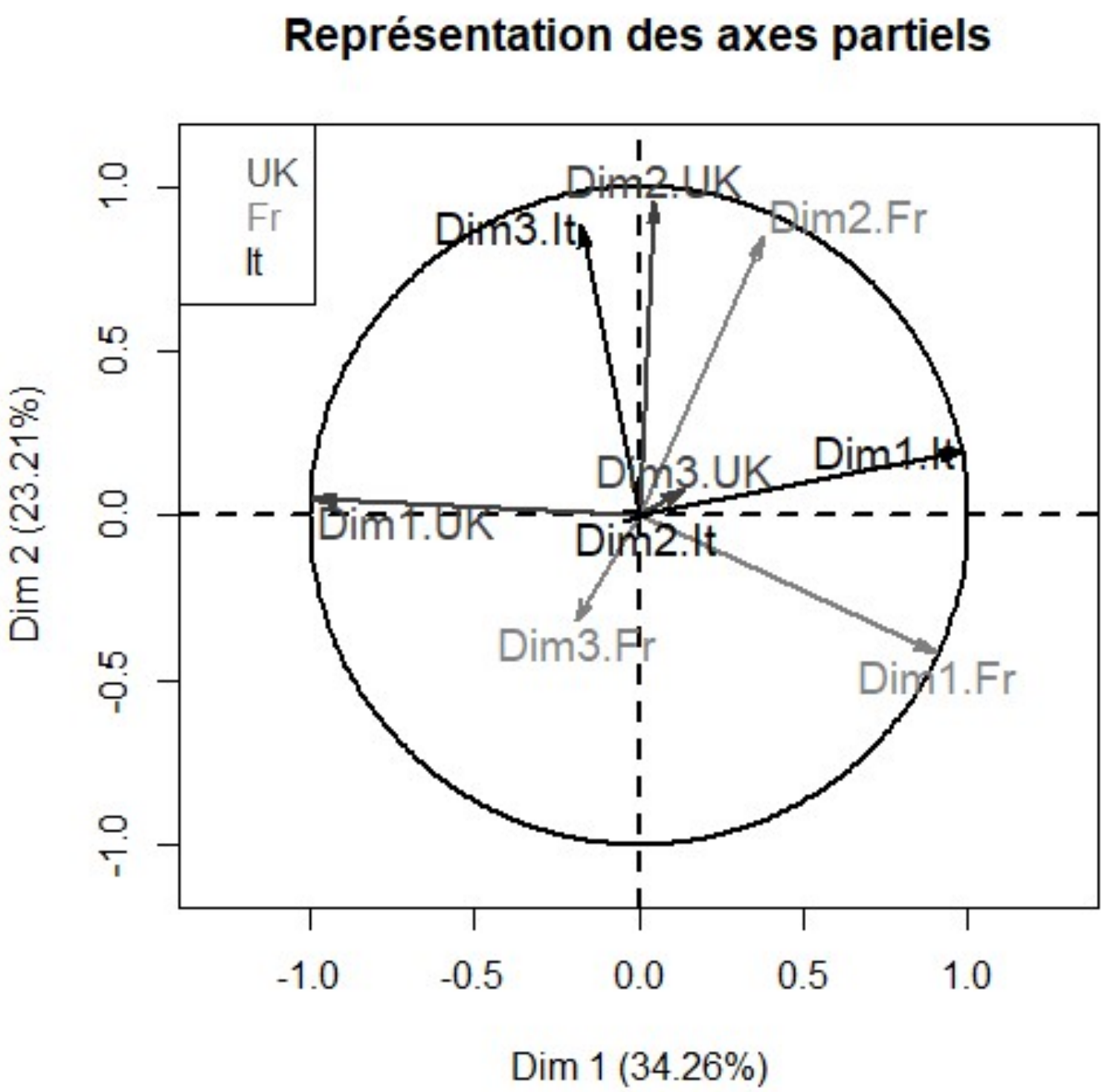
Mots relatifs à des préoccupations générales



Représentation superposée des catégories




Représentation des groupes
Dim 2 (23.21%)
0.0
0.2
0.4
0.6
0.8
1.0
UK
Fr
It


Représentation des axes partiels
UK
Fr
It
Dim2.UK
Dim3.It
Dim2.Fr
Dim1.It
Dim3.UK
Dim1.UK
Dim2.It
Dim3.Fr
Dim1.Fr
1.0
0.5
0.0
-0.5
-1.0
Dim 2 (23.21%)
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (34.26%)

5. Extension à des tableaux agrégés généralisés multiples

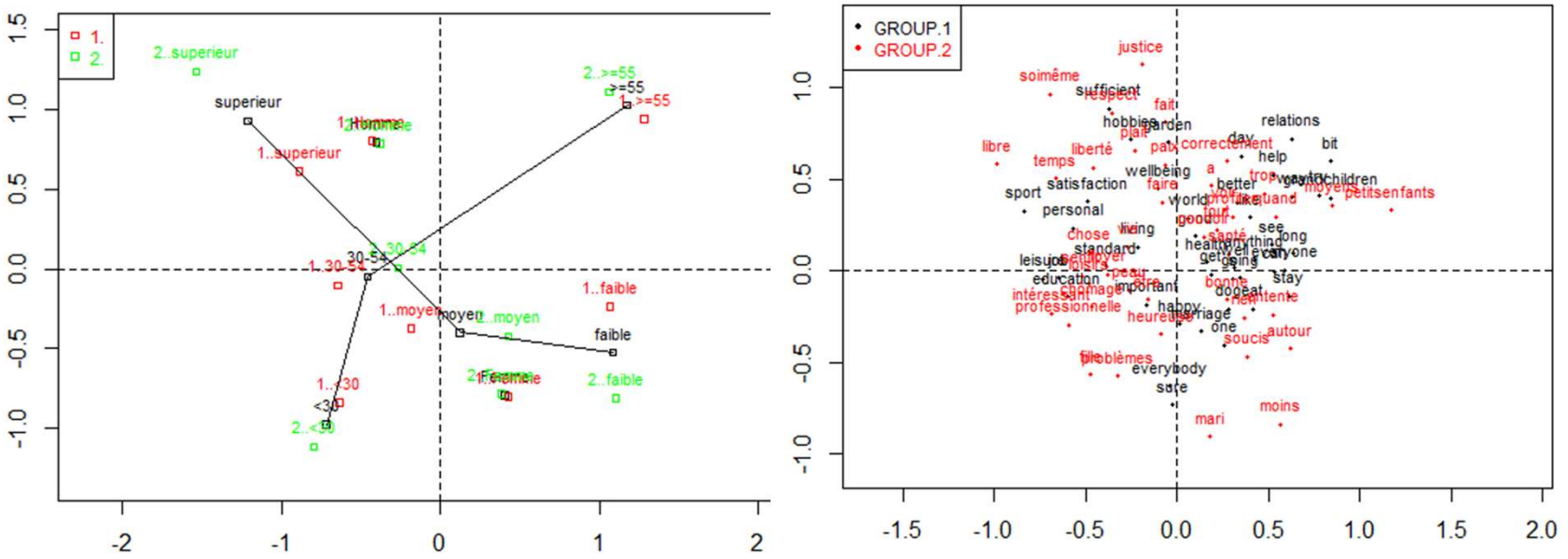
Travaux en cours avec B. Kostov, R. Alvarez-Esteban et F. Husson

Permet d'utiliser plusieurs variables quantitatives ou qualitatives. Analyses séparées qui sont des CA-GALT. Cette méthode permet de démêler l'influence de plusieurs variables contextuelles sur les choix lexicaux des répondants.

On analyse alors des tableaux lexicaux agrégés généralisés au moyen de la méthode AFM-TLAG (MFA_GALT).

Exemples de résultats....

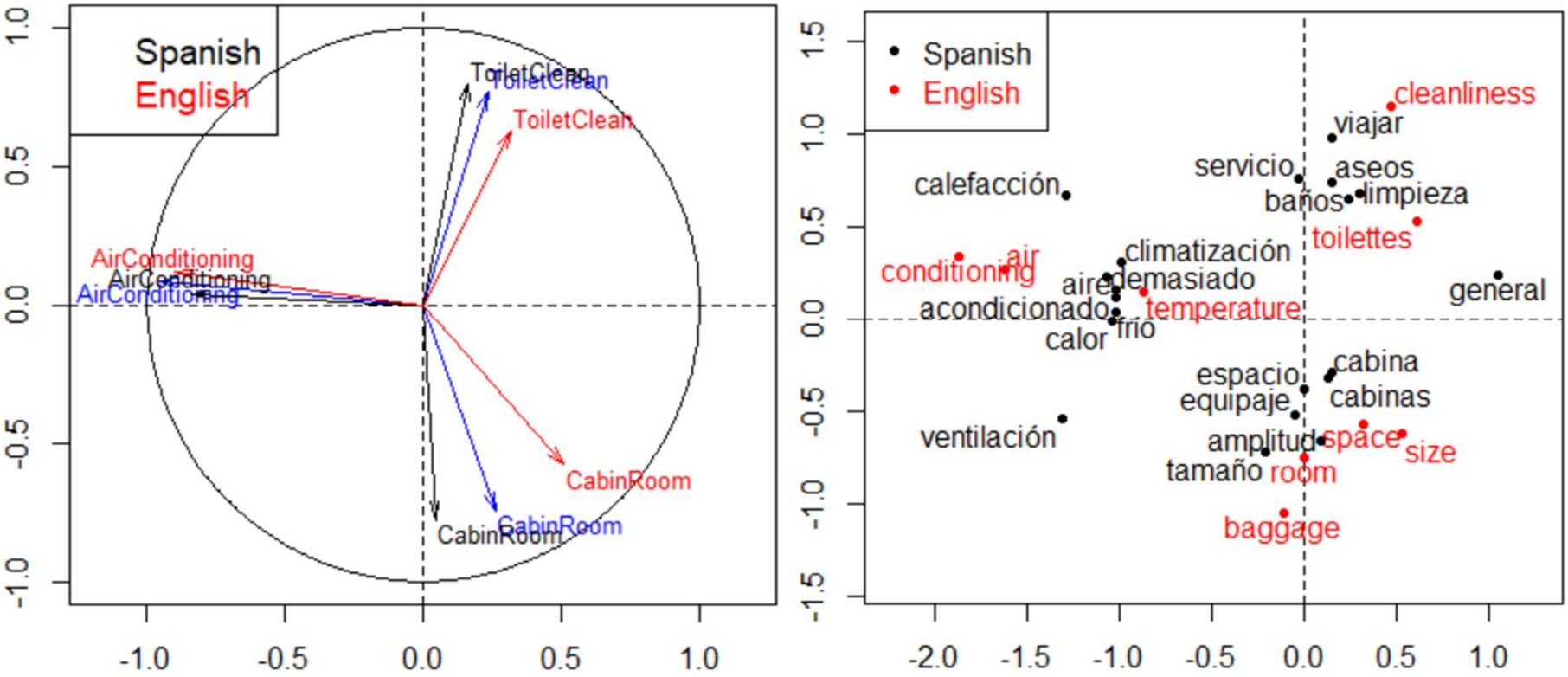
Données “*Life*” (variables qualitatives)



Analyses séparées= CA-GALT~ACC

Permet de démêler l’influence des différentes variables sur les choix lexicaux

Données “*Trains de nuit*” (variables quantitatives)



6. Conclusion

1) l'AFMTC est une méthode factorielle
adaptée au traitement de tableaux de contingence multiple

2) Cette méthode prend en compte
les différences entre les marges-colonne
ainsi qu'entre les intensités de structure d'un tableau à l'autre (comme en AFM)

3) L'AFMTC fournit
Les résultats classiques de toute analyse factorielle
- coordonnées, contributions et qualité de representations des lignes et des colonnes …

- des outils pour comparer les tableaux de contingence
comme la représentation superposée des lignes partielles

4) On peut étendre la méthode à plusieurs variables quantitatives ou qualitatives

5) On traiter simultanément des questions ouvertes et des questions fermées

Références

BÉCUE BERTAUT M., PAGÈS J. 2004 - A principal axes method for comparing contingency tables: MFACT. Comp. Statistics & Data Análisis CSDA, Vol 45/3 pp 481-503

BÉCUE-BERTAUT, M. & PAGÈS J. 2008 - Analysis of a mixture of quantitative, categorical and frequency data through an extension of multiple factor analysis. Application to survey *data."* aceptado para publicación por *Computational Statistics and Data Analysis*

BÉCUE-BERTAUT, M. & PAGÈS J. 2015 – Correspondence analysis of textual data involving contextual information: CA-GALT on principal components. Adv Data Anal Classif. pp. 125-142.

BÉCUE-BERTAUT, M. , PAGÈS J. & KOSTOV B. 2014 -Untangling the influence of several contextual variables on the respondents' lexical choices. A statistical approach. *SORT-Statistics and Operations Research Transactions*, 38 (2) pp. 285-02

BÉCUE-BERTAUT, M - 2018 Analyse Textuelle avec R Presses Universitaires de Rennes.

BÉCUE-BERTAUT, M - 2018 Textual Data Science with R. Taylor & Francis.

ALVAREZ- ESTEBAN, BÉCUE-BERTAUT, KOSTOV, SÁNCHEZ-ESPIGARES - CRAN **package Xplortext** – https://cran.r project.org/web/packages/Xplortext/index.html

https://xplortext.unileon.es/ WEB avec scripts et bases correspondant aux manuels