

Analyses Factorielles

UT2J - Frédéric Ferraty

L3MIASHS 2021 - 2022

Contenu de ce cours

- Analyse en Composantes Principales (ACP)
- Analyse Factorielle des Correspondances (AFC)
- Analyse Factorielle des Correspondances Multiples (AFCM)

Document en ligne

- Polycopié du cours sur IRIS
- Introduction à l'analyse d'enquêtes avec R et RStudio (J. Barnier, J. Biaudet, F. Briatte, M. Bouchet-Valat, E. Gallic, F. Giraud, J. Gombin, M. Kauffmann, C. Lalanne, J. Larmarange, N. Robette)
<https://larmarange.github.io/analyse-R/analyse-R.pdf>
- Ce diaporama sur IRIS

Analyse en Composantes Principales

Données de températures (Y)

	Janv	Fevr	Mars	Avri	Mai	Juin	juil	Aout	Sept	Octo	Nove	Dece
Bordeaux	5,6	6,6	10,3	12,8	15,8	19,3	20,9	21	18,6	13,8	9,1	6,2
Brest	6,1	5,8	7,8	9,2	11,6	14,4	15,6	16	14,7	12	9	7
Clermont	2,6	3,7	7,5	10,3	13,8	17,3	19,4	19,1	16,2	11,2	6,6	3,6
Grenoble	1,5	3,2	7,7	10,6	14,5	17,8	20,1	19,5	16,7	11,4	6,5	2,3
Lille	2,4	2,9	6	8,9	12,4	15,3	17,1	17,1	14,7	10,4	6,1	3,5
Lyon	2,1	3,3	7,7	10,9	14,9	18,5	20,7	20,1	16,9	11,4	6,7	3,1
Marseille	5,5	6,6	10	13	16,8	20,8	23,3	22,8	19,9	15	10,2	6,9
Montpellier	5,6	6,7	9,9	12,8	16,2	20,1	22,7	22,3	19,3	14,6	10	6,5
Nantes	5	5,3	8,4	10,8	13,9	17,2	18,8	18,6	16,4	12,2	8,2	5,5
Nice	7,5	8,5	10,8	13,3	16,7	20,1	22,7	22,5	20,3	16	11,5	8,2
Paris	3,4	4,1	7,6	10,7	14,3	17,5	19,1	18,7	16	11,4	7,1	4,3
Rennes	4,8	5,3	7,9	10,1	13,1	16,2	17,9	17,8	15,7	11,6	7,8	5,4
Strasbourg	0,4	1,5	5,6	9,8	14	17,2	19	18,3	15,1	9,5	4,9	1,3
Toulouse	4,7	5,6	9,2	11,6	14,9	18,7	20,9	20,9	18,3	13,3	8,6	5,5
Vichy	2,4	3,4	7,1	9,9	13,6	17,1	19,3	18,8	16	11	6,6	3,4

15 villes (individus)

12 variables (moyennes mensuelles sur 30 ans des températures)

⇒ tableau de données rectangulaire 15 lignes × 12 colonnes

Notations

K variables

	y^1	\cdots	y^k	\cdots	y^K	
1	y_{11}	\cdots	y_{1k}	\cdots	y_{1K}	
\vdots	\vdots		\vdots		\vdots	
i	y_{i1}	\cdots	y_{ik}	\cdots	y_{iK}	$= \mathbf{Y}$
\vdots	\vdots		\vdots		\vdots	
n	y_{n1}	\cdots	y_{nk}	\cdots	y_{nK}	

Analyse en composantes principales ?

- contexte : K variables quantitatives sur n individus
- Tableau = ensemble d'individus (lignes)
 - → mise en valeur des différences/ressemblances
 - → construction de groupes d'individus homogènes
- Tableau = ensemble de variables (colonnes)
 - → mise en valeur en valeur des liaisons (corrélation linéaire) entre variables
 - → construction de variables synthétiques résumant un groupe de variables homogènes

Analyse en composantes principales ?

Objectifs de l'ACP

- outil descriptif et exploratoire → visualisation 2D voire 3D des données
- outil de synthèse fournissant un résumé du tableau des données

Deux nuages de points

$$\begin{array}{c} \text{Étude des individus} \\ \begin{matrix} & 1 & \cdots & k & \cdots & K \\ \begin{matrix} 1 & y_{11} & \cdots & y_{1k} & \cdots & y_{1K} \\ \vdots & \vdots & & \vdots & & \vdots \\ i & y_{i1} & \cdots & y_{ik} & \cdots & y_{iK} \\ \vdots & \vdots & & \vdots & & \vdots \\ n & y_{n1} & \cdots & y_{nk} & \cdots & y_{nK} \end{matrix} \end{matrix} = \mathbf{X} \end{array} \quad \begin{array}{c} \text{Étude des variables} \\ \begin{matrix} & 1 & \cdots & k & \cdots & K \\ \begin{matrix} 1 & y_{11} & \cdots & y_{1k} & \cdots & y_{1K} \\ \vdots & \vdots & & \vdots & & \vdots \\ i & y_{i1} & \cdots & y_{ik} & \cdots & y_{iK} \\ \vdots & \vdots & & \vdots & & \vdots \\ n & y_{n1} & \cdots & y_{nk} & \cdots & y_{nK} \end{matrix} \end{matrix} = \mathbf{X} \end{array}$$

1 individu = K coordonnées 1 variable = n coordonnées

Nuage des individus \mathcal{N}_{ind}

1 individu = 1 ligne du tableau des données
= 1 point d'un espace à K dimensions

- $K = 1$ (i.e. 1 variable) :
1 axe pour représenter les individus
- $K = 2$ (i.e. 2 variables) :
2 axes pour représenter les individus
- $K = 3$ (i.e. 3 variables) :
représentation 3D (3 axes)
- $K \geq 4$ (i.e. ≥ 4 variables) :
difficile de représenter parfaitement les individus.

Représentation des individus quand $K \geq 4$?

Solution : définir un système d'axes (appelés **axes factoriels**) permettant de représenter au mieux les différences/ressemblances entre individus dans un sous-espace de petite dimension (≤ 3)

Distance entre 2 individus i et i' ?

	1	\cdots	k	\cdots	K
\vdots	\vdots		\vdots		\vdots
i	y_{i1}	\cdots	y_{ik}	\cdots	y_{iK}
\vdots	\vdots		\vdots		\vdots
i'	$y_{i'1}$	\cdots	$y_{i'k}$	\cdots	$y_{i'K}$
\vdots	\vdots		\vdots		\vdots

$$\begin{aligned} d(i, i')^2 &= (y_{i1} - y_{i'1})^2 + (y_{i2} - y_{i'2})^2 + \cdots + (y_{iK} - y_{i'K})^2 \\ &= \sum_{k=1}^K (y_{ik} - y_{i'k})^2 \end{aligned}$$

Distance entre 2 individus

Propriété

2 individus se ressemblent



prennent des valeurs proches pour les K variables

2 principes pour la représentation des individus

2 individus ressemblants doivent être proches

2 individus différents doivent être éloignés

L'étude des individus revient à analyser la forme du nuage des individus \mathcal{N}_{ind} à partir des distances entre individus

Centrage et réduction du tableau de données

centrage = positionne le centre de gravité du nuage de individus $(\bar{y}^1, \dots, \bar{y}^K)$ à l'origine du repère $(0, \dots, 0)$.

réduction = ramène les écart-types des variables à 1.

variable y^k : moyenne, variance et écart-type

- $\bar{y}^k \stackrel{\text{déf}}{=} \frac{1}{n} \sum_{i=1}^n y_{ik}$
- $Var(y^k) \stackrel{\text{déf}}{=} \frac{1}{n} \sum_{i=1}^n (y_{ik} - \bar{y}^k)^2$
- $s_k \stackrel{\text{déf}}{=} \sqrt{Var(y^k)}$

Centrage et réduction du tableau de données

Standardisation = centrage + réduction

standardiser le tableau de données = transformer l'élément y_{ik} du tableau correspondant à l'individu i et la variable k de la façon suivante :

$$x_{ik} = \frac{y_{ik} - \bar{y^k}}{s_k}$$

$$\mathbf{x}^k = (x_{1k}, \dots, x_{nk})^T : \bar{\mathbf{x}^k} = 0 \text{ et } \text{Var}(\mathbf{x}^k) = 1$$

exemple : 47 communes de la Gironde

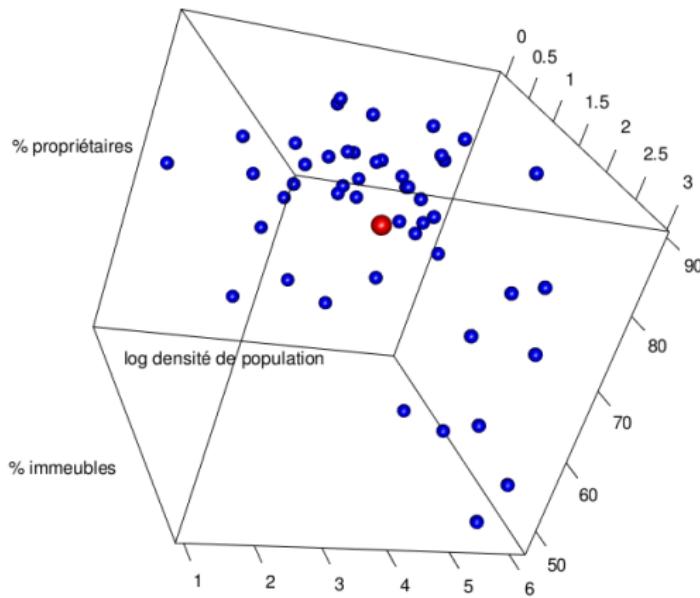
Données : 47 communes (indvidus) du département de la Gironde

	immeuble (%) (y ¹)	log(densite_pop) (y ²)	proprietaire (%) (y ³)
DAIGNAC	0.85	4.44	78.29
HURE	1.14	4.24	75.47
PELLEGRUE	0.59	3.27	61.80
SAUMOS	0.08	2.23	77.20
CAUMONT	0.47	3.04	77.05
SALIGNAC	1.26	4.65	77.44
TAILLECAVAT	0.70	3.45	80.87
RUCH	0.60	3.64	75.76
PUJOLS-SUR-CIRON	1.10	4.57	84.52
BAURECH	1.13	4.58	74.06
:	:	:	:

TABLE: Tableau brut

exemple : 47 communes de la Gironde

Nuage des individus à partir du **tableau brut**



point rouge = barycentre du nuage = point de coordonnées $(\bar{y}^1, \bar{y}^2, \bar{y}^3) = (1.03, 4.09, 73.91)$
où \bar{y}^j = moyenne empirique de la variable y^j pour $j = 1, 2$ ou 3

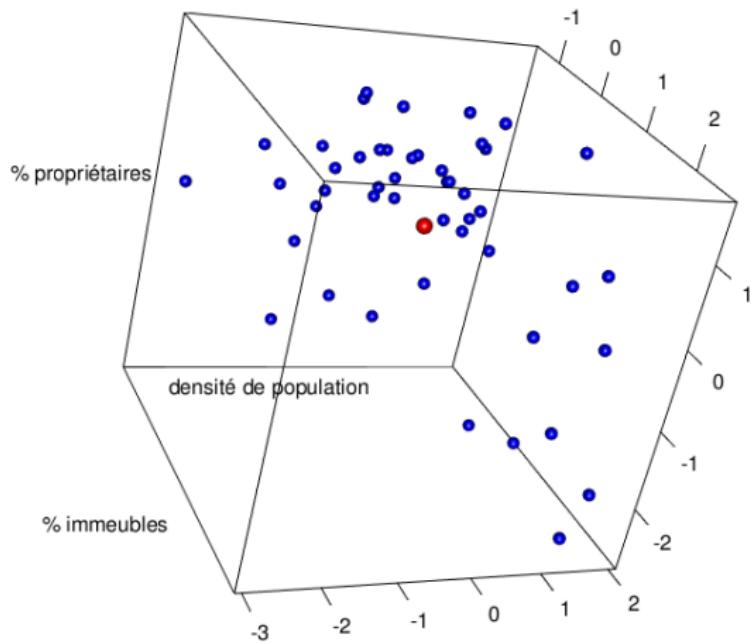
exemple : 47 communes de la Gironde

	immeuble (%) x^1	densite_pop x^2	proprietaire (%) x^3
DAIGNAC	-0.25	0.34	0.45
HURE	0.16	0.15	0.16
PELLEGRUE	-0.59	-0.79	-1.23
SAUMOS	-1.29	-1.79	0.34
CAUMONT	-0.75	-1.01	0.32
SALIGNAC	0.32	0.54	0.36
TAILLECAVAT	-0.44	-0.62	0.71
RUCH	-0.58	-0.43	0.19
:	:	:	:

TABLE: Tableau centré-réduit

exemple : 47 communes de la Gironde

Nuage de points centré-réduit



point rouge = barycentre du nuage =
point de coordonnées
 $(\bar{x}^1, \bar{x}^2, \bar{x}^3) = (0, 0, 0)$

Standardisation = centrage + réduction

- **centrage** \Rightarrow barycentre du nuage de points de coordonnées $(0, \dots, 0)$; la forme du nuage de points ne change pas
- **réduction** \Rightarrow changement des unités (variance de chaque variable centrée-réduite = 1)

ACP : mécanisme intuitif

Recherche intuitive d'une représentation 2D du nuage des individus
"la moins déformante possible"

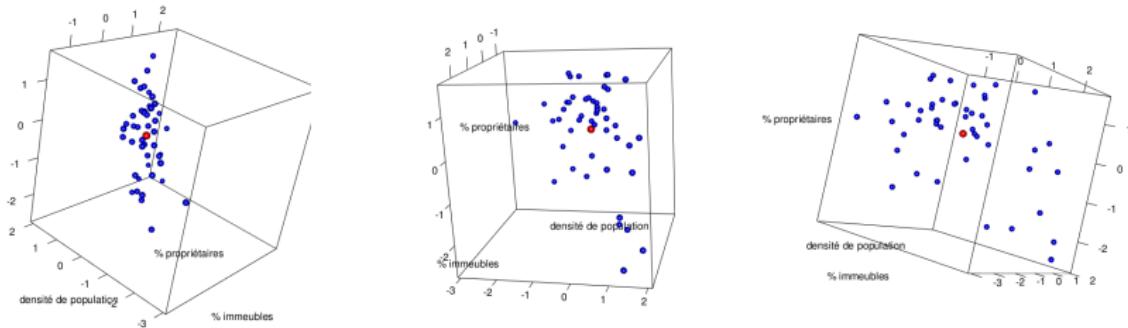
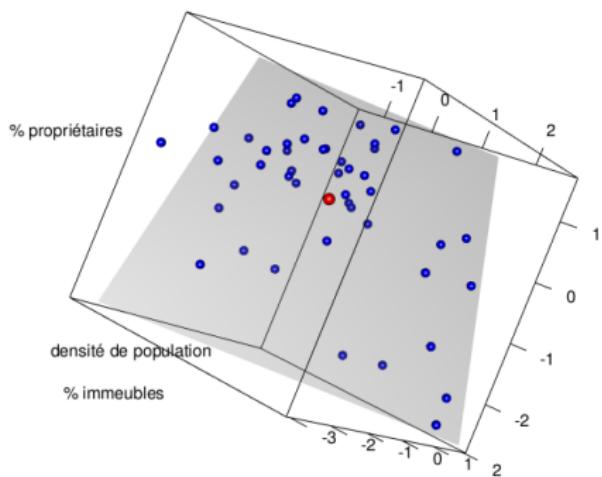


FIGURE: 3 vues du même nuage de points.

ACP : mécanisme intuitif

Recherche intuitive d'une représentation 2D du nuage des individus
"la moins déformante possible"



L'approximation
d'un nuage de points
par un plan est
d'autant meilleure
que la dispersion des
points dans ce plan
est grande.

ACP : recherche intuitive de l'axe factoriel 1

ACP
=
recherche
séquentielle
d'axes
(factoriels)
orthogonaux

ACP : recherche intuitive de l'axe factoriel 2

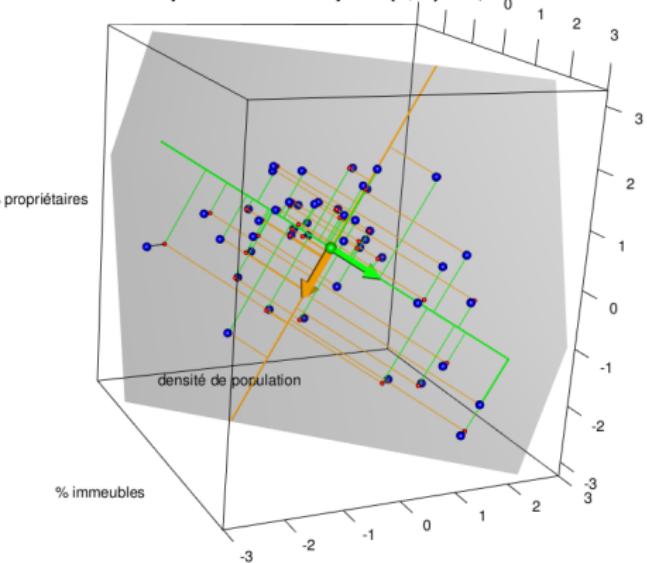
ACP
=
recherche
séquentielle
d'axes
(factoriels)
orthogonaux

ACP : représentation 2D optimale des individus

ANALYSE EN COMPOSANTES PRINCIPALES

Plan principal formé par les axes 1 et 2

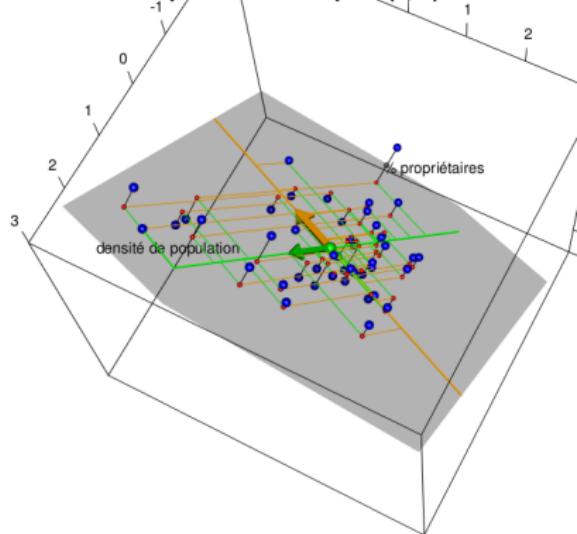
$$\text{Variance exprimée dans le plan } (1, 2) \approx 2.02 + 0.84 = 2.86$$



ANALYSE EN COMPOSANTES PRINCIPALES

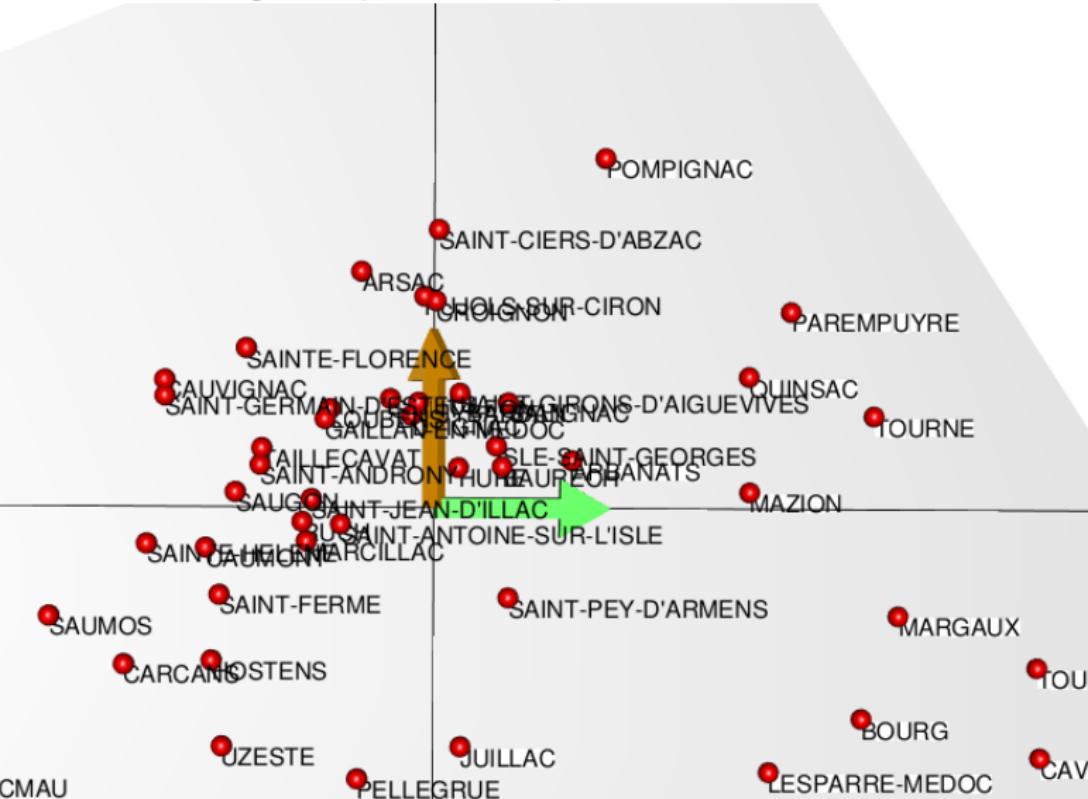
Plan principal formé par les axes 1 et 2

$$\text{Variance exprimée dans le plan } (1, 2) = 2.02 + 0.84$$



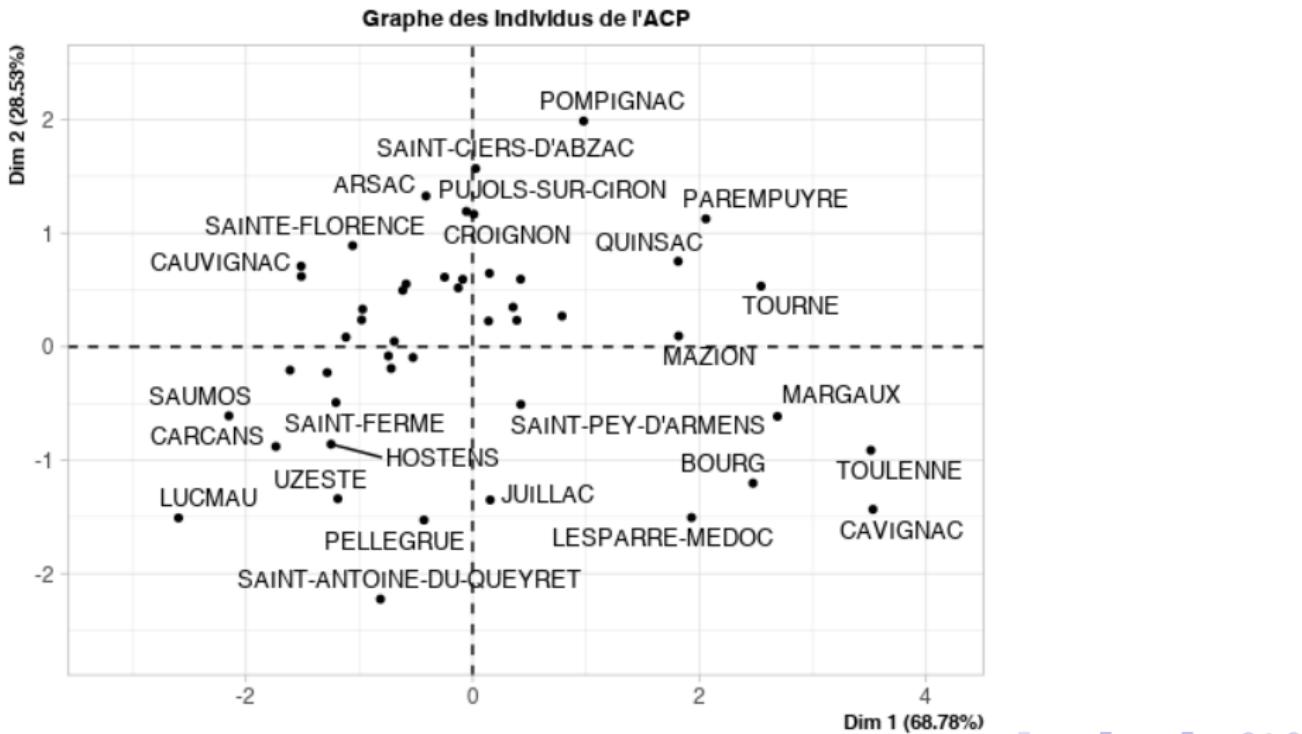
ACP : représentation 2D optimale des individus

Représentation 2D du nuage de points "la plus fidèle"



ACP : représentation 2D optimale des individus

Sortie logiciel R (package FactoMineR)



Composantes principales

- **1ère composante principale** F_1 = vecteur des coordonnées des individus sur l'axe factoriel 1 = $\mathbf{X} \mathbf{u}_1$
- **2ème composante principale** F_2 = vecteur des coordonnées des individus sur l'axe factoriel 2 = $\mathbf{X} \mathbf{u}_2$
-

Plan factoriel et principal et plan

- **plan factoriel** (j, k) = plan formé par \mathbf{u}_j et \mathbf{u}_k
- **plan principal** = plan factoriel (1, 2) = plan formé par \mathbf{u}_1 et \mathbf{u}_2

Retour à notre exemple (47 communes de Gironde)

1ère composante principale F_1

$$X \ u_1 = \underbrace{\begin{bmatrix} -0.25 & 0.34 & 0.45 \\ 0.16 & 0.15 & 0.16 \\ -0.59 & -0.79 & -1.23 \\ -1.29 & -1.79 & 0.34 \\ \vdots & \vdots & \vdots \\ -0.28 & 0.44 & 1.29 \\ 0.78 & 0.88 & -2.17 \\ -1.00 & -0.17 & 0.47 \\ 1.37 & 1.38 & 0.00 \end{bmatrix}}_X \underbrace{\begin{bmatrix} 0.68 \\ 0.63 \\ -0.38 \end{bmatrix}}_{u_1} = \underbrace{\begin{bmatrix} -0.13 \\ 0.14 \\ -0.42 \\ -2.13 \\ \vdots \\ -0.41 \\ 1.91 \\ -0.97 \\ 1.79 \end{bmatrix}}_{F_1}$$

Rappel : $-0.13 = (-0.25 \times 0.68) + (0.34 \times 0.63) + (0.45 \times -0.38)$

Comp. principales = variables "synthétiques"

	x^1	\dots	x^k	\dots	x^K	F_1	F_2	\dots
1	x_{11}	\cdots	x_{1k}	\cdots	x_{1K}	F_{11}	F_{12}	\cdots
\vdots	\vdots		\vdots		\vdots	\vdots	\vdots	\cdots
i	x_{i1}	\cdots	x_{ik}	\cdots	x_{iK}	F_{i1}	F_{i2}	\cdots
\vdots	\vdots		\vdots		\vdots	\vdots	\vdots	\cdots
n	x_{n1}	\cdots	x_{nk}	\cdots	x_{nK}	F_{n1}	F_{n2}	\cdots

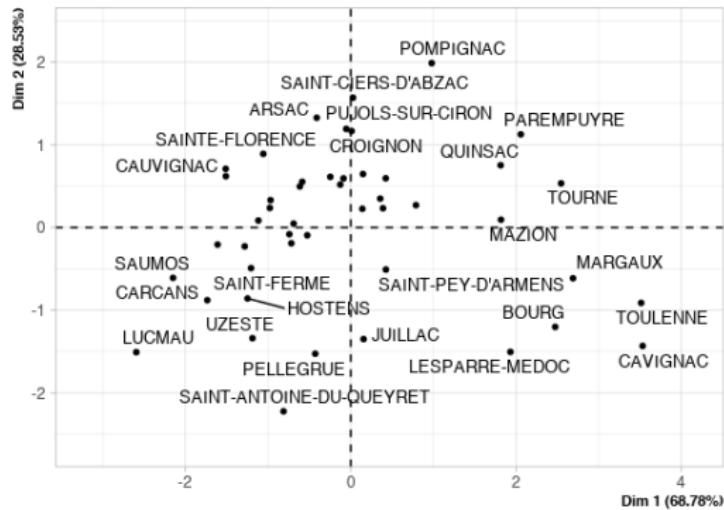
- F_1, F_2, \dots possèdent n coordonnées
- composante principale
= combinaison linéaire de x^1, x^2, \dots, x^K
= nouvelle variable dite "synthétique"

Comp. principales = variables "synthétiques"

	x^1	\dots	x^k	\dots	x^K	F_1	F_2	\dots
1	x_{11}	\cdots	x_{1k}	\cdots	x_{1K}	F_{11}	F_{12}	\cdots
\vdots	\vdots		\vdots		\vdots	\vdots	\vdots	\cdots
i	x_{i1}	\cdots	x_{ik}	\cdots	x_{iK}	F_{i1}	F_{i2}	\cdots
\vdots	\vdots		\vdots		\vdots	\vdots	\vdots	\cdots
n	x_{n1}	\cdots	x_{nk}	\cdots	x_{nK}	F_{n1}	F_{n2}	\cdots

- F_1, F_2, \dots possèdent n coordonnées
- composante principale
= combinaison linéaire de x^1, x^2, \dots, x^K
= nouvelle variable dite "synthétique"

Ex. 47 communes de Gironde



	F_1	F_2
POMPIGNAC	$F_{39,1} = 0.98$	$F_{39,2} = 1.99$
SAINT-FERME	$F_{40,1} = -1.21$	$F_{40,2} = -0.49$
MAZION	$F_{41,1} = 1.82$	$F_{41,2} = 0.09$
MARGAUX	$F_{42,1} = 2.69$	$F_{42,2} = -0.61$
	⋮	⋮

Qualité de représentation et inertie

- $x_1, \dots, x_i, \dots, x_n$ n individus où $x_i = (x_{i1}, \dots, x_{iK})^T = i$ ème ligne du tableau centré-réduit X
- $\|x_i\|^2 = x_{i1}^2 + \dots + x_{iK}^2 =$ distance au carré entre l'individu i et le barycentre du nuage des individus

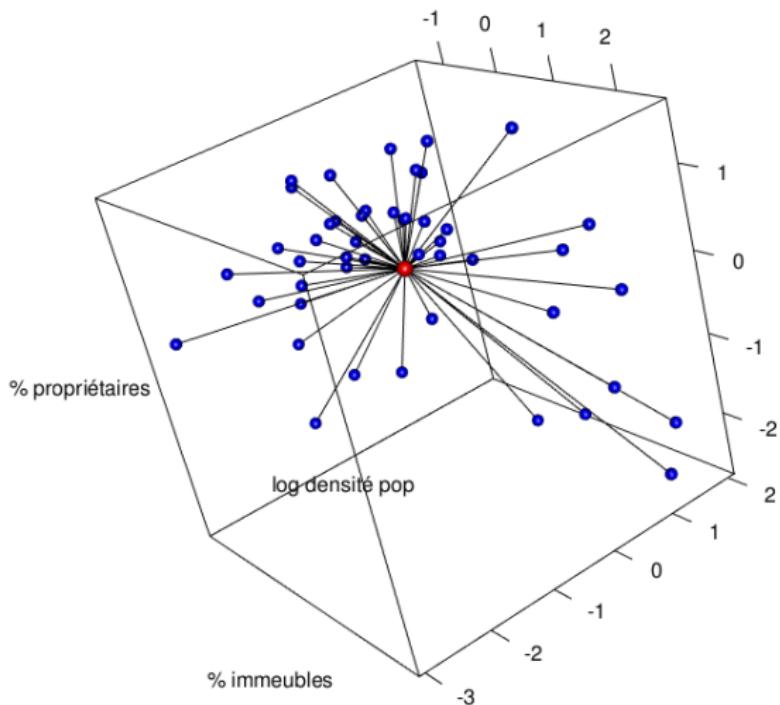
Inertie totale

Inertie totale = inertie du nuage des individus

$$= (\|x_1\|^2 + \dots + \|x_i\|^2 + \dots + \|x_n\|^2) / n$$

= moyenne des carrés des distances entre les individus et le barycentre du nuage des individus

Qualité de représentation et inertie



Inertie totale =
moyenne des carrés
des longueurs des
segments = 2.94

Qualité de représentation et inertie

Inertie associée aux sous-espaces factoriels

sous-espace factoriel	inertie
1er axe factoriel	= variance de la 1ère comp. principale F_1 = $\text{Var}(X u_1)$ = var. des coord. des ind. proj. sur l'axe 1
2ème axe factoriel	= variance de la 2ème comp. principale F_2 = $\text{Var}(X u_2)$ = Var. des coord. des ind. proj. sur l'axe 2
kème axe factoriel	= variance de la k ème comp. principale F_k = $\text{Var}(X u_k)$ = Var. des coord. des ind. proj. sur l'axe k
plan principal	= inertie axe factoriel 1 + inertie axe factoriel 2 = $\text{Var}(X u_1) + \text{Var}(X u_2)$

Qualité de représentation d'un sous-espace factoriel = inertie associée

Ex : 47 communes de Gironde

	inertie	% inertie
1er axe factoriel	2.02	68.7
2ème axe factoriel	0.84	28.5
3ème axe factoriel	0.08	2.8
inertie totale	$2.02 + 0.84 + 0.08 = 2.94$	100.0
plan principal	$2.02 + 0.84 = 2.86$	97.2

Remarque sorties logiciel R :

inertie axe factoriel 1 → Dim 1 (68.78%)

inertie axe factoriel 1 → Dim 1 (28.53%)

Ex : 47 communes de Gironde

	inertie	% inertie
1er axe factoriel	2.02	68.7
2ème axe factoriel	0.84	28.5
3ème axe factoriel	0.08	2.8
inertie totale	$2.02 + 0.84 + 0.08 = 2.94$	100.0
plan principal	$2.02 + 0.84 = 2.86$	97.2

Remarque sorties logiciel R :

inertie axe factoriel 1 → Dim 1 (68.78%)

inertie axe factoriel 1 → Dim 1 (28.53%)

Un peu de théorie

Rappel : $\text{Cov}(y^j, y^k) \stackrel{\text{déf}}{=} \frac{1}{n} \sum_{i=1}^n (y_{ij} - \bar{y}^j)(y_{ik} - \bar{y}^k)$

Matrice des covariances

Matrice des covariances V = matrice K lignes et K colonnes des variances et covariances :

$$V \stackrel{\text{déf}}{=} \begin{pmatrix} Var(y^1) & \dots & Cov(y^1, y^k) & \dots & Cov(y^1, y^K) \\ \vdots & & \vdots & & \vdots \\ Cov(y^j, y^1) & \dots & Cov(y^j, y^k) & \dots & Cov(y^j, y^K) \\ \vdots & & \vdots & & \vdots \\ Cov(y^K, y^1) & \dots & Cov(y^K, y^k) & \dots & Var(y^K) \end{pmatrix}.$$

Un peu de théorie

Rappel : $\text{Corr}(y^j, y^k) \stackrel{\text{déf}}{=} \text{Cov}(y^j, y^k) / (s_j s_k)$

Matrice des corrélations

Matrice des corrélations C = matrice K lignes et K colonnes des corrélations :

$$C \stackrel{\text{déf}}{=} \begin{pmatrix} 1 & \dots & \text{Corr}_n(y^1, y^k) & \dots & \text{Corr}_n(y^1, y^K) \\ \vdots & & \vdots & & \vdots \\ \text{Corr}_n(y^j, y^1) & \dots & \text{Corr}_n(y^j, y^k) & \dots & \text{Corr}_n(y^j, y^K) \\ \vdots & & \vdots & & \vdots \\ \text{Corr}_n(y^K, y^1) & \dots & \text{Corr}_n(y^K, y^k) & \dots & 1 \end{pmatrix}$$

Matriciellement, $C = \frac{1}{n} \mathbf{X}^T \mathbf{X} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$ avec \mathbf{X} tableau des données centrées-réduites

Ex : 47 communes de Gironde

	% immeubles	log(densite.pop)	% proprietaires
% immeubles	0.54	0.67	-3.09
log(densite.pop)	0.67	1.08	-1.90
% proprietaires	-3.09	-1.90	96.46

TABLE: Matrice des covariances

	% immeubles	log(densite.pop)	% proprietaires
% immeubles	1.00	0.88	-0.43
log(densite.pop)	0.88	1.00	-0.19
% proprietaires	-0.43	-0.19	1.00

TABLE: Matrice des corrélations (valeurs entre - 1 et 1)

Un peu de théorie

On peut montrer les résultats suivants (voir polycopié de cours)

Détermination du 1er axe factoriel

- Le 1er axe factoriel est la droite (sous-espace de dimension 1) pour laquelle la projection des individus (les points rouges dans les représentations 3D précédentes) fournit la meilleure approximation du nuage des individus. Ce 1er axe est engendré par le vecteur propre \mathbf{u}_1 de la matrice des corrélations \mathbf{C} associé à sa plus grande valeur propre λ_1 : $\mathbf{C} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$.
- λ_1 représente l'inertie du nuage des individus associée au 1er axe factoriel. Cette quantité est par définition la variance du nuage des individus projeté sur l'axe 1.

Dans notre exemple, $\lambda_1 = 2.02$

Un peu de théorie

Détermination du 2ème axe factoriel

- Le 2ème axe factoriel est la droite (sous-espace de dimension 1) parmi les perpendiculaires à l'axe 1 pour laquelle la projection des individus (les points rouges dans les représentations 3D précédentes) fournit la meilleure approximation du nuage des individus. Ce 2ème axe est engendré par le vecteur propre u_2 de la matrice C associé à la 2ème plus grande valeur propre λ_2 :
$$C u_2 = \lambda_2 u_2.$$
- λ_2 représente l'inertie du nuage des individus associée au 2ème axe factoriel. Cette quantité est par définition la variance du nuage des individus projeté sur l'axe 2.

Dans notre exemple, $\lambda_2 = 0.84$

Plan principal

- il est formé par les deux premiers axes factoriels engendrés par (u_1, u_2) ,
- le plan principal maximise la dispersion des individus projetés sur ce plan ; ce plan permet d'offrir l'image 2D la plus fidèle dans laquelle les individus sont le mieux séparés.
- $\lambda_1 + \lambda_2$ représente l'inertie (i.e. la dispersion) du nuage des individus projeté sur le plan principal (dans notre exemple $2.02 + 0.84 = 2.86$).

Un peu de théorie

Détermination du k ème axe factoriel

- Le q ème axe factoriel engendré par \mathbf{u}_q est la droite (sous-espace de dimension 1) pour laquelle la projection des individus (les points rouges dans les représentations 3D précédentes) fournit la meilleure approximation du nuage des individus sous contrainte d'orthogonalité avec le sous-espace engendré par $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{q-1}$. \mathbf{u}_q est le vecteur propre de la matrice \mathbf{C} associé à la q ème plus grande valeur propre λ_q : $\mathbf{C} \mathbf{u}_q = \lambda_q \mathbf{u}_q$.
- λ_q représente l'inertie du nuage des individus associée au q ème axe factoriel. Cette quantité est par définition la variance du nuage des individus projeté sur le q ème axe factoriel.

Remarque : si le tableau des données n'est pas réduit, on remplace la matrice des corrélations \mathbf{C} par la matrice des covariances \mathbf{V} .

Inertie et sorties logiciel

- Dans la plupart des logiciels, les valeurs propres $\lambda_1, \lambda_2, \dots$ se trouvent sous l'intitulé Eigenvalues
- les inerties, % d'inertie et leurs versions cumulées sont en général disponibles et identifiées en terme de variance

Exemple 47 communes de la Gironde

Eigenvalues

	Dim.1	Dim.2	Dim.3
Variance	2.021	0.838	0.081
% of var.	68.707	28.503	2.755
Cumulative % of var.	68.707	97.210	100.000

Vocabulaire - Terminologie

Les 4 assertions suivantes sont équivalentes :

- L'**axe (factoriel) 1** exprime 68.7% de l'inertie totale
- L'**axe (factoriel) 1** représente 68.7% de l'inertie totale
- La part de variance exprimée par l'**axe (factoriel) 1** est 68.7%
- La part de variance représentée par l'**axe (factoriel) 1** est 68.7%

Dans ces assertions, on peut utiliser la même terminologie pour parler de n'importe quel autre axe factoriel ou bien du plan principal.

Exemple : on peut dire que le plan principal représente 97.2% de l'inertie totale. Cela signifie que d'un point de vue de l'inertie, la qualité de représentation des individus dans le plan principal est très bonne (la représentation 2D est très fidèle au tableau des données)

Vocabulaire - Terminologie

Les 4 assertions suivantes sont équivalentes :

- L'**axe (factoriel) 1** exprime 68.7% de l'inertie totale
- L'**axe (factoriel) 1** représente 68.7% de l'inertie totale
- La part de variance exprimée par l'**axe (factoriel) 1** est 68.7%
- La part de variance représentée par l'**axe (factoriel) 1** est 68.7%

Dans ces assertions, on peut utiliser la même terminologie pour parler de n'importe quel autre axe factoriel ou bien du plan principal.

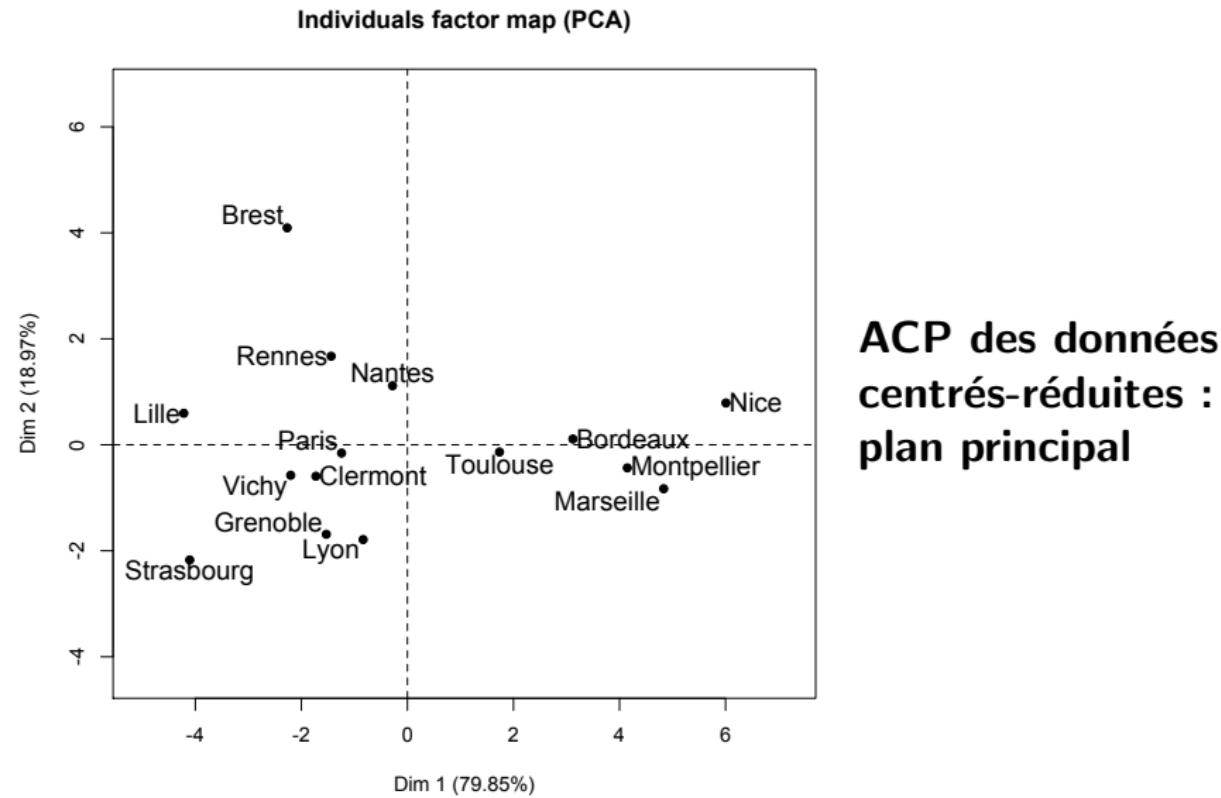
Exemple : on peut dire que le plan principal représente 97.2% de l'inertie totale. Cela signifie que d'un point de vue de l'inertie, la qualité de représentation des individus dans le plan principal est très bonne (la représentation 2D est très fidèle au tableau des données)

Données températures

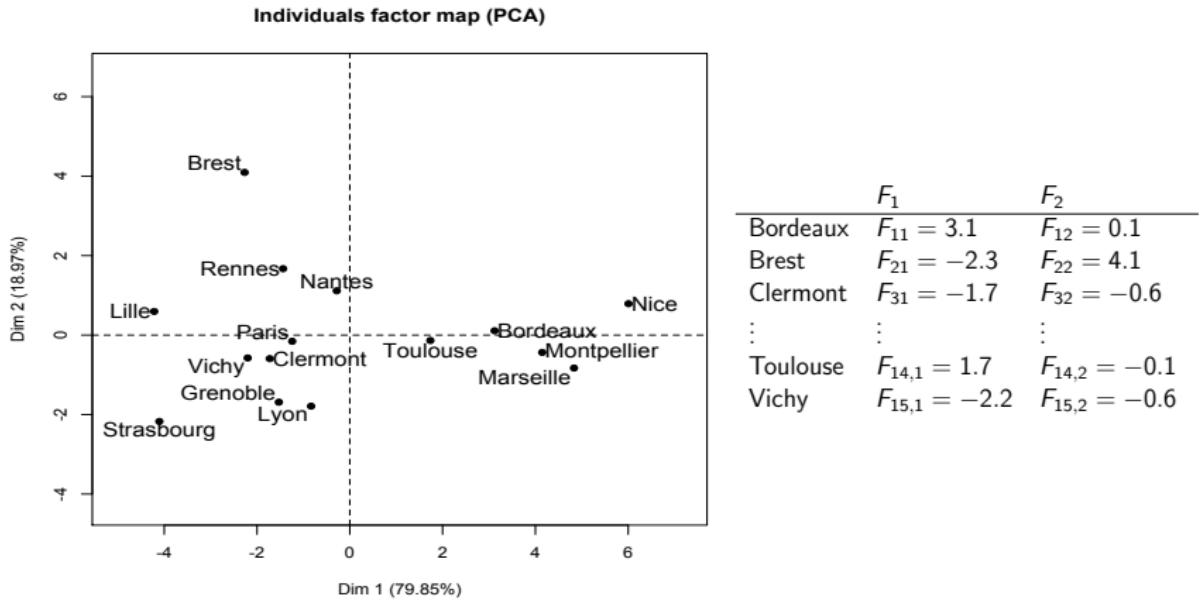
	Janv	Fevr	Mars	Avri	Mai	Juin	juil	Aout	Sept	Octo	Nove	Dece
Bordeaux	5,6	6,6	10,3	12,8	15,8	19,3	20,9	21	18,6	13,8	9,1	6,2
Brest	6,1	5,8	7,8	9,2	11,6	14,4	15,6	16	14,7	12	9	7
Clermont	2,6	3,7	7,5	10,3	13,8	17,3	19,4	19,1	16,2	11,2	6,6	3,6
Grenoble	1,5	3,2	7,7	10,6	14,5	17,8	20,1	19,5	16,7	11,4	6,5	2,3
Lille	2,4	2,9	6	8,9	12,4	15,3	17,1	17,1	14,7	10,4	6,1	3,5
Lyon	2,1	3,3	7,7	10,9	14,9	18,5	20,7	20,1	16,9	11,4	6,7	3,1
Marseille	5,5	6,6	10	13	16,8	20,8	23,3	22,8	19,9	15	10,2	6,9
Montpellier	5,6	6,7	9,9	12,8	16,2	20,1	22,7	22,3	19,3	14,6	10	6,5
Nantes	5	5,3	8,4	10,8	13,9	17,2	18,8	18,6	16,4	12,2	8,2	5,5
Nice	7,5	8,5	10,8	13,3	16,7	20,1	22,7	22,5	20,3	16	11,5	8,2
Paris	3,4	4,1	7,6	10,7	14,3	17,5	19,1	18,7	16	11,4	7,1	4,3
Rennes	4,8	5,3	7,9	10,1	13,1	16,2	17,9	17,8	15,7	11,6	7,8	5,4
Strasbourg	0,4	1,5	5,6	9,8	14	17,2	19	18,3	15,1	9,5	4,9	1,3
Toulouse	4,7	5,6	9,2	11,6	14,9	18,7	20,9	20,9	18,3	13,3	8,6	5,5
Vichy	2,4	3,4	7,1	9,9	13,6	17,1	19,3	18,8	16	11	6,6	3,4

- 15 individus (villes)
- 12 variables (moyennes par mois)

Données températures : représentation 2D optimale



Coordonnées des individus sur le plan (1,2)



Rappel : F_1 et F_2 = composantes principales

ACP : représentation optimale des variables ?

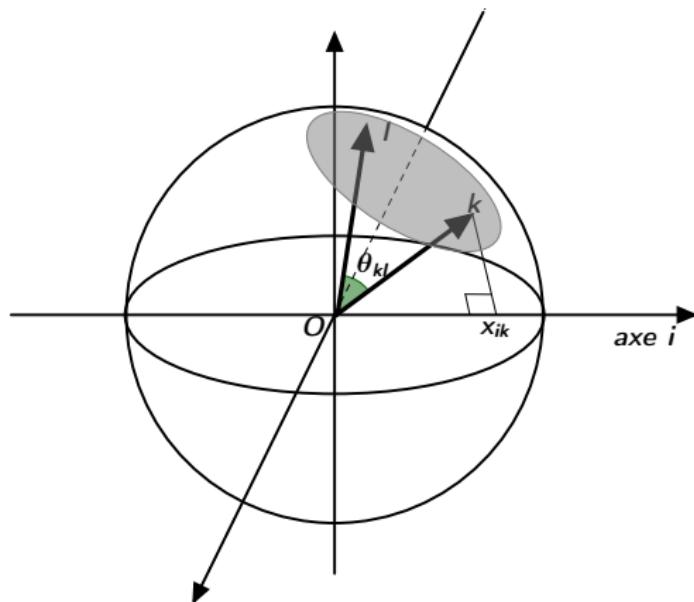
$$\begin{matrix} & \boldsymbol{x}^1 & \cdots & \boldsymbol{x}^k & \cdots & \boldsymbol{x}^K \\ \begin{matrix} 1 \\ \vdots \\ i \\ \vdots \\ n \end{matrix} & \left| \begin{matrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1K} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{iK} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{nK} \end{matrix} \right| & = & \boldsymbol{X} \end{matrix}$$

ACP comme outil de représentation optimale des K variables
 $\boldsymbol{x}^1, \boldsymbol{x}^2, \dots, \boldsymbol{x}^K$

Géométrie du nuage des variables

Variables réduites

⇒ sphère unité



Géométrie du nuage des variables

cosinus d'angle formé par 2 variables centrées-réduites

$$\begin{aligned}\cos \theta_{kl} &= \frac{\text{produit scalaire entre les 2 variables}}{(O k) \times (O l)} \\ &= \frac{\sum_{i=1}^n x_{ik} x_{il}}{\sqrt{\sum_{i=1}^n x_{ik}^2} \sqrt{\sum_{i=1}^n x_{il}^2}} \\ &= \text{Corr}(\mathbf{x}^l, \mathbf{x}^k).\end{aligned}$$

$\cos \theta_{kl}$ = représentation géométrique du coefficient de corrélation linéaire entre \mathbf{x}^k et \mathbf{x}^l

ACP = approximation du nuage des variables

Objectif : visualiser le nuage de variables vivant dans un espace de dimension n en minimisant les déformations

Solution : construction séquentielle de variables synthétiques les plus corrélées avec toutes les variables

ACP = approximation du nuage des variables

On peut montrer que les vecteurs propres $\textcolor{green}{u}_1, \textcolor{red}{u}_2, \textcolor{purple}{u}_3, \dots$ sont t.q. :

- $\textcolor{green}{u}_1$ maximise $\sum_{k=1}^K \text{Corr}(\mathbf{X}\mathbf{u}, \mathbf{x}^k)^2$ pour tout \mathbf{u}
- $F_1 = \mathbf{X}\textcolor{green}{u}_1$ est la c.p. la plus corrélées à l'ensemble des variables

- $\textcolor{red}{u}_2$ maximise $\sum_{k=1}^K \text{Corr}(\mathbf{X}\mathbf{u}, \mathbf{x}^k)^2$ pour tout $\mathbf{u} \perp \textcolor{green}{u}_1$
- $F_2 = \mathbf{X}\textcolor{red}{u}_2$ est la c.p. qui résume au mieux l'information qui n'a pas été synthétisée par F_1

ACP = approximation du nuage des variables

On peut montrer que les vecteurs propres $\textcolor{green}{u}_1, \textcolor{red}{u}_2, \textcolor{purple}{u}_3, \dots$ sont t.q. :

- $\textcolor{green}{u}_1$ maximise $\sum_{k=1}^K \text{Corr}(\mathbf{X}\mathbf{u}, \mathbf{x}^k)^2$ pour tout \mathbf{u}
- $F_1 = \mathbf{X}\textcolor{green}{u}_1$ est la c.p. la plus corrélées à l'ensemble des variables

- $\textcolor{red}{u}_2$ maximise $\sum_{k=1}^K \text{Corr}(\mathbf{X}\mathbf{u}, \mathbf{x}^k)^2$ pour tout $\mathbf{u} \perp \textcolor{green}{u}_1$
- $F_2 = \mathbf{X}\textcolor{red}{u}_2$ est la c.p. qui résume au mieux l'information qui n'a pas été synthétisée par F_1

ACP = approximation du nuage des variables

- \mathbf{u}_3 maximise $\sum_{k=1}^K \text{Corr}(\mathbf{X}\mathbf{u}, x^k)^2$ pour tout $\mathbf{u} \perp \mathbf{u}_1$ et $\mathbf{u} \perp \mathbf{u}_2$
- $\mathbf{F}_3 = \mathbf{X}\mathbf{u}_3$ est la c.p. qui résume au mieux l'information qui n'a pas été synthétisée par \mathbf{F}_1 et \mathbf{F}_2

et ainsi de suite...

Représentation des variables

	x^1	\dots	x^k	\dots	x^K	F_1	F_2	\dots
1	x_{11}	\cdots	x_{1k}	\cdots	x_{1K}	F_{11}	F_{12}	\cdots
\vdots	\vdots		\vdots		\vdots	\vdots	\vdots	\cdots
i	x_{i1}	\cdots	x_{ik}	\cdots	x_{iK}	F_{i1}	F_{i2}	\cdots
\vdots	\vdots		\vdots		\vdots	\vdots	\vdots	\cdots
n	x_{n1}	\cdots	x_{nk}	\cdots	x_{nK}	F_{n1}	F_{n2}	\cdots

Les c.p. F_1, F_2, \dots forment un système de vecteurs orthogonaux permettant une représentation optimale des variables

Coordonnées des variables

On peut montrer (voir polycopié de cours) :

Coordonnées des variables = corrélations

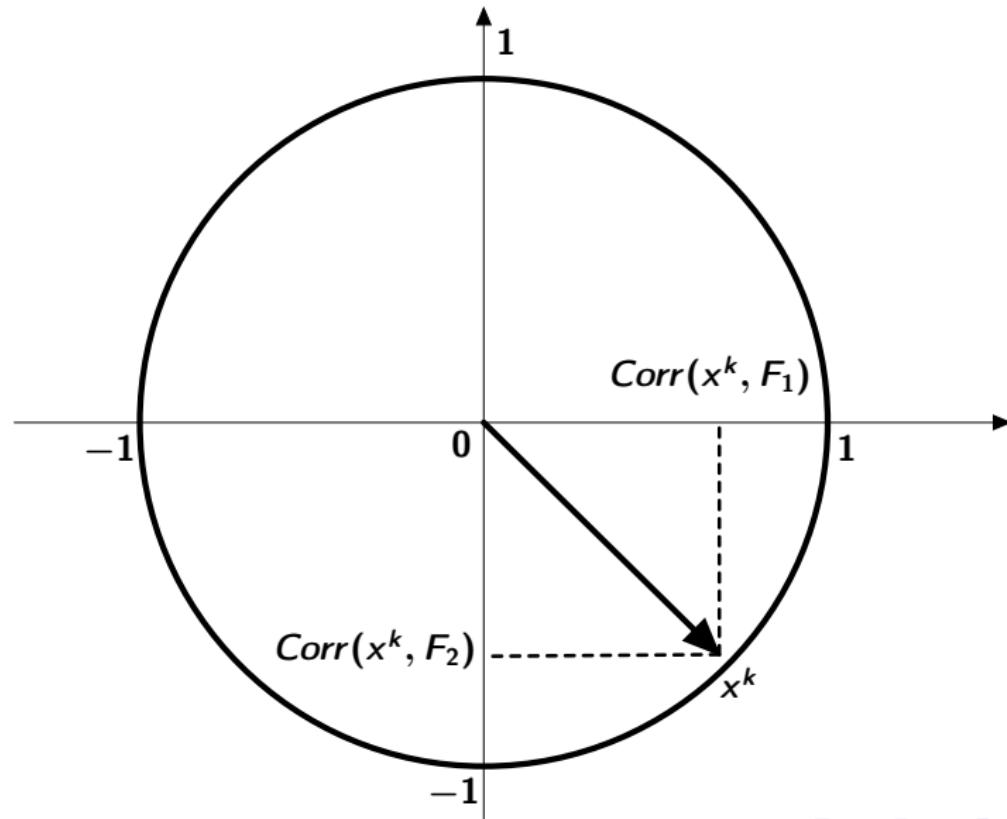
- $\text{Corr}(x^k, F_1)$ = coord. variable x^k sur l'axe 1 engendré par F_1
- $\text{Corr}(x^k, F_2)$ = coord. variable x^k sur l'axe 2 engendré par F_2
- $\vdots \quad \vdots \quad \vdots$

où

$\text{Corr}(x^k, F_1)$ = corrélation entre variable x^k et c.p. F_1 (axe 1),
 $\text{Corr}(x^k, F_2)$ = corrélation entre variable x^k et F_2 (axe 2),

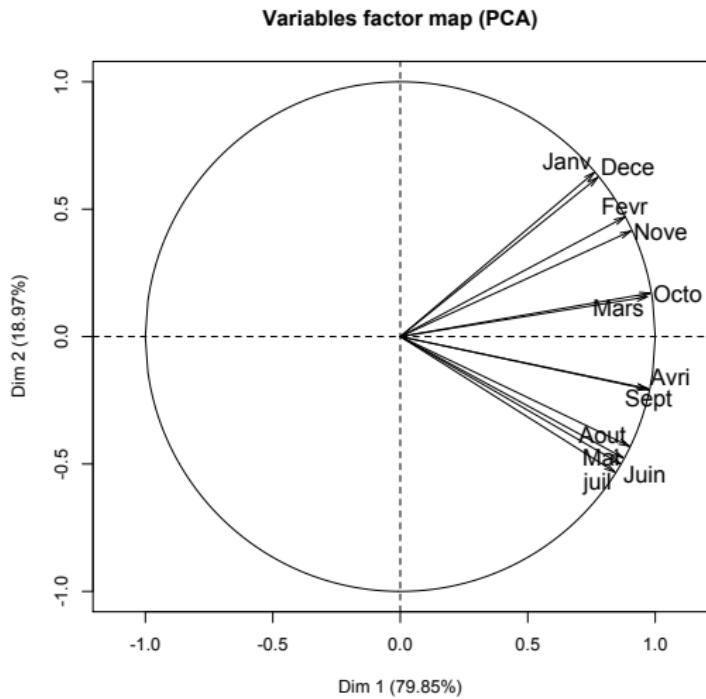
$\vdots \quad \vdots \quad \vdots$

Coordonnées des variables : cercle des corrélations



Données de température : cercle des corrélations

Cercle des corrélations = représentation 2D optimale des variables



- toutes les variables corrélées positivement avec F_1
- Janv et Dece = 2 variables les plus corrélées positivement avec F_2
- Mai, Juin, Juil et Aout = 4 variables les plus corrélées négativement avec F_2

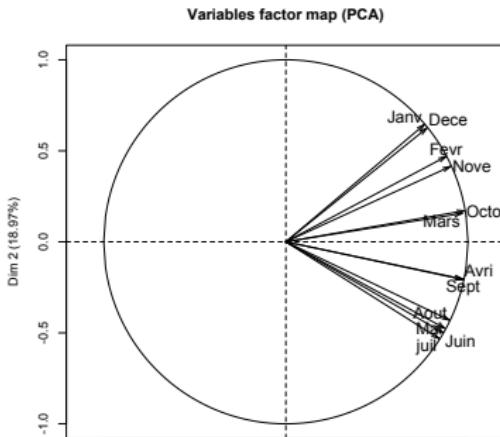
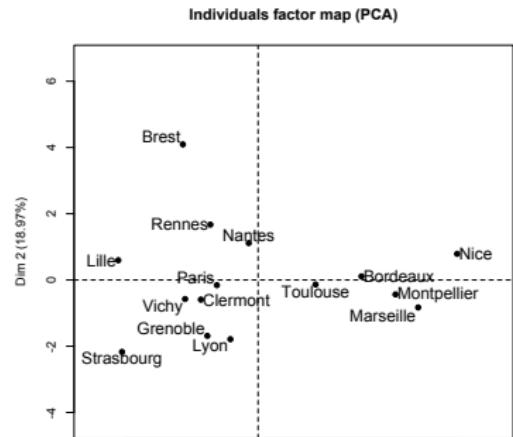
Interprétation

$\text{Corr}(\text{Janv}, F_1) > 0 \Rightarrow \text{temp. de janvier progressent dans le même sens que les coord. sur l'axe 1 :}$

- les villes où il fait le plus froid en janvier = faible coordonnée sur l'axe 1 ; elles sont représentées à gauche
- celles où il fait le plus chaud en janvier = grande coordonnée sur l'axe 1 ; elles sont représentées à droite

Idem pour les autres mois de l'année avec l'axe 1

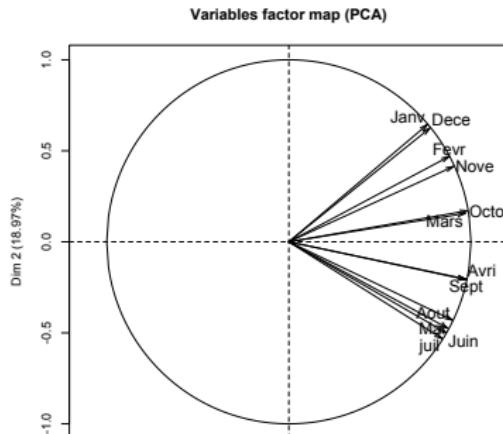
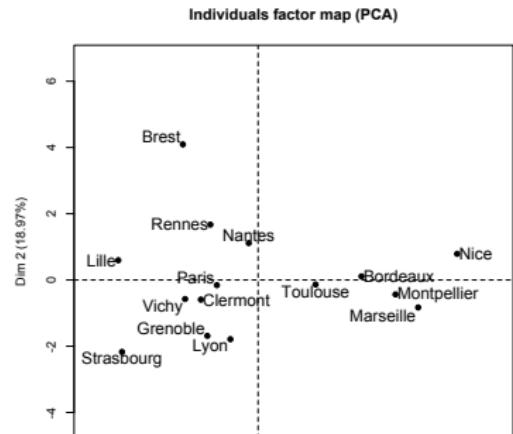
Données de température : interprétation



Axe 1 : villes plus chaudes tte l'année (à droite) opposées aux villes plus froides tte l'année (à gauche)

Axe 2 : villes de faible amplitude thermique (hivers doux et été frais) en haut opposées aux villes de grande amplitude thermique (hivers froid et été chaud) en bas

Données de température : interprétation



Axe 1 : villes plus chaudes tte l'année (à droite) opposées aux villes plus froides tte l'année (à gauche)

Axe 2 : villes de faible amplitude thermique (hivers doux et été frais) en haut opposées aux villes de grande amplitude thermique (hivers froid et été chaud) en bas

Cercle des corrélations

En résumé

- aide pour interpréter les individus (villes froides/chaudes ; villes d'amplitude thermique faible/forte)
- représentation optimale du nuage des variables en dimension 2 (image la moins déformante)

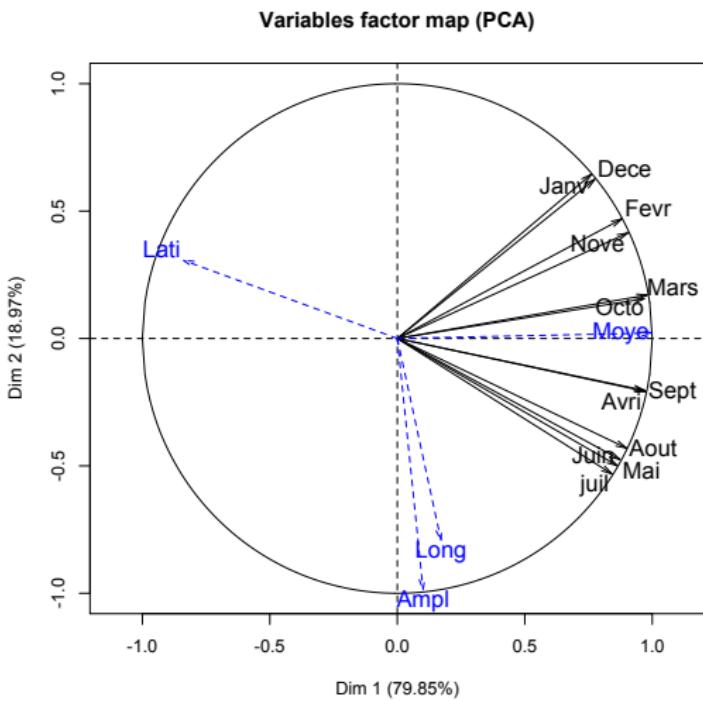
Variables supplémentaires ?

Définition

variables supplémentaires = variables qui **ne participent pas** à la construction des axes

- variables suppl. **quantitatives** : projection sur le cercle des corrélation
- variables suppl. **qualitatives** : modalités positionnées au barycentre des individus qui les prennent

Variables suppl. quantitatives : interprétation



Variables suppl. quantitatives : interprétation

4 var. suppl. quant. :

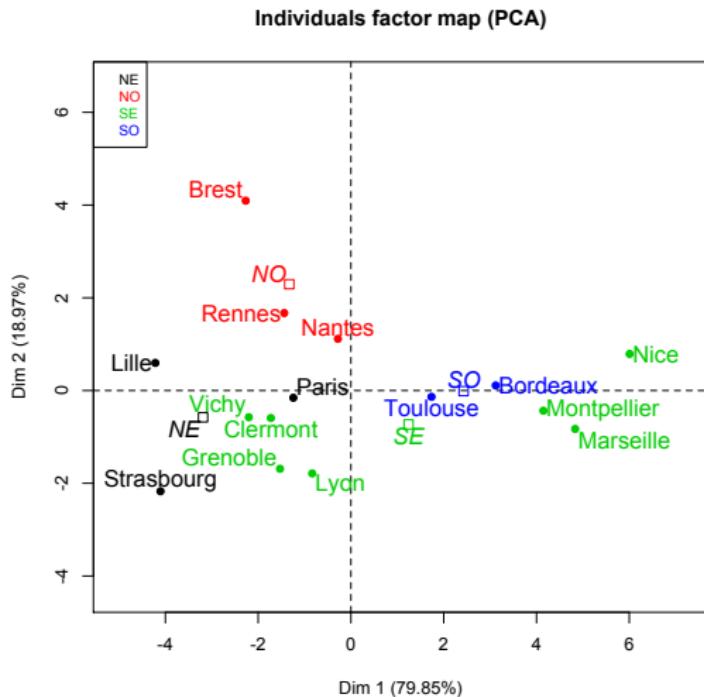
- **Ampl** (amplitude thermique) : très fortement corrélée avec l'axe 2 (confirme l'opposition villes d'amplitude thermique faible/élevée)
- **Lati** (latitude) : corrélée avec l'axe 1 (opposition des villes nord/sud)
- **Long** (longitude) : corrélée avec l'axe 2 (opposition des villes est/ouest)
- **Moye** (moyenne annuelle des température) : très fortement corrélée avec l'axe 1 (confirme l'opposition villes froides/chaudes)

Variable suppl. qualitative : interprétation

1 var. **région**

d'appartenance :

- NE (Nord-Est)
- NO (Nord-Ouest)
- SE (Sud-Est)
- SO (Sud-Ouest)



Variable suppl. REGION : interprétation

Confirmation des conclusions précédentes (cercle des corrélations) :

- **axe 1** : opposition des villes du sud (**SE** et **SO**) aux villes du nord (**NE** et **NO**)
- **axe 2** : opposition des villes de l'ouest (**NO** et **SO**) aux villes de l'est (**NE** et **SE**)

Cosinus carré entre 1 variable et sa projection

Données de températures :

	Dim. 1	Dim. 2	Dim. 3	...
Janv	0.58	0.42	0.00	...
Févr.	0.78	0.22	0.00	...
Mars	0.94	0.02	0.02	...
:	:	:	...	

variable "Mars" très bien représentée sur l'axe 1 (cosinus carré 0.94 proche de 1) et très mal représentée sur l'axe 2 (cosinus carré 0.02 proche de 0)

Cosinus carré entre 1 individu et sa projection

Données de températures :

	Dim. 1	Dim. 2	Dim. 3	...
Bordeaux	0.95	0.00	0.05	...
Brest	0.23	0.76	0.00	...
Clermont	0.88	0.10	0.00	...
:	:	:	...	

"Bordeaux" est très bien représentée sur l'axe 1 (cosinus carré 0.95) et inexiste sur l'axe 2 (cosinus carré 0)

"Brest" est très mal représentée sur l'axe 1 (cosinus carré 0.23) alors que sa représentation sur l'axe 2 est satisfaisante (cosinus carré 0.76)

Qualité de représentation sur 1 plan

Qualité de représentation sur le plan principal (i.e. axes 1 et 2) = sommes cosinus carrés associés à chacun de ces axes

Données de températures :

- bien que l'individu "Brest" soit mal représentée sur l'axe 1, on obtient une qualité de représentation presque parfaite (0.99) sur le plan principal
- variable "Janvier" : qualité de représentation parfaite sur le plan principal

Contribution d'1 var. à la construction d'1 axe

Contribution de la variable k à l'axe s

$$Contrib_s(k) = \frac{Corr(x^k, F_s)^2}{\sum_{k'=1}^K Corr(x^{k'}, F_s)^2}.$$

Propriété

Somme des contributions des variables à un axe $s = 1$:

$$\sum_{k=1}^K Contrib_s(k) = 1$$

Contribution d'1 var. à la construction d'1 axe

Conséquence : les variables x^k qui contribuent le plus à un axe s sont celles qui possèdent les coordonnées $\text{Corr}(x^k, F_s)$ (voir cercle des corrélations) les plus élevées en valeur absolue.

Données de températures :

	Dim. 1	Dim. 2	Dim. 3	...
Janv	6.05	18.24	0.66	...
Févr.	8.09	9.67	1.61	...
Mars	9.79	1.07	34.03	...
⋮	⋮	⋮	⋮	⋮
Somme	100	100	100	...

Contribution d'un individu à la construction d'un axe

Contribution de l'individu i à l'axe s

$$Contrib_s(i) = \frac{F_{is}^2}{\sum_{i'=1}^n F_{i's}^2}.$$

Propriété

Somme des contributions des individus à un axe $s = 1$:

$$\sum_{i=1}^n Contrib_s(i) = 1$$

Contribution d'1 ind. à la construction d'1 axe

Conséquence : les individus i qui contribuent le plus à un axe s sont ceux qui possèdent les coordonnées F_{is} les plus élevées en valeur absolue.

Données de températures :

	Dim. 1	Dim. 2	Dim. 3	...
Bordeaux	6.78	0.03	49.48	...
Brest	3.58	49.07	1.26	...
Clermont	2.07	1.03	0.03	...
⋮	⋮	⋮	⋮	⋮
Somme	100	100	100	...

Descrip. dim. : variables quantitatives

Il est possible d'obtenir les variables ayant un coefficient de corrélation avec l'axe **s** significativement différent de zéro

Une corrélation est significativement différente de zéro dès que la valeur de la "p.value" correspondante (calculée par le logiciel) est plus petite que 5% (soit 0.05).

Descrip. dim. : variables quantitatives

Axe 1

	correlation	p.value
Moye	0.9997097	2.001976e-22
Octo	0.9801599	1.609672e-10
Sept	0.9740289	9.130415e-10
Avri	0.9693357	2.657670e-09
Mars	0.9687704	2.988670e-09
Nove	0.9037531	3.834950e-06
Aout	0.8986059	5.312317e-06
Fevr	0.8804578	1.480634e-05
Mai	0.8727646	2.178413e-05
Juin	0.8635747	3.349417e-05
juil	0.8415346	8.385040e-05
Dece	0.7743349	7.017832e-04
Janv	0.7612384	9.784512e-04
Lati	-0.8389348	9.259113e-05

Axe 2

\$Dim.2	correlation	p.value
\$Dim.2\$quanti		
Janv	0.6443379	9.519348e-03
Dece	0.6242957	1.285835e-02
juil	-0.5314197	4.148657e-02
Long	-0.7922192	4.298867e-04
Ampl	-0.9856753	1.963381e-11

Descrip. dim. : variables quantitatives

Axe 1

- "Moyenne" = variable la plus corrélée positivement avec l'axe 1
- variable supplémentaire "Latitude" corrélée négativement avec l'axe 1
- de nombreuses variables caractérisent l'axe 1

Axe 2

- "Longitude" et "Amplitude" fortement corrélées avec l'axe 2
- "Janvier", "Décembre" et "Juillet" jouent un rôle important pour décrire l'axe 2

Descrip. dim. : variables qualitatives (suppl.)

En présence de variables qualitatives (nécessairement supplémentaires), on peut savoir pour chaque axe, quelles variables sont globalement significatives et si c'est le cas, quelles modalités de ces variables ont un impact significatif

Données de température :

```
$Dim.2  
$Dim.2$quali
```

	R2	p.value
Region	0.6009012	0.01467946

```
$Dim.2  
$Dim.2$category
```

	Estimate	p.value
NO	2.050365	0.001011474
SE	-0.9738852	0.047120253

Descrip. dim. : variables qualitatives (suppl.)

\$Dim.2

\$Dim.2\$quali

	R2	p.value
Region	0.6009012	0.01467946

- variable "Region" : rapport de corrélation de 0.6 (noté ici R2) significativement $\neq 0$ ($p.value \approx 0.0147 < 0.05$)
- $\Rightarrow 60\%$ de la variabilité des coordonnées des individus sur l'axe 2 expliquée par la variable "Region"

Descrip. dim. : variables qualitatives (suppl.)

\$Dim.2

\$Dim.2\$category

	Estimate	p.value
NO	2.050365	0.001011474
SE	-0.9738852	0.047120253

- "Estimate" = moyenne des coord. des villes NO (resp. SE) sur l'axe 2
- coordonnées des villes du nord-ouest significativement positive sur l'axe 2 ; celles des villes du sud-est de la France significativement négative (toujours sur l'axe 2)

Méthodologie de l'ACP en bref

- ① Choisir les variables quantitatives actives (participent à la construction des axes) ; les autres variables (supplémentaires) peuvent servir à les interpréter
- ② Choisir de réduire ou non les variables (i.e. s'il existe ou non des variations d'échelle entre les variables)
- ③ Réaliser l'ACP
- ④ Choisir le nombre de dimensions à interpréter
- ⑤ Interpréter simultanément le graphe des individus et celui des variables
- ⑥ Utiliser les indicateurs pour enrichir l'interprétation (contributions, cosinus carré, etc)

Reconstruction des données

K axes factoriels engendrés par $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_K$
(vecteurs propres de la matrice des corrélations)
= base orthonormée de l'espace des individus

↪ $\mathbf{x}_i = P_{\mathbf{u}_1} \mathbf{x}_i + P_{\mathbf{u}_2} \mathbf{x}_i + \cdots + P_{\mathbf{u}_K} \mathbf{x}_i$ avec

$$\begin{aligned}P_{\mathbf{u}_k} &= \text{projection orthogonale sur } \mathbf{u}_k \\&= \underbrace{\langle \mathbf{x}_i, \mathbf{u}_k \rangle}_{\substack{\text{coord. de } \mathbf{x}_i \\ \text{sur } \mathbf{u}_k}} \mathbf{u}_k\end{aligned}$$

↪ $\mathbf{x}_i = \langle \mathbf{x}_i, \mathbf{u}_1 \rangle \mathbf{u}_1 + \langle \mathbf{x}_i, \mathbf{u}_2 \rangle \mathbf{u}_2 + \cdots + \langle \mathbf{x}_i, \mathbf{u}_K \rangle \mathbf{u}_K$

Reconstruction des données

$$\hookrightarrow \mathbf{x}_i^T = \langle \mathbf{x}_i, \mathbf{u}_1 \rangle \mathbf{u}_1^T + \langle \mathbf{x}_i, \mathbf{u}_2 \rangle \mathbf{u}_2^T + \cdots + \langle \mathbf{x}_i, \mathbf{u}_K \rangle \mathbf{u}_K^T$$

où le vecteur ligne \mathbf{u}_k^T est le transposé du vecteur colonne \mathbf{u}_k .

Matriciellement, on a

$$\mathbf{x}_i^T = [\langle \mathbf{x}_i, \mathbf{u}_1 \rangle, \langle \mathbf{x}_i, \mathbf{u}_2 \rangle, \dots, \langle \mathbf{x}_i, \mathbf{u}_K \rangle] \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_K^T \end{bmatrix}$$

Reconstruction des données

$$\mathbf{X} = \text{matrice des données} = \begin{bmatrix} \frac{\mathbf{x}_1^T}{\| \mathbf{x}_1 \|} \\ \frac{\mathbf{x}_2^T}{\| \mathbf{x}_2 \|} \\ \vdots \\ \frac{\mathbf{x}_n^T}{\| \mathbf{x}_n \|} \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} \frac{\mathbf{x}_1^T}{\| \mathbf{x}_1 \|} \\ \frac{\mathbf{x}_2^T}{\| \mathbf{x}_2 \|} \\ \vdots \\ \frac{\mathbf{x}_n^T}{\| \mathbf{x}_n \|} \end{bmatrix} = \underbrace{\begin{bmatrix} \langle \mathbf{x}_1, \mathbf{u}_1 \rangle & \langle \mathbf{x}_1, \mathbf{u}_2 \rangle & \dots & \langle \mathbf{x}_1, \mathbf{u}_K \rangle \\ \langle \mathbf{x}_2, \mathbf{u}_1 \rangle & \langle \mathbf{x}_2, \mathbf{u}_2 \rangle & \dots & \langle \mathbf{x}_2, \mathbf{u}_K \rangle \\ \vdots & \vdots & \dots & \vdots \\ \langle \mathbf{x}_n, \mathbf{u}_1 \rangle & \langle \mathbf{x}_n, \mathbf{u}_2 \rangle & \dots & \langle \mathbf{x}_n, \mathbf{u}_K \rangle \end{bmatrix}}_F \begin{bmatrix} \frac{\mathbf{u}_1^T}{\| \mathbf{u}_1 \|} \\ \frac{\mathbf{u}_2^T}{\| \mathbf{u}_2 \|} \\ \vdots \\ \frac{\mathbf{u}_K^T}{\| \mathbf{u}_K \|} \end{bmatrix}$$

où les K colonnes de la matrice F sont les K c.p. (coord. des ind. sur chaque axe factoriel) $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_K$.

Reconstruction des données

La matrice des données \mathbf{X} se décompose de la façon suivante :

$$\mathbf{X} = [\mathbf{F}_1 \mid \mathbf{F}_2 \mid \cdots \mid \mathbf{F}_K] \begin{bmatrix} \mathbf{u}_1^T \\ \hline \mathbf{u}_2^T \\ \vdots \\ \hline \mathbf{u}_K^T \end{bmatrix}$$

$$\widehat{\mathbf{X}}^1 = \mathbf{F}_1 \mathbf{u}_1^T = \text{meilleure approx. 1D de } \mathbf{X}$$

$$\widehat{\mathbf{X}}^2 = \mathbf{F}_1 \mathbf{u}_1^T + \mathbf{F}_2 \mathbf{u}_2^T = \text{meilleure approx. 2D de } \mathbf{X}$$

$$\widehat{\mathbf{X}}^k = \mathbf{F}_1 \mathbf{u}_1^T + \mathbf{F}_2 \mathbf{u}_2^T + \cdots + \mathbf{F}_k \mathbf{u}_k^T = \text{meilleure approx. } k\text{D de } \mathbf{X}$$

Reconstruction des données par l'exemple

5 villes parmi les 47 communes de Gironde :

	immeubles	densite.pop	proprietaires
HAUX	-0.49	0.12	-0.18
MOMBRIER	0.45	0.20	1.19
PREIGNAC	0.90	0.86	-0.73
SAVIGNAC-DE-L'ISLE	0.15	0.52	0.84
SIGALENS	-0.98	-1.31	0.54

Reconstruction des données par l'exemple

- Coord. des ind. sur les axes factoriels 1 et 2

	Dim. 1	Dim. 2
Haux	-0.19	-0.24
Mombrier	0.07	1.29
Preignac	1.42	-0.32
Savignac-de-l'Isle	0.19	0.96
Sigalens	-1.72	0.01

- $\mathbf{u}_1 = [0.67, 0.67, -0.31]^T$ et $\mathbf{u}_2 = [0.20, 0.23, 0.95]^T$

Exercice

Pour chacune de ces communes, calculer la meilleure approximation 1D puis 2D.

Correction

meilleure approximation 1D = $\underbrace{\begin{bmatrix} -0.19 \\ 0.07 \\ 1.42 \\ 0.19 \\ -1.72 \end{bmatrix}}_{\text{coord. sur axe 1}} \underbrace{[0.67, 0.67, -0.31]}_{u_1}$

= $\begin{bmatrix} -0.13 & -0.13 & 0.06 \\ 0.05 & 0.05 & -0.02 \\ 0.95 & 0.95 & -0.44 \\ 0.13 & 0.13 & -0.06 \\ -1.15 & -1.15 & 0.53 \end{bmatrix}$

Par exemple, $-0.13 \simeq -0.19 \times 0.67$ et $-0.44 \simeq 1.42 \times -0.31$

Reconstruction des données par l'exemple

Extrait des données initiales

	immeubles	densite.pop	proprietaires
HAUX	-0.49	0.12	-0.18
MOMBRIER	0.45	0.20	1.19
PREIGNAC	0.90	0.86	-0.73
SAVIGNAC-DE-L'ISLE	0.15	0.52	0.84
SIGALENS	-0.98	-1.31	0.54

Meilleure approximation 1D

	immeubles	densite.pop	proprietaires
HAUX	-0.13	-0.13	0.06
MOMBRIER	0.05	0.05	-0.02
PREIGNAC	0.95	0.95	-0.44
SAVIGNAC-DE-L'ISLE	0.13	0.13	-0.06
SIGALENS	-1.15	-1.15	0.53

Correction

meilleure approximation 2D

= meilleure approximation 1D

$$+ \underbrace{\begin{bmatrix} -0.24 \\ 1.29 \\ -0.32 \\ 0.96 \\ 0.01 \end{bmatrix}}_{\text{coord. sur axe 2}} \underbrace{[0.20, 0.23, 0.95]}_{u_2}$$

$$= \begin{bmatrix} -0.13 & -0.13 & 0.06 \\ 0.05 & 0.05 & -0.02 \\ 0.95 & 0.95 & -0.44 \\ 0.13 & 0.13 & -0.06 \\ -1.15 & -1.15 & 0.53 \end{bmatrix} + \begin{bmatrix} -0.05 & -0.05 & -0.23 \\ 0.26 & 0.30 & 1.23 \\ -0.06 & -0.07 & -0.30 \\ 0.19 & 0.22 & 0.91 \\ 0.00 & 0.00 & 0.01 \end{bmatrix}$$

Reconstruction des données par l'exemple

Extrait des données initiales

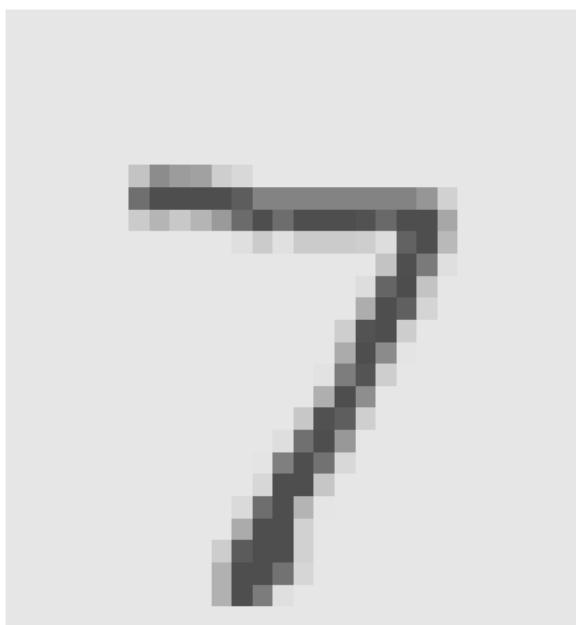
	immeubles	densite.pop	proprietaires
HAUX	-0.49	0.12	-0.18
MOMBRIER	0.45	0.20	1.19
PREIGNAC	0.90	0.86	-0.73
SAVIGNAC-DE-L'ISLE	0.15	0.52	0.84
SIGALENS	-0.98	-1.31	0.54

Meilleure approximation 2D

	immeubles	densite.pop	proprietaires
HAUX	-0.18	-0.18	-0.17
MOMBRIER	0.30	0.34	1.20
PREIGNAC	0.89	0.88	-0.75
SAVIGNAC-DE-L'ISLE	0.32	0.35	0.85
SIGALENS	-1.15	-1.15	0.54

ACP : outil pour la reconstruction d'images

1er chiffre écrit à la main



1 image = 1 individu

300 chiffres écrits à la main

72104149590690159784
96654074013134727121
17423512446355604195
78937464307029173297
16278473613693141769
60549921948739744492
54767905856657810164
67317182029955156034
46546545144723271818
18508925011109031642
36111395294593903655
72271284173388792241
59872304424195772826
85779181803019941821
29759264154292040028

ACP : outil pour la reconstruction d'images

1 image = 28 pixels \times 28 pixels (784 pixels)

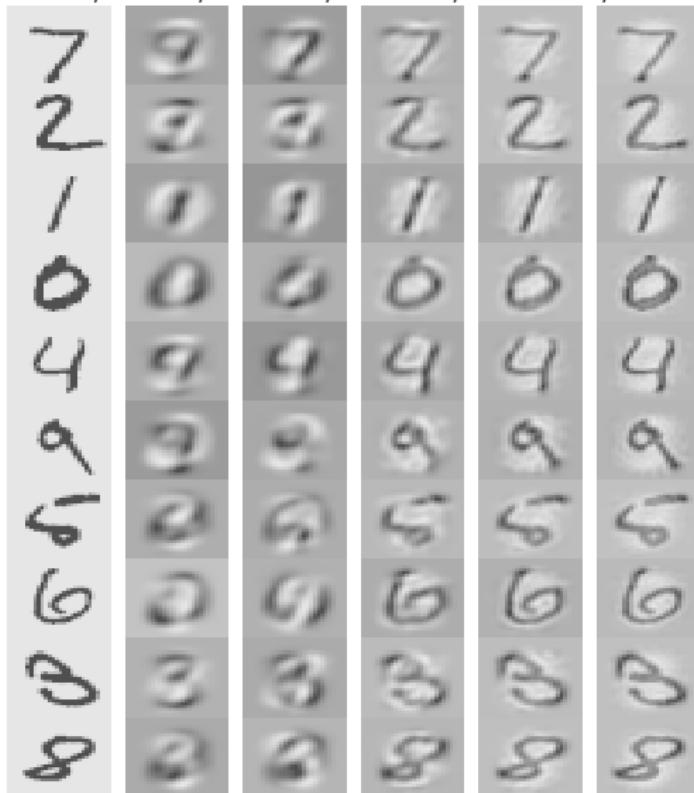
1 pixel = 1 niveau de gris (0-255)



- 1 image vectorisée = vecteur de longueur 784 (ici colonne par colonne)
- 1 image vectorisée = 784 niveaux de gris (0-255)
- 10,000 images = 10,000 individus
- tableau des données = 10,000 lignes \times 784 colonnes

ACP : reconstruction des images 3D, 10D, ...

obs. / 3D / 10D / 50D / 100D / 200D



- tableau initial = $10,000 \times 784 = 7,840,000$ éléments
- approx. 50D nécessite F_1, \dots, F_{50} et u_1, \dots, u_{50} soit $10,000 \times 50 + 784 \times 50 = 539,200$ éléments
- approx. 50D nécessite 15 fois moins d'espace mémoire !

Analyse Factorielle des Correspondances

Reconnaissance de 3 saveurs (sucré, acide, amer)

	Perçu_sucré	Perçu_acide	Perçu_amer
Sucré	10	0	0
Acide	0	9	1
Amer	0	3	7

Croisement de 2 variables

- 1 variable **saveur proposée X**
 → 3 modalités : *Sucré*, *Acide* et *Amer*
- 1 variable **saveur perçue Y**
 → 3 modalités : *Perçu_sucré*, *Perçu_acide* et *Perçu_amer*

Peut-on décrire/représenter la relation entre ces 2 variables ?

Données "Orientation universitaire et les CSP"

Enquête sociologique réalisée auprès de 10000 étudiants en France

Croisement de 2 variables

- 1 variable CSP du père X
 → 9 modalités : *Exp. Agricole, Salarié, Patron, Libéral, Cadre Moyen, Employé, Ouvrier, Service, Autres*
- 1 variable orientation universitaire Y
 → 8 modalités : *Droit, Sc. Eco., Lettres, Sciences, Med., Pharma., Pluridis., IUT*

Choix des études universitaires vs CSP du père

	Droit	Sc. Eco.	Lettres	Sciences	Med.	Pharma.	Pluridis.	IUT
Exp. Agricole	80	36	134	99	65	28	11	58
Salarié	6	2	15	6	4	1	1	4
Patron	168	74	312	137	208	53	21	62
Libéral	470	191	806	400	876	164	45	79
Cadre Moyen	236	99	493	264	281	56	36	87
Employé	145	52	281	133	135	30	20	54
Ouvrier	166	64	401	193	127	23	28	129
Service	16	6	27	11	8	2	2	8
Autres	305	115	624	247	301	47	42	90

Peut-on décrire/représenter la relation entre ces 2 variables ?

Enquête sociologique sur le travail des femmes

Ces données relèvent de l'histoire contemporaine et sont issues d'une enquête du CREDOC (N. Tabard, 1974).

1724 femmes ont répondu simultanément aux deux questions à choix multiples suivantes :

- Pour vous, la famille idéale est celle où :
 - les deux conjoints travaillent également ?
 - le mari a un métier plus absorbant que celui de sa femme ?
 - seul le mari travaille ?
- Pour vous, quelle activité convient 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 ?

Travail des femmes et vie de famille en 1974

Croisement de 2 variables

- 1ère question = variable X

↪ 3 modalités :

1 = *les deux conjoints travaillent également*

2 = *le mari a un métier plus absorbant que sa femme*

3 = *seul le mari travaille*

- 2ème question = variable Y

↪ 3 modalités :

1 = *rester au foyer*

2 = *travailler à mi-temps*

3 = *travailler à plein-temps*

Travail des femmes et vie de famille en 1974

	rester au foyer	travailler à mi-temps	travailler à plein- temps
les deux conjoints travaillent également	13	142	106
le mari a un métier plus absorbant que celui de sa femme	30	408	117
seul le mari travaille	241	573	94

Peut-on décrire/représenter la relation entre ces 2 variables ?

Table de contingence et tableau des fréquences

Des données brutes à la **table de contingence**

The diagram shows a transition from individual data to a frequency matrix. On the left, a vertical axis labeled "individus" lists individuals from 1 to n . Above them are two columns labeled X and Y , each containing two entries: a blue "X" and a red "X". To the right of this is a downward-pointing arrow. Next to it, the text "modalités de X" is written vertically. To the right of the arrow is a large bracket spanning the entire width of the diagram, labeled "modalités de Y" in red at the top. Below this bracket is a large rectangular box representing a matrix. The matrix has n rows, labeled $1, \dots, i, \dots, L$ in blue, and C columns, labeled $1, \dots, c, \dots, C$ in red. The entries in the matrix are labeled n_{ij} for row i and column j . The entire matrix is labeled $= T$ at the bottom right.

n_{Ic} = nb d'individus prenant les modalités (I , c) de (X , Y)

$$n_{11} + n_{12} + \cdots + n_{LC} = n \text{ (i.e. taille d'échantillon)}$$

Table de contingence et tableau des fréquences

Tableau des fréquences

modalités de Y

	1	...	c	...	C
1	f_{11}	...	f_{1c}	...	f_{1C}
⋮	⋮		⋮		⋮
I	f_{I1}	...	f_{Ic}	...	f_{IC}
⋮	⋮		⋮		⋮
L	f_{L1}	...	f_{Lc}	...	f_{LC}

= F avec $f_{Ic} = n_{Ic} / n$

f_{Ic} = fréquence du couple de modalités (I , c) de (X , Y)

$$f_{11} + f_{12} + \cdots + f_{LC} = 1$$

Table de contingence et tableau des fréquences

Fréquences marginales

		modalités de Y					fréquences marginales de X	
		1	...	c	...	C		
modalités de X	1	f_{11}	...	f_{1c}	...	f_{1C}	$f_{1\bullet}$	
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
	I	f_{I1}	...	f_{Ic}	...	f_{IC}	$f_{I\bullet}$	
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
	L	f_{L1}	...	f_{Lc}	...	f_{LC}	$f_{L\bullet}$	
		$f_{\bullet 1}$...	$f_{\bullet c}$...	$f_{\bullet C}$		

fréquences marginales de Y

$$f_{I\bullet} = f_{I1} + f_{I2} + \cdots + f_{IC} \text{ et } f_{\bullet c} = f_{1c} + f_{2c} + \cdots + f_{Lc}$$

Objectif de l'AFC

Étudier la relation entre les 2 variables catégorielles X et Y en examinant l'écart entre

- les données observées
- celles que l'on devraient observer s'il n'existe aucune liaison/relation entre variables X et Y . On parle alors d'indépendance entre les 2 variables X et Y

Indépendance entre 2 variables catégorielles

En probabilité, A et B sont dits **indépendants** SSI

$$Prob(A \text{ et } B) = Prob(A) \times Prob(B)$$

Traduction empirique de l'indépendance

- A = "prend la modalité i de la variable X "
- B = "prend la modalité c de la variable Y "
- f_{Ic} = proba. empirique d'avoir A et B
- $f_{I\bullet}$ = proba. empirique d'avoir A
- $f_{\bullet c}$ = proba. empirique d'avoir B

$$X \text{ et } Y \text{ indépendantes} \Leftrightarrow f_{Ic} = f_{I\bullet} \times f_{\bullet c}$$

Terminologie : **probabilité empirique = fréquence**

Indépendance entre 2 variables catégorielles

indépendance SSI

fréquences conjointes = produit des fréquences marginales

De façon équivalente, on a **indépendance SSI**

$$\frac{f_{Ic}}{f_{I\bullet}} = f_{\bullet c} \text{ ou bien } \frac{f_{Ic}}{f_{\bullet c}} = f_{I\bullet}$$

$f_{Ic}/f_{I\bullet}$ = proba. qu'1 individu prenne la modalité c de Y sachant qu'il possède la modalité I de X

$f_{Ic}/f_{\bullet c}$ = proba. qu'1 individu prenne la modalité I de X sachant qu'il possède la modalité c de Y

	rester foyer	au	travailler à mi-temps	travailler à plein-temps	Fréq. de X	marg.
les deux conjoints tra- vaillent également	0.008		0.082	0.061		0.151
le mari a un métier plus absorbant que celui de sa femme	0.017		0.24	0.068		0.322
seul le mari travaille	0.140		0.332	0.055		0.527
Fréq. marg. de Y	0.165		0.651	0.184	1	

Proba qu'1 femme réponde "rester au foyer" sachant qu'elle a choisi la modalité "seul le mari travaille" ? (rép. : $0.140 / \textcolor{blue}{0.527} = 0.266$)

Proba qu'1 femme choisisse "les 2 conjoints travaillent également" sachant qu'elle a répondu "travailler à mi-temps" ?
(rép. $0.061 / \textcolor{red}{0.184} = 0.332$)

Intensité de liaison entre 2 variables catégorielles

Principe

Mesurer la liaison entre 2 variables catégorielles X et Y revient à mesurer l'écart entre les fréquences observées f_{Ic} et les fréquences théoriques $f_{I\bullet} \times f_{\bullet c}$ que l'on devrait obtenir si X et Y étaient indépendantes

Interprétation

Plus les fréquences observées f_{Ic} seront proches des fréquences théoriques $f_{I\bullet} \times f_{\bullet c}$, plus l'intensité de la liaison entre X et Y sera faible.

Indice de liaison : phi-deux

$$\begin{aligned}\Phi_{obs}^2 &\stackrel{\text{déf}}{=} \sum_{l=1}^L \sum_{c=1}^C \frac{(\text{fréq. observé} - \text{fréq. théorique})^2}{\text{fréq. théorique}} \\ &= \sum_{l=1}^L \sum_{c=1}^C \frac{(f_{lc} - f_{l\bullet} f_{\bullet c})^2}{f_{l\bullet} f_{\bullet c}}\end{aligned}$$

Indice de liaison : phi-deux

- Φ_{obs}^2 mesure par définition l'écart entre les fréq. obs. et les fréq. théoriques obtenues sous hypothèse d'indépendance
- Φ_{obs}^2 quantifie l'intensité de la liaison entre 2 variables catégorielles
- Φ_{obs}^2 ne dépend pas de la taille de l'échantillon mais uniquement des effectifs

Remarque : khi-deux = $n \Phi_{obs}^2$

L'AFC travaille sur le tableau des fréquences (ou probabilités empiriques) et a pour objectif de visualiser la liaison entre les 2 variables afin de pouvoir la décrire

Profil-lignes et profil-colonnes

Analyse des lignes : on étudie les écarts entre $\frac{f_{Ic}}{f_{I\bullet}}$ et $f_{\bullet c}$

Analyse des colonnes : on étudie les écarts entre $\frac{f_{Ic}}{f_{\bullet c}}$ et $f_{I\bullet}$

Profil-lignes

Les **profil-lignes** sont obtenus en divisant chaque ligne du tableau des fréquences (ou probabilités) observées par la fréquence (ou probabilité) marginale

On appelle *I*ème **profil-ligne** le profil de la ligne *I* défini par les probabilités conditionnelles empiriques :

$$\left(\frac{f_{I1}}{f_{I\bullet}}, \dots, \frac{f_{Ic}}{f_{I\bullet}}, \dots, \frac{f_{IC}}{f_{I\bullet}} \right) = \left(\frac{n_{I1}}{n_{I\bullet}}, \dots, \frac{n_{Ic}}{n_{I\bullet}}, \dots, \frac{n_{IC}}{n_{I\bullet}} \right)$$

*I*ème profil-ligne = fréq. conditionnelles des modalités de **Y** sachant que l'individu prend la modalité *I* de **X**

Tableau des profil-lignes

Définition à l'aide des **fréquences**

		modalités de Y					
		1	...	c	...	C	somme des lignes
modalités de X	1	f_{11}		f_{1c}		f_{1C}	1
		$\frac{f_{11}}{f_{1\bullet}}$		$\frac{f_{1c}}{f_{1\bullet}}$		$\frac{f_{1C}}{f_{1\bullet}}$	
I	1	f_{I1}		f_{Ic}		f_{IC}	1
		$\frac{f_{I1}}{f_{I\bullet}}$		$\frac{f_{Ic}}{f_{I\bullet}}$		$\frac{f_{IC}}{f_{I\bullet}}$	
L	1	f_{L1}		f_{Lc}		f_{LC}	1
		$\frac{f_{L1}}{f_{L\bullet}}$		$\frac{f_{Lc}}{f_{L\bullet}}$		$\frac{f_{LC}}{f_{L\bullet}}$	

Tableau des profil-lignes

Définition à l'aide des **effectifs**

		modalités de Y					somme des lignes
		1	...	c	...	C	
modalités de X	1	$\frac{n_{11}}{n_{1\bullet}}$...	$\frac{n_{1c}}{n_{1\bullet}}$...	$\frac{n_{1C}}{n_{1\bullet}}$	1
	\vdots	\vdots		\vdots		\vdots	
I	I	$\frac{n_{I1}}{n_{I\bullet}}$...	$\frac{n_{Ic}}{n_{I\bullet}}$...	$\frac{n_{IC}}{n_{I\bullet}}$	1
	\vdots	\vdots		\vdots		\vdots	
L	L	$\frac{n_{L1}}{n_{L\bullet}}$...	$\frac{n_{Lc}}{n_{L\bullet}}$...	$\frac{n_{LC}}{n_{L\bullet}}$	1
	\vdots	\vdots		\vdots		\vdots	

Profil-ligne moyen \overline{PL}

Profil-ligne moyen = fréq. marginales des modal. de Y :

$$\overline{PL} = (f_{\bullet 1}, \dots, f_{\bullet c}, \dots, f_{\bullet L})$$

$f_{\bullet c}$ = moyenne des quantités $\frac{f_{1c}}{f_{1\bullet}}, \dots, \frac{f_{lc}}{f_{l\bullet}}, \dots, \frac{f_{Lc}}{f_{L\bullet}}$ formant la colonne c pondérée par les poids $f_{1\bullet}, \dots, f_{l\bullet}, \dots, f_{L\bullet}$

L'objectif de l'AFC est de visualiser l'écart à l'indépendance :

- ⇒ comparer l'écart entre fréq. cond. et fréq. marg.
- ⇒ étudier l'écart entre profil-lignes et profil-ligne moyen \overline{PL}

Profil-colonnes

Les **profil-colonnes** sont obtenus en divisant chaque colonne du tableau des fréquences (ou probabilités) observées par la fréquence (ou probabilité) marginale

On appelle **cème profil-colonne** le profil de la colonne **c** défini par les probabilités conditionnelles empiriques :

$$\left(\frac{f_{1c}}{f_{\bullet c}}, \dots, \frac{f_{Ic}}{f_{\bullet c}}, \dots, \frac{f_{Lc}}{f_{\bullet c}} \right) = \left(\frac{n_{1c}}{n_{\bullet c}}, \dots, \frac{n_{Ic}}{n_{\bullet c}}, \dots, \frac{n_{Lc}}{n_{\bullet c}} \right)$$

cème profil-colonne = fréq. conditionnelles des modalités de **X**
sachant que l'individu prend la modalité **c** de **Y**

Tableau des profil-colonnes à l'aide des fréquences

		modalités de Y				
		1	...	C	...	C
modalités de X	1	$\frac{f_{11}}{f_{\bullet 1}}$...	$\frac{f_{1c}}{f_{\bullet c}}$...	$\frac{f_{1C}}{f_{\bullet C}}$
	2	\vdots	\vdots	\vdots	\vdots	\vdots
	I	$\frac{f_{I1}}{f_{\bullet 1}}$...	$\frac{f_{Ic}}{f_{\bullet c}}$...	$\frac{f_{IC}}{f_{\bullet C}}$
	3	\vdots	\vdots	\vdots	\vdots	\vdots
	L	$\frac{f_{L1}}{f_{\bullet 1}}$...	$\frac{f_{Lc}}{f_{\bullet c}}$...	$\frac{f_{LC}}{f_{\bullet C}}$

somme des colonnes

1 ... 1 ... 1

Tableau des profil-colonnes à l'aide des effectifs

		modalités de Y				
		1	...	c	...	C
modalités de X	1	$\frac{n_{11}}{n_{\bullet 1}}$...	$\frac{n_{1c}}{n_{\bullet c}}$...	$\frac{n_{1C}}{n_{\bullet C}}$
	2	\vdots		\vdots		\vdots
	I	$\frac{n_{I1}}{n_{\bullet 1}}$...	$\frac{n_{Ic}}{n_{\bullet c}}$...	$\frac{n_{IC}}{n_{\bullet C}}$
	2	\vdots		\vdots		\vdots
	L	$\frac{n_{L1}}{n_{\bullet 1}}$...	$\frac{n_{Lc}}{n_{\bullet c}}$...	$\frac{n_{LC}}{n_{\bullet C}}$
		1	...	1	...	1
somme des colonnes						

Profil-colonne moyen \overline{PC}

Profil-colonne moyen = fréq. marginales des modal. de X :

$$\overline{PC} = (f_{1\bullet}, \dots, f_{I\bullet}, \dots, f_{L\bullet})$$

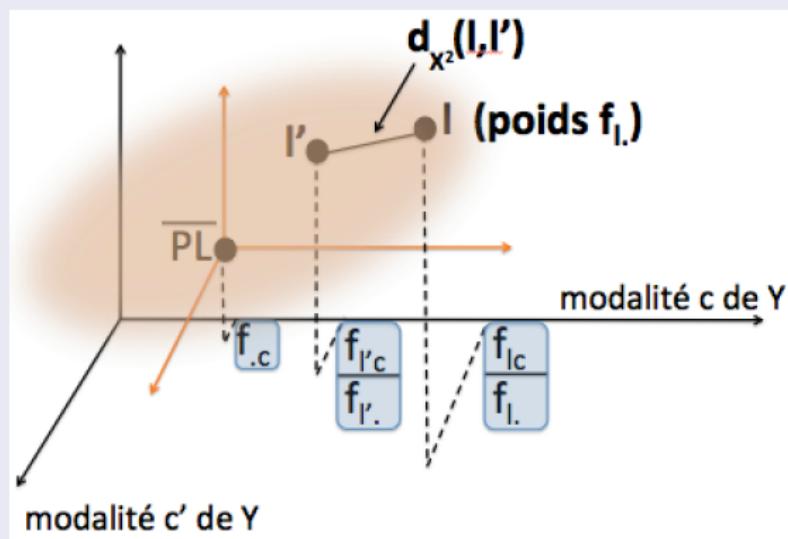
$f_{I\bullet}$ = moyenne des quantités $\frac{f_{I1}}{f_{\bullet 1}}, \dots, \frac{f_{Ic}}{f_{\bullet c}}, \dots, \frac{f_{IC}}{f_{\bullet C}}$ formant la ligne I pondérée par les poids $f_{\bullet 1}, \dots, f_{\bullet c}, \dots, f_{\bullet C}$

L'objectif de l'AFC est de visualiser l'écart à l'indépendance :

- ⇒ comparer l'écart entre fréq. cond. et fréq. marg.
- ⇒ étudier l'écart entre profil-colonnes et profil-colonne moyen \overline{PC}

Nuage des profil-lignes

Les L profil-lignes appartiennent à un espace de dimension C (chaque profil-ligne est identifié par C coordonnées)



Distance du khi-deux entre 2 profil-lignes I et I'

$$d_{\chi^2}(I, I') = \sum_{c=1}^C \frac{1}{f_{\bullet c}} \left(\frac{f_{lc}}{f_{l\bullet}} - \frac{f_{l'c}}{f_{l'\bullet}} \right)^2$$

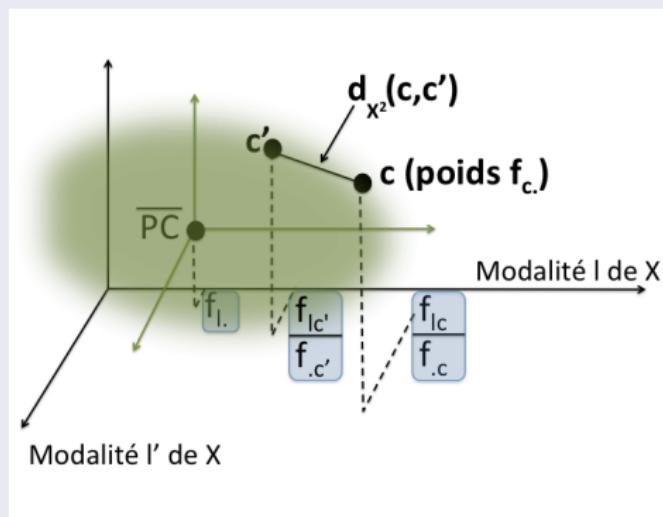
Interprétation : $d_{\chi^2}(I, I')$ = somme des carrés des écarts entre les 2 profil-lignes I et I' , chaque terme de cette somme étant pondéré par l'inverse du poids de la modalité de Y correspondante

Distance du khi-deux entre profil-ligne et \overline{PL}

$$d_{\chi^2}(I, \overline{PL}) = \sum_{c=1}^C \frac{1}{f_{\bullet c}} \left(\frac{f_{lc}}{f_{l\bullet}} - f_{\bullet c} \right)^2$$

Nuage des profil-colonnes

Les C profil-colonnes appartiennent à un espace de dimension L
(chaque profil-colonne est identifié par L coordonnées)



Distance du khi-deux entre 2 profil-colonnes I et I'

$$d_{\chi^2}(c, c') = \sum_{l=1}^L \frac{1}{f_{l\bullet}} \left(\frac{f_{lc}}{f_{\bullet c}} - \frac{f_{lc'}}{f_{\bullet c'}} \right)^2$$

Interprétation : $d_{\chi^2}(c, c')$ = somme des carrés des écarts entre les 2 profil-colonnes c et c' , chaque terme de cette somme étant pondéré par l'inverse du poids de la modalité de X correspondante

Distance du khi-deux entre profil-colonne et \overline{PC}

$$d_{\chi^2}(c, \overline{PC}) = \sum_{l=1}^L \frac{1}{f_{l\bullet}} \left(\frac{f_{lc}}{f_{\bullet c}} - f_{l\bullet} \right)^2$$

Inertie du nuage des profil-lignes

Inertie du profil I

$$\text{inertie}(I) \stackrel{\text{déf.}}{=} f_{I\bullet} d_{\chi^2}^2(I, \overline{PL})$$

Inertie du profil-ligne I = distance au carré entre profil-ligne et \overline{PL} pondéré par le poids $f_{I\bullet}$ de la modalité I de X

Inertie du nuage des profil-lignes

$$\text{inertie}(PL) = \sum_{I=1}^L \text{inertie}(I) = \sum_{I=1}^L f_{I\bullet} d_{\chi^2}^2(I, \overline{PL})$$

Inertie du nuage des profil-colonnes

Inertie du profil c

$$\text{inertie}(c) \stackrel{\text{déf.}}{=} f_{\bullet c} d_{\chi^2}^2(c, \overline{PC})$$

Inertie du profil-colonne c = dist. au carré entre profil-colonne et \overline{PC} pondéré par le poids $f_{\bullet c}$ de la modalité c de Y

Inertie du nuage des profil-colonnes

$$\text{inertie}(PC) = \sum_{c=1}^C \text{inertie}(c) = \sum_{c=1}^C f_{\bullet c} d_{\chi^2}^2(c, \overline{PC})$$

Particularité des inerties en AFC

$$inertie(PC) = \Phi_{obs}^2 = inertie(PL)$$

En AFC, l'inertie mesure l'intensité de la liaison entre les 2 variables catégorielles

En AFC, **lignes et colonnes jouent des rôles symétriques** contrairement à l'ACP

Que se passe-t-il lorsqu'il y a indépendance entre les 2 variables ?

Du point de vue des profil-lignes

indépendance entre X et $Y \Leftrightarrow d_{\chi^2}(I, \overline{PL}) = \mathbf{0}$ pour tout I

Du point de vue des profil-colonnes

indépendance entre X et $Y \Leftrightarrow d_{\chi^2}(c, \overline{PC}) = \mathbf{0}$ pour tout c

Du point de vue de l'inertie

X et Y indép $\Leftrightarrow \text{inertie}(PL) = \mathbf{0} = \text{inertie}(PC)$

L'AFC a pour objectif de représenter les profils-lignes et profils-colonnes dans des sous-espaces de petites dimensions (2D voire 3D)

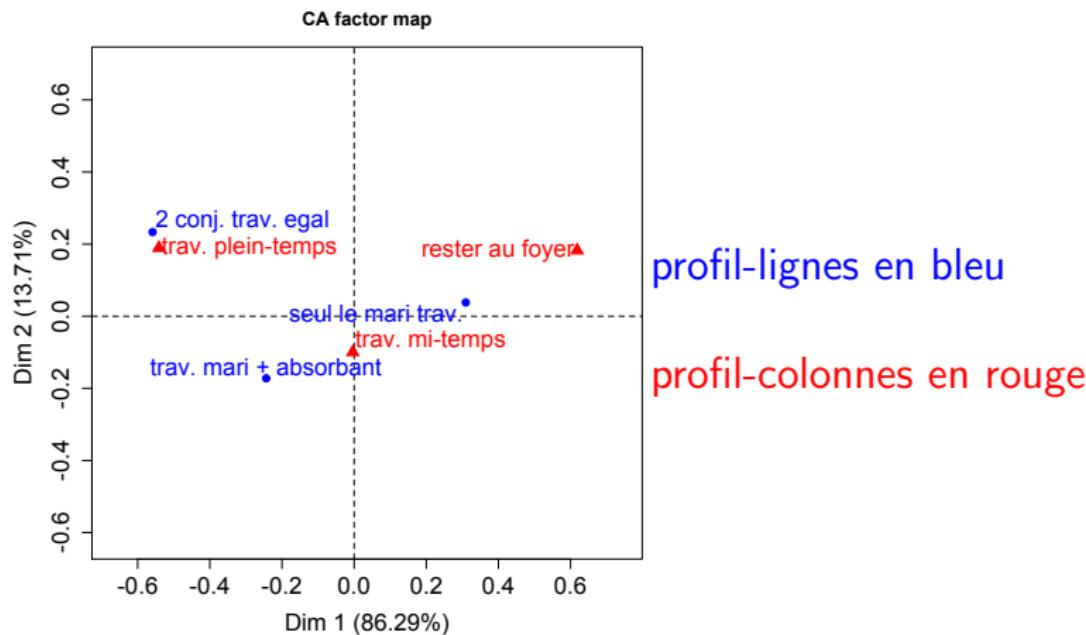
L'AFC opère une décomposition de l'inertie du nuage des profils (i.e. profil-lignes ou profil-colonnes)

L'AFC projette le nuage des profils sur une suite d'axes orthogonaux $\mathbf{u}_1, \mathbf{u}_2, \dots$, la projection sur chaque axe étant d'inertie maximale :

- \mathbf{u}_1 d'inertie maximale notée λ_1 ,
- \mathbf{u}_2 d'inertie maximale notée $\lambda_2 (< \lambda_1)$ avec $\mathbf{u}_2 \perp \mathbf{u}_1$,
- ainsi de suite.

Exemple : condition féminine en 1974

Représentation simultanée des modalités des 2 variables



Exemple : condition féminine en 1974

Profil-colonnes (%)

	rester au foyer	trav. mi-temps	trav. plein-temps	\overline{PC}
2 conj. trav. égal	4.58	12.64	33.44	15.14
trav. mari + absorbant	10.56	36.33	36.91	32.19
seul le mari trav.	84.86	51.02	29.65	52.67

trav. mi-temps proche de l'origine car similaire à \overline{PC}

trav. plein-temps éloigné de l'origine car très différent de \overline{PC}

Exemple : condition féminine en 1974

Interprétation axe 1

- modalités de la variable *famille idéale* ordonnées de la plus favorable au travail des femmes à la moins favorable
- modalités de la variable *activité professionnelle mère de famille* ordonnées de la modalité la plus favorable au travail des femmes à la moins favorable

Conclusion : l'axe 1 traduit l'attitude à l'égard du travail féminin, du plus favorable (à gauche) au moins favorable (à droite)

Qualité de représentation du nuage des profils

L'AFC construit une suite d'axes orthogonaux u_1, u_2, \dots associés à des inerties décroissantes $\lambda_1 > \lambda_2 > \dots$

Qualité de représentation du nuage sur l'axe s

$$\frac{\text{inertie du nuage projeté sur } u_s}{\text{inertie totale du nuage}} = \frac{\lambda_s}{\lambda_1 + \lambda_2 + \dots}$$

Exemple : condition féminine en 1974

	inertie	inertie (%)
axe 1	0.117	86.292
axe 2	0.019	13.708
Somme	0.135	100

Qualité de représentation du nuage des profils

Exemple : condition féminine en 1974

	inertie	inertie (%)
axe 1	0.117	86.292
axe 2	0.019	13.708
Somme	0.135	100

Interprétation

le 1er axe représente environ 86% de l'inertie totale

→ écart à l'indépendance bien résumé par l'axe 1

→ l'interprétation de l'AFC peut se contenter de l'étude du 1er axe

Propriétés des inerties

Propriété d'additivité des inerties

somme des λ_s = inertie du nuage des profils

Exemple : condition féminine en 1974

inertie du nuage des profils = $0.117 + 0.019 = 0.135$

Rappel : inertie totale = intensité de lien entre les deux variables (i.e.
 $0.135 = \Phi^2$)

Choix du nombre d'axes à retenir pour l'analyse

Nombre maximum d'axes d'inertie non nulle

- L = nb de modalités de X
- C = nb de modalités de Y
- $N_{max} = \min \{L - 1, C - 1\}$

N_{max} = nb maxi d'axes d'inertie non nulle

Conséquence : $\Phi^2 \leq \min \{L - 1, C - 1\}$

→ définir l'intensité d'une liaison entre 2 variables catégorielles par un indicateur borné (appelé V de Cramér) entre 0 et 1 :

$$V \text{ de Cramér} = \frac{\Phi^2}{\min\{L - 1, C - 1\}} \in [0 ; 1]$$

Intensité de liaison : V de Cramér

Travail féminin		
	inertie	%
axe 1	0.117	86.292
axe 2	0.019	13.708
somme	0.135	100

$$V = 0.135/2 = 0.0675$$

Trois saveurs (faible confusion)		
	inertie	%
axe 1	1	72.727
axe 2	0.375	27.273
somme	1.375	100

$$V = 1.375/2 = 0.6875$$

Trois saveurs (forte confusion)		
	inertie	%
axe 1	1	96
axe 2	0.042	4
somme	1.042	100

$$V=1.042/2 = 0.521$$

Choix du nombre d'axes à retenir pour l'analyse

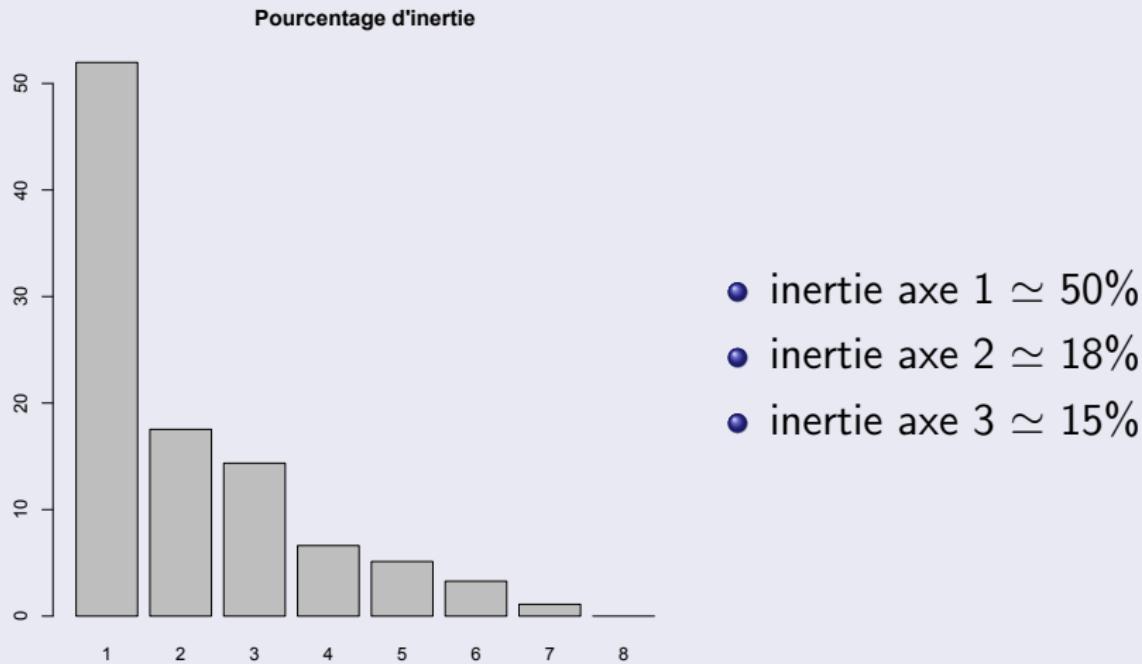
La décroissance des inerties peut suggérer le nombre d'axes à conserver dans l'analyse

Exemple : enquête portant sur les raisons pouvant faire hésiter une femme ou un couple pour avoir un enfant

2 variables catégorielles :

- raisons influençant négativement sur le choix d'avoir un enfant (situation économique, peur d'avoir un enfant, pb de santé, futur incertain, chomage, etc) → 18 modalités
- niveau d'étude auquel s'ajoute des tranches d'âge → 8 modalités

Choix du nombre d'axes à retenir pour l'analyse



Inertie et structure de la table de contingence

En AFC, l'inertie associée à chaque axe est toujours comprise entre 0 et 1 (i.e. $0 \leq \lambda_s \leq 1$ pour tout s)

cas limite $\lambda_s = 1$

Interprétation :

		modalités de Y	
		C_1	C_2
modalités de X	L_1		0
	L_2	0	

- axe s oppose parfaitement le bloc de modalités L_1 au bloc L_2 ainsi que le bloc C_1 au bloc C_2
- toutes les modalités des deux blocs L_1 et C_1 (resp. L_2 et C_2) sont confondues sur l'axe s

Inertie et structure de la table de contingence

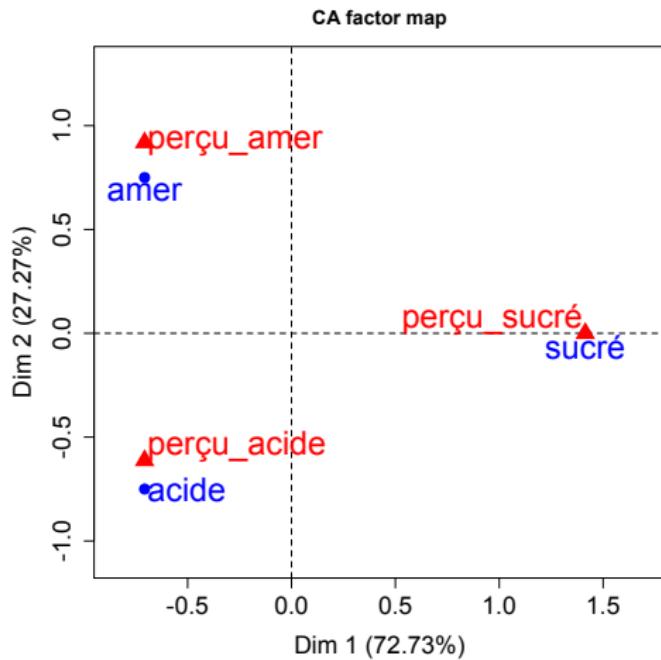
Exemple : perception trois saveurs (sucré, acide, amer)

	perçu sucré	perçu acide	perçu amer
sucré	10	0	0
acide	0	9	1
amer	0	3	7

	inertie	inertie (%)
axe 1	1	72.727
axe 2	0.375	27.273
Somme	1.375	100

Inertie et structure de la table de contingence

Exemple : perception trois saveurs (sucré, acide, amer)



On retrouve les groupes de modalités confondues sur l'axe 1 :

- sucré et perçu_sucré
- amer/acide et perçu_amer/perçu_acide

Inertie et structure de la table de contingence

Augmentation de la confusion acide/amer ?

Confusion *acide/amer* faible

	perçu sucré	perçu acide	perçu amer
sucré	10	0	0
acide	0	9	1
amer	0	3	7

Confusion *acide/amer* augmentée

	perçu sucré	perçu acide	perçu amer
sucré	10	0	0
acide	0	7	3
amer	0	5	5

Inertie et structure de la table de contingence

Inertie avec confusion *acide/amer faible*

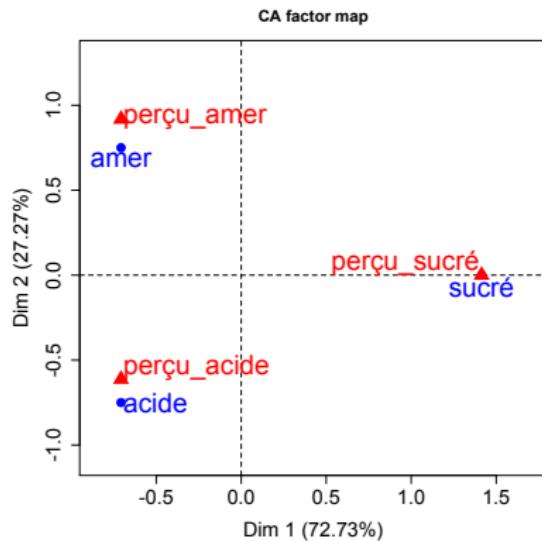
	inertie	inertie (%)
axe 1	1	72.727
axe 2	0.375	27.273
Somme	1.375	100

Inertie avec confusion *acide/amer augmentée*

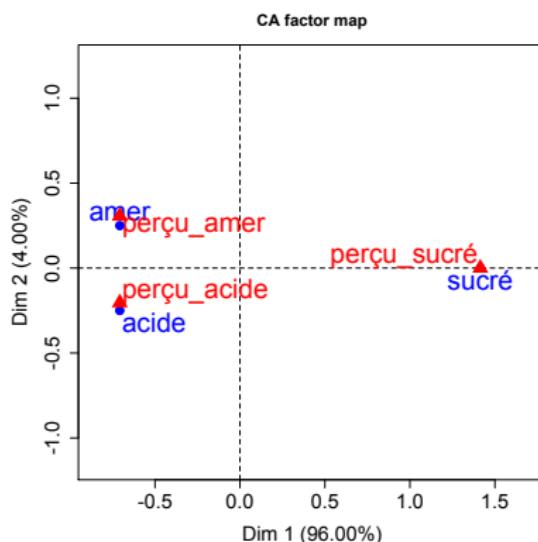
	inertie	inertie (%)
axe 1	1	96
axe 2	0.042	4
Somme	1.042	100

Inertie et structure de la table de contingence

Confusion acide/amer
faible



Confusion acide/amer
augmentée



Inertie et structure de la table de contingence

Exemple : opinion sur le travail des femmes

	inertie	inertie (%)
axe 1	0.117	86.292
axe 2	0.019	13.708
Somme	0.135	100

- $\lambda_1 = 0.117 << 1$: on est loin d'une association exclusive (i.e. liaison parfaite) entre des groupes de modalités des 2 variables
- $\Phi^2 = 0.135 << 2$: on est loin d'une liaison parfaite entre des groupes de modalités des 2 variables

Représentation barycentrique

Formule de transition de Y vers X

$$Mod_s^X(l) = \frac{1}{\sqrt{\lambda_s}} \sum_{c=1}^C \frac{f_{lc}}{f_{l\bullet}} Mod_s^Y(c) \text{ avec}$$

- $Mod_s^X(l)$ = coordonnée de la modalité l de X sur l'axe de rang s
- $Mod_s^Y(c)$ = coordonnée de la modalité c de Y sur l'axe de rang s
- $\frac{f_{lc}}{f_{l\bullet}}$ = c ème élément du profil-ligne l
- λ_s = inertie associé à l'axe s ($\lambda_s \leq 1$)

$$\sum_{c=1}^C \frac{f_{lc}}{f_{l\bullet}} Mod_s^Y(c) \text{ barycentre des modal. de } Y \text{ avec les poids } \frac{f_{lc}}{f_{l\bullet}}$$

Représentation barycentrique

coordonnées d'1 modalité de $\textcolor{blue}{X}$ = barycentre dilaté des coordonnées des modalités de $\textcolor{red}{Y}$ car pour tout axe $\textcolor{green}{s}$, $\frac{1}{\sqrt{\lambda_s}} > 1$

Si la proba. conditionnelle $f_{Ic}/f_{I\bullet}$ est très grande, alors la coordonnée de la modal. c de $\textcolor{red}{Y}$ sur l'axe $\textcolor{green}{s}$ compte beaucoup pour calculer la coordonnée de la modalité I de $\textcolor{blue}{X}$

Une modalité de $\textcolor{blue}{X}$ sera représentée du côté des modalités de $\textcolor{red}{Y}$ auxquelles elle s'associe le plus.

Représentation barycentrique

Formule de transition de X vers Y

$$Mod_s^Y(c) = \frac{1}{\sqrt{\lambda_s}} \sum_{l=1}^L \frac{f_{lc}}{f_{\bullet c}} Mod_s^X(l) \text{ avec}$$

- $Mod_s^Y(c)$ = coordonnée de la modalité c de Y sur l'axe de rang s
- $Mod_s^X(l)$ = coordonnée de la modalité l de X sur l'axe de rang s
- $\frac{f_{lc}}{f_{\bullet c}}$ = l ème élément du profil-colonne c
- λ_s = inertie associé à l'axe s ($\lambda_s \leq 1$)

$$\sum_{l=1}^L \frac{f_{lc}}{f_{\bullet c}} Mod_s^X(l) \text{ barycentre des modal. de } X \text{ avec les poids } \frac{f_{lc}}{f_{\bullet c}}$$

Représentation barycentrique

coordonnées d'1 modalité de $\textcolor{red}{Y}$ = barycentre dilaté des coordonnées des modalités de $\textcolor{blue}{X}$ car pour tout axe $\textcolor{green}{s}$, $\frac{1}{\sqrt{\lambda_s}} > 1$

Si la proba. conditionnelle $f_{Ic}/f_{\bullet c}$ est très grande, alors la coordonnée de la modal. I de $\textcolor{blue}{X}$ sur l'axe $\textcolor{green}{s}$ compte beaucoup pour calculer la coordonnée de la modalité c de $\textcolor{blue}{X}$

Une modalité de $\textcolor{red}{Y}$ sera représentée du côté des modalités de $\textcolor{blue}{X}$ auxquelles elle s'associe le plus.

Représentation barycentrique

Exemple des 3 saveurs avec peu de confusions :

	perçu sucré	perçu acide	perçu amer
sucré	10	0	0
acide	0	9	1
amer	0	3	7

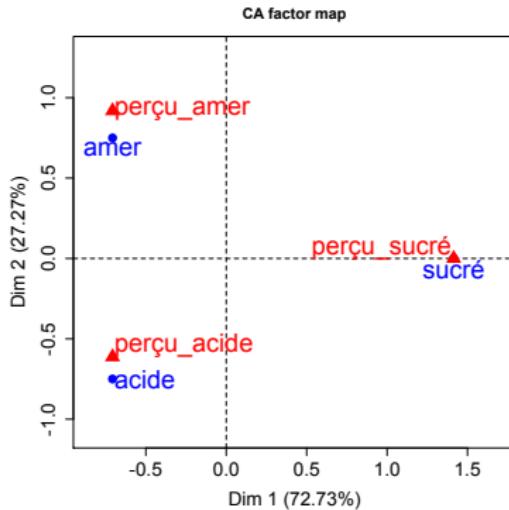
	inertie	inertie (%)
axe 1	1	72.727
axe 2	0.375	27.273
Somme	1.375	100

profil-ligne 1 = (1, 0, 0) implique que :

$$\begin{aligned}Mod_1^X(1) &= \frac{1}{\sqrt{1}} \{1 \times Mod_1^Y(1) + 0 \times Mod_1^Y(2) + 0 \times Mod_1^Y(3)\} \\&= Mod_1^Y(1)\end{aligned}$$

coord. de *sucré* sur l'axe 1 = coord. de *perçu_sucré* sur l'axe 1

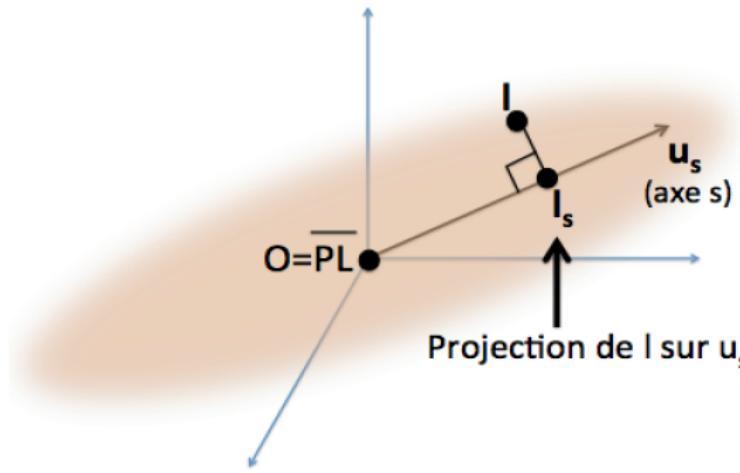
Représentation simultanée



En conclusion, l'AFC visualise la **nature** de la liaison. Pour obtenir des précisions sur l'**intensité** de la liaison, il faudra consulter les **inerties** associées à chaque axe.

Qualité de représentation : cosinus carré

Nuage des profil-lignes



La qualité de représentation de la modalité I sur l'axe s est définie par le rapport

$$\frac{\text{inertie projetée de } I \text{ sur } u_s}{\text{inertie totale de } I} = \frac{f_{I \bullet} \times d_{\chi^2}^2(I_s, \overline{PL})}{f_{I \bullet} \times d_{\chi^2}^2(I, \overline{PL})} = \cos^2(\overrightarrow{OI}, u_s)$$

Qualité de représentation : cosinus carré

Exemple des trois saveurs avec une confusion élevée

	perçu sucré	perçu acide	perçu amer
sucré	10	0	0
acide	0	7	3
amer	0	5	5

	Qualité de représentation (\cos^2)	
	Axe 1	Axe 2
sucré	1.000	0.000
acide	0.889	0.111
amer	0.889	0.111
perçu sucré	1.000	0.000
perçu acide	0.923	0.000
perçu amer	0.842	0.152

Lorsqu'on a beaucoup de profils (modalités), pour commencer une interprétation, on sélectionne quelques profils qui ont à la fois des coordonnées remarquables (i.e. éloignées le long de l'axe que l'on étudie) et à la fois qui ont une bonne qualité de représentation

La contribution d'une modalité l est définie par :

$$\text{contrib}(l) = \frac{\text{inertie projetée de } l \text{ sur } u_s}{\text{inertie de l'axe } s} == \frac{f_{l\bullet} \times d_{\chi^2}^2(l_s, \overline{PL})}{\lambda_s}$$

Interprétation

- si les effectifs marginaux sont équilibrés, la consultation des contributions apporte très peu par rapport aux coordonnées (ces dernières représentant bien les contributions dans ce cas)
- si on constate des différences d'effectifs marginaux importantes, alors la consultation des contributions est indispensable

AFC : un outil d'analyse textuelle

Équivalence distributionnelle

Si plusieurs lignes ayant le même profil sont regroupées en une seule, les résultats de l'AFC sont strictement équivalents (idem pour le regroupement de colonnes)

Application à l'analyse textuelle

Analyse textuelle = étude des occurrences de mots (corpus de textes) :
→ si 2 mots (ou plus) sont employés dans les mêmes circonstances, leurs coordonnées sont proches et faire l'analyse avec les deux termes ou avec un terme unique qui regroupe ces deux notions conduit strictement aux mêmes résultats. Cela peut être notamment très utile pour réaliser des regroupements de singuliers et pluriels, de conjugaison de verbes, de synonymes, etc

Analyse Factorielle des Correspondances

Multiples

Analyse Factorielle des Correspondances Multiples (AFCM)

AFCM = outil graphique de la statistique descriptive multidimensionnelle

Objectif

Représenter et décrire les liaisons qui peuvent exister entre plusieurs variables catégorielles (variables dont les modalités sont des catégories)

→ extension de l'AFC à 3, 4, ... variables catégorielles

Micro-organismes de financement

492 individus

→ 492 micro-organismes de financement

5 variables catégorielles

- ① **Region** → 6 modalités : Africa / East Asia and the Pacific / Eastern Europe and Central Asia / Latin America and The Caribbean / Middle East and North Africa / South Asia,
- ② **Age** → 3 modalités : Mature / Young / New,
- ③ **legal status** → 5 modalités : Bank / Credit-Union-Cooperative / NBFI / NGO / Rural Bank,
- ④ **Profit** → 2 modalités : no / yes,
- ⑤ **Scale** → 3 modalités : Large / Medium / Small.

Micro-organismes de financement

Région	Age	legal status	profit	scale	...
South Asia	mature	NGO	no	medium	...
Africa	mature	bank	yes	large	...
South Asia	young	NBFI	yes	medium	...
Africa	mature	NBFI	yes	medium	...
South Asia	mature	NGO	no	small	...
South Asia	mature	NGO	no	large	...
Africa	mature	NGO	no	small	...
Africa	mature	NGO	no	medium	...
Africa	mature	NBFI	yes	large	...
Africa	young	bank	yes	large	...
South Asia	young	bank	yes	large	...
:	:	:	:	:	:

individus = patientes atteintes d'un cancer du sein

4 variables catégorielles

- ① **Centre hospitalier** → 3 modalités : Boston / Glamorgan / Tokyo
- ② **Age** → 3 modalités : < 50 / $50 - 69$ / ≥ 70 ,
- ③ **Survie** → 2 modalités : oui / non
- ④ **Type inflammation** → 4 modalités : Petite bénigne / Petite maligne / Grande bénigne / Grande maligne

Centre	Age	Survie	Petite Inflammation		Grande Inflammation	
			Maligne	Bénigne	Maligne	Bénigne
Tokyo	< 50	non	9	7	4	3
		oui	26	68	25	9
Tokyo	50 - 69	non	9	9	11	2
		oui	20	46	18	5
Tokyo	≥ 70	non	2	3	1	0
		oui	1	6	5	1
Boston	< 50	non	6	7	6	0
		oui	11	24	4	0
Boston	50 - 69	non	8	20	3	2
		oui	18	58	10	3
Boston	≥ 70	non	9	18	3	0
		oui	15	26	1	1
Glamorgan	< 50	non	16	7	3	0
		oui	16	20	8	1
Glamorgan	50 - 69	non	14	12	3	0
		oui	27	39	10	4
Glamorgan	> 70	non	3	7	3	0

Activités de loisir des français

8403 individus

22 variables catégorielles

4 variables signalétiques (Sexe, Age, Situation Matrimoniale, CSP) +
18 variables sur les activités de loisir (ci-dessous les 6 1ères) :

- ① Lecture (Oui / Non)
- ② Ecouter de la musique (Oui / Non)
- ③ Aller au cinéma (Oui / Non)
- ④ Aller au spectacle (théâtre, concert, danse, ...) (Oui / Non)
- ⑤ Visiter une exposition, un musée, ... (Oui / Non)
- ⑥ Utiliser un ordinateur, une console de jeux (Oui / Non)

+ 1 variables quantitatives = nombre total d'acitivités

Activités de loisir des français

Lecture	Ecouter musique	Cinema	Spectacle	...	Cuisine	Peche	TV	Sexe	Age	Situation	Profession	Nb activites
O	O	O	O	...	N	N	2	F	(55,65]	Marié	cadre	11
O	N	N	N	...	N	N	4	M	(45,55]	Marié	NA	9
O	O	N	N	...	N	N	4	F	(25,35]	Remarié	cadre	5
O	N	N	N	...	N	N	1	M	(75,85]	Marié	NA	5
O	N	O	O	...	N	O	4	F	(35,45]	Seul	cadre	8
N	O	N	N	...	O	O	2	M	(25,35]	Seul	agent de maîtrise	9
N	N	N	N	...	O	N	2	F	(55,65]	Marié	employé	7
:	:	:	:	:	:	:	:	:	:	:	:	:

Codage disjonctif complet

Exemple artificiel

		variables		
		V1	V2	V3
individus	1	A1	B2	C3
	2	B1	A2	A3
	3	B1	B2	B3
	4	C1	B2	A3
	5	C1	A2	B3
	6	B1	B2	C3

V1 → A1 / B1 / C1 (3 modalités)

V2 → A2 / B2 (2 modalités)

V3 → A3 / B3 / C3 (3 modalités)

Codage disjonctif complet

Codage de V1

V1	A1	B1	C1
A1	1	0	0
B1	0	1	0
B1	0	1	0
C1	0	0	1
C1	0	0	1
B1	0	1	0

Codage disjonctif complet

Codage de V2 et V3

V2	A2	B2	V3	A3	B3	C3
B2	0	1	B3	0	0	1
A2	1	0	A3	1	0	0
B2	0	1	B3	0	1	0
B2	0	1	B3	1	0	0
A2	1	0	A3	0	1	0
B2	0	1	C3	0	0	1

Codage disjonctif complet

En reprenant ces 3 transformations, on obtient le **tableau disjonctif complet** (TDC)

individus	variables			modalités							
	V1	V2	V3	A1	B1	C1	A2	B2	A3	B3	C3
1	A1	B2	C3	1	1	0	0	1	0	0	1
2	B1	A2	A3	2	0	1	0	1	0	1	0
3	B1	B2	B3	3	0	1	0	0	1	0	0
4	C1	B2	A3	4	0	0	1	0	1	1	0
5	C1	A2	B3	5	0	0	1	1	0	0	1
6	B1	B2	C3	6	0	1	0	0	1	0	0

Tableau Disjonctif Complet (TDC)

		modalités								
		1	2	3	4	5	6	7	8	Σ
individus	1	1	0	0	0	1	0	0	1	3
	2	0	1	0	1	0	1	0	0	3
	3	0	1	0	0	1	0	1	0	3
	4	0	0	1	0	1	1	0	0	3
	5	0	0	1	1	0	0	1	0	3
	6	0	1	0	0	1	0	0	1	3
	Σ	1	3	2	2	4	2	2	2	
$\sum = 6$				$\sum = 6$			$\sum = 6$			

Tableau Disjonctif Complet (TDC)

p variables catégorielles :

- $V^1 \rightarrow m_1$ modalités
- $V^2 \rightarrow m_2$ modalités
-
- $V^p \rightarrow m_p$ modalités

- TDC contient n lignes et $m = \sum_{k=1}^p m_k$ colonnes
- somme de chaque ligne du TDC = p
- somme de chaque colonne du TDC = effectif de la modalité

Notations

- y_{ij} élément du TDC corresp. à l'individu i et la modalité j :

$$y_{ij} = \begin{cases} 0 & \text{si l'individu } i \text{ ne prend pas la modalité } j \\ 1 & \text{si l'individu possède la modalité } j \end{cases}$$

- $y_{i\bullet} \stackrel{\text{déf}}{=} \sum_{j=1}^m y_{ij} =$ somme des éléments de la ligne $i = p$
- $\sum_{j \in \mathcal{M}_k} y_{ij} = 1$ où \mathcal{M}_k = indices des modalités de la variable k
- $y_{\bullet j} \stackrel{\text{déf}}{=} \sum_{j=1}^n y_{ij} =$ nb individus prenant la modalité j
- $\sum_{j \in \mathcal{M}_k} y_{\bullet j} = n =$ nb total individus (i.e. taille échantillon)

TDC en résumé

		modalités des p variables					
		1	\cdots	j	\cdots	m	Σ
individus	1	y_{11}	\cdots	y_{1j}	\cdots	y_{1m}	$y_{1\bullet} = p$
	\vdots	\vdots		\vdots		\vdots	\vdots
	i	y_{i1}	\cdots	y_{ij}	\cdots	y_{im}	$y_{i\bullet} = p$
	\vdots	\vdots		\vdots		\vdots	\vdots
	n	y_{n1}	\cdots	y_{nj}	\cdots	y_{nm}	$y_{n\bullet} = p$
		Σ	$y_{\bullet 1}$	\cdots	$y_{\bullet j}$	\cdots	$y_{\bullet m}$

$$\text{où } np = \sum_{j=1}^m \sum_{i=1}^n y_{ij} = \sum_{j=1}^m y_{\bullet j} = \sum_{i=1}^n y_{i\bullet}$$

- 1 individu = 1 ligne du TDC = 1 ensemble de modalités
- 2 individus se ressemblent s'ils ont choisi majoritairement les mêmes modalités
- 2 individus sont différents s'il y a peu de modalités en commun dans leurs réponses

Objectif de l'AFCM

L'ensemble des **ressemblances** et **différences** entre les individus constitue ce qu'on appelle la **variabilité des individus** :

- l'AFCM explore cette **variabilité** d'un point de vue multidimensionnel
- la façon d'étudier cette **variabilité** est d'en extraire les principales **dimensions** sur lesquelles elle s'exprime
- mise en évidence des **dimensions** qui séparent des individus moyens et des individus extrêmes
- ces **dimensions** seront décrites en relation avec les modalités

Comme ici les variables sont catégorielles, on va s'intéresser aux **associations entre modalités**

- 1 variable = 1 colonne du TDC = 1 ensemble de modalités
- 2 variables catégorielles sont liées si les modalités de l'une s'associent aux modalités de l'autre de façon particulière

Objectif de l'AFCM

L'un des objectifs de l'AFCM est de fournir une visualisation d'ensemble de ces associations entre modalités :

- l'AFCM va construire des **variables synthétiques** qui résument au mieux les variables catégorielles observées
- l'AFCM va fournir des indicateurs quantitatifs fondés sur ces **variables synthétiques**

Rappel TDC

modalités des p variables

	1	\cdots	j	\cdots	m	Σ
1	y_{11}	\cdots	y_{1j}	\cdots	y_{1m}	$y_{1\bullet} = p$
\vdots	\vdots		\vdots		\vdots	\vdots
i	y_{i1}	\cdots	y_{ij}	\cdots	y_{im}	$y_{i\bullet} = p$
\vdots	\vdots		\vdots		\vdots	\vdots
n	y_{n1}	\cdots	y_{nj}	\cdots	y_{nm}	$y_{n\bullet} = p$

$$\Sigma \boxed{y_{\bullet 1} \quad \cdots \quad y_{\bullet j} \quad \cdots \quad y_{\bullet m}} \quad np$$

où $y_{ij} = \begin{cases} 0 & \text{si l'individu } i \text{ ne prend pas la modalité } j \\ 1 & \text{si l'individu possède la modalité } j \end{cases}$

Principes de l'AFCM

- on affecte à chaque individu le même poids $\frac{1}{n}$
- si un individu possède une modalité rare, cela le caractérise beaucoup plus qu'une modalité fréquente
↪ on divise l'élément y_{ij} du TDC par la fréquence
$$f_{\bullet j} \stackrel{\text{déf}}{=} \frac{1}{n} \sum_{i=1}^n y_{ij} \text{ de la modalité } j : z_{ij} \stackrel{\text{déf}}{=} \frac{y_{ij}}{f_{\bullet j}}$$
- on affecte à chaque modalité j le poids $p_j \stackrel{\text{déf}}{=} \frac{y_{\bullet j}}{n p} = \frac{f_{\bullet j}}{p}$

Exercice

- Montrer que $\sum_{j=1}^m p_j = 1$
- Pour toute modalité j , montrer que : $(1/n) \sum_{i=1}^n z_{ij} = 1$

Principes de l'AFCM (suite)

- En AFCM, on s'intéresse au TDC **standardisé** :

		modalités des p variables					Σ
		1	...	j	...	m	
individus	1	x_{11}	...	x_{1j}	...	x_{1m}	$x_{1\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	i	x_{i1}	...	x_{ij}	...	x_{im}	$x_{i\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	n	x_{n1}	...	x_{nj}	...	x_{nm}	$x_{n\bullet}$
		Σ	0	...	0	...	0

$$\text{avec } x_{ij} = z_{ij} - 1 = \frac{y_{ij}}{f_{\bullet j}} - 1$$

Nuage des individus

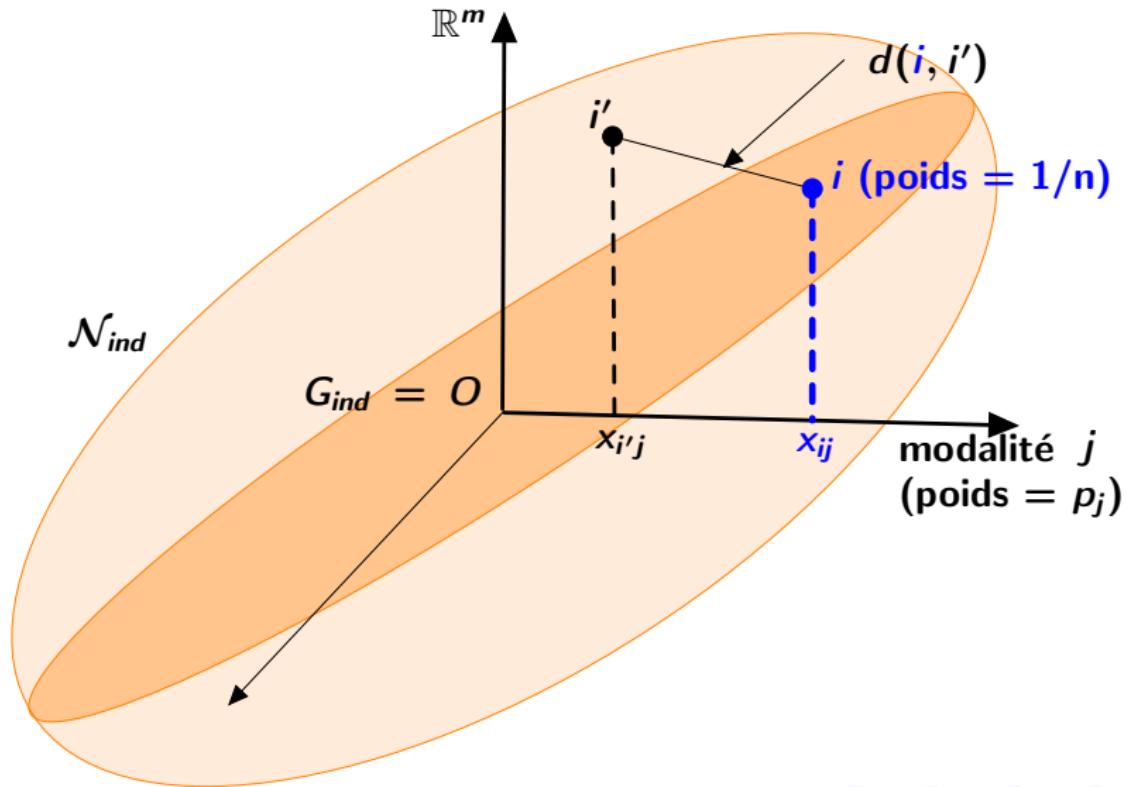
On appelle **nuage des individus** (\mathcal{N}_{ind}) l'ensemble des points-lignes du TDC **standardisé**

Chaque point-ligne (i.e. individu) possède m coordonnées :

		modalités des p variables					Σ
		1	\dots	j	\dots	m	
individus	1	x_{11}	\dots	x_{1j}	\dots	x_{1m}	$x_{1\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	i	x_{i1}	\dots	x_{ij}	\dots	x_{im}	$x_{i\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	n	x_{n1}	\dots	x_{nj}	\dots	x_{nm}	$x_{n\bullet}$

$$\Sigma \quad \boxed{0 \quad \dots \quad 0 \quad \dots \quad 0}$$

Géométrie du nuage des individus



Distance entre 2 individus

$$d(i, i')^2 \stackrel{\text{déf}}{=} \sum_{j=1}^m p_j (x_{ij} - x_{i'j})^2 = \frac{1}{p} \sum_{j=1}^m \frac{1}{f_{\bullet j}} (y_{ij} - y_{i'j})^2$$

- si i et i' prennent les mêmes modalités, alors $d(i, i') = 0$
- si i et i' ont en commun beaucoup de modalités, alors $d(i, i')$ petite
- si on dispose de 2 individus dont l'un possède une modalité **rare** (i.e. $f_{\bullet j}$ **petit**), alors leur distance est relativement grande
- si 2 individus ont en commun une modalité rare, alors leur distance est relativement petite

Géométrie du nuage des individus

Distance d'un individu à l'origine

$$d(i, G_{ind})^2 \stackrel{\text{déf}}{=} \sum_{j=1}^m p_j x_{ij}^2$$

Exercice. Montrer que $d(i, G_{ind})^2 = \frac{1}{p} \sum_{j=1}^m \frac{y_{ij}^2}{f_{\bullet j}} - 1$

Plus un individu possède des modalités rares, plus il sera éloigné de l'origine du nuage des points

Inertie d'un individu i

$$Inertie(i) \stackrel{\text{déf}}{=} \frac{1}{n} d(i, G_{ind})^2$$

Inertie du nuage des individus

$$Inertie(\mathcal{N}_{ind}) \stackrel{\text{déf}}{=} \frac{1}{n} \sum_{i=1}^n d(i, G_{ind})^2$$

Propriété - Exercice

$$Inertie(\mathcal{N}_{ind}) = \frac{m}{p} - 1$$

En AFCM, l'inertie totale dépend uniquement du format des données contrairement à l'AFC

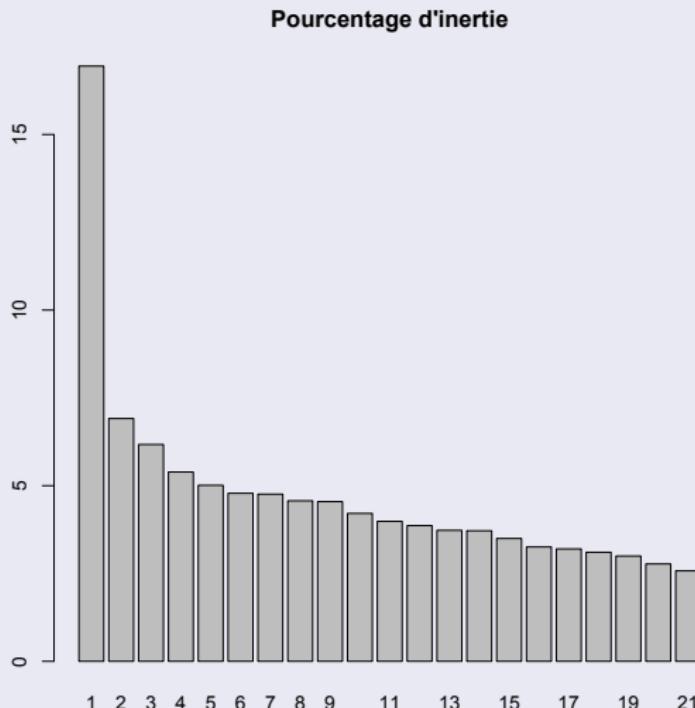
Ajustement du nuage des individus

Comme toute analyse factorielle, on **projette** ce nuage des individus sur une **suite d'axes orthogonaux** pour le visualiser dans un sous-espace de plus **petite dimension** (2D voire 3D)

- **1er axe factoriel** = direction qui explique la plus grande part de variabilité (i.e. inertie)
- **2ème axe factoriel** = direction orthogonale au 1er axe factoriel qui représente la plus grande part de variabilité restante
- et ainsi de suite.

Exemple : Activités de loisir des français

% d'inertie expliquée par chaque axe



plus l'inertie représentée par un axe est petite, plus son interprétation sera difficile

Exemple : Activités de loisir des français

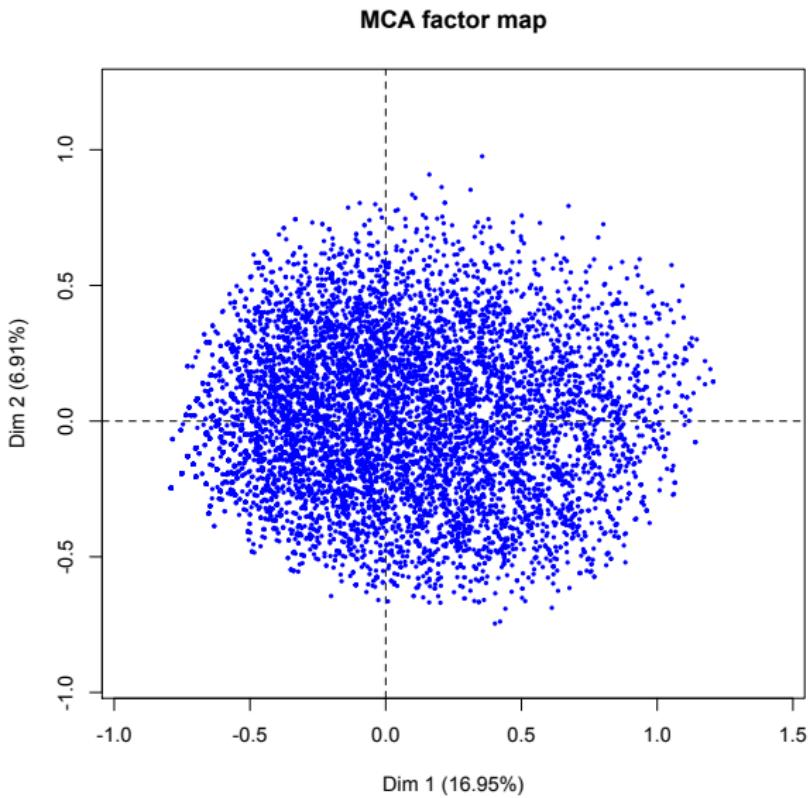
- 1er axe factoriel $\simeq 17\%$ d'inertie $>> 7\%$ (2ème axe)
- \rightarrow interpréter le 1er plan factoriel (i.e. axes 1 et 2) en priorité qui totalise $17 + 7 = 24\%$ de l'inertie

Remarque : **18 variables**

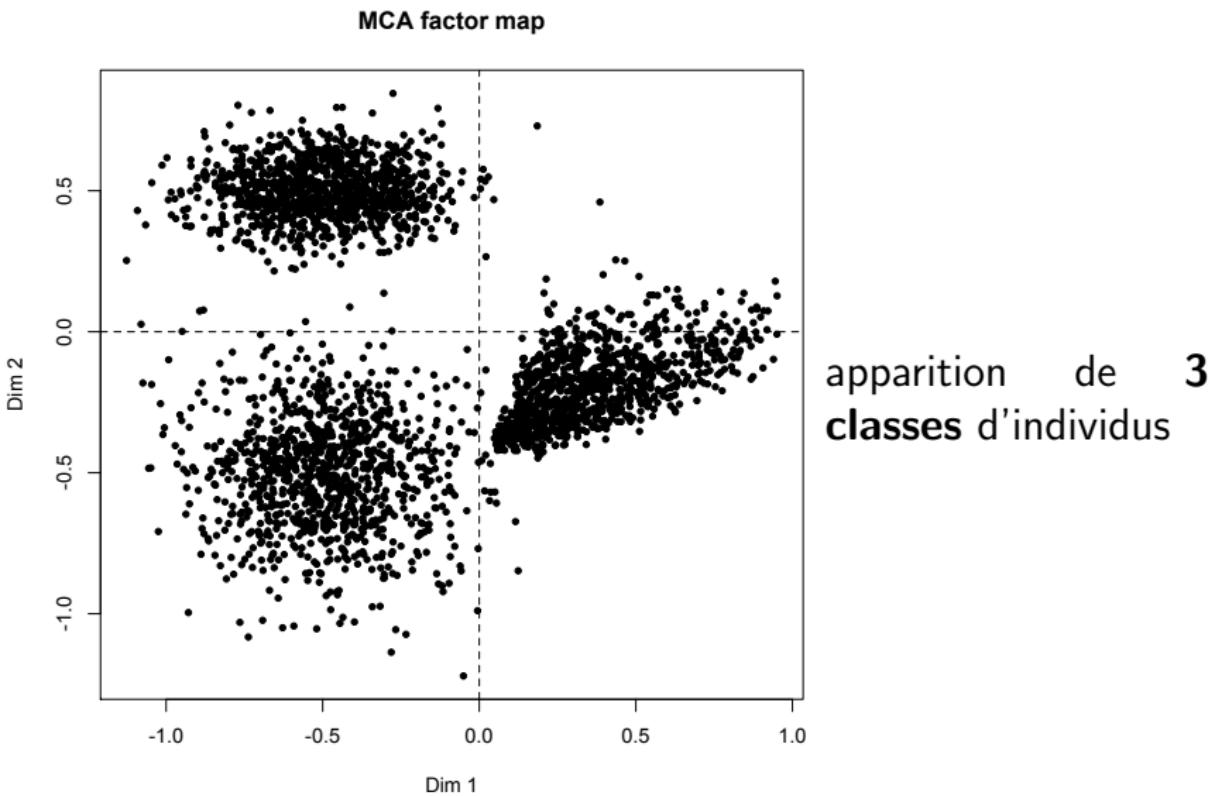
- \rightarrow profils d'activités de loisir complexes
- \rightarrow la variabilité ne peut pas s'exprimer uniquement sur les 2 1ers axes factoriels

Dans ce contexte, obtenir 24 % d'inertie sur le 1er plan factoriel est satisfaisant

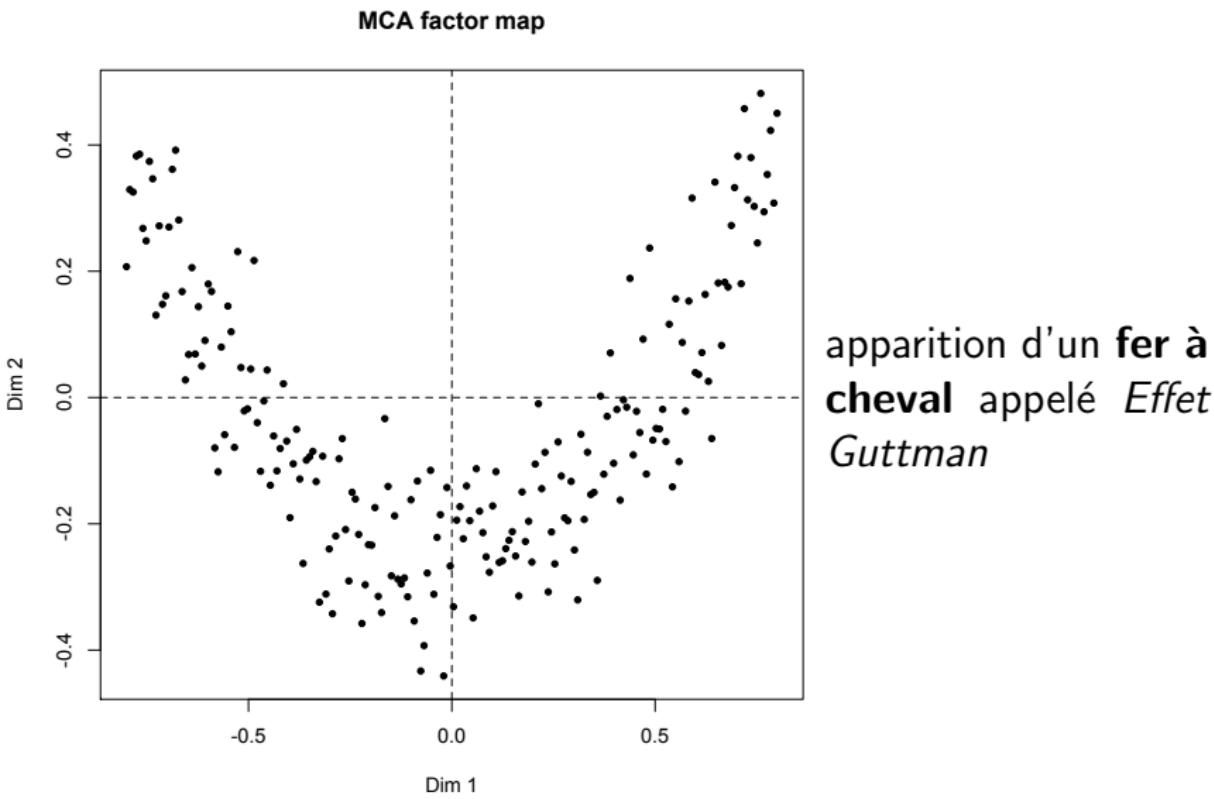
Activités de loisir des français : nuage des individus



Nuage des individus particulier ?

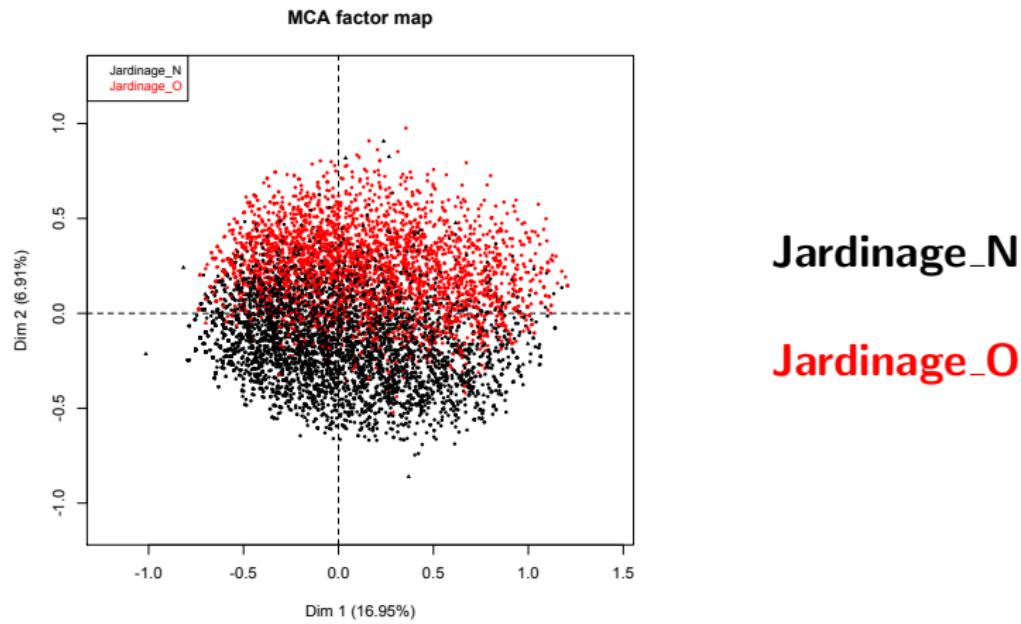


Nuage des individus particulier ?



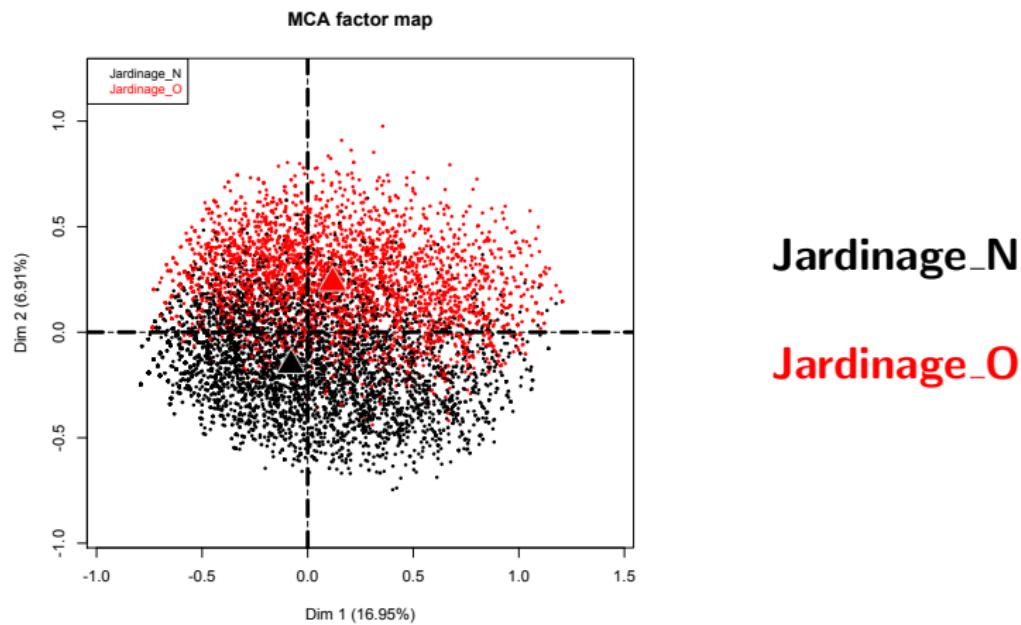
Interpréter les facteurs principaux de variabilités

Objectif : donner un sens au nuage des individus obtenus par l'AFCM
→ **colorier les individus en fonction des modalités qu'ils prennent sur une variable** (ici la variable jardinage) :



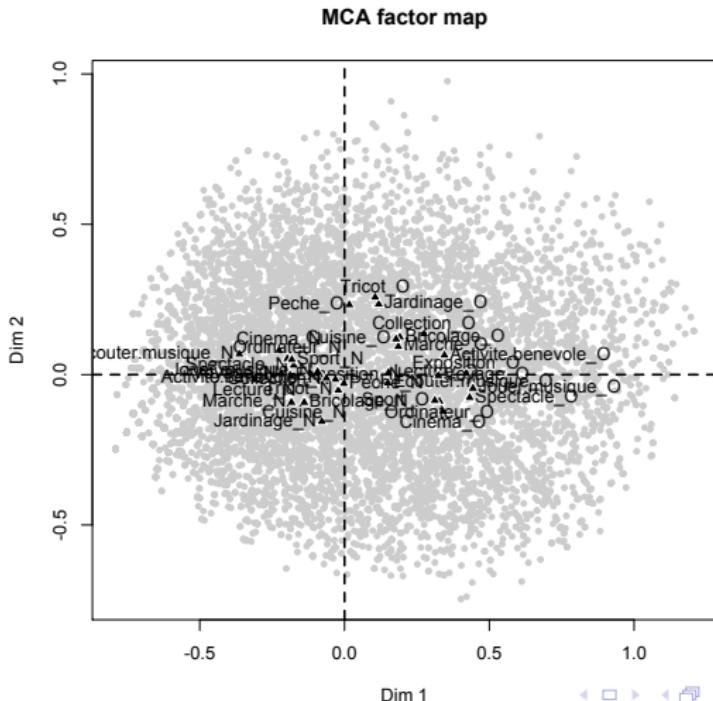
Interpréter les facteurs principaux de variabilités

→ Représenter les modalités de la variable **Jardinage** au barycentre des individus qui les prennent :



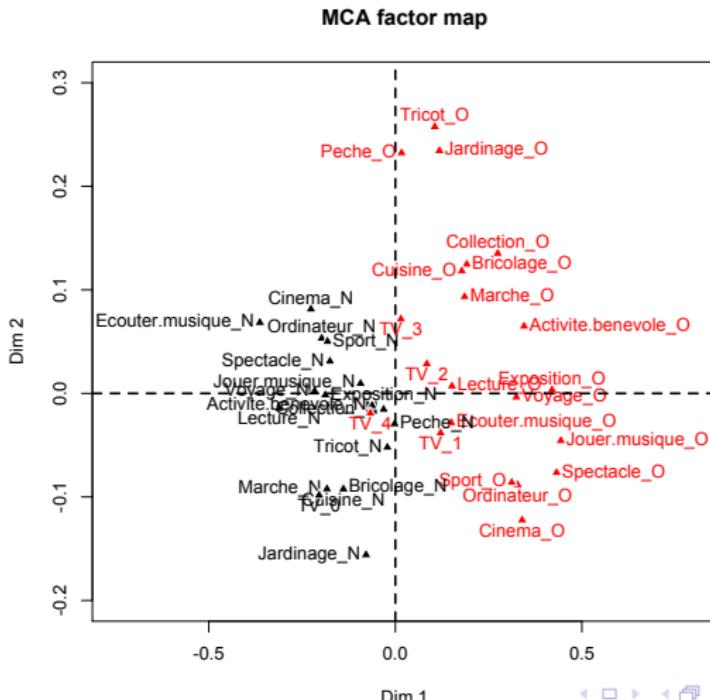
Interpréter les facteurs principaux de variabilités

→ Représenter **toutes les modalités** au barycentre des individus qui les prennent :



Interpréter les facteurs principaux de variabilités

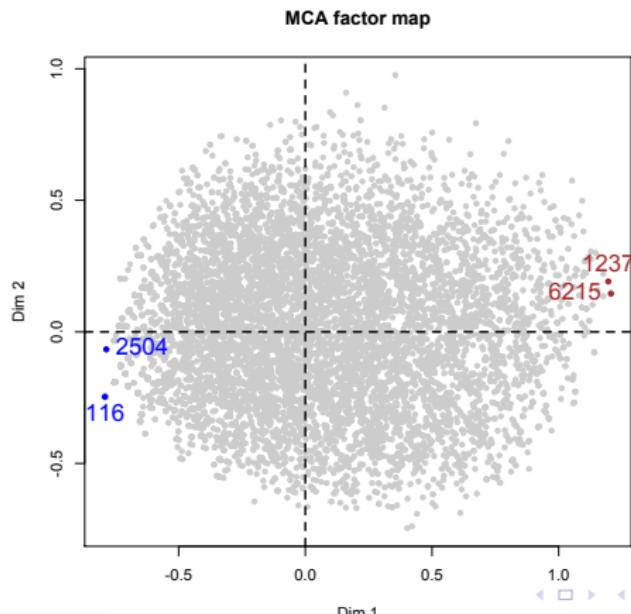
→ Effacement des individus + zoom sur les modalités colorées ("activités OUI" en rouge) :



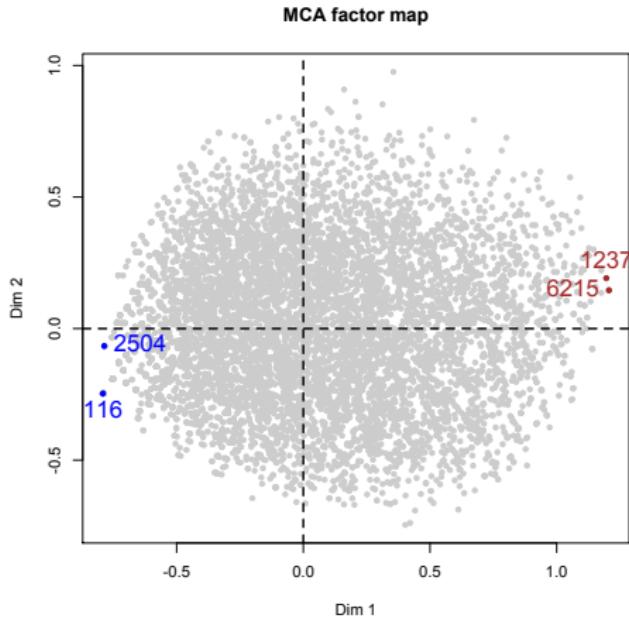
Interprétation de l'axe 1

1er axe oppose les individus qui pratiquent beaucoup d'activités à ceux qui en pratiquent peu

Confirmation en prenant des individus extrêmes sur l'axe 1 :



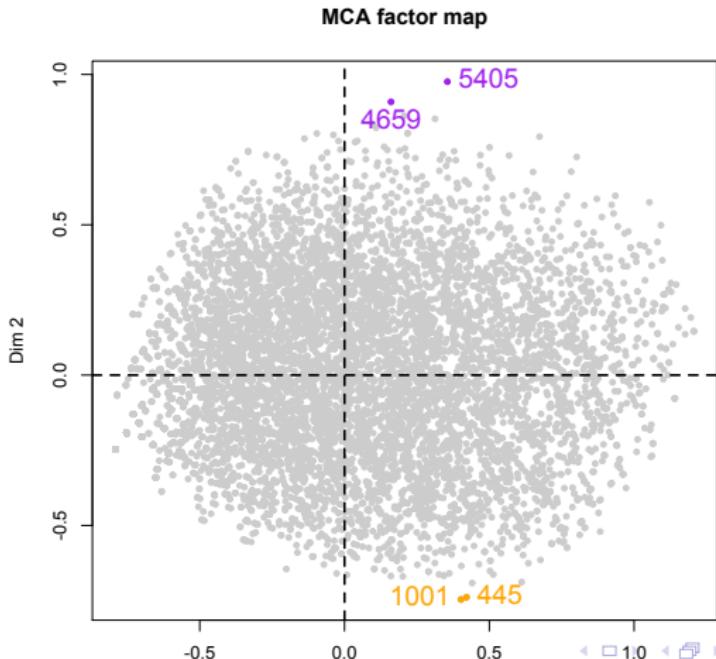
Interprétation de l'axe 1 : retour aux données



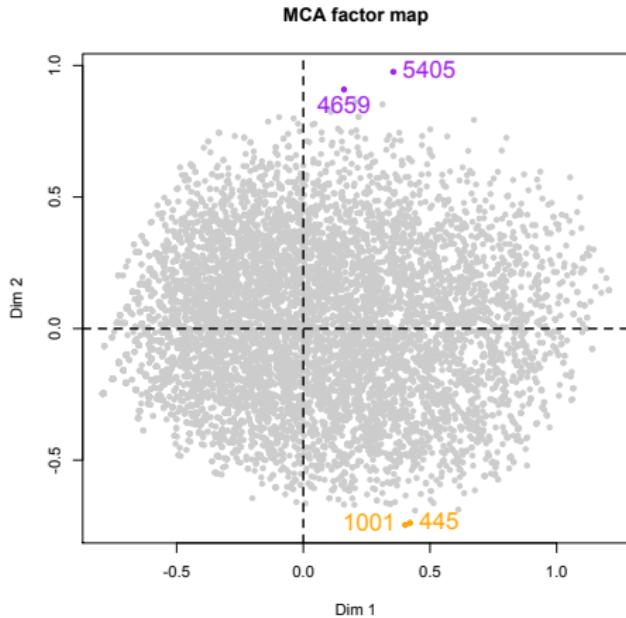
Lect.	Ecou.mus	Ciné	Spect	Expo	Ordi	Sport	Marche	Voyage	Jouer.mus	Collec	Benevole	Brico	Jardin	Tricot	Cuisine	Peche	TV
116	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	0
2504	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	0	0
1237	O	O	O	O	O	O	O	O	O	O	O	O	O	O	N	O	2
6215	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	N	1

Interprétation de l'axe 2

2ème axe oppose les individus pratiquant des activités "tranquilles" à ceux qui pratiquent des activités "jeunes"



Interprétation de l'axe 2 : individus extrêmes



	Lect.	Ecou.mus	Ciné	Spect	Expo	Ordi	Sport	Marche	Voyage	Jouer.mus	Collec	Benevole	Brico	Jardin	Tricot	Cuisine	Pêche	TV
4659	O	O	N	N	N	O	N	O	N	N	O	N	O	O	O	O	O	3
5405	O	O	N	N	O	N	N	O	O	N	O	N	O	O	O	O	O	2
445	O	O	O	O	O	O	O	N	N	O	N	N	N	N	N	N	N	0
1001	O	O	O	O	N	O	O	N	O	O	N	N	N	N	N	N	N	0

Interprétation des axes factoriels 1 et 2

- **axe 1** : une dimension d'"intensité" (1er axe factoriel) de l'activité de loisir qui oppose les individus pratiquant beaucoup d'activités à ceux qui en déclarent peu
- **axe 2** une dimension "type d'activité" (2ème axe factoriel) qui sépare les individus ayant des activités plutôt tranquilles à ceux qui pratiquent des activités plutôt "jeunes"

Principe de représentation des variables

On considère les coordonnées des projetés des individus sur un axe puis on calcule un indicateur de liaison (carré du rapport de corrélation) entre ces coordonnées et chaque variable catégorielle

Rapport de corrélation

Le rapport de corrélation mesure le lien entre une variable quantitative F et une variable catégorielle Q et est noté $\eta(F, Q)$. Cet indicateur vaut :

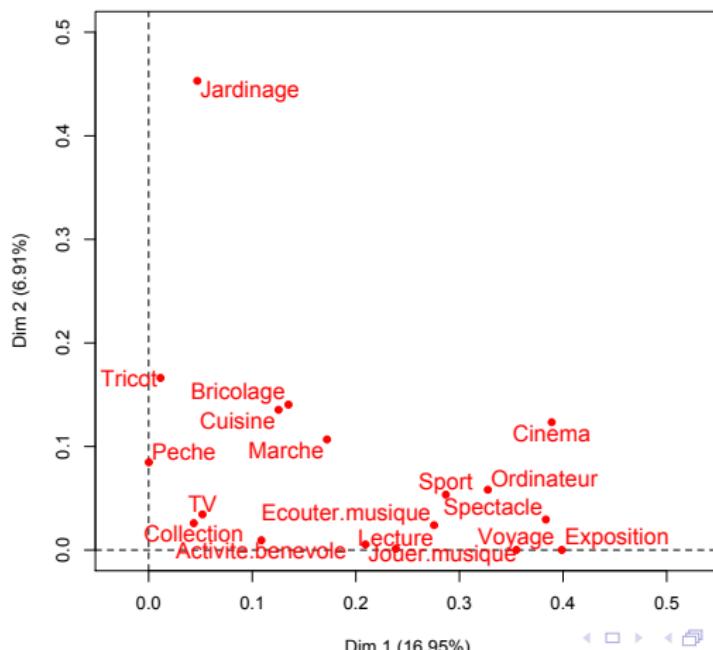
- 0 si les classes d'individus définies par la variable catégorielle Q ont toute la même moyenne relativement à la variable quantitative F (liaison nulle)
- 1 si les individus d'une même classe prennent exactement la même valeur pour la variable quantitative F , cette valeur étant différente d'une classe à l'autre (liaison maximale)

Représentation des variables

F_1 = coordonnées des individus sur l'axe 1

F_2 = coordonnées des individus sur l'axe 2

Variable catégorielle Q de coordonnées ($\eta^2(F_1, Q)$, $\eta^2(F_2, Q)$) :



Interprétation des axes factoriels

L'axe factoriel s est orthogonal à tout axe t ($t < s$) et est le plus lié aux variables catégorielles au sens du carré du rapport de corrélation :

$$F_s = \max_F \sum_{j=1}^p \eta^2(F, V_j)$$

où la variable F_s contient les coordonnées associées à l'axe factoriel s et V_1, V_2, \dots, V_p sont les p variables catégorielles

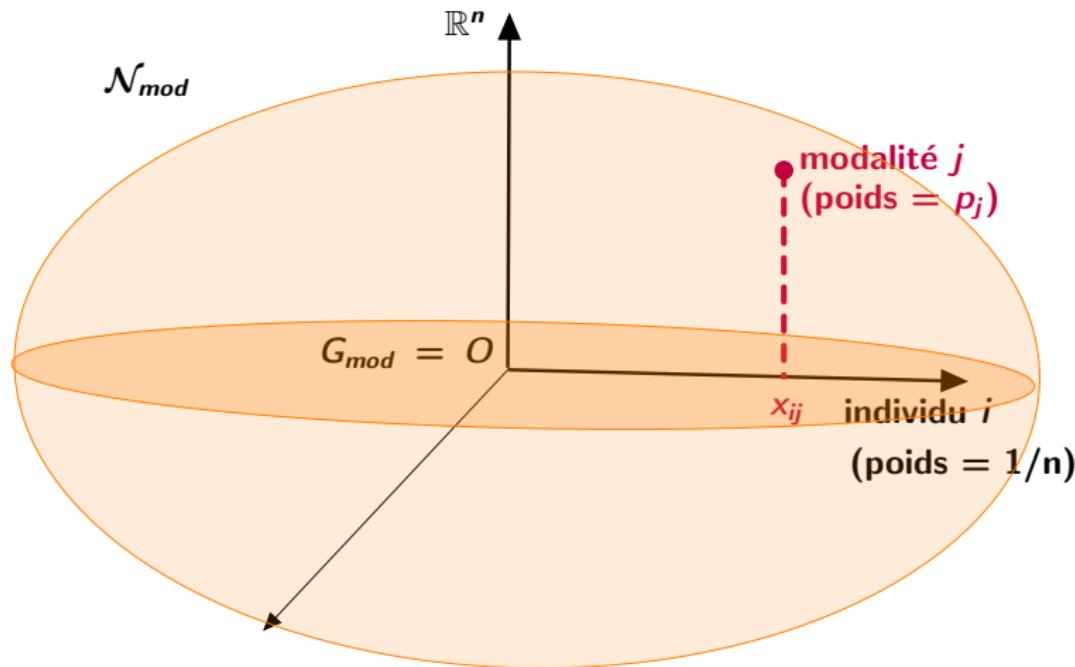
Nuage des modalités

On appelle **nuage des modalités**, noté \mathcal{N}_{mod} , l'ensemble des points-colonnes du TDC **standardisé**

Chaque point-colonne (i.e. modalité) possède n coordonnées :

		modalités des p variables					
		1	\dots	j	\dots	m	Σ
individus	1	x_{11}	\dots	x_{1j}	\dots	x_{1m}	$x_{1\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	i	x_{i1}	\dots	x_{ij}	\dots	x_{im}	$x_{i\bullet}$
	\vdots	\vdots		\vdots		\vdots	\vdots
	n	x_{n1}	\dots	x_{nj}	\dots	x_{nm}	$x_{n\bullet}$
Σ		0	\dots	0	\dots	0	

Géométrie du nuage des modalités



Rappel : $x_{ij} = v_{ij}/f_{ij} - 1$ où f_{ij} = fréquence de la modalité i

Distance d'une modalité à l'origine

La distance au carré $d^2(j, G_{mod})$ entre une modalité j et l'origine G_{mod} est inversement proportionnelle à la fréq. de la modalité j :

$$d^2(j, G_{mod}) \stackrel{\text{déf}}{=} \sum_{i=1}^n \frac{1}{n} x_{ij}^2 = \frac{1}{f_{\bullet j}} - 1$$

Plus une modalité est rare, plus elle s'éloigne du centre du nuage

Exemple

$f_{\bullet j}$	1/2	1/5	1/10	1/101
$d(j, G_{mod})$	1	2	3	10

Inertie d'une modalité

Inertie de la modalité j

$$\text{Inertie}(j) \stackrel{\text{déf}}{=} \frac{f_{\bullet j}}{p} d^2(j, G_{mod})$$

Plus une modalité est rare, plus son inertie est grande

Montrer que $\text{Inertie}(j) = \frac{1 - f_{\bullet j}}{p}$

Exemple

$f_{\bullet j}$	1/2	1/5	1/10	1/101
$d(j, G_{mod})$	1	2	3	10
$(p = 10) \text{ Inertie}(j)$	0.05	0.08	0.09	0.099

Inertie d'une variable et inertie totale

V_k variable catégorielle possédant m_k modalités

Inertie de la variable V_k

$$\text{Inertie}(V_k) = \frac{m_k - 1}{p}$$

L'inertie d'une variable est proportionnelle au nombre de ses modalités (moins 1)

Inertie totale

$$\text{Inertie totale} = \frac{m}{p} - 1$$

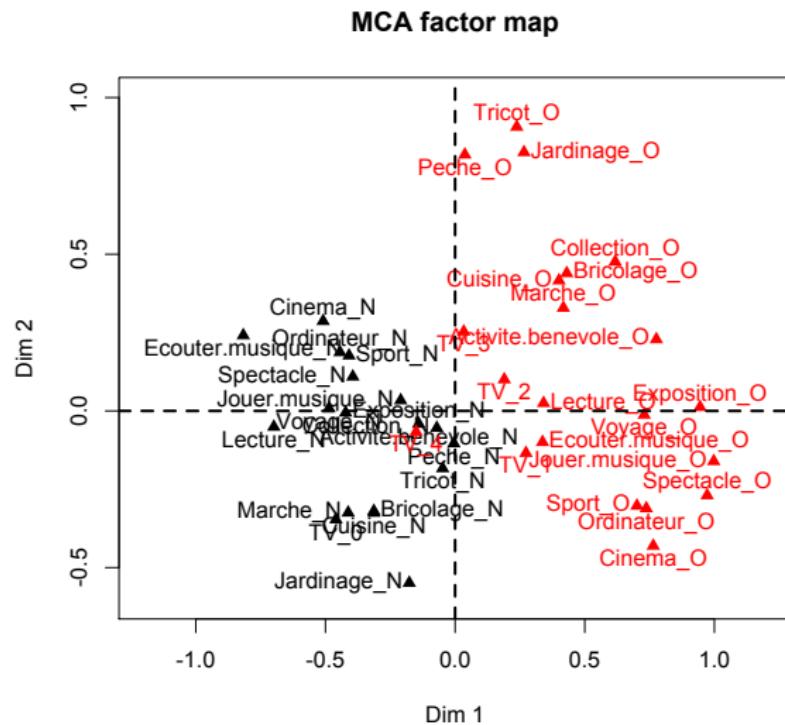
Rappel : $m = \text{nb total de modalités}$ et $p = \text{nb de variables}$

Ajustement du nuage des modalités

Comme toute analyse factorielle, on **projette** ce nuage des modalités sur une **suite d'axes orthogonaux** d'inertie maximum pour visualiser en 2D (voire 3D)

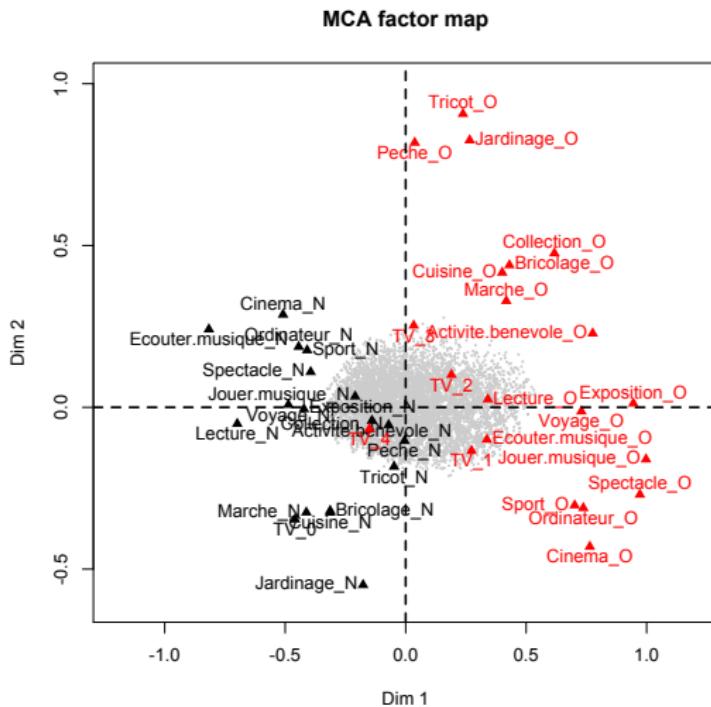
- **1er axe factoriel** = direction qui explique la plus grande part de variabilité (i.e. inertie)
- **2ème axe factoriel** = direction orthogonale au 1er axe factoriel qui représente la plus grande part de variabilité restante
- et ainsi de suite

Représentation du nuage des modalités



Représentation simultanée individus/modalités

Représentation des individus au barycentre des modalités possédées



Représ. optim. des ind

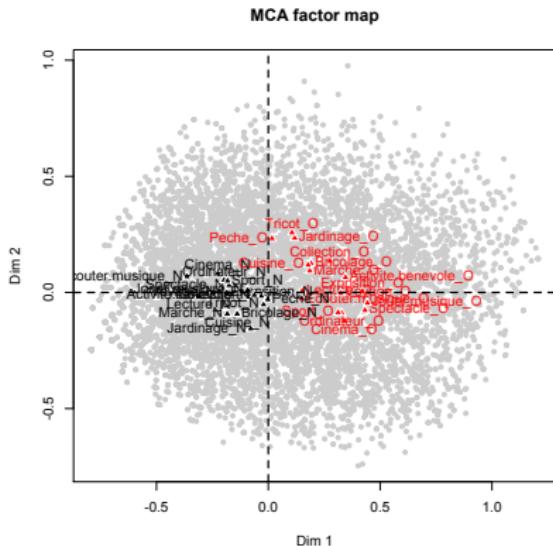
+

représ. **barycentrique** des mod.

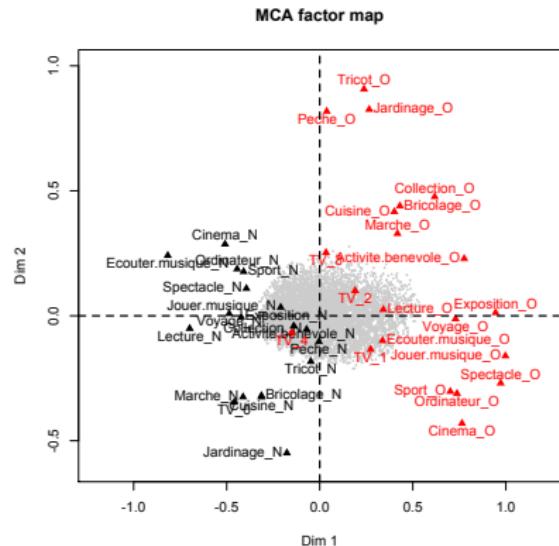
Représ. optim. des modalités

+

représ. **barycentrique** des ind.



$$Mod_s^{bary}(j) = \sum_{i=1}^n \frac{y_{ij}}{y_{\bullet j}} Ind_s(i)$$



$$Ind_s^{bary}(i) = \sum_{j=1}^m \frac{y_{ij}}{p} Mod_s(j)$$

Inertie revisitée

L'inertie associée à l'axe factoriel s est la variabilité totale des modalités exprimée par l'axe factoriel s :

$$\lambda_s = \sum_{j=1}^m p_j Mod_s(j)^2$$

où $p_j = y_{\bullet j}/p$ = poids de la modalité j

L'inertie associée à l'axe factoriel s est la variabilité totale des individus exprimée par l'axe factoriel s :

$$\lambda_s = \frac{1}{n} \sum_{i=1}^n Ind_s(i)^2$$

Conséquence : $\sum_{j=1}^m p_j Mod_s(j)^2 = \frac{1}{n} \sum_{i=1}^n Ind_s(i)^2$

Représentation pseudo-barycentrique

des modalités dans le graphe des individus

$$\widetilde{Mod}_s^{bary}(j) \stackrel{\text{déf}}{=} \frac{1}{\sqrt{\lambda_s}} Mod_s^{bary}(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^n \frac{y_{ij}}{y_{\bullet j}} Ind_s(i)$$

des individus dans le graphe des modalités

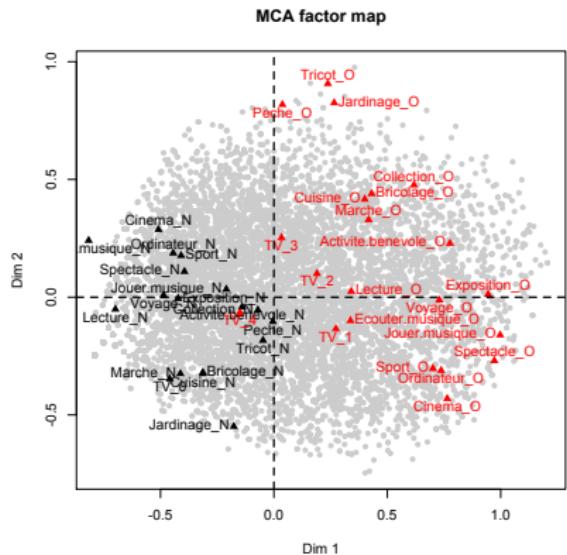
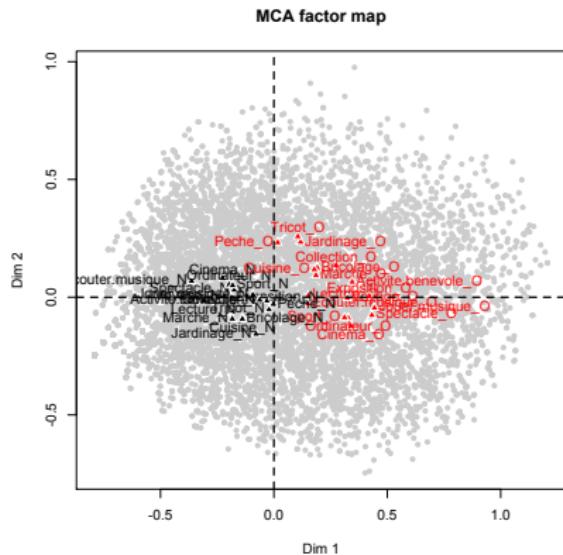
$$\widetilde{Ind}_s^{bary}(i) \stackrel{\text{déf}}{=} \frac{1}{\sqrt{\lambda_s}} Ind_s^{bary}(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^m \frac{y_{ij}}{p} Mod_s(j)$$

Remarque : comme λ_s est compris entre 0 et 1, multiplier par $1/\sqrt{\lambda_s}$ revient à opérer une **dilatation**

Représ. pseudo-barycentrique des modalités

$$Mod_s^{bary}(j) = \sum_{i=1}^n \frac{y_{ij}}{y_{\bullet j}} Ind_s(i)$$

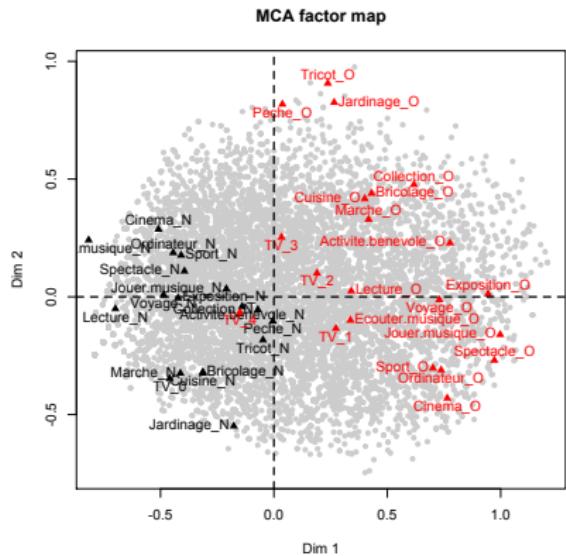
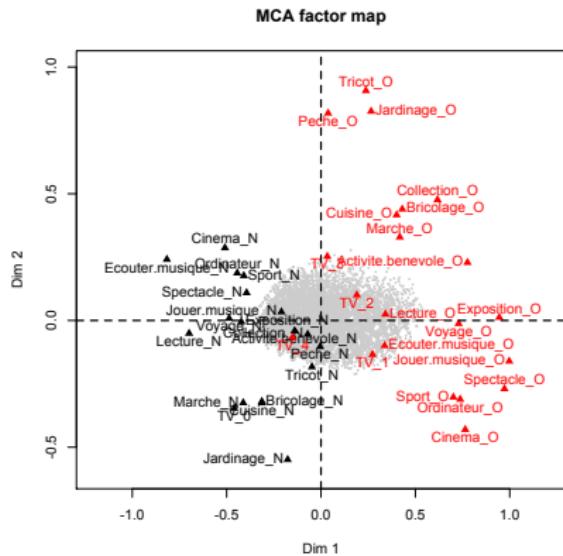
$$\widetilde{Mod}_s^{bary}(j) = \frac{\mathbf{1}}{\sqrt{\lambda_s}} \sum_{i=1}^n \frac{y_{ij}}{y_{\bullet j}} Ind_s(i)$$



Représ. pseudo-barycentrique des individus

$$Ind_s^{bary}(i) = \sum_{j=1}^m \frac{y_{ij}}{p} Mod_s(j)$$

$$\widetilde{Ind}_s^{bary}(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^m \frac{y_{ij}}{p} Mod_s(j)$$



Interprétation des coord. pseudo-baycentriques

- une modalité sera positionnée du côté des individus qui la possèdent et à l'opposé des individus qui ne la prennent pas
- un individu sera placé du côté des modalités qu'il possède et à l'opposé des modalités qu'il ne prend pas

Inertie en AFCM

F_s = coord. des individus sur l'axe s

V_1, \dots, V_p les p variables catégorielles

Inertie associée à l'axe s : $\lambda_s = \frac{1}{p} \sum_{j=1}^p \eta^2(F_s, V_j)$

Part d'inertie maxi. ass. à un axe : $\frac{\lambda_s}{\lambda_1 + \lambda_2 + \dots} \leq \frac{p}{m-p}$

Inertie moyenne ass. à un axe : $\frac{1}{m-p} \sum_{s=1}^{m-p} \lambda_s = \frac{1}{p}$

En pratique

Exercice 1. Montrer que $\frac{\lambda_s}{\lambda_1 + \lambda_2 + \dots} \leq \frac{p}{m - p}$

Exercice 2. Si on dispose de 10 variables possédant chacune 10 modalités, montrer que $\lambda_s \leq 11.1\%$ de l'inertie totale

En AFCM, le % d'inertie ass. à un axe est souvent faible

Remarque : si on considère les loisirs des français, on a $p = 18$ variables catégorielles ; l'inertie moyenne ass. à chaque axe = 0.056

→ **En AFCM, on interprétera uniquement les axes qui expriment plus d'inertie que la moyenne**

Contribution

Contrib. abs. d'une modalité j à l'axe s :

$$\text{contrib}(j) = p_j \text{Mod}_s(j)^2$$

Contrib. abs d'une variable V_k à l'axe s :

$$\text{contrib}(V_k) = \frac{\eta^2(F_s, V_k)}{p}$$

Contrib. relative d'une modalité j à l'axe s :

$$\text{contrib}(j) = \frac{p_j \text{Mod}_s(j)^2}{\lambda_s}$$

Contrib. relative d'une variable V_k à l'axe s :

$$\text{contrib}(V_k) = \frac{\eta^2(F_s, V_k)}{p \lambda_s}$$

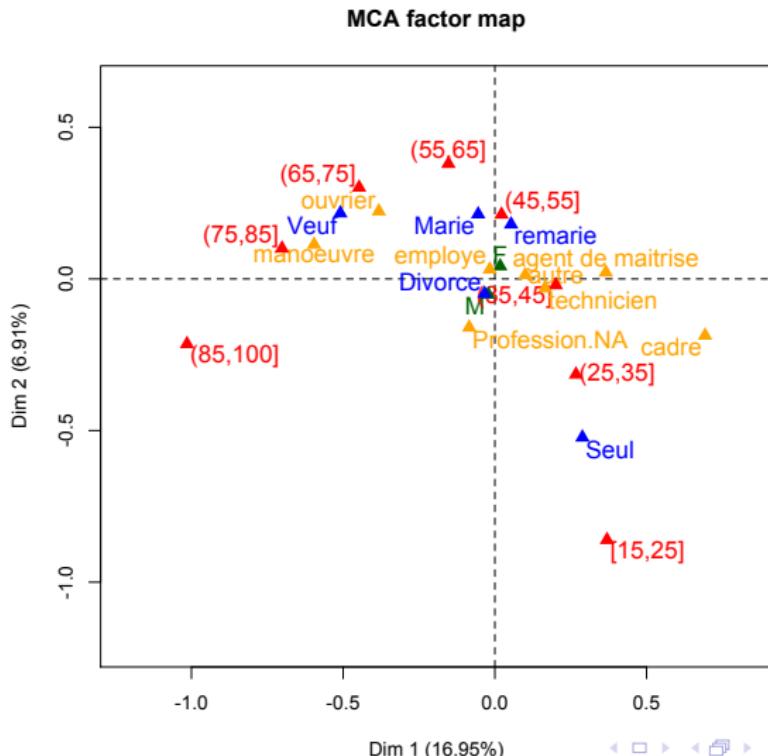
Variables catégorielles supplémentaires

- **variables actives** = variables utilisées pour construire les axes factoriels
- **variables supplémentaires** = variables fournissant des éléments sur le contexte de l'analyse, permettant d'affiner l'interprétation des résultats de l'AFCM

Les variables supplémentaires sont simplement rajoutées sur les graphes visualisant les modalités des variables actives en se basant sur le principe de la représentation pseudo-barycentrique

Exemple : loisirs des français

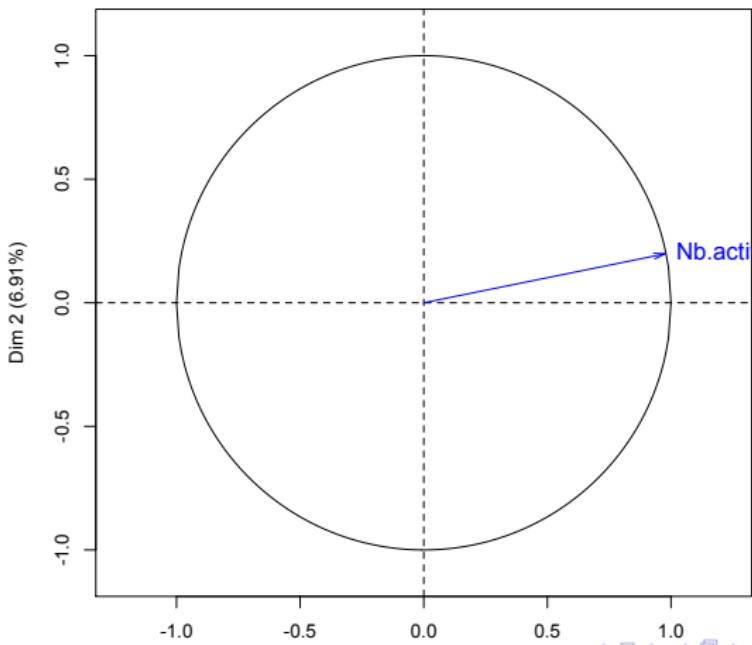
Représentation des 4 var. signalétiques supplémentaires



Variables quantitatives supplémentaires

On représente une variable quantitative par ses coefficients de corrélation avec les facteurs

Supplementary variables on the MCA factor map



Variables quantitatives supplémentaires

Comment rendre active une variable quantitative ?

il suffit de procéder à leur découpage en classes afin de les transformer en une variable catégorielle