

# Analyse Multivariée des données

Dr Mory Ouattara  
Data Science

INPHB

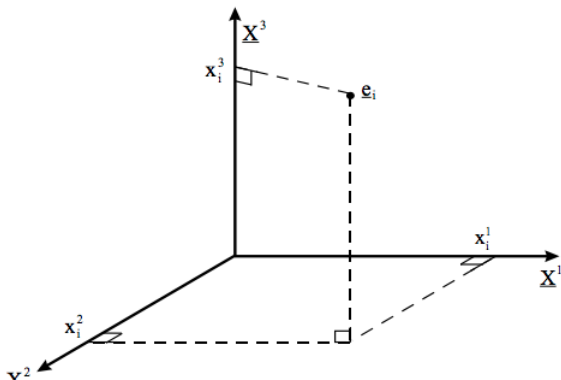
2018

- 1 Données Multivariées
- 2 Analyse en composantes principales (ACP)
  - Lien entre  $\eta_j$  et  $\xi_i$
  - Disque de corrélation
  - Individus et variables supplémentaires
- 3 L'Analyse des Correspondances Simples (AFC)
  - Test Khi2
  - L'analyse des correspondances simples
  - ACP du nuage des profils lignes-profils colonnes
  - Cas pratique
- 4 Analyse des correspondances multiples (ACM)
  - Présentation Formelle
  - Aspects pratiques

# Données Multivariées

A chaque individu noté  $e_i$ , on peut associer un point dans  $\mathbb{R}^p$  = espace des individus.

A chaque variable du tableau  $X$  est associé un axe de  $\mathbb{R}^p$ .



**Impossible à  
visualiser dès  
que  $p > 3$ .**

# Données Multivariées

Soit  $x \in \mathbb{R}^p$  un vecteur aléatoire :

$$x = (\xi_1, \xi_2, \dots, \xi_p)^T$$

où  $v^T$  désigne la transposée du vecteur  $v$  avec  $\xi_j$  est une variable aléatoire

Un **échantillon multidimensionnel** est une suite  $x_1, \dots, x_n$  de réalisations aléatoires du vecteur  $x$ .

$x_{ij}$  désignera la  $j$  ème composante du vecteur  $x_i$

$$\mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$$

# Statistiques

## 1 Les moyennes empiriques

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ki} \quad k = 1, \dots, p$$

qui forment le vecteur

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{pmatrix} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \mathbf{X}^T \mathbf{1} \quad \text{avec} \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n.$$

## 2 Les covariances empiriques

$$s_{jk} = \frac{1}{n} \sum_i x_{ij} x_{ik} - \bar{x}_j \bar{x}_k \quad k, j = 1, \dots, p$$

qui forment la matrice covariance empirique  $S = (S_{kj})$

# Statistiques

## 3 Les corrélations empiriques

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} \quad k, j = 1, \dots, p$$

qui forment la matrice de corrélation empirique

$$R = (r_{jk})_{k,j=1,\dots,p}$$

# Statistiques

Il est facile de voir que

$$S = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T - \bar{\mathbf{x}} \bar{\mathbf{x}}^T = \frac{1}{n} \mathbf{X} \mathbf{X}^T - \frac{1}{n^2} \mathbf{X} \mathbf{1} \mathbf{1}^T \mathbf{X}^T = \frac{1}{n} \mathbf{X}^T \mathbf{H} \mathbf{X}$$

où

$$\mathbf{H} = \mathbf{I}_n - \frac{1}{n} \mathbf{1} \mathbf{1}^T$$

est la matrice de centrage

- ❶ *Montrer que  $\mathbf{H}$  est un projecteur, i. e.  $\mathbf{H} = \mathbf{H}^2$  et  $\mathbf{H}^T = \mathbf{H}$ . Sur quel sous-espace vectoriel de  $\mathbb{R}^n$  projette-t-il ?*
- ❷ *Montrer la matrice de covariance empirique  $\mathbf{S}$  est positive, en effet pour tout vecteur  $\mathbf{R}^p$ .*

# L'Analyse en Composantes Principales (ACP)

L'Analyse en composantes principales (ACP) est une méthode de traitement des données multidimensionnelles qui poursuit les deux objectifs suivants :

- 1 Visualiser les données (**Notion de distances entre individus**)
- 2 Réduire la dimension effective des données (**en fonction de leurs corrélations**).

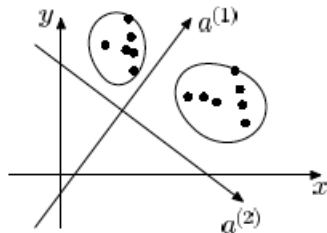


# L'idée de l'Analyse en composantes principales (ACP)

si  $a = (a_1, \dots, a_p)^T \in R^p$  est une direction de projection.

- Les données projetées  $(a^T x_1, \dots, a^T x_n)$  forment un échantillon de dimension 1.
- que l'on peut visualiser et qui est donc plus facile à interpréter que l'échantillon de départ  $(x_1, \dots, x_n)$ .

# L'idée de l'Analyse en composantes principales (ACP)

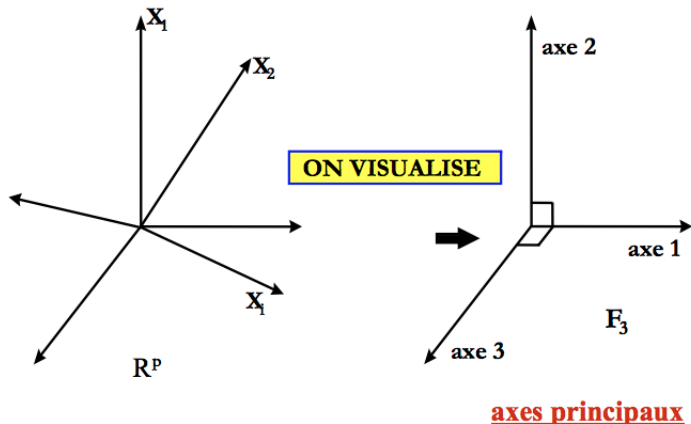


Bonne et mauvaise directions de projection.

L'ACP a pour objectif de trouver un sous-espace linéaire de  $\mathbb{R}^p$  de dimension  $p^* \ll p$  tel que la projection sur ce sous-espace "capte" presque toute la structure des données.

# L'idée de l'Analyse en composantes principales (ACP)

L'Analyse en composantes principales (ACP) est une méthode de traitement des données multidimensionnelles qui poursuit les deux objectifs suivants :



# L'idée de l'Analyse en composantes principales (ACP)

L'Analyse en composantes principales (ACP) est une méthode de traitement des données multidimensionnelles qui poursuit les deux objectifs suivants :

①

$F_k$  devra être « ajusté » le mieux possible au nuage des individus: la somme des carrés des distances des individus à  $F_k$  doit être minimale.



②

$F_k$  est le sous-espace tel que le nuage projeté ait une **inertie** (dispersion) maximale.

① et ② sont basées sur les notions de :

**distance**

# Critère de l'ACP

L'idée de base de l'ACP est de chercher la direction  $\mathbf{a} \in \mathbb{R}^p$  qui maximise en  $\mathbf{a}$  la variance empirique de l'échantillon unidimensionnel  $(\mathbf{a}^T \mathbf{x}_1, \dots, \mathbf{a}^T \mathbf{x}_n)$

$$\begin{aligned} s_a^2 &\stackrel{\text{déf}}{=} \frac{1}{n} \sum_{i=1}^n (\mathbf{a}^T \mathbf{x}_i)^2 - \left( \frac{1}{n} \sum_{i=1}^n (\mathbf{a}^T \mathbf{x}_i) \right)^2 \\ &= \frac{1}{n} \mathbf{a}^T \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{a} - \frac{1}{n^2} \mathbf{a}^T \left( \sum_{i=1}^n \mathbf{x}_i \sum_{i=1}^n \mathbf{x}_i^T \right) \mathbf{a} = \mathbf{a}^T \mathbf{S} \mathbf{a}, \end{aligned}$$

la direction la plus intéressante  $\hat{\mathbf{a}}$  est une solution de

$$\max_{\mathbf{a} \in \mathbb{R}^p: \|\mathbf{a}\|=1} \mathbf{a}^T \mathbf{S} \mathbf{a} = \hat{\mathbf{a}}^T \mathbf{S} \hat{\mathbf{a}}$$

ou

$$\hat{\mathbf{a}} = \arg \max_{\mathbf{a} \in \mathbb{R}^p: \|\mathbf{a}\|=1} \mathbf{a}^T \mathbf{S} \mathbf{a}$$

# Décomposition spectrale

Nous nous intéressons à la solution du problème suivant :

$$a^* = \arg \max_{a \in \mathbb{R}^p: \|a\|=1} \text{Var}(a^T X) \text{ avec } E(\|X\|^2) < \infty$$

Soit

$$\Sigma = \Gamma \Lambda \Gamma^T$$

une décomposition spectrale de covariance

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}, \quad \Gamma = (\gamma_{(1)}, \dots, \gamma_{(p)})$$

Où les  $\lambda_i$  sont les valeurs propres de  $\Sigma$  rangées dans l'ordre croissant et

$$\|\gamma_{(i)}\| = 1 \text{ avec } \gamma_{(i)}^T \gamma_{(k)} = 0 \text{ pour } i \neq k$$

# Les Composantes principales $\eta_{(j)}$

## Composante principale $\eta_{(j)}$

La variable aléatoire  $\eta_{(j)} = \gamma_{(j)}^T(\mathbf{x} - \mu)$  est dite  $j$  ème composante principale du vecteur aléatoire  $\mathbf{x} \in R^p$ .

Les  $\gamma_{(j)}$  sont les vecteurs propres de la matrice de covariance  $\Sigma$  du vecteur aléatoire  $\mathbf{x}$ , on obtient :

$$\text{Var}[\eta_j] = E[\gamma_{(j)}^T(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \gamma_{(j)}] = \gamma_{(j)}^T \Sigma \gamma_{(j)} = \gamma_{(j)}^T \lambda_j \gamma_{(j)} = \lambda_j,$$

$$\text{Cov}(\eta_j, \eta_k) = E[\gamma_{(j)}^T(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \gamma_{(k)}] = \gamma_{(j)}^T \Sigma \gamma_{(k)} = \gamma_{(j)}^T \lambda_k \gamma_{(k)} = 0,$$

## ACP

Soit  $x$  un vecteur aléatoire de  $\mathbb{R}^2$  de moyenne nulle et de matrice de covariance

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad 0 \leq \rho \leq 1.$$



# Exemple ACP

EXEMPLE 7.1. Soit  $x$  un vecteur aléatoire de  $\mathbb{R}^2$  de moyenne nulle et de matrice de covarian

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad 0 \leq \rho \leq 1.$$

Considérons les vecteurs propres orthonormés de cette matrice

$$\gamma_{(1)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \gamma_{(2)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Donc si les coordonnées de  $x$  sont  $\xi_1$  et  $\xi_2$ , les composantes principales de  $x$  valent

$$\eta_1 = \frac{\xi_1 + \xi_2}{\sqrt{2}}, \quad \eta_2 = \frac{\xi_1 - \xi_2}{\sqrt{2}}.$$

# Exemple ACP

## Théorème

Soit  $x \in \mathbb{R}^p$  un vecteur aléatoire tel que  $E(\|x\|) < \infty$ . Alors  $\hat{a} = \gamma_{(1)}$  est vérifié :

$$\text{Var}(\hat{a}^T X) = \max_{a \in \mathbb{R}^p: \|a\|=1} (a^T X) = \max_{a \in \mathbb{R}^p: \|a\|=1} (a^T (X - \mu))$$

## ACP

*Preuve.* La décomposition spectrale de la matrice  $\Sigma$  est de la forme

$$\Sigma = \Gamma \Lambda \Gamma^T = \sum_{j=1}^p \lambda_j \gamma_{(j)} \gamma_{(j)}^T.$$

On a donc

$$\text{Var}[a^T \mathbf{x}] = \sum_{j=1}^p \lambda_j (a^T \gamma_{(j)}) (\gamma_{(j)}^T a) = \sum_{j=1}^p \lambda_j c_j^2,$$

où  $c_j = a^T \gamma_{(j)}$  est la projection du vecteur  $a$  sur la direction  $\gamma_{(j)}$ . Puisque les vecteurs  $\gamma_{(j)}$  forment une base orthonormée de  $\mathbb{R}^p$ , on a  $c_1^2 + \dots + c_p^2 = \|a\|^2$ . Comme  $\lambda_j \leq \lambda_1$ , on en déduit que

$$\text{Var}[a^T \mathbf{x}] = \sum_{j=1}^p \lambda_j c_j^2 \leq \lambda_1 \sum_{j=1}^p c_j^2 = \lambda_1 \|a\|^2 = \lambda_1.$$

Par ailleurs, si  $a = \hat{a} = \gamma_{(1)}$ , les coefficients  $c_j$  sont tous nuls sauf le premier  $c_1 = 1$ . On a donc  $\text{Var}[\hat{a}^T \mathbf{x}] = \lambda_1$ . Par conséquent,  $\hat{a}$  est une solution du problème de maximisation (7.2) et  $\text{Var}[\hat{a}^T \mathbf{x}] = \lambda_1 = \text{Var}[\eta_1]$ . ■

# Étude des corrélations

On définit la variance totale de  $x$  par

$$E(\|X - \mu\|^2) = E(X - \mu)^T (X - \mu) = E(X - \mu)^T \Gamma \Gamma^T (X - \mu)$$

avec

$$\Gamma^T (\mathbf{x} - \mu) = \begin{pmatrix} \gamma_{(1)}^T (\mathbf{x} - \mu) \\ \vdots \\ \gamma_{(p)}^T (\mathbf{x} - \mu) \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_p \end{pmatrix} \stackrel{\text{déf}}{=} \mathbf{y}.$$

Compte tenu de ces notations et de l'égalité  $E(\eta_i^2) = \lambda_i$ , on obtient

$$E(\|X - \mu\|^2) = E(\eta_1^2 + \dots + \eta_p^2)$$

# ACP

Donc : Comment détermine t-on le meilleur sous espace de projection ?

# Part de variance expliquée

## Part de variance expliquée

On appelle part de la variance totale de  $x$  expliquée par les  $k$  premières composantes principales  $(\eta_1, \dots, \eta_k)$  la quantité

$$\frac{\sum_{j=1}^K \lambda_j}{\sum_{j=1}^p \lambda_j}$$

# Choix du bon nombre d'axe

## Combien d'axes ?

Différentes procédures sont complémentaires:

① **Pourcentage d'inertie souhaité : a priori**

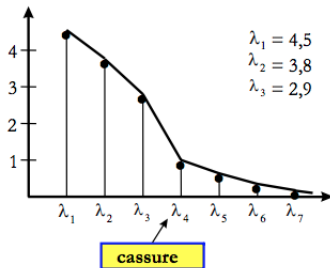
② Diviser l'inertie totale par le nombre de variables initiales

⇒ inertie moyenne par variable : I.M.

**Conserver tous les axes apportant une inertie supérieure à cette valeur (inertie > 1 si variables centrées réduites).**

### ③ Histogramme

Conserver les axes associés aux valeurs propres situées avant la cassure.



Lien entre  $\eta_j$  et  $\xi_i$ 

Calculons d'abord la matrice de covariance des vecteurs aléatoires  $\mathbf{x}$  et  $\mathbf{y}$ .

$$C(\mathbf{x}, \mathbf{y}) = E[(\mathbf{x} - \mu)\mathbf{y}^T] = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \Gamma] = \Sigma \Gamma = \Gamma \Lambda \Gamma^T \Gamma = \Gamma \Lambda.$$

Comme  $\text{Cov}(\xi_i, \eta_j)$  est le  $(i, j)$ <sup>ème</sup> élément de cette matrice, on obtient

$$\text{Cov}(\xi_i, \eta_j) = \gamma_{ij} \lambda_j.$$

La corrélation  $\tilde{\rho}_{ij} = \text{Corr}(\xi_i, \eta_j)$  entre  $\xi_i$  et  $\eta_j$  vaut

$$\tilde{\rho}_{ij} = \frac{\text{Cov}(\xi_i, \eta_j)}{\sqrt{\text{Var}(\xi_i) \text{Var}(\eta_j)}} = \gamma_{ij} \sqrt{\frac{\lambda_j}{\sigma_{ii}}}.$$



## Lien entre $\eta_j$ et $\xi_i$

### Proposition

Soit  $x \in R^p$  un vecteur aléatoire, tel que  $E(\|x\|^2) < \infty$  et  $\sigma_{ii} > 0$  pour tout  $i = 1, \dots, p$ .

Alors,

$$\sum_{j=1}^p \tilde{\rho}_{ij}^2 = 1$$

On appelle  $\tilde{\rho}_{ij}^2$  part de variance de la variable  $\xi_i$  expliquée par la  $j$  ème composante principale  $\eta_j$ .

Pour tout sous-ensemble  $J$  de  $1, \dots, p$ ,

$$\sum_{j \in J} \lambda_j = \sum_{j=1}^p \sigma_{ii} \tilde{\rho}_{ij}^2 \text{ avec } \tilde{\rho}_{iJ}^2 = \sum_{j \in J} \rho_{ij}^2$$

# Lien entre $\eta_j$ et $\xi_i$

*Preuve.*

$$\sum_{i=1}^p \sigma_{ii} \tilde{\rho}_{iJ}^2 = \sum_{i=1}^p \sigma_{ii} \sum_{j \in J} \gamma_{ij}^2 \frac{\lambda_j}{\sigma_{ii}} = \sum_{j \in J} \lambda_j \sum_{i=1}^p \gamma_{ij}^2.$$

Le résultat de la proposition découle du fait que la dernière somme vaut 1, car  $\|\gamma_{(j)}\|^2 = \sum_{i=1}^p \gamma_{ij}^2 = 1$ . ■

# Disque des corrélations

## Proposition

Soient  $\xi_i$  et  $\xi_k$  deux variables entièrement expliquées par les deux premières composantes principales, i.e.

$$\tilde{\rho}_{i1}^2 + \tilde{\rho}_{i2}^2 = 1 \text{ et } \tilde{\rho}_{k1}^2 + \tilde{\rho}_{k2}^2 = 1$$

Alors, la corrélation de  $\xi_i$  et  $\xi_k$  est donnée par la formule

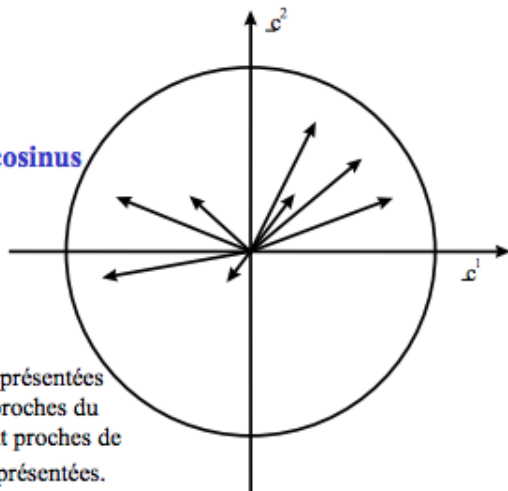
$$\rho_{ik} = \tilde{\rho}_{i1}\tilde{\rho}_{k1} + \tilde{\rho}_{i2}\tilde{\rho}_{k2} = \cos(\varphi),$$

où  $\varphi$  est l'angle formé par les vecteurs  $(\tilde{\rho}_{i1}, \tilde{\rho}_{i2})$  et  $(\tilde{\rho}_{k1}, \tilde{\rho}_{k2})$ .

# Disque des corrélations

Le cercle des corrélations est la projection du nuage des variables sur le plan des composantes principales.

**corrélation = cosinus**



Les variables bien représentées  
sont celles qui sont proches du  
cercle, celles qui sont proches de  
l'origine sont mal représentées.

# Variables

- On calcule le coefficient de corrélation entre la variable supplémentaire et les composantes principales.
- Ceci permet sa représentation sur le cercle des corrélations.

# Individus

- Individu de poids nul ne participant pas à l'analyse
- Appliquer aux coordonnées de l'individu les expressions définissant les composantes principales.

# Cas pratique : Campagne CNL 2005

VARIABLE QUANTITATIVES		VARIABLES QUALITATIVES	
HABITUDE DU MENAGE			
ICOS1	Utilisation de déodorant dans le ménage	QOM1	Vous sortez les ordures à l'extérieur
ICOS2	Utilisation d'eau de toilette dans le ménage	TMG6	Bricolage du mois
ICOS3	Utilisation des produits de soins dans les cheveux	TMG7	Jardinage du mois
ICOS4	Utilisation des produits de soins visage et corps	DGG3n	Nombre de Voitures dans le garage
ICOS5	Utilisation des vernis à ongle et dissolvants	FUMEUR	Dans le logement
QPD2b	Utilisation d'autre type de désodorisant	QPV	Nombre de Plantes à l'intérieur
CARACTERISTIQUES DU LOGEMENT			
HSRF	surface du logement en dehors des annexes	CHEM1	Equipé d'une cheminée ou pas
		DGG2b	Existence d'un garage
NIACe1	Ancienneté de l'immeuble ou du logement	FC3	Type de logement
		DCa3	Existence d'une cave
MENAGE			
Age	Personnes	DIPLOM2	Diplômes
REV3	revenus du ménage	PROFES	Profession
FC10	Nombre de personnes dans le ménage	NOCCUA	Occupation actuelle
FC11	Nombre d'enfant de moins de 10 ans dans le ménage	REV1	Ressource principales du ménage
FC11b	Nombre d'enfants de plus de 10 ans dans le ménage	STRMEN	Structure du ménage

# Cas pratique : Corrélation entre variables

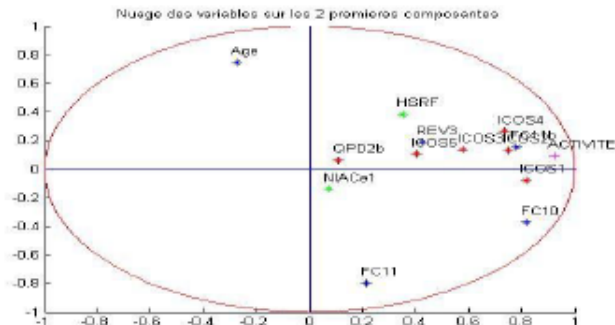
	IC0S1	IC0S2	IC0S3	IC0S4	IC0S5	OP02b	Age	REV3	FC10	FC11	FC11b	HSRF	NIACe1
IC0S1	1,00												
IC0S2	0,62	1,00											
IC0S3	0,38	0,40	1,00										
IC0S4	0,56	0,52	0,42	1,00									
IC0S5	0,28	0,20	0,20	0,27	1,00								
OP02b	0,06	0,05	0,06	0,07	0,04	1,00							
Age	-0,33	-0,12	-0,12	-0,04	-0,06	-0,01	1,00						
REV3	0,23	0,27	0,15	0,25	0,24	0,03	0,02	1,00					
FC10	0,60	0,50	0,34	0,45	0,22	0,06	-0,34	0,29	1,00				
FC11	0,10	0,00	0,00	-0,06	0,05	0,01	-0,42	0,07	0,59	1,00			
FC11b	0,62	0,54	0,37	0,54	0,20	0,05	-0,07	0,20	0,68	-0,07	1,00		
HSRF	0,15	0,25	0,19	0,33	0,17	0,09	0,25	0,41	0,36	0,07	0,28	1,00	
NIACe1	0,08	0,08	0,09	0,05	-0,01	0,05	-0,01	0,12	0,13	0,06	0,08	0,02	1



# Cas pratique : Inertie des axes



# Cas pratique : Variables dans le 1er Plan

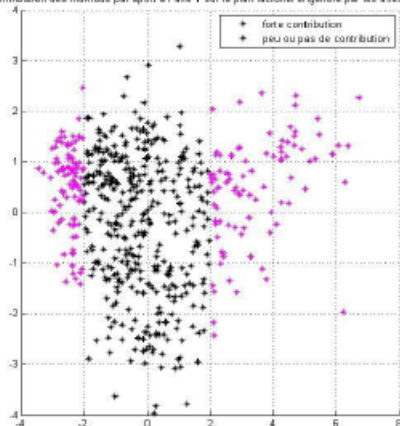


Cas pratique :  $\tilde{\rho}_{ij}$ 

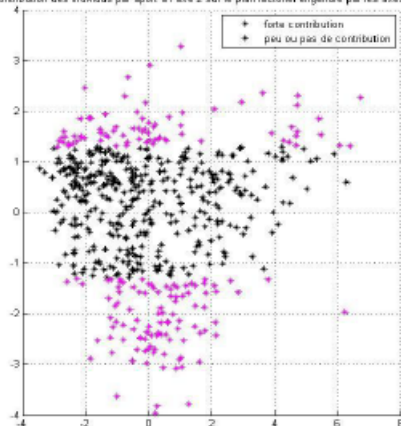
		CP1=30.70	CP2=12.97	CP3=9.51	CP4=7.78	CP5=7.52	CP6=7.02	CP7=5.9
habitude	ICOS1	0,817	-0,080	-0,213	0,115	-0,092	-0,011	0,196
	ICOS2	0,751	0,134	-0,192	0,049	0,011	0,079	0,147
	ICOS3	0,579	0,135	-0,193	-0,032	-0,087	-0,048	-0,679
	ICOS4	0,734	0,265	-0,150	0,036	-0,044	-0,006	-0,009
	ICOS5	0,404	0,105	0,212	0,165	-0,104	-0,768	-0,095
	QPD2b	0,110	0,062	0,115	-0,715	-0,661	-0,051	0,126
Logement	HSRF	0,351	0,382	0,620	-0,105	0,082	0,293	-0,216
	NIACe1	0,071	-0,140	-0,433	-0,626	0,557	-0,195	-0,059
ménage	Age	-0,275	0,748	0,106	-0,162	0,233	0,121	-0,027
	REV3	0,425	0,189	0,470	-0,097	0,342	-0,276	0,295
	FC10	0,820	-0,373	0,177	-0,060	0,099	0,203	0,015
	FC11	0,216	-0,797	0,408	-0,112	0,114	0,071	-0,142
	FC11b	0,780	0,155	-0,190	0,032	0,011	0,227	0,152

# Cas pratique : Contribution des individus

Contribution des individus par rapport à l'axe 1 sur le plan factoriel engendré par les axes 1 et 2



Contribution des individus par rapport à l'axe 2 sur le plan factoriel engendré par les axes 1 et 2



# L'Analyse des Correspondances Simples A F C

# Structure de base des données

Objective : mesurer des liaisons entre deux variables qualitatives : **Khi-deux**

exemple : Il s'agit de tester l'indépendance de deux variables qualitatives. Y a-t-il indépendance entre :

- la catégorie socioprofessionnelle et le vote à l'élection présidentielle ?
- le niveau d'études et les journaux lus ?
- Que pensez vous de cette affirmation : *On en a assez de ceux qui bloquent la vie du pays par leurs revendications par les armes.*
  - 1 pas du tout d'accord
  - 2 pas tellement d'accord
  - 3 bien d'accord
  - 4 entièrement d'accord

# Structure de base des données

Existe-t- il un lien entre les réponses et la tendance politique ?

Tendance Politique	1	2	3	4	5	Total
PIT	714	71	0	143	71	1000
FPI	284	216	199	174	127	1000
MFA	87	106	228	335	244	1000
RDR	16	86	156	271	471	1000
PDCI	71	71	0	214	643	1000
Indifférent	82	120	244	301	263	1000
Non Reponse	88	124	269	285	233	1000

# Le tableau de contingence

Croisement de deux variables qualitatives I et J à p et q modalités.

	1	2	...	j	...	q	
1							
2							
...							
i				$n_{ij}$			$n_{i\bullet}$
...							
p							
				$n_{\bullet j}$			$n$

$$n_{i\bullet} = \sum_{j=1}^q n_{ij} \quad (\text{total ligne})$$

$$n_{\bullet j} = \sum_{i=1}^p n_{ij} \quad (\text{total colonne})$$

$$n = \sum_{i=1}^p \sum_{j=1}^q n_{ij} \quad (\text{total})$$



# Profils lignes - profils-colonnes - profils marginaux

► **p Profils des lignes**  $\frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}$

profil de la ligne i noté  $\ell_i$

$$\left( \frac{n_{i1}}{n_{i\bullet}} \quad \frac{n_{i2}}{n_{i\bullet}} \quad \dots \quad \frac{n_{iq}}{n_{i\bullet}} \right) \Leftrightarrow \left( \frac{f_{i1}}{f_{i\bullet}} \quad \frac{f_{i2}}{f_{i\bullet}} \quad \dots \quad \frac{f_{iq}}{f_{i\bullet}} \right)$$

► **q Profils des colonnes**

profil de la colonne j  
noté  $c_j$

$$\begin{pmatrix} \frac{n_{1j}}{n_{\bullet j}} \\ \frac{n_{2j}}{n_{\bullet j}} \\ \frac{n_{pj}}{n_{\bullet j}} \end{pmatrix} \Leftrightarrow \begin{pmatrix} \frac{f_{1j}}{f_{\bullet j}} \\ \frac{f_{2j}}{f_{\bullet j}} \\ \frac{f_{pj}}{f_{\bullet j}} \end{pmatrix}$$

# Profils lignes - profils-colonnes - profils marginaux

Si les deux variables qualitatives I et J étaient indépendantes, les profils lignes seraient tous identiques, et donc identiques au profil marginal correspondant.

$$\text{Indépendance} \Rightarrow \frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n} \Rightarrow n_{ij} = \frac{n_{\bullet j} * n_{i\bullet}}{n}$$

# Profils lignes - profils-colonnes - profils marginaux

- On pouvait établir la relation précédente en raisonnant sur les profils colonnes.

$$\boxed{f_{ij} = f_{i\bullet} f_{\bullet j}}$$

$$\text{avec } f_{ij} = \frac{n_{ij}}{n} \text{ et } f_{i\bullet} = \frac{n_{i\bullet}}{n}$$

$$f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

Elle exprime clairement que dans le cas de l'indépendance le tableau de contingence est entièrement déterminé par ses marges

# Khi-deux

- Pour chaque case, on peut donc calculer le nombre de cas attendus (sous hypothèse d'indépendance)  $n_{ij} = \frac{n_{\bullet j} * n_{i \bullet}}{n}$
- On peut comparer les nombres de cas attendus  $E_{ij}$  aux nombres observés.

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{\left( n_{ij} - \frac{n_{i \bullet} n_{\bullet j}}{n} \right)^2}{\frac{n_{i \bullet} n_{\bullet j}}{n}}$$

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

# Test Khi-deux

Si les deux variables sont réellement indépendantes, cette expression suit une distribution du Khi-deux avec un nombre de degrés de liberté égal à :

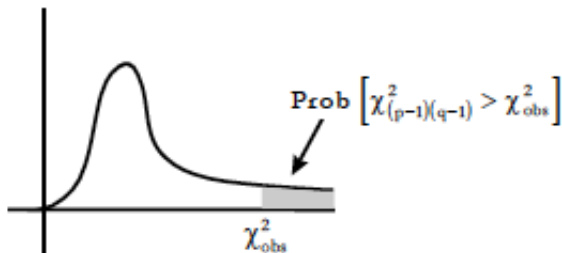
$$(p - 1)(q - 1)$$

Dans une table on lit  $\chi^2_{\alpha,k}$  valeur ayant une probabilité  $\alpha$  d'être dépassée pour une distribution du khi-deux avec  $k = (p - 1)(q - 1)$  degrés de liberté.

- ❶ Si  $\chi^2 \leq \chi^2_{\alpha,k}$  On accepte  $H_o$  : independance
- ❷ Si  $\chi^2 > \chi^2_{\alpha,k}$  On rejette  $H_o$  : independance

# Pratique sous logiciel statistique

- Calcul du  $\chi^2$  associé au tableau de contingence noté  $\chi_{obs}^2$ .
- Probabilité pour une v.a. suivant une loi du khi-deux à  $(p-1)(q-1)$  d.d.l. de dépasser  $\chi_{obs}^2$ .



Si cette probabilité est faible (en général  $< 5\%$ ), on rejette l'hypothèse d'indépendance entre les deux variables qualitatives.

# Représentation des profils lignes

- Les profils lignes sont considérés comme des individus.
- Les  $p$  profils-lignes forment un nuage de  $p$  points dans  $R^q$
- A chaque profil-ligne est associé un poids égal à sa fréquence marginale profil ligne poids  $f_{i\bullet}$ .

On note  $N(I)$  le nuage de points formé des profils-lignes pondérés :  $(I_i; f_{i\bullet})$   
Le centre de gravité  $g$  est défini par :

$$g_I = \sum_{i=1}^p f_{i\bullet} \cdot I_i$$

La  $j$ ème coordonnée de  $g_I$  vaut  $f_{\bullet j}$

$g_I$  = profil marginal de la variable  $J$  (à  $q$  modalités)  $g_I = f_J$

# Représentation des profils colonnes

$N(J)$  = nuage de points formé des  $q$  profils - colonnes pondérés  $(c_j, f_{\bullet j})$

Le centre de gravité  $g_c$  est le profil marginal  
de la variable  $I$  à  $p$  modalités.

$$g_c = f_{\bullet}$$



Le problème qui se pose est **l'étude de la dépendance** entre les deux variables qualitatives.

Dans le cas où les deux variables sont **indépendantes**, on a identité des profils :

$$(1) \quad \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j} \quad \text{profil - ligne}$$

$$(2) \quad \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet} \quad \text{profil - colonne}$$

$$\boxed{f_{ij} = f_{i\bullet} f_{\bullet j}}$$

Dans le cas de l'indépendance, le nuage des profils-lignes se réduit à un point  $\underline{g}_\ell$

De même, le nuage des profils-colonnes se réduit à un point  $\underline{g}_c$ .

⇒ L'étude de la dépendance consiste à étudier la **forme des nuages**.

⇒ **Problème d'analyse en composantes principales.**

**Quelle métrique ?**

# Métrique du $\chi^2$

- Pour les profils lignes :

$$d_{\chi^2}^2(l_i, l'_i) = \sum_{j=1}^q \frac{n}{n_{\bullet j}} \left( \frac{n_{ij}}{n_{i\bullet}} - \frac{n_{i'j}}{n_{i'\bullet}} \right)^2$$

- Donne un poids important aux différences portant sur les petits pourcentages.
- Vérifie le principe d'équivalence distributionnelle : si deux colonnes ont le même profil, on les réunit en une seule d'effectif somme sans modifier les distances entre profils lignes.
- Pour les profils-colonnes :

$$d_{\chi^2}^2(c_j, c'_j) = \sum_{i=1}^q \frac{n}{n_{i\bullet}} \left( \frac{n_{ij}}{n_{\bullet j}} - \frac{n_{ij'}}{n_{\bullet j'}} \right)^2$$

# Inertie du nuage $N(I)$

$I_{N(I)}$  l'inertie du nuage  $N(I)$  calculée par rapport au centre de gravité  $f_J$  vaut

$$\frac{\chi^2}{n}$$

où  $\chi^2$  = Khi-deux associé au tableau de contingence étudié.

On obtient le même résultat pour l'inertie du nuage  $N(J)$ .

## l'A.C.P. du nuage des profils-lignes :

- Les données profils-lignes jouent le rôle d'individus

$$X = D_1^{-1}N$$

- La métrique utilisée pour le calcul des distances entre individus

$$M = nD_2^{-1} \text{ et le poids } D = \frac{D_1}{n}$$

- Les Facteurs principaux  $u_k$  du nuage des profils-lignes  $N(I)$  sont vecteurs propres de

$$MX'DX = (nD_2^{-1})(D_1^{-1}N)'(\frac{D_1}{n})(D_1^{-1}N) = D_2^{-1}N'D_1^{-1}N$$

On a donc pour chaque axe principal  $k$

$$D_2^{-1}N'D_1^{-1}Nu_k = \lambda_k u_k$$

- La composante principale  $a_k$  associée au facteur  $u_k$  est vecteur propre de :

$$D_1^{-1}N'D_2^{-1}N$$

# l'A.C.P. du nuage des profils-colonnes :

Analyse des profils-colonnes on échange les indices 1 et 2 et on transpose N.  
On notera  $b_k$  les composantes principales.

## Comparaison Lignes-Colonnes

	ACP profils-lignes	ACP profils-colonnes
Facteurs principaux	Vecteurs propres de $D_2^{-1}N'D_1^{-1}N$	Vecteurs propres de $D_1^{-1}ND_2^{-1}N'$
Composantes principales	Vecteurs propres de $D_1^{-1}ND_2^{-1}N'$ normalisés par $\text{var } a_k = a'_k \frac{D_1}{n} a_k = \lambda_k$	Vecteurs propres de $D_2^{-1}N'D_1^{-1}N$ normalisés par $\text{var } b_k = b'_k \frac{D_2}{n} b_k = \lambda_k$

# Abandon du principe barycentrique

Les modalités de chaque ensemble sont représentées par les :

$$a_i^k, i = 1, \dots, p$$

$$b_j^k, j = 1, \dots, q$$

Cette représentation permet de déterminer les proximités entre certains éléments de I et certains éléments de J (**compte tenu de la qualité de la représentation**).

# Formules de transition

En notant  $b_j$  et  $a_i$  les  $j^{eme}$  et  $i^{eme}$  coordonnées des composantes principales  $b$  et  $a$  associées à la même valeur propre  $\lambda$  :

$$\sqrt{\lambda} b_j = \sum_{i=1}^p \frac{n_{ij}}{n_{\bullet j}} a_i$$

$$\sqrt{\lambda} a_i = \sum_{j=1}^q \frac{n_{ij}}{n_{i\bullet}} b_j$$

À  $\lambda$  près, la coordonnée d'une modalité  $i$  d'une variable est la moyenne des coordonnées des catégories de l'autre variable pondérées par les fréquences conditionnelles du profil de  $i$ .

# Contributions

de la ligne  $i$  à l'axe  $k$

$$\frac{f_{i\bullet} (a_i^k)^2}{\lambda_k} \quad \text{avec } f_{i\bullet} = \frac{n_{i\bullet}}{n}$$

de la colonne  $j$  à l'axe  $k$

$$\frac{f_{\bullet j} (a_j^k)^2}{\lambda_k} \quad \text{avec } f_{\bullet j} = \frac{n_{\bullet j}}{n}$$



# Cosinus carrés

Modalité  $i$  représentée sur l'axe  $k$

$$\frac{(a_i^k)^2}{d^2(i, G)}$$

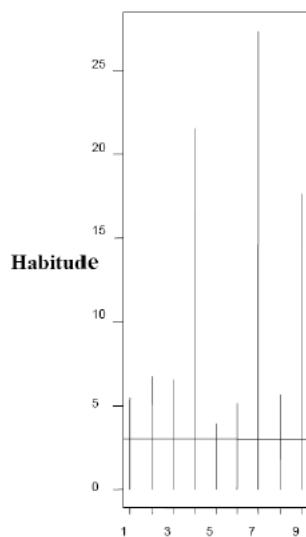
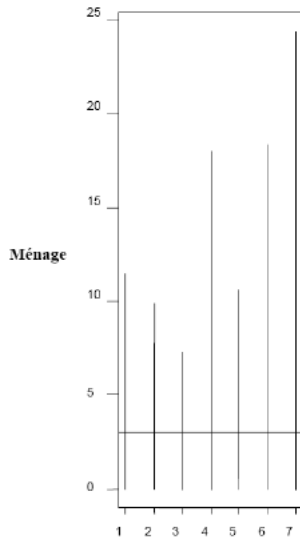
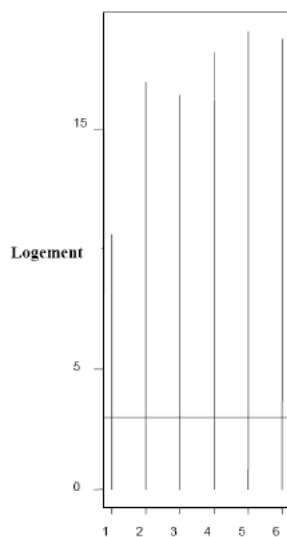
Modalité  $j$  représentée sur l'axe  $k$

$$\frac{(b_j^k)^2}{d^2(j, G)}$$

# Aspects pratiques de l'interprétation

- L'interprétation peut se faire à partir des représentations graphiques (en s'assurant de la qualité de représentation de chaque modalité à l'aide des  $\cos^2$ ).
- Quand le nombre de modalités est élevé, il est conseillé d'éditer d'abord le graphique des profils-lignes, puis celui des profils colonnes, enfin la représentation simultanée.
- Les profils ayant des poids différents la lecture de leurs contributions à l'inertie de chaque axe s'avère très utile.
- On peut repérer les profils dont la contribution est supérieure au poids

# Exemples : Données CNL 2005



Modalités

# Exemples : Test du Khi2

	$\alpha$	p.value
Logement-ménage	114.4779	$(8.462)^{-12}$
Logement-habitude	172.6236	$(2.2)^{-16}$
Ménage-habitude	164.3131	$(1.214)^{-14}$

S'en suit l'étude des liens entre les modalités.

# Analyse des Correspondances Multiples

- Étendre l'AFC au cas de  $p \geq 2$  variables  $\xi_1, \xi_2, \dots, \xi_p$  à  $m_1, \dots, m_p$  modalités

$$\begin{array}{cccc} \xi_1 & \dots & \xi_p & \text{variables} \\ m_1 & \dots & m_p & \text{modalités} \end{array}$$

Utile pour l'exploration d'enquêtes où les questions sont à réponses multiples.

- L'analyse des correspondances utilise une table de contingence qui est difficilement généralisable au cas  $p \geq 2$ .

Trouver un moyen différent d'analyser  $p > 2$  variables et vérifier que les résultats sont comparables à l'AFC pour  $p = 2$ .

# Données

- **Données brutes** : chaque individu est décrit par les numéros des modalités qu'il possède pour chacune des  $p$  variables  $\xi_j$ .

Impossible de faire des calculs sur ce tableau : valeurs arbitraires..

- Une forme mathématique utile pour les calculs est alors **le tableau disjonctif des indicatrices des  $p$  variables** obtenu en juxtaposant les  $p$  tableaux d'indicatrices de chaque variable  $\mathcal{X}_j$
- **Tableau disjonctif** : Remplacer la  $j$ -ième colonne par  $m_j$  colonnes d'indicatrices : mettre un zéro dans chaque colonne, sauf celle correspondant à la modalité de l'individu  $i$  qui reçoit 1.

# Technique de description de données qualitatives

$n$  individus décrits par  $p$  variables qualitatives

$\mathcal{X}_1$	...	$\mathcal{X}_p$	variables
$m_1$	...	$m_p$	modalités

- L'A.C.M. décrit les relations deux à deux entre  $p$  variables qualitatives à travers une représentation des groupes d'individus correspondant aux diverses modalités. .
- Cette méthode est particulièrement bien adaptée à l'exploration d'enquêtes.

# Données

## Exemple

On interroge 6 personnes sur la couleur de leurs cheveux (CB, CC et CR pour blond, châtain et roux), la couleur de leurs yeux (YB, YV et YM pour bleu, vert et marron) et leur sexe (H/F). Les tableaux brut (ci-dessous à gauche) sont équivalents aux tableaux disjonctifs (à droite).

$$\begin{pmatrix} \text{CB} \\ \text{CB} \\ \text{CC} \\ \text{CC} \\ \text{CR} \\ \text{CB} \end{pmatrix}
 \begin{pmatrix} \text{YB} \\ \text{YV} \\ \text{YB} \\ \text{YM} \\ \text{YV} \\ \text{YB} \end{pmatrix}
 \begin{pmatrix} \text{H} \\ \text{H} \\ \text{F} \\ \text{H} \\ \text{F} \\ \text{F} \end{pmatrix}
 \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}
 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}
 \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$



# Tableau disjonctif et tableau de contingence

A chaque variable  $\xi_j$  est associée un tableau disjonctif  $X_j(n \times m_j)$ .  
 Pour 2 variables  $\xi_j$  et  $\xi_l$  le tableau de contingence est donné par :

$$\mathcal{X} = X_j | X_l$$

Disjonctifs

$$N_{j,l} = X_j' X_l$$

Contingence

$$X_j' X_j$$

Marge  $\xi_j$

$$X_l' X_l$$

Marge  $\xi_l$

$$\mathbf{N}_{12} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{D}_1 = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

# Tableau disjonctif joint

$$\mathcal{X}(n \times m) = X_1 | X_2 \dots | X_p$$

$$m = m_1 + \dots + m_p$$

**Exemple** Pour les variables précédentes, on a le tableau disjonctif joint suivant

$$\mathbf{X} = \left( \begin{array}{ccc|ccc|cc} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array} \right)$$

Chaque somme de lignes vaut 3. Les sommes de colonnes valent

$$(3 \ 2 \ 1 | 3 \ 2 \ 1 | 3 \ 3)$$

# Le tableau de Burt

C'est un super-tableau de contingence des variables  $X_1, \dots, X_p$ , formé de tableaux de contingence et de matrices d'effectifs marginaux. :

$$B = X'X = \begin{bmatrix} X'_1X_1 & X'_1X_2 & \cdots & X'_1X_p \\ X'_2X_1 & X'_2X_2 & & \\ \vdots & & \ddots & \vdots \\ X'_pX_1 & \cdots & & X'_pX_p \end{bmatrix}$$

$$= \begin{bmatrix} D_1 & N_{12} & \cdots & N_{1p} \\ N_{21} & D_2 & & \\ \vdots & & \ddots & \vdots \\ N_{p1} & \cdots & & D_p \end{bmatrix}$$

**Exemple** Toujours pour les mêmes variables

$$B = \left( \begin{array}{ccc|ccc|cc} 3 & 0 & 0 & 2 & 1 & 0 & 2 & 1 \\ 0 & 2 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ \hline 2 & 1 & 0 & 3 & 0 & 0 & 1 & 2 \\ 1 & 0 & 1 & 0 & 2 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ \hline 2 & 1 & 0 & 1 & 1 & 1 & 3 & 0 \\ 1 & 1 & 1 & 2 & 1 & 0 & 0 & 3 \end{array} \right)$$

# L'ACM : une AFC sur tableau disjonctif

Chercher une représentation des  $m_1 + \dots + m_p$  catégories comme points d'un espace de faible dimension.

**Méthode** Faire une AFC sur le tableau disjonctif joint

$$\mathcal{X}(n \times m) = X_1 | X_2 | \dots | X_p$$

# L'ACM : une AFC sur tableau disjonctif

- **Les lignes** : La somme des éléments de chaque ligne de  $\mathcal{X}$  est égale à  $p$ . Le **tableau des profils-lignes** est donc  $\frac{1}{p}\mathcal{X}$
- **Les colonnes** : la somme des éléments de chaque colonne de  $\mathcal{X}$  est égale à l'effectif marginal de la catégorie correspondante.

Le tableau des profils colonnes est donc  $\mathcal{X}D^{-1}$  où  $D$  est la matrice diagonale par blocs.

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{D}_p \end{pmatrix}$$

# Les coordonnées factorielles des catégories

On note  $a_k = (a_{1k}, \dots, a_{pk})$  vecteur à  $m_1 + \dots + m_p$  composantes des coordonnées factorielles des catégories sur l'axe  $k$ .

## Calcul de l'AFC sur $\mathcal{X}$

La matrice des profils lignes est

$$\frac{1}{p} \mathcal{X}$$

et celle des profils colonnes

$$\mathcal{X} D^{-1}$$

.

# Les coordonnées factorielles des catégories

On note  $a_k = (a_{1k}, \dots, a_{pk})$  vecteur à  $m_1 + \dots + m_p$  composantes des coordonnées factorielles des catégories sur l'axe  $k$ .

$a_k$  est vecteur propre de

$$(\mathcal{X}D^{-1})' \frac{1}{p} \mathcal{X} = \frac{1}{p} D^{-1} \mathcal{X}' \mathcal{X} = \frac{1}{p} D^{-1} B$$

l'équation des coordonnées des catégories est donc

$$\frac{1}{p} D^{-1} B a_k = \mu_k a_k$$

Avec la convention de normalisation suivantes

$$\frac{1}{np} a_k' D a_k = \mu_k$$

# Resolution cas $p=2$

On note  $a = (a_1, a_2)$  vecteur à  $m_1 + m_2$  composantes factorielles des catégorie et  $\mu_k$  la valeur propre correspondante.

**Calcul de l'AFC sur  $\mathcal{X}$**

$$\frac{1}{2} \mathbf{D}^{-1} \mathbf{B} \begin{bmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \mathbf{I}_{m_1} & \mathbf{D}_1^{-1} \mathbf{N} \\ \mathbf{D}_2^{-1} \mathbf{N}' & \mathbf{I}_{m_2} \end{bmatrix} \begin{bmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{bmatrix} = \mu_k \begin{bmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{bmatrix}$$



## Resolution cas $p=2$

On note  $a = (a_1, a_2)$  vecteur à  $m_1 + m_2$  composantes factorielles et  $\mu_k$  la valeur propre correspondante.

On obtient les équations

$$\begin{cases} \mathbf{D}_1^{-1} \mathbf{N} \mathbf{b}_k = (2\mu_k - 1) \mathbf{a}_k \\ \mathbf{D}_2^{-1} \mathbf{N}' \mathbf{a}_k = (2\mu_k - 1) \mathbf{b}_k \end{cases}$$

et donc on retrouve les coordonnées des modalités de lignes et de colonnes dans l'AFC classique (avec  $\mu_k = (2\lambda_k - 1)^2$ )

$$\begin{cases} \mathbf{D}_2^{-1} \mathbf{N}' \mathbf{D}_1^{-1} \mathbf{N} \mathbf{b}_k = (2\mu_k - 1)^2 \mathbf{b}_k \\ \mathbf{D}_1^{-1} \mathbf{N} \mathbf{D}_2^{-1} \mathbf{N}' \mathbf{a}_k = (2\mu_k - 1)^2 \mathbf{a}_k \end{cases}$$

## Différences ACM/AFC pour $p = 2$

- **Nombre de valeurs propres** : on a a priori  $m_1 + m_2 - 2$  valeurs propres non nulles, En particulier pour chaque  $\lambda_k$ , on a deux  $\mu_k$  possibles

$$\begin{cases} \mu_k = \frac{1+\sqrt{\lambda_k}}{2} & \text{associée à } \begin{bmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{bmatrix} \\ \mu'_k = \frac{1-\sqrt{\lambda_k}}{2} & \text{associée à } \begin{bmatrix} \mathbf{a}_k \\ -\mathbf{b}_k \end{bmatrix} \end{cases}$$

On ne garde donc que les valeurs  $\mu_k > 0.5$

- **Inertie** : l'interprétation de la part d'inertie expliquée par les valeurs propres est maintenant très différente.

En particulier les valeurs propres qui étaient très séparées dans l'AFC de N le sont beaucoup moins dans celle de X.

# Formules barycentriques

- Les coordonnées des individus

$$\mathbf{c}_k = \frac{1}{\sqrt{\mu_k}} \frac{1}{p} \mathbf{X} \mathbf{a}_k \quad \text{et donc} \quad c_{ik} = \frac{1}{\sqrt{\mu_k}} \frac{1}{p} \sum_{j \text{ catégorie de } i} a_{jk}$$

Avec variance

$$\text{var } \mathbf{c}_k = \frac{1}{n} \mathbf{c}_k' \mathbf{c}_k = \mu_k$$

- Les coordonnées des catégories

$$\mathbf{a}_k = \frac{1}{\sqrt{\mu_k}} \mathbf{D}^{-1} \mathbf{X}' \mathbf{c}_k \quad \text{c-à-d} \quad a_{jk} = \frac{1}{\sqrt{\mu_k}} \frac{1}{n_j} \sum_{i \text{ de catégorie } j} c_{ik}$$

# Barycentres et représentation

- Les points représentatifs des catégories sont barycentres des groupes d'individus.
- Moyennes comme  $c_k$  est une variable de moyenne nulle, la formule de barycentre indique que pour chaque variable  $X_i$  les coordonnées de ses catégories sont de moyenne nulle.
- Pour que les catégories se trouvent visuellement au barycentre des individus qui les représentent on peut remplacer  $a_k$

$$\alpha_k = \mathbf{D}^{-1} \mathbf{X}' \mathbf{c}_k = \sqrt{\mu_k} \mathbf{a}_k$$

# Sélection des axes

- règle courante : garder les axes tels que  $\mu_k > \frac{1}{p}$  (la moyenne des valeurs propres est  $\frac{1}{p}$ ).
- les axes intéressants sont ceux que l'on peut interpréter, en regardant les contributions des variables actives et les valeurs-tests associées aux variables supplémentaires.
- En pratique on se contente souvent d'interpréter le premier plan principal.

# Sélection des axes

- Si  $n_j$  est l'effectif de la catégorie  $j$  et  $a_{jk}$  sa coordonnée sur l'axe factoriel  $k$ , alors

$$\text{var } \mathbf{a}_k = \sum_{j \in \text{catégories}} \frac{n_j}{np} (a_{jk})^2 = \mu_k$$

- **Catégorie** La contribution de la catégorie  $j$  à l'axe factoriel est :

$$\frac{n_j}{np} \frac{(a_{jk})^2}{\mu_k},$$

- **Variable** : la contribution totale de la variable  $\xi_v$  à l'axe factoriel est

$$\frac{1}{\mu_k} \frac{1}{np} \sum_{j \text{ modalité de } \mathcal{X}_v} n_j (a_{jk})^2$$

# Contribution d'un individu

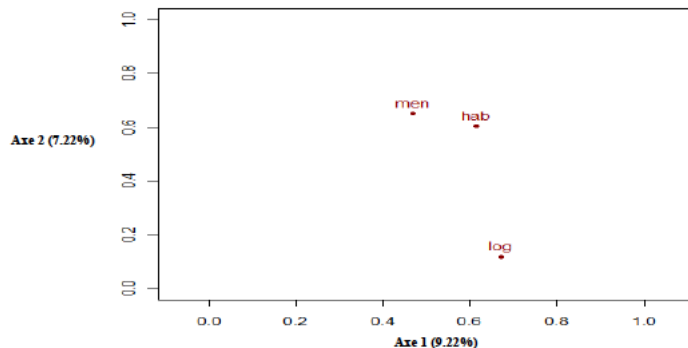
- Elle est égale pour l'individu  $i$  à

$$\frac{1}{n} \frac{(c_{ik})^2}{\mu_k}$$

- Qualité de la représentation pour le sous-espace formé par les premier axes, la qualité de la représentation de l'individu  $i$  est le cosinus carré habituel

$$\frac{\sum_{k=1}^{\ell} (c_{ik})^2}{\sum_{k=1}^q (c_{ik})^2}$$

# Exemple

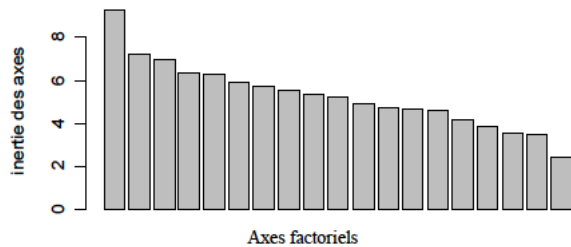


Projection des variables qualitatives sur le premier plan

Que pensez vous de l'inertie portée par le premier plan ?



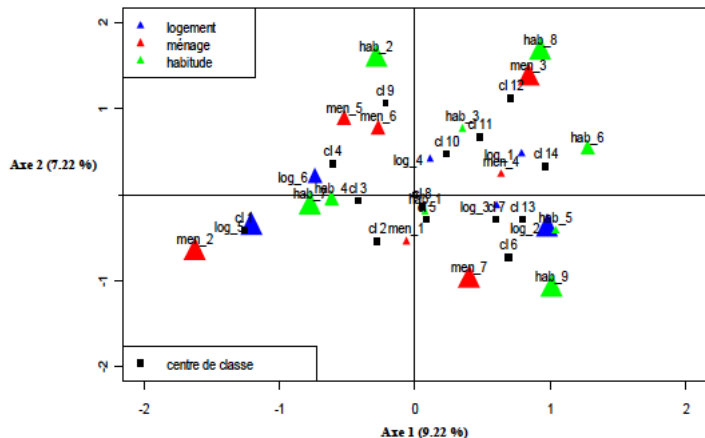
# Exemple



Inertie des différents

Que pensez vous du nombre d'axes ?

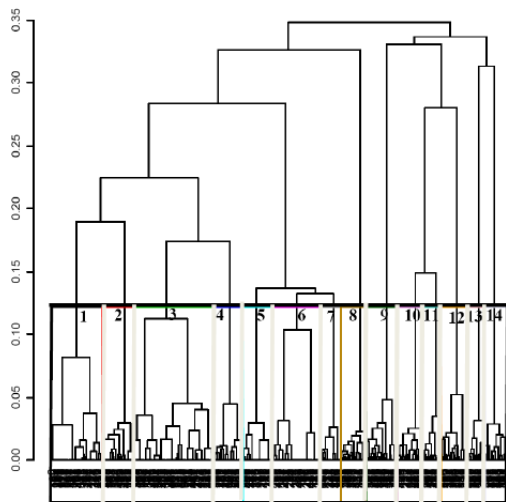
# Exemple



Projection des modalités des variables qualitatives

Interprétez ?

# Exemple : Utilisation des composantes factorielles



CAH sur les composantes factorielles