

# Panorama sur les méthodes d'analyse exploratoire des données

Magalie Houée-Bigot & François Husson

Unité de mathématiques appliquées, Agrocampus Ouest, Rennes

11-13 mars 2019

# Plan

Introduction

Analyse en Composantes Principales



Analyse des correspondances

Analyse des correspondances multiples

Analyse Factorielle Multiple

Classification

Conclusion

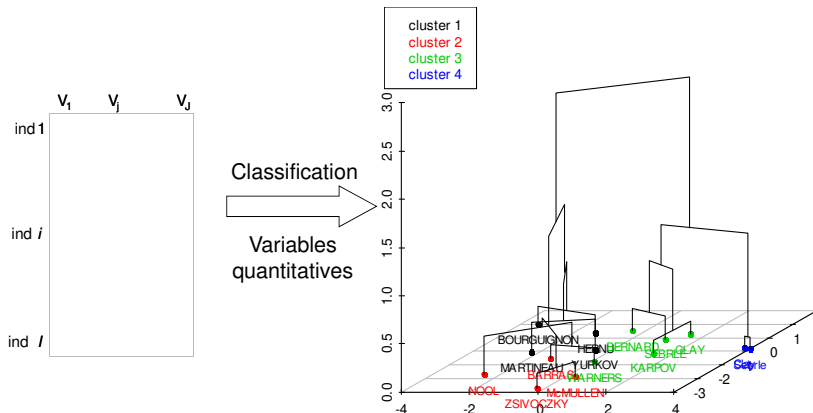
# Unité de mathématiques appliquées

- Recherche
  - Analyse factorielle, multi-tableaux, données manquantes
  - modélisation en grande dimension - tests multiples
  - Application : génomique, écologie, sensoriel
- Enseignement
  - L3 : modèle linéaire, analyse de données, plan d'expériences
  - Spécialisation et Master en science des données : sensométrie, tableaux multiples, données génomiques
  - Livres : Analyse de données avec R, R pour la statistique et la science des données, Statistique avec R, Statistique générale
  - MOOC Analyse de données, Sensométrie
- Autres activités
  - Packages R : FactoMineR, missMDA, FAMT, SensoMineR
  - Formation continue : statistique avec R, analyse de données





# Les méthodes d'analyse de données



## Les méthodes d'analyse de données

- Analyse en composantes principales (variables quantitatives)
- Analyse des correspondances (tableau de contingence)
- Analyse des correspondances multiples (variables qualitatives)
- Analyse factorielle de données mixtes (variables quanti et qualitatives)
- Analyse factorielle multiple (groupes de variables quantitatives et/ou qualitatives)
- Classification

⇒ Statistique descriptive multidimensionnelle

⇒ Objectifs communs : résumer, visualiser les données

# Plan

Introduction

Analyse en Composantes Principales



Analyse des correspondances

Analyse des correspondances multiples

Analyse Factorielle Multiple

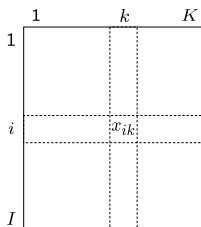
Classification

Conclusion



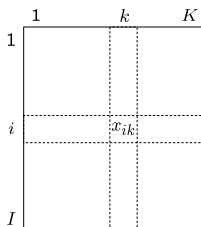
## Quel type de données ?

L'ACP s'intéresse à des tableaux de données rectangulaires avec des **individus** en lignes et des **variables quantitatives** en colonnes



## Quel type de données ?

L'ACP s'intéresse à des tableaux de données rectangulaires avec des **individus** en lignes et des **variables quantitatives** en colonnes



- Écologie : concentration du **polluant**  $k$  dans la **rivière**  $i$
- Écologie : **caractéristique physique**  $k$  du **sol**  $i$
- Biologie : **mesure morphologique**  $k$  pour l'**animal**  $i$
- Génétique : expression du **gène**  $k$  pour le **patient**  $i$
- Sociologie : **tps passé à l'activité**  $k$  par les individus de la **CSP**  $i$

## Les données vins

- 10 individus : vins blancs du Val de Loire
- 30 variables :
  - 27 variables quantitatives : descripteurs sensoriels
  - 2 variables quantitatives : appréciation de l'odeur et générale
  - 1 variable qualitative : label des vins (Vouvray - Sauvignon)



## Les données vins

- 10 individus : vins blancs du Val de Loire
- 30 variables :
  - 27 variables quantitatives : descripteurs sensoriels
  - 2 variables quantitatives : appréciation de l'odeur et générale
  - 1 variable qualitative : label des vins (Vouvray - Sauvignon)

	O.fruity	O.passion	O.citrus	...	Sweetness	Acidity	Bitterness	Astringency	Aroma.intensity	Aroma.persistency	Visual.intensity	Odor.preference	Overall.preference	Label
S Michaud	4.3	2.4	5.7	...	3.5	5.9	4.1	1.4	7.1	6.7	5.0	6.0	5.0	Sauvignon
S Renaudie	4.4	3.1	5.3	...	3.3	6.8	3.8	2.3	7.2	6.6	3.4	5.4	5.5	Sauvignon
S Trotignon	5.1	4.0	5.3	...	3.0	6.1	4.1	2.4	6.1	6.1	3.0	5.0	5.5	Sauvignon
S Buisse Domaine	4.3	2.4	3.6	...	3.9	5.6	2.5	3.0	4.9	5.1	4.1	5.3	4.6	Sauvignon
S Buisse Cristal	5.6	3.1	3.5	...	3.4	6.6	5.0	3.1	6.1	5.1	3.6	6.1	5.0	Sauvignon
V Aub Silex	3.9	0.7	3.3	...	7.9	4.4	3.0	2.4	5.9	5.6	4.0	5.0	5.5	Vouvray
V Aub Marigny	2.1	0.7	1.0	...	3.5	6.4	5.0	4.0	6.3	6.7	6.0	5.1	4.1	Vouvray
V Font Domaine	5.1	0.5	2.5	...	3.0	5.7	4.0	2.5	6.7	6.3	6.4	4.4	5.1	Vouvray
V Font Brûlés	5.1	0.8	3.8	...	3.9	5.4	4.0	3.1	7.0	6.1	7.4	4.4	6.4	Vouvray
V Font Coteaux	4.1	0.9	2.7	...	3.8	5.1	4.3	4.3	7.3	6.6	6.3	6.0	5.7	Vouvray

# Problèmes - objectifs

Tableau = ensemble de lignes ou ensemble de colonnes

## Etude des individus

- construction de groupes d'individus se ressemblant du point de vue de l'ensemble des variables
- bilan des ressemblances, une partition des individus

## Etude des variables

- recherche des ressemblances, liaisons (linéaires) entre variables
- bilan des liaisons : visualisation de la matrice des corrélations
- recherche d'indicateurs synthétiques résumant les variables

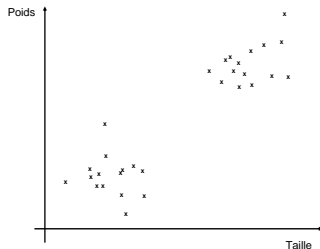
Lien entre les deux études

- caractérisation des classes d'individus par les variables
- individus spécifiques pour comprendre les liaisons entre variables

## Objectifs de l'ACP :

- Descriptif - exploratoire : visualisation de données
- Synthèse - résumé de grands tableaux individus  $\times$  variables

## Nuage des individus



- Les individus vivent dans  $\mathbb{R}^p$
- Etudier la forme du nuage des individus

- Notion de distance entre individus : **Quelle distance ? question cruciale !!!**

Doit-on normer les variables ? Transformer les variables (par ex. passage au log) ?

## Ajustement du nuage

Trouver le sous-espace qui fournit la meilleure représentation des données

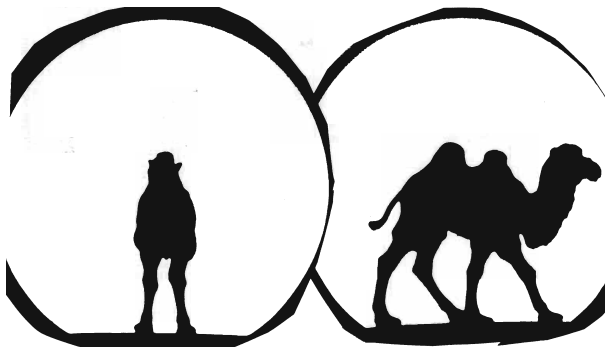
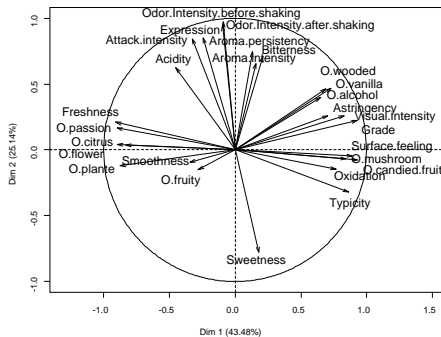
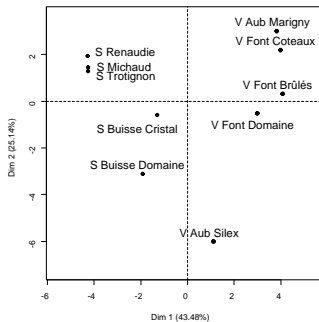


FIGURE – Chameau ou dromadaire ? source J.P. Fenelon

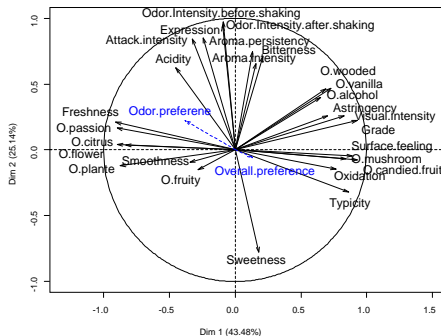
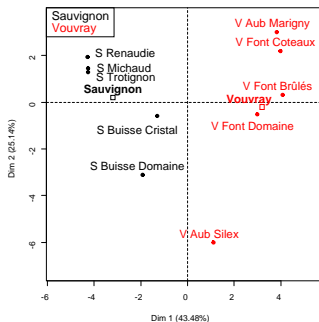
- ⇒ Meilleure approximation par projection
- ⇒ Meilleure représentation de la diversité, de la variabilité

# Représentation des individus et des variables





# Représentation des individus et des variables

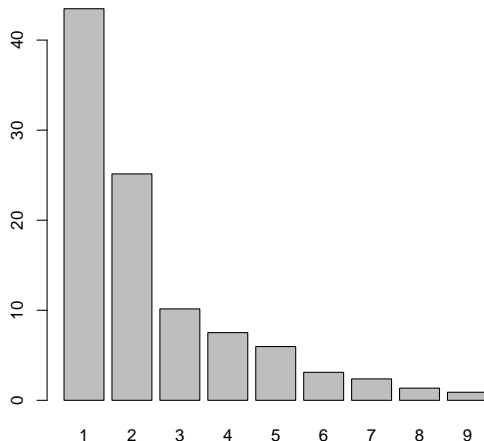


⇒ Utiliser de l'information supplémentaire

- Variables qualitatives : modalités au barycentre des individus qui prennent cette modalité
- Variables quantitatives : projection des variables

## Pourcentage d'inertie

- Pourcentage d'information (d'inertie) expliqué par chaque axe



⇒ Choix d'un nombre de dimensions à interpréter

# Pourcentage d'inertie si indépendance entre variables

nbind	Nombre de variables												
	4	5	6	7	8	9	10	11	12	13	14	15	16
5	96.5	93.1	90.2	87.6	85.5	83.4	81.9	80.7	79.4	78.1	77.4	76.6	75.5
6	93.3	88.6	84.8	81.5	79.1	76.9	75.1	73.2	72.2	70.8	69.8	68.7	68.0
7	90.5	84.9	80.9	77.4	74.4	72.0	70.1	68.3	67.0	65.3	64.3	63.2	62.2
8	88.1	82.3	77.2	73.8	70.7	68.2	66.1	64.0	62.8	61.2	60.0	59.0	58.0
9	86.1	79.5	74.8	70.7	67.4	65.1	62.9	61.1	59.4	57.9	56.5	55.4	54.3
10	84.5	77.5	72.3	68.2	65.0	62.4	60.1	58.3	56.5	55.1	53.7	52.5	51.5
11	82.8	75.7	70.3	66.3	62.9	60.1	58.0	56.0	54.4	52.7	51.3	50.1	49.2
12	81.5	74.0	68.6	64.4	61.2	58.3	55.8	54.0	52.4	50.9	49.3	48.2	47.2
13	80.0	72.5	67.2	62.9	59.4	56.7	54.4	52.2	50.5	48.9	47.7	46.6	45.4
14	79.0	71.5	65.7	61.5	58.1	55.1	52.8	50.8	49.0	47.5	46.2	45.0	44.0
15	78.1	70.3	64.6	60.3	57.0	53.9	51.5	49.4	47.8	46.1	44.9	43.6	42.5
16	77.3	69.4	63.5	59.2	55.6	52.9	50.3	48.3	46.6	45.2	43.6	42.4	41.4
17	76.5	68.4	62.6	58.2	54.7	51.8	49.3	47.1	45.5	44.0	42.6	41.4	40.3
18	75.5	67.6	61.8	57.1	53.7	50.8	48.4	46.3	44.6	43.0	41.6	40.4	39.3
19	75.1	67.0	60.9	56.5	52.8	49.9	47.4	45.5	43.7	42.1	40.7	39.6	38.4
20	74.1	66.1	60.1	55.6	52.1	49.1	46.6	44.7	42.9	41.3	39.8	38.7	37.5
25	72.0	63.3	57.1	52.5	48.9	46.0	43.4	41.4	39.6	38.1	36.7	35.5	34.5
30	69.8	61.1	55.1	50.3	46.7	43.6	41.1	39.1	37.3	35.7	34.4	33.2	32.1
35	68.5	59.6	53.3	48.6	44.9	41.9	39.5	37.4	35.6	34.0	32.7	31.6	30.4
40	67.5	58.3	52.0	47.3	43.4	40.5	38.0	36.0	34.1	32.7	31.3	30.1	29.1
45	66.4	57.1	50.8	46.1	42.4	39.3	36.9	34.8	33.1	31.5	30.2	29.0	27.9
50	65.6	56.3	49.9	45.2	41.4	38.4	35.9	33.9	32.1	30.5	29.2	28.1	27.0
100	60.9	51.4	44.9	40.0	36.3	33.3	31.0	28.9	27.2	25.8	24.5	23.3	22.3

TABLE – Quantile à 95 % du pourcentage d'inertie des 2 premières dimensions de 10000 PCA obtenue avec des variables indépendantes

## Pourcentage d'inertie si indépendance entre variables

nbind	Nombre de variables												
	17	18	19	20	25	30	35	40	50	75	100	150	200
5	74.9	74.2	73.5	72.8	70.7	68.8	67.4	66.4	64.7	62.0	60.5	58.5	57.4
6	67.0	66.3	65.6	64.9	62.3	60.4	58.9	57.6	55.8	52.9	51.0	49.0	47.8
7	61.3	60.7	59.7	59.1	56.4	54.3	52.6	51.4	49.5	46.4	44.6	42.4	41.2
8	57.0	56.2	55.4	54.5	51.8	49.7	47.8	46.7	44.6	41.6	39.8	37.6	36.4
9	53.6	52.5	51.8	51.2	48.1	45.9	44.4	42.9	41.0	38.0	36.1	34.0	32.7
10	50.6	49.8	49.0	48.3	45.2	42.9	41.4	40.1	38.0	35.0	33.2	31.0	29.8
11	48.1	47.2	46.5	45.8	42.8	40.6	39.0	37.7	35.6	32.6	30.8	28.7	27.5
12	46.2	45.2	44.4	43.8	40.7	38.5	36.9	35.5	33.5	30.5	28.8	26.7	25.5
13	44.4	43.4	42.8	41.9	39.0	36.8	35.1	33.9	31.8	28.8	27.1	25.0	23.9
14	42.9	42.0	41.3	40.4	37.4	35.2	33.6	32.3	30.4	27.4	25.7	23.6	22.4
15	41.6	40.7	39.8	39.1	36.2	34.0	32.4	31.1	29.0	26.0	24.3	22.4	21.2
16	40.4	39.5	38.7	37.9	35.0	32.8	31.1	29.8	27.9	24.9	23.2	21.2	20.1
17	39.4	38.5	37.6	36.9	33.8	31.7	30.1	28.8	26.8	23.9	22.2	20.3	19.2
18	38.3	37.4	36.7	35.8	32.9	30.7	29.1	27.8	25.9	22.9	21.3	19.4	18.3
19	37.4	36.5	35.8	34.9	32.0	29.9	28.3	27.0	25.1	22.2	20.5	18.6	17.5
20	36.7	35.8	34.9	34.2	31.3	29.1	27.5	26.2	24.3	21.4	19.8	18.0	16.9
25	33.5	32.5	31.8	31.1	28.1	26.0	24.5	23.3	21.4	18.6	17.0	15.2	14.2
30	31.2	30.3	29.5	28.8	26.0	23.9	22.3	21.1	19.3	16.6	15.1	13.4	12.5
35	29.5	28.6	27.9	27.1	24.3	22.2	20.7	19.6	17.8	15.2	13.7	12.1	11.1
40	28.1	27.3	26.5	25.8	23.0	21.0	19.5	18.4	16.6	14.1	12.7	11.1	10.2
45	27.0	26.1	25.4	24.7	21.9	20.0	18.5	17.4	15.7	13.2	11.8	10.3	9.4
50	26.1	25.3	24.6	23.8	21.1	19.1	17.7	16.6	14.9	12.5	11.1	9.6	8.7
100	21.5	20.7	19.9	19.3	16.7	14.9	13.6	12.5	11.0	8.9	7.7	6.4	5.7

TABLE – Quantile à 95 % du pourcentage d'inertie des 2 premières dimensions de 10000 PCA obtenue avec des variables indépendantes

## Qualité de représentation – contribution

- Qualité de représentation d'une **variable** et d'un **individu**  
 $\cos^2$  entre une var. et sa projection       $\cos^2$  entre  $Oi$  et  $OH_i$

```
round(res.pca$var$cos2,2)
      Dim.1 Dim.2 Dim.3
0.fruity  0.08  0.02  0.33
0.passion 0.80  0.03  0.01
```

```
round(res.pca$ind$cos2,2)
      Dim.1 Dim.2 Dim.3
S Michaud  0.62  0.07  0.07
S Renaudie 0.73  0.15  0.00
```

⇒ Seuls les éléments bien projetés peuvent être interprétés

- Contribution d'1 **var.** et d'1 **individu** à la construction de l'axe s :

$$Ctr_s(k) = \frac{r(x.k, v_s)^2}{\sum_{k=1}^K r(x.k, v_s)^2} (\times 100)$$

$$Ctr_s(i) = \frac{F_{is}^2}{\sum_{i=1}^I F_{is}^2} (\times 100)$$

```
round(res.pca$var$contrib,2)
      Dim.1 Dim.2 Dim.3
0.fruity   0.67  0.34 12.05
0.passion  6.84  0.40  0.30
```

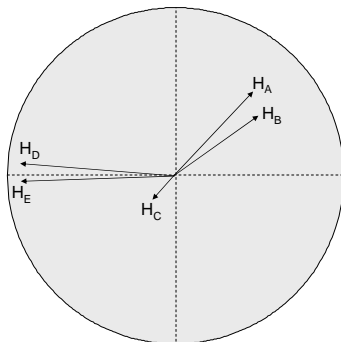
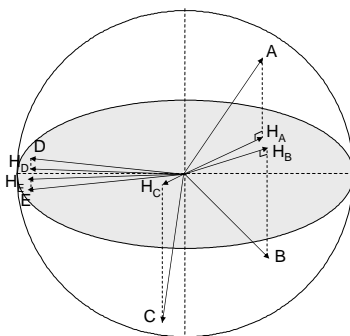
```
round(res.pca$ind$contrib,2)
      Dim.1 Dim.2 Dim.3
S Michaud  15.49  3.10  7.37
S Renaudie 15.56  5.56  0.26
```

⇒ Éléments avec une forte coordonnée contribuent le plus

## Projections...

$$r(A, B) = \cos(\theta_{A,B})$$

$\cos(\theta_{A,B}) \approx \cos(\theta_{H_A, H_B})$  si les variables sont bien projetées



Seules les variables bien projetées peuvent être interprétées !

## Description des dimensions

Par des variables quantitatives :

- calcul du coefficient de corrélation entre chaque variable et les coordonnées des individus sur un axe
  - tri des coefficients de corrélation
  - les coefficients de corrélation significativement différents de 0 sont fournis

```
> dimdesc(res.pca)
```

	\$Dim.1\$quanti			\$Dim.2\$quanti	
	corr	p.value		corr	p.value
0.candied.fruit	0.93	9.5e-05	Odor.Intensity.before.shaking	0.97	3.1e-06
Grade	0.93	1.2e-04	Odor.Intensity.after.shaking	0.95	3.6e-05
Surface.feeling	0.89	5.5e-04	Attack.intensity	0.85	1.7e-03
Typicity	0.86	1.4e-03	Expression	0.84	2.2e-03
0.mushroom	0.84	2.3e-03	Aroma.persistency	0.75	1.3e-02
...	...	...	Bitterness	0.71	2.3e-02
0.plante	-0.87	1.0e-03			
0.flower	-0.89	4.9e-04			
0.passion	-0.90	4.5e-04			
Freshness	-0.91	2.9e-04	Sweetness	-0.78	8.0e-03

## Description des dimensions

Par des variables qualitatives :

- réalisation d'une analyse de variance avec les coordonnées des individus en fonction de la variable qualitative
  - un F-test par variable
  - un test  $t$  de Student par modalité pour comparer la moyenne de la modalité à la moyenne générale

```
> dimdesc(res.pca)
```

```
Dim.1$quali
```

	R2	p.value
Label	0.874	7.30e-05

```
Dim.1$category
```

	Estimate	p.value
Vouvray	3.203	7.30e-05
Sauvignon	-3.203	7.30e-05



## Pratique de l'ACP

1. Choisir les variables actives
2. Choisir une transformation des variables (ou non)
3. Choisir de réduire ou non les variables
4. Réaliser l'ACP
5. Choisir le nombre de dimensions à interpréter
6. Interpréter simultanément le graphe des individus et celui des variables
7. Utiliser les indicateurs pour enrichir l'interprétation
8. Revenir aux données brutes pour interpréter

# Plan

Introduction

Analyse en Composantes Principales



Analyse des correspondances

Analyse des correspondances multiples

Analyse Factorielle Multiple

Classification

Conclusion

# FACTOMINER<sup>R</sup> en quelques mots

## Le package

- propose des méthodes d'analyses factorielles et de classification
- de nombreux indicateurs (qualité de représentation, contribution, description automatique des axes, ...)
- possibilité d'ajouter des éléments supplémentaires
- interface graphique (en français et en anglais)
- gestion des données manquantes (package missMDA)
- module graphique (package Factoshiny)
- rapport automatisé (package FactoInvestigate)
- aides à l'utilisateur (site internet, vidéos, livres, MOOC)

# FACTOMINER<sup>®</sup> en quelques mots

Différentes méthodes pour différents formats de données :

Données	Méthodes	Fonction
Variables quantitatives	An. en composantes principales	PCA
Table de contingence	An. des correspondances	CA
Variables qualitatives	An. des correspondances multiples	MCA
Données mixtes	An. factorielle de données mixtes	FAMD
Groupes de variables	An. factorielle multiple	MFA
Hierarchie sur les variables	An. factorielle multiple hiérarchique	HMFA
Groupes d'individus	An. factorielle multiple duale	DMFA
Tableau de contingence et variables contextuelles	An. des correspondances généralisée sur tableaux lexicaux agrégés	CaGalt

Méthodes de classification et méthodes outils complémentaires :

Méthodes	Fonction
Classification ascendante hiérarchique	HCPC
Description d'une variable qualitative (ex. var. de classe)	catdes
Description d'une variable quantitative (ex. d'une dimension)	condes, dimdesc

## Menu déroulant – Interface graphique – Package complémentaire

- `RcmdrPlugin.FactoMineR` : menu déroulant
- `Factoshiny` : graphes interactifs

⇒ faciliter l'utilisation des packages pour les utilisateurs

- `FactoInvestigate` : rapport automatisé

⇒ propose une interprétation des résultats

- `missMDA` : gestion des données manquantes

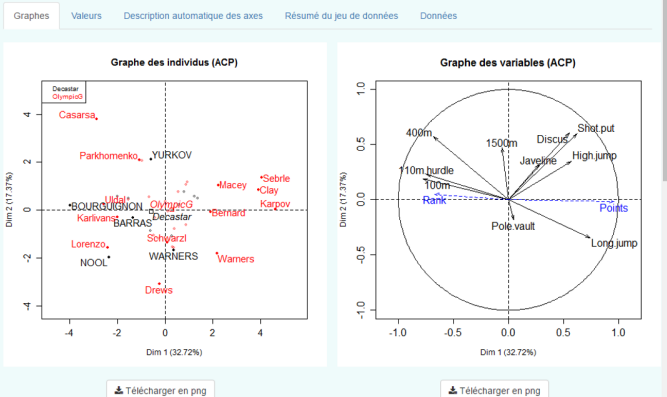
⇒ aller plus loin que les méthodes standards du package

# Graphiques interactifs avec le package Factoshiny

- Réaliser des analyses sans besoin de maîtriser le code
- Visualisation en temps réel des modifications apportées

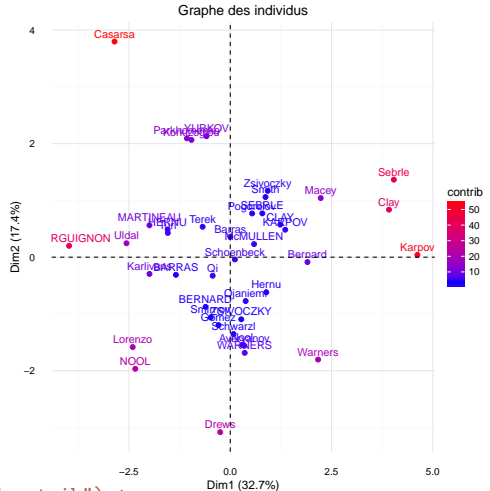
```
> res <- PCAshiny(decathlon) ## analyse factorielle sur les données
> res <- PCAshiny(res.pca)   ## graphe sur un objet résultat de FactoMineR
> res2 <- PCAshiny(res)      ## objet résultat de Factoshiny
```

## ACP sur le jeu de données decathlon



## De nouveaux packages graphiques

- le package **explor**
  - graphes interactifs
  - possibilité de bouger les libellés
- le package **factoextra**
  - basé sur **ggplot2**
  - construction séquentielle des graphes en ajoutant des couches



```
> library(factoextra)
> fviz_pca_ind(res.pca, col.ind="contrib") +
  labs(title="Graphe des individus") +
  scale_color_gradient2(mid="blue",high="red") +
  theme_minimal()
```

# Rapport automatisé avec le package FactoInvestigate

Propose une interprétation des résultats basée sur l'objet résultat

```
> res.pca <- PCA(MesDonnees, ...)
> library(FactoInvestigate)
> Investigate(res.pca)
```

<http://factominer.free.fr/reporting>

## Analyse en Composantes Principales

### Jeu de données decathlon

Ce jeu de données contient 41 individus et 13 variables, 2 variables quantitatives sont illustratives, 1 variable qualitative est illustrative.

#### 1. Observation d'individus extrêmes

L'analyse des graphes ne révèle aucun individu singulier.

#### 2. Distribution de l'inertie

L'inertie des deux facteurs indique d'une part si les variables sont structurées et suggère d'autre part le nombre judicieux de composantes principales à étudier.

Les 2 premiers axes de l'ACP expliquent **80.09%** de l'inertie totale du jeu de données ; cela signifie que 80.09% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. C'est un pourcentage assez important, et le premier plan représente donc convenablement la variabilité contenue dans une grande part du jeu de données ACP. Cette valeur est supérieure à la valeur référentielle de **37.75%** (la variabilité expliquée par ce plan est donc significative (cette inertie de référence est le quartile 0.95 de la distribution des pourcentages d'inertie obtenus en simulant 1000 jeux de données aléatoires de dimensions comparables à la base d'une distribution normale).

Du fait de ces observations, il serait tout de même probablement préférable de considérer également dans l'analyse les dimensions supérieures ou égales à la troisième.

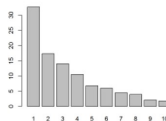
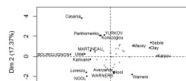


Figure 2 - Décomposition de la totale inertie on the components of the ACP

Une estimation du nombre pertinent d'axes à interpréter suggère de restreindre l'analyse à la description des 3 premiers axes. Ces composantes révèlent un taux d'inertie supérieur à celle du quartile 0.95 de distributions aléatoires (84.14% contre 51.44%). Cette observation suggère que seuls ces axes sont porteurs d'une véritable information. En conséquence, la description de l'analyse sera restreinte à ces seuls axes.

#### 3. Description du plan 1:2





# missMDA : package de gestion des données manquantes

```
> library(missMDA)
> data(orange)
```

	Color intensity	Odor intensity	Attack intensity	Sweet	Acid	Bitter	Pulp	Typicity
1	4.79	5.29	NA	NA	NA	2.83	NA	5.21
2	4.58	6.04	4.42	5.46	4.13	3.54	4.62	4.46
3	4.71	5.33	NA	NA	4.29	3.17	6.25	5.17
4	6.58	6.00	7.42	4.17	6.75	NA	1.42	3.42
5	NA	6.17	5.33	4.08	NA	4.38	3.42	4.42
6	6.33	5.00	5.38	5.00	5.50	3.63	4.21	4.88
7	4.29	4.92	5.29	5.54	5.25	NA	1.29	4.33
8	NA	4.54	4.83	NA	4.96	2.92	1.54	3.96
9	4.42	NA	5.17	4.62	5.04	3.67	1.54	3.96
10	4.54	4.29	NA	5.79	4.38	NA	NA	5.00
11	4.08	5.13	3.92	NA	NA	NA	7.33	5.25
12	6.50	5.88	6.13	4.88	5.29	4.17	1.50	3.50

## Mise en œuvre logiciel avec missMDA

```
> library(missMDA)
> data(orange)
> nb <- estim_ncpPCA(orange, scale=TRUE)      ## Estime le nb de dimensions
> comp <- imputePCA(orange, ncp=2, scale=TRUE) ## Complète le tableau
> res.pca <- PCA(comp$completeObs)           ## Effectue l'ACP
```

## Mise en œuvre logiciel avec missMDA

```
> library(missMDA)
> data(orange)
> nb <- estim_ncpPCA(orange, scale=TRUE)      ## Estime le nb de dimensions
> comp <- imputePCA(orange, ncp=2, scale=TRUE) ## Complète le tableau
> res.pca <- PCA(comp$completeObs)           ## Effectue l'ACP

> orange
  Sweet Acid Bitter Pulp Typicity
   NA   NA  2.83   NA   5.21
5.46 4.13  3.54 4.62   4.46
   NA 4.29  3.17 6.25   5.17
...
4.88 5.29  4.17 1.50   3.50
```

## Mise en œuvre logiciel avec missMDA

```

> library(missMDA)
> data(orange)
> nb <- estim_ncpPCA(orange, scale=TRUE)      ## Estime le nb de dimensions
> comp <- imputePCA(orange, ncp=2, scale=TRUE) ## Complète le tableau
> res.pca <- PCA(comp$completeObs)           ## Effectue l'ACP

```

> orange					> comp\$completeObs				
Sweet	Acid	Bitter	Pulp	Typicity	Sweet	Acid	Bitter	Pulp	Typicity
NA	NA	2.83	NA	5.21	5.54	4.13	2.83	5.89	5.21
5.46	4.13	3.54	4.62	4.46	5.46	4.13	3.54	4.62	4.46
NA	4.29	3.17	6.25	5.17	5.45	4.29	3.17	6.25	5.17
...					...				
4.88	5.29	4.17	1.50	3.50	4.88	5.29	4.17	1.50	3.50

# Mise en œuvre logiciel avec missMDA

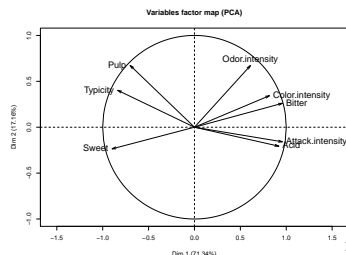
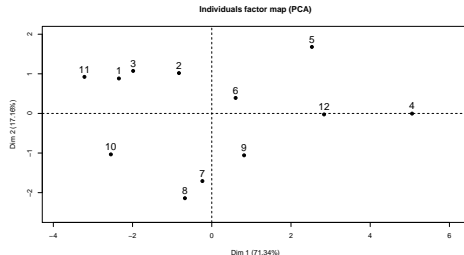
```
> library(missMDA)
> data(orange)
> nb <- estim_ncpPCA(orange, scale=TRUE) ## Estime le nb de dimensions
> comp <- imputePCA(orange, ncp=2, scale=TRUE) ## Complète le tableau
> res.pca <- PCA(comp$completeObs) ## Effectue l'ACP
```

```
> orange
```

Sweet	Acid	Bitter	Pulp	Typicity
NA	NA	2.83	NA	5.21
5.46	4.13	3.54	4.62	4.46
NA	4.29	3.17	6.25	5.17
...				
4.88	5.29	4.17	1.50	3.50

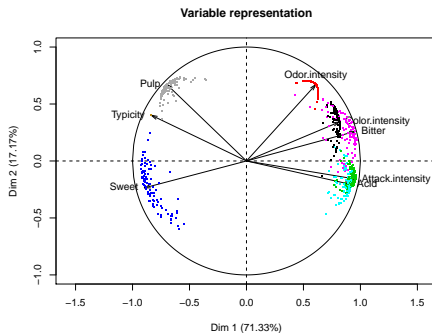
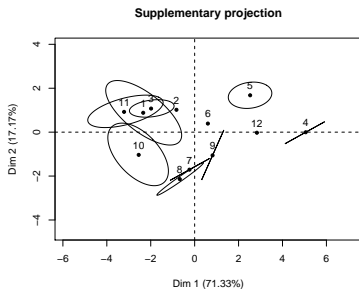
```
> comp$completeObs
```

Sweet	Acid	Bitter	Pulp	Typicity
5.54	4.13	2.83	5.89	5.21
5.46	4.13	3.54	4.62	4.46
5.45	4.29	3.17	6.25	5.17
...				
4.88	5.29	4.17	1.50	3.50



# Visualisation de l'incertitude liée aux données manquantes

```
> library(missMDA)
> mi <- MIPCA(orange, scale = TRUE, ncp=2)
> mi$res.MI          ## sortie pour les tableaux imputés
> plot(mi)
```



Permet de ne pas analyser de tableaux avec trop de données manquantes

## Aides à l'utilisateur : site internet

- <http://factominer.free.fr>
- en anglais et en français
- exemples, aides sur les fonctions, références, etc.

FactoMineR : analyse de

← → ↻ ① factominer.free.fr/index\_fr.html ☆ ⓘ ⋮

Accueil Méthodes FactoMineR Enseignement MOOC, livres Améliorations graphiques Valeurs manquantes missMDA Rapport automatique Google group Plus

# FACTOMINER

FactoMineR en quelques mots

FactoMineR est un package R dédié à l'analyse exploratoire multidimensionnelle de données (à la Française). Il a été développé et il est maintenu par François Husson, Julie Josse, Sébastien Lê, d'Agrocampus Rennes, et J. Mazet.

Pourquoi utiliser FactoMineR?

1. Il permet de mettre en oeuvre des méthodes analyses de données telles que l'analyse en composantes principales (ACP), l'analyse des correspondances (AC), l'analyse des correspondances multiples (ACM) ainsi que des analyses plus avancées.
2. Il permet l'ajout d'information supplémentaire telle que des individus et/ou des variables supplémentaires.
3. Il fournit un point de vue géométrique et de nombreuses sorties graphiques.
4. Il fournit de nombreuses aides à l'interprétation (description automatique des axes, nombreux indicateurs, ...).
5. Il peut prendre en compte diverses structures sur les données (structure sur les variables, hiérarchie sur les variables, structure sur les individus).
6. Beaucoup de matériel pédagogique (MOOC, livres, etc.) est disponible pour

Menu de l'accueil

- Présentation de FactoMineR
- Nouvelles
- Installation de FactoMineR
- Comment citer FactoMineR?
- Historique de FactoMineR

Liens utiles

- Département de Mathématiques d'Agrocampus Rennes
- R Project

## Aides à l'utilisateur : un Google group

- <https://groups.google.com/group/factominer-users/>
- possibilité de poser des questions et/ou répondre
- en français ou en anglais

The screenshot shows a web browser window displaying the Google Groups page for 'FactoMineR users'. The URL in the address bar is <https://groups.google.com/forum/?hl=fr#forum/factominer-users>. The page header includes the Google logo, a search bar, and navigation links. Below the header, the group name 'FactoMineR users' is displayed, along with the number of subjects (32 sur 405) and a 'G+' icon. A table of recent posts is shown, with columns for subject, number of messages, and date.

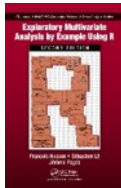
Subject	Messages	Date
La 2ème session du MOOC "Analyse de données multidimensionnelles" débute le 1er mars	1 message	19 janv.
Nouveau module graphique	11	17/01/2015
Select the best sample using a reference	1	16 mai
salut	1	15 mai
PCAshiny scale.unit=F impossible ?	2	12 mai
ACM et questions à choix multiples (plusieurs modalités dans la même question)	6	2 mai
Plot CA neatly (1)	1	30 avr.
Interpreting MCA results	11	22 avr.
Installing FactoMineR on Linux (1)	1	21 avr.
Estimation of PC for MFA	4	21 avr.
General questions for FAMD	1	1 avr.



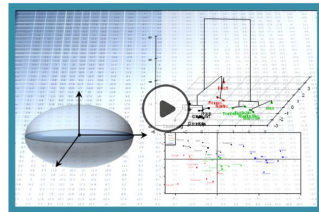
# Aides à l'utilisateur : diffusion scientifique

*Analyse de données avec R (2<sup>e</sup> ed)*

*R pour la stat. et sc. des données*



MOOC analyse de données multidimensionnelles



- 2 articles dans R journal ([CA-galt](#), [MFACT](#))
- 2 articles dans J. of stat. software ([FactoMineR](#), [missMDA](#))

# Plan

Introduction

Analyse en Composantes Principales



Analyse des correspondances

Analyse des correspondances multiples

Analyse Factorielle Multiple

Classification

Conclusion

## Tableau de correspondances

		Ensemble $J$		
		1	$j$	$J$
Ensemble $I$	1			
	$i$		$x_{ij}$	
	$I$			

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

Personnages de Mots

Phèdre (Racine)

Milieus

Espèces

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

Abondance de l'espèce  $j$  dans le  
milieu  $i$

Parfums

Descripteur

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

⇒ Exemples où le test d'indépendance du  $\chi^2$  peut être appliqué

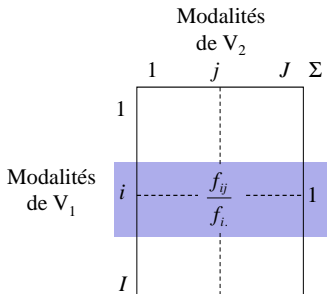
## Données sur les prix Nobel

	Chimie	Economie	Littérature	Médecine	Paix	Physique	Somme
Allemagne	24	1	8	18	5	24	80
Canada	4	3	2	4	1	4	18
France	8	3	11	12	10	9	53
GB	23	6	7	26	11	20	93
Italie	1	1	6	5	1	5	19
Japon	6	0	2	3	1	11	23
Russie	4	3	5	2	3	10	27
USA	51	43	8	70	19	66	257
Somme	121	60	49	140	51	149	570

Y a-t'il un lien entre les pays et les catégories de prix ? Certains pays ont-ils des spécificités ?

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

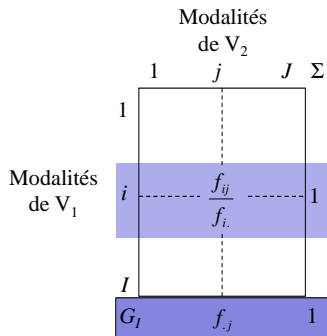
Analyse par lignes :  $\frac{f_{ij}}{f_{i.}} = f_{.j}$



Profil ligne  $i$  = distribution conditionnelle de  $V_2$  sachant que l'on possède la modalité  $i$  de  $V_1$

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

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

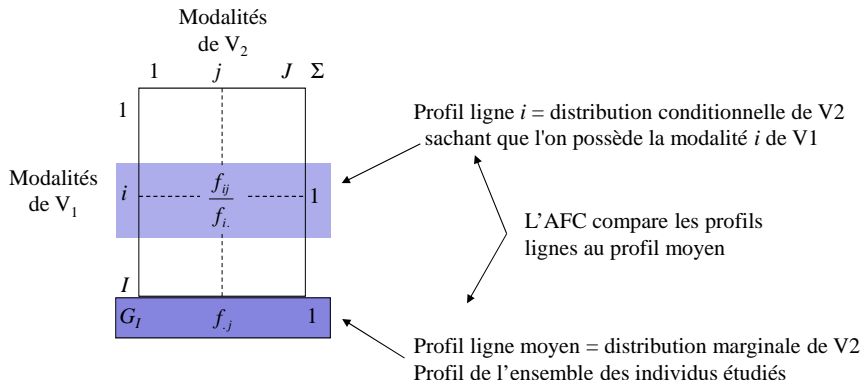


Profil ligne  $i$  = distribution conditionnelle de  $V_2$  sachant que l'on possède la modalité  $i$  de  $V_1$

Profil ligne moyen = distribution marginale de  $V_2$   
Profil de l'ensemble des individus étudiés

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

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



Approche multidimensionnelle de l'écart à l'indépendance

## Comparaison du profil ligne au profil moyen

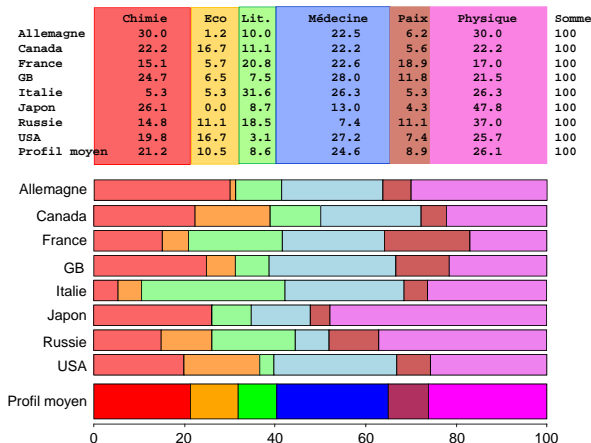
	Chimie	Eco	Lit.	Médecine	Paix	Physique	Somme
Allemagne	30.0	1.2	10.0	22.5	6.2	30.0	100
Canada	22.2	16.7	11.1	22.2	5.6	22.2	100
France	15.1	5.7	20.8	22.6	18.9	17.0	100
GB	24.7	6.5	7.5	28.0	11.8	21.5	100
Italie	5.3	5.3	31.6	26.3	5.3	26.3	100
Japon	26.1	0.0	8.7	13.0	4.3	47.8	100
Russie	14.8	11.1	18.5	7.4	11.1	37.0	100
USA	19.8	16.7	3.1	27.2	7.4	25.7	100
Profil moyen	21.2	10.5	8.6	24.6	8.9	26.1	100



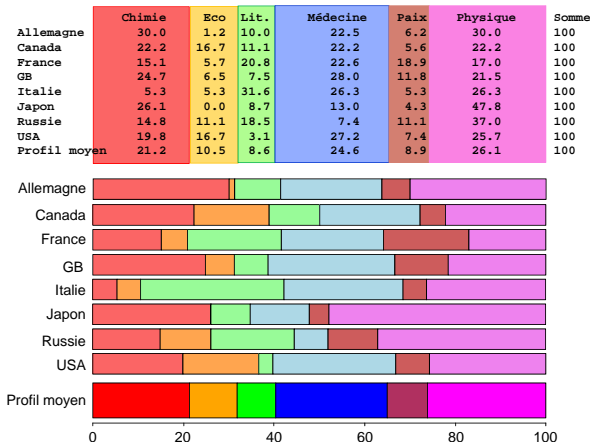
## Comparaison du profil ligne au profil moyen

	Chimie	Eco	Lit.	Médecine	Paix	Physique	Somme
Allemagne	30.0	1.2	10.0	22.5	6.2	30.0	100
Canada	22.2	16.7	11.1	22.2	5.6	22.2	100
France	15.1	5.7	20.8	22.6	18.9	17.0	100
GB	24.7	6.5	7.5	28.0	11.8	21.5	100
Italie	5.3	5.3	31.6	26.3	5.3	26.3	100
Japon	26.1	0.0	8.7	13.0	4.3	47.8	100
Russie	14.8	11.1	18.5	7.4	11.1	37.0	100
USA	19.8	16.7	3.1	27.2	7.4	25.7	100
Profil moyen	21.2	10.5	8.6	24.6	8.9	26.1	100

# Comparaison du profil ligne au profil moyen



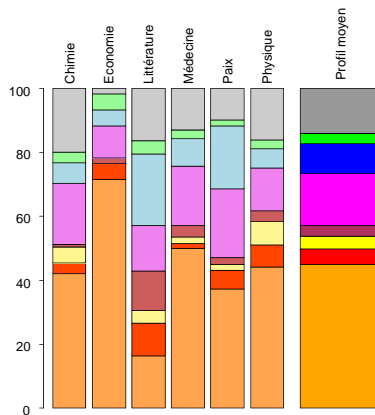
## Comparaison du profil ligne au profil moyen



Les Italiens obtiennent-ils des prix Nobel dans des disciplines particulières ?

## Comparaison du profil colonne au profil moyen

	Chimie	Eco	Lit	Méd	Paix	Phys	Profil moyen
Allemagne	19.8	1.7	16.3	12.9	9.8	16.1	14.0
Canada	3.3	5.0	4.1	2.9	2.0	2.7	3.2
France	6.6	5.0	22.4	8.6	19.6	6.0	9.3
GB	19.0	10.0	14.3	18.6	21.6	13.4	16.3
Italie	0.8	1.7	12.2	3.6	2.0	3.4	3.3
Japon	5.0	0.0	4.1	2.1	2.0	7.4	4.0
Russie	3.3	5.0	10.2	1.4	5.9	6.7	4.7
USA	42.1	71.7	16.3	50.0	37.3	44.3	45.1
Somme	100	100	100	100	100	100	100



La répartition par pays des prix Nobel en littérature est elle la même que la répartition de l'ensemble des prix Nobel ?

## Représentation simultanée des lignes et colonnes

Relation de transition = propriétés barycentriques

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^J \underbrace{\frac{f_{ij}}{f_{i.}} G_s(j)}_{\text{barycentre}} \quad \begin{array}{l} F_s(i) : \text{coord. de la ligne } i \text{ sur l'axe de rang } s \\ \frac{f_{ij}}{f_{i.}} : \text{jème élément du profil } i \\ G_s(j) : \text{coord. de la colonne } j \text{ sur l'axe de rang } s \\ \lambda_s : \text{inertie associée à l'axe } s \text{ (en AFC } \lambda_s \leq 1) \end{array}$$

Le long de l'axe de rang  $s$ , on calcule le barycentre de toutes les colonnes, chaque colonne  $j$  étant affectée du poids  $f_{ij}/f_{i.}$

## Représentation simultanée des lignes et colonnes

Relation de transition = propriétés barycentriques

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^J \underbrace{\frac{f_{ij}}{f_{i.}} G_s(j)}_{\text{barycentre}} \quad \begin{array}{l} F_s(i) : \text{coord. de la ligne } i \text{ sur l'axe de rang } s \\ \frac{f_{ij}}{f_{i.}} : \text{jème élément du profil } i \\ G_s(j) : \text{coord. de la colonne } j \text{ sur l'axe de rang } s \\ \lambda_s : \text{inertie associée à l'axe } s \text{ (en AFC } \lambda_s \leq 1) \end{array}$$

Le long de l'axe de rang  $s$ , on calcule le barycentre de toutes les colonnes, chaque colonne  $j$  étant affectée du poids  $f_{ij}/f_{i.}$

Le barycentre est ensuite d'autant plus écarté de l'origine que  $\lambda_s$  est petit :  $1/\sqrt{\lambda_s} \geq 1$

## Représentation simultanée des lignes et colonnes

Relation de transition = propriétés barycentriques

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^J \underbrace{\frac{f_{ij}}{f_{i.}} G_s(j)}_{\text{barycentre}} \quad \begin{array}{l} F_s(i) : \text{coord. de la ligne } i \text{ sur l'axe de rang } s \\ \frac{f_{ij}}{f_{i.}} : \text{jème élément du profil } i \\ G_s(j) : \text{coord. de la colonne } j \text{ sur l'axe de rang } s \\ \lambda_s : \text{inertie associée à l'axe } s \text{ (en AFC } \lambda_s \leq 1) \end{array}$$

Le long de l'axe de rang  $s$ , on calcule le barycentre de toutes les colonnes, chaque colonne  $j$  étant affectée du poids  $f_{ij}/f_{i.}$

Le barycentre est ensuite d'autant plus écarté de l'origine que  $\lambda_s$  est petit :  $1/\sqrt{\lambda_s} \geq 1$

Ligne  $i$  du côté des colonnes avec lesquelles elle s'associe le plus (et à l'opposé des colonnes avec lesquelles elle s'associe le moins)

## Représentation simultanée des lignes et colonnes

Relation de transition = propriétés barycentriques

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^J \underbrace{\frac{f_{ij}}{f_{i.}} G_s(j)}_{\text{coord. de la ligne } i \text{ sur l'axe de rang } s}$$

$F_s(i)$  : coord. de la ligne  $i$  sur l'axe de rang  $s$   
 $\frac{f_{ij}}{f_{i.}}$  : jème élément du profil  $i$   
 $G_s(j)$  : coord. de la colonne  $j$  sur l'axe de rang  $s$   
 $\lambda_s$  : inertie associée à l'axe  $s$  (en AFC  $\lambda_s \leq 1$ )

Le long de l'axe de rang  $s$ , on calcule le barycentre de toutes les colonnes, chaque colonne  $j$  étant affectée du poids  $f_{ij}/f_{i.}$

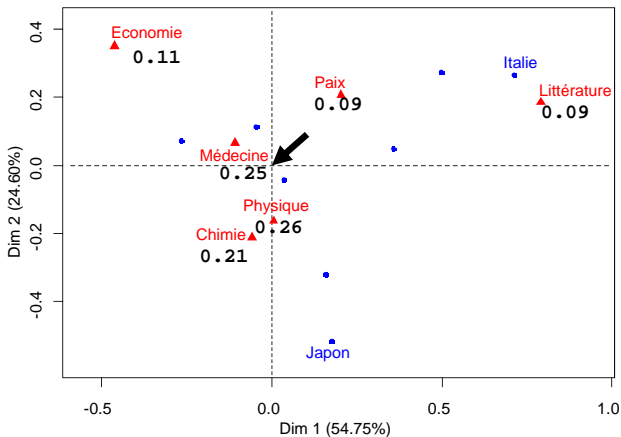
Le barycentre est ensuite d'autant plus écarté de l'origine que  $\lambda_s$  est petit :  $1/\sqrt{\lambda_s} \geq 1$

Ligne  $i$  du côté des colonnes avec lesquelles elle s'associe le plus (et à l'opposé des colonnes avec lesquelles elle s'associe le moins)

Et symétriquement :  $G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^I \frac{f_{ij}}{f_{.j}} F_s(i)$

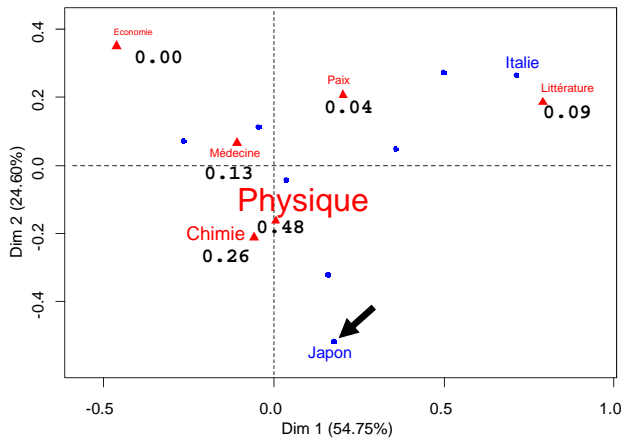


## Propriété barycentrique



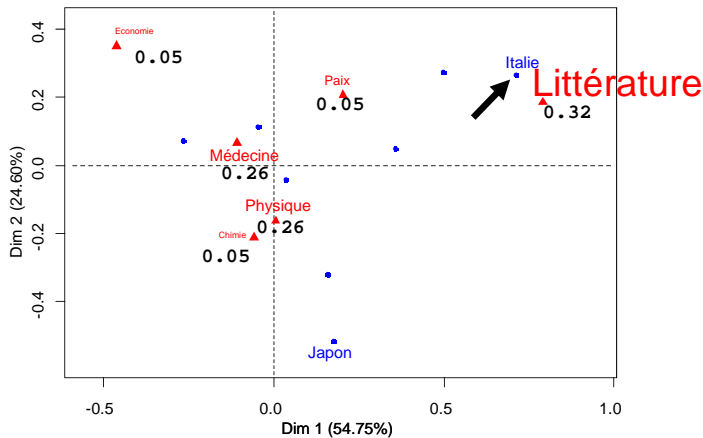
	Chimie	Economie	Littérature	Médecine	Paix	Physique
Italie	5.26	5.26	31.58	26.32	5.26	26.32
Japon	26.09	0.00	8.70	13.04	4.35	47.83
Profil moyen	21.23	10.53	8.60	24.56	8.95	26.14

# Propriété barycentrique



	Chimie	Economie	Littérature	Médecine	Paix	Physique
Italie	5.26	5.26	31.58	26.32	5.26	26.32
Japon	26.09	0.00	8.70	13.04	4.35	47.83
Profil moyen	21.23	10.53	8.60	24.56	8.95	26.14

## Propriété barycentrique

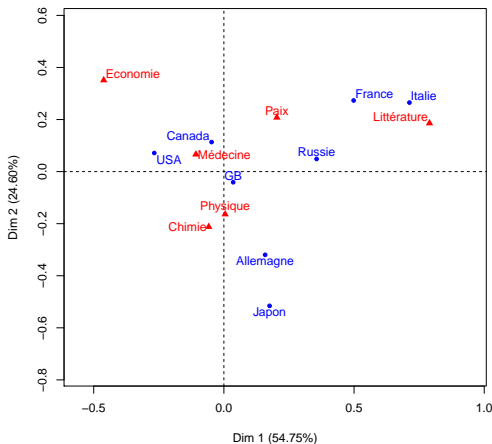


	Chimie	Economie	Littérature	Médecine	Paix	Physique
Italie	5.26	5.26	31.58	26.32	5.26	26.32
Japon	26.09	0.00	8.70	13.04	4.35	47.83
Profil moyen	21.23	10.53	8.60	24.56	8.95	26.14

## Représentation superposée

- Le barycentre représente l'indépendance
- La distance entre niveaux d'une même variable peut être interprétée
- La représentation est pseudo-barycentrique (dilatation) : formule de transition
- Il n'est pas possible d'interpréter la distance entre les modalités de deux variables mais ...
- ... c'est un barycentre pondéré de toutes les modalités  $\Rightarrow$  la direction est interprétable

## Interprétation sur l'exemple



- opposition sciences - autres dans une moindre mesure, opposition physique/chimie - science économique
- positions des pays illustrent leur spécificité dans l'obtention des prix Nobel

AFC donne une visualisation synthétique de l'écart à l'indépendance qui aide la compréhension du tableau (a fortiori avec de grands tableaux)

# Plan

Introduction

Analyse en Composantes Principales



Analyse des correspondances

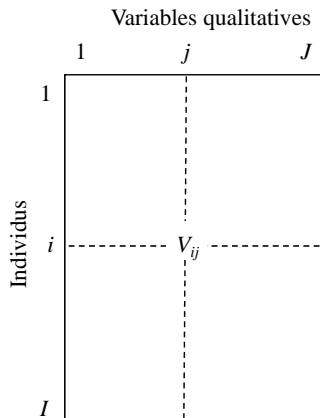
Analyse des correspondances multiples

Analyse Factorielle Multiple

Classification

Conclusion

## Les données

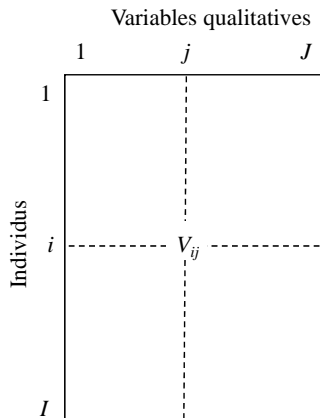


$I$  individus

$J$  variables qualitatives

$v_{ij}$  : modalité de la variable  $j$   
possédée par l'individu  $i$

## Les données



$I$  individus

$J$  variables qualitatives

$v_{ij}$  : modalité de la variable  $j$   
possédée par l'individu  $i$

Exemple : enquête où  $I$  personnes  
sont interrogées sur  $J$  questions à  
choix multiples



# Codage des données

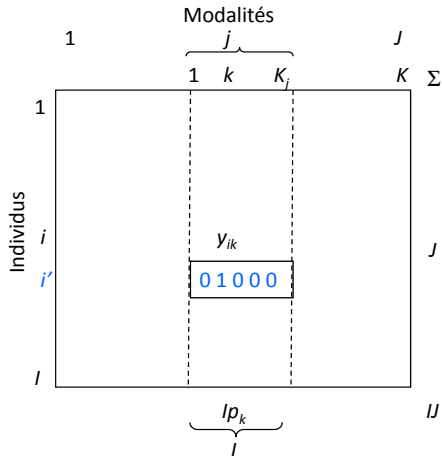
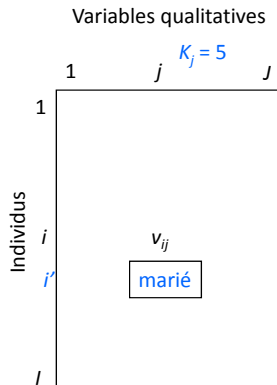


Tableau disjonctif complet (TDC)

# Objectifs – problématique

## 1. Etude des individus

Un individu = une ligne du TDC = ensemble de ses modalités

Ressemblance des individus    Variabilité des individus

Principales dimensions de la variabilité des individus

(en relation avec les modalités)

## Objectifs – problématique

### 1. Etude des individus

Un individu = une ligne du TDC = ensemble de ses modalités  
Ressemblance des individus    Variabilité des individus  
Principales dimensions de la variabilité des individus  
(en relation avec les modalités)

### 2. Etude des variables

Liaisons entre variables qualitatives  
(en relation avec les modalités)  
Visualisation d'ensemble des associations entre modalités  
Variable synthétique  
(Indicateur quantitatif fondé sur des variables qualitatives)

## Objectifs – problématique

### 1. Etude des individus

Un individu = une ligne du TDC = ensemble de ses modalités

Ressemblance des individus    Variabilité des individus

Principales dimensions de la variabilité des individus  
(en relation avec les modalités)

### 2. Etude des variables

Liaisons entre variables qualitatives  
(en relation avec les modalités)

Visualisation d'ensemble des associations entre modalités

Variable synthétique

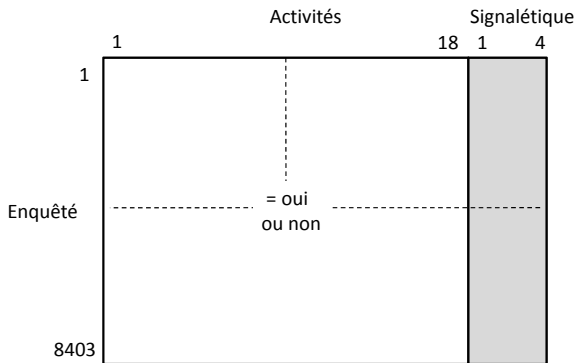
(Indicateur quantitatif fondé sur des variables qualitatives)

⇒ Problématique voisine de celle de l'ACP

## Les données loisirs

- Extrait d'une enquête de l'Insee de 2003 sur la construction des identités, appelée « Histoire de vie »
- 8403 individus
- 2 sortes de variables :
  - *Parmi les loisirs suivants, indiquez ceux que vous pratiquez régulièrement* : Lecture, Ecouter de la musique, Cinéma, Spectacle, Exposition, Ordinateur, Sport, Marche, Voyage, Jouer de la musique, Collection, Activité bénévole, Bricolage, Jardinage, Tricot, Cuisine, Pêche, nombre d'heures moyen par jour à regarder la TV
  - le signalétique (4 questions) : sexe, âge, profession, statut matrimonial

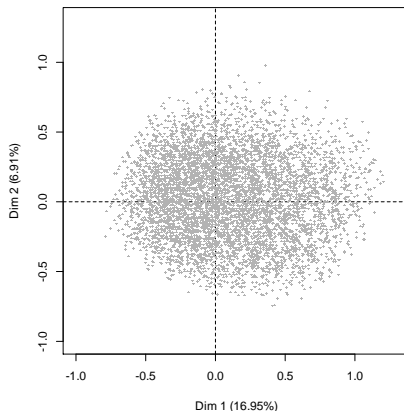
## Les données loisirs



ACM : loisirs en actif, signalétique en supplémentaire

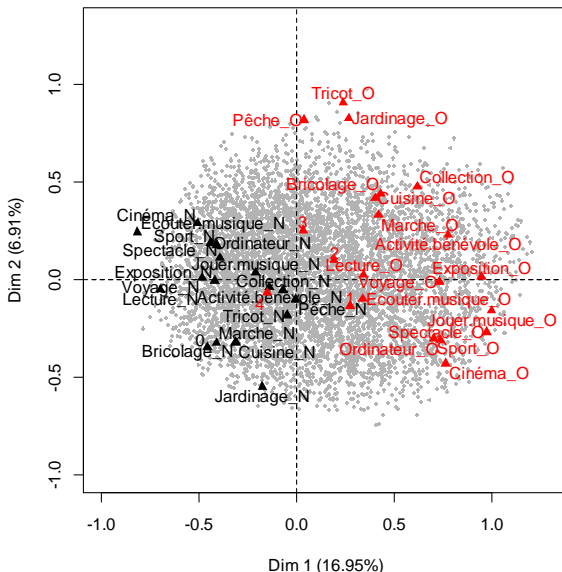
- 1 individu = profil d'activités
- Principales dimensions de variabilité des profils d'activités
- Liaisons entre ces dimensions et le signalétique

## Représentation du nuage des individus



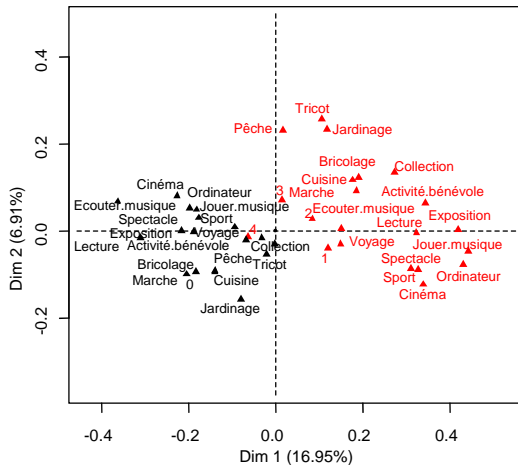
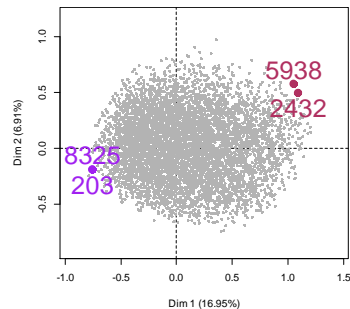
- 2 individus sont superposés s'ils prennent les mêmes modalités
- 2 individus ont en commun beaucoup de modalités : distance petite
- 2 individus dont l'un des 2 possède une modalité rare : distance grande pour prendre en compte la spécificité d'un des 2
- 2 individus ont en commun une modalité rare : distance petite pour prendre en compte leur spécificité commune

# Représentations barycentriques – représentation simultanée



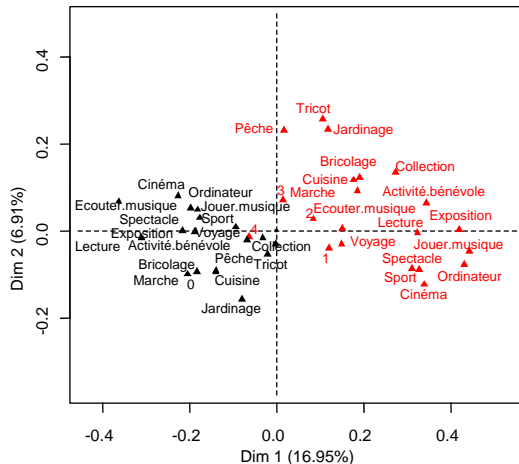
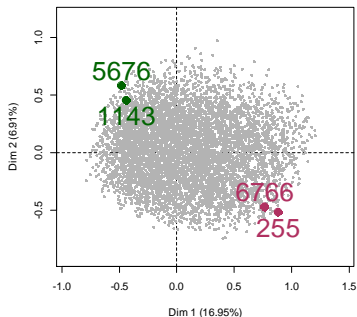


# Représentation des modalités dans le nuage des individus



	Ecouter				Jouer				Activité									
	Lecture	musique	Ciné	Spectacle	Expo	Ordi	Sport	Marche	Voyage	musique	Collec	bénévole		Bricol	Jardin	Tricot	Cuisine	Pêche
5938	O	O	N	O	O	O	O	O	O	O	O	O	O	O	O	O	N	3
2432	O	O	O	O	O	O	N	O	O	O	O	O	O	O	O	O	N	2
8325	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	4
203	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	4

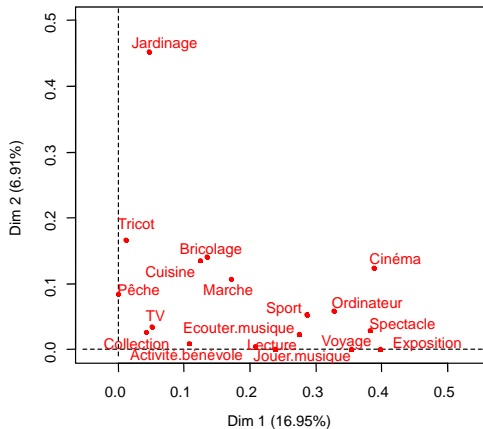
# Représentation des modalités dans le nuage des individus



	Ecouter					Jouer					Activité							
	Lecture	musique	Ciné	Spectacle	Expo	Ordi	Sport	Marche	Voyage	musique	Collec	bénévole	Bricol	Jardin	Tricot	Cuisine	Pêche	TV
255	O	O	O	O	O	O	O	O	O	O	N	O	N	N	N	N	N	1
6766	O	O	O	O	O	O	O	O	O	O	N	N	N	N	N	O	N	0
5676	N	N	N	N	N	N	N	N	N	N	N	N	O	O	O	O	N	4
1143	O	N	N	N	N	N	N	N	N	N	N	N	O	O	O	N	N	4

# Représentation des variables pour interpréter les dimensions

Utilisation des rapports  
de corrélation au carré



## Aides à l'interprétation

- Contribution et  $\cos^2$  pour les individus et les modalités  $\Rightarrow$  Modalités extrêmes ne contribuent pas nécessairement beaucoup (cela dépend de leur fréquence)  
 $\Rightarrow \cos^2$  petits ... ce qui est attendu car bcp de dimensions
- Contribution d'une variable
- Représentation de modalités supplémentaires
- Représentation de variables quantitatives supplémentaires

## Conclusion

- L'ACM est la méthode factorielle adaptée aux tableaux individus  $\times$  variables qualitatives
- Les pourcentages d'inertie et les qualités de représentation sont faibles
- Il faut souvent interpréter plus de 2 dimensions : fortes différences entre individus qui ne peuvent être expliquées par deux dimensions
- Revenir aux données en analysant des tableaux de contingence par AFC
- L'ACM comme pré-traitement d'une classification

## Gestion des données manquantes : exemple en ACM

```
> library(missMDA)
> data(vnf)
> nb <- estim_ncpMCA(vnf,ncp.max=5)      ## Estime le nb de dimensions
> imp <- imputeMCA(vnf, ncp=nb)          ## Complète le tableau disjonctif
> res <- MCA(vnf,tab.disj=imp$tab.disj)  ## ACM utilisant tab disj complété
```

	V1	V2	V3	...	V14
ind 1	a	NA	g	...	u
ind 2	NA	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	NA		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	0.71	0.29	1	0	...
ind 2	0.12	0.29	0.59	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	0.37	0.63	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

⇒ Même principe avec FAMD et MFA

# Plan

Introduction

Analyse en Composantes Principales



Analyse des correspondances

Analyse des correspondances multiples

**Analyse Factorielle Multiple**

Classification

Conclusion

## Description sensorielle de vins : comparaison de jurys

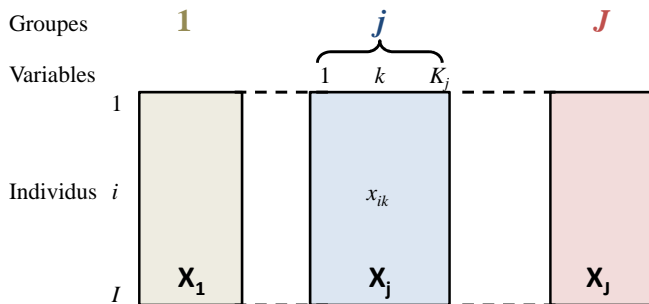
- 10 vins blancs du Val de Loire : 5 Vouvray - 5 Sauvignon
- description sensorielle de 3 jurys : œnologue, conso., étudiant

	Expert (27)	Conso (15)	Etudiant (15)
Vin 1			
Vin 2			
...			
Vin 10			

- Comment caractériser les vins ?
- Les vins sont-ils décrits de la même façon par les différents jurys ? Y-a t'il des spécificités par jury ?

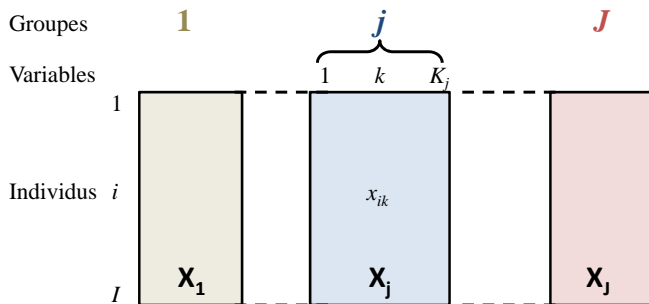


## Tableaux multiples



Exemples avec des variables **quantitatives et/ou qualitatives**  
**et/ou des tableaux de contingence** :

## Tableaux multiples



Exemples avec des variables **quantitatives et/ou qualitatives**  
**et/ou des tableaux de contingence** :

- rivières et quantités de polluants : mesures à plusieurs dates
- sols caractérisés par des mesures physiques et chimiques
- tableaux milieux - espèces mesurés sur plusieurs années
- eaux : mesures physico-chimiques, nombre d'animaux

# Objectifs

- Etudier les ressemblances entre individus du point de vue de l'ensemble des variables ET les relations entre variables

# Objectifs

- Etudier les ressemblances entre individus du point de vue de l'ensemble des variables ET les relations entre variables

## Prendre en compte la structure en groupes

- Etudier globalement les ressemblances et les différences entre groupes (voir les spécificités de chaque groupe)
- Etudier les ressemblances et les différences entre groupes du point de vue individuel
- Comparer les typologies issues des analyses séparées

# Objectifs

- Etudier les ressemblances entre individus du point de vue de l'ensemble des variables ET les relations entre variables

## Prendre en compte la structure en groupes

- Etudier globalement les ressemblances et les différences entre groupes (voir les spécificités de chaque groupe)
- Etudier les ressemblances et les différences entre groupes du point de vue individuel
- Comparer les typologies issues des analyses séparées

⇒ Equilibrer l'influence de chaque groupe dans l'analyse

## Equilibrer l'influence des groupes de variables

*"Doing a data analysis, in good mathematics, is simply searching eigenvectors, all the science of it (the art) is just to find the right matrix to diagonalize" (Benzécri)*

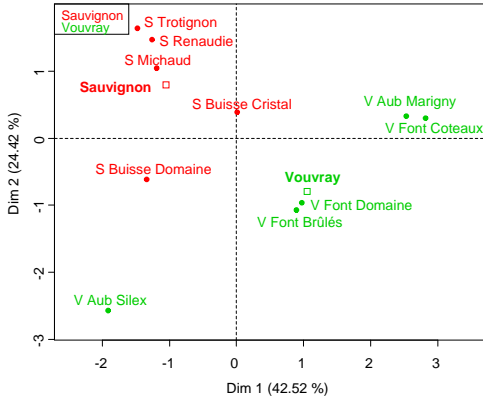
## Equilibrer l'influence des groupes de variables

*"Doing a data analysis, in good mathematics, is simply searching eigenvectors, all the science of it (the art) is just to find the right matrix to diagonalize"* (Benzécri)

L'AFM est une ACP pondérée qui équilibre la contribution des groupes tel que :

- Même poids pour toutes les variables d'un même groupe : la structure du groupe est préservée
- Pour chaque groupe, la variance de la principale dimension de variabilité (première valeur propre) est égale à 1
- Aucun groupe ne peut générer à lui seul la première dimension
- Un groupe multidimensionnel contribue à plus de dimensions qu'un groupe uni-dimensionnel

## Représentation des individus

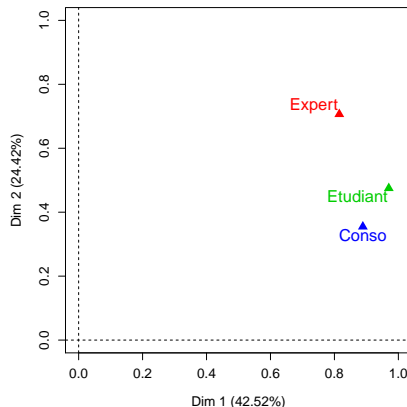


- Les deux cépages sont bien séparés
- Les Vouvray sont plus différents du point de vue sensoriel
- Plusieurs groupes de vins, ...





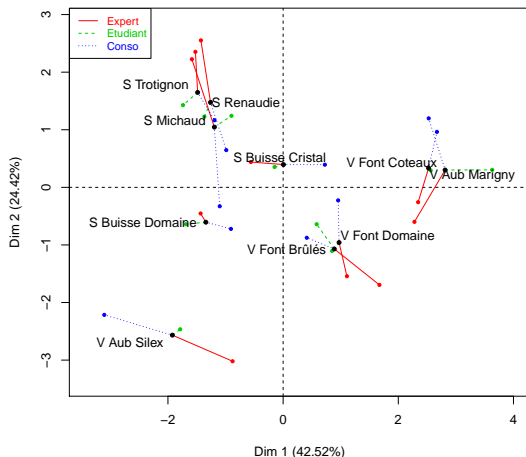
## Représentation des groupes



- 1ère dimension commune à tous les groupes
- 2ème dimension due au groupe Expert
- 2 groupes sont proches quand ils induisent la même structure

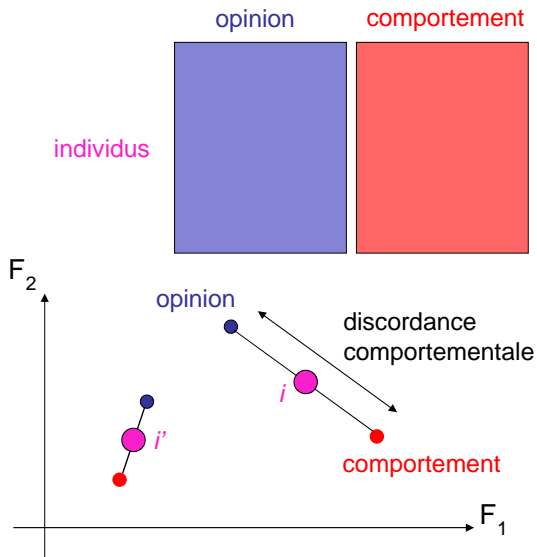
⇒ Ce graphe fournit une comparaison synthétique des groupes  
⇒ Les positions relatives des individus sont-elles similaires d'un groupe à l'autre ?

## Représentation des points partiels

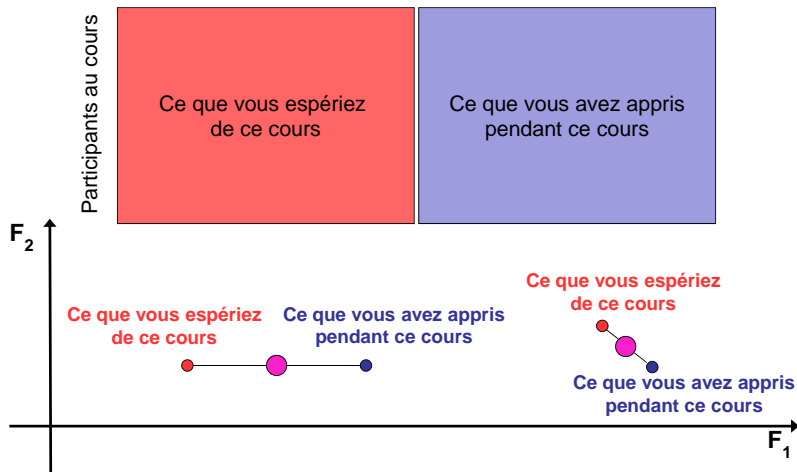


- Point partiel = représentation d'un individu vu par un groupe
- Un individu est au barycentre de ses points partiels
- Un individu est homogène si ses points partiels sont proches

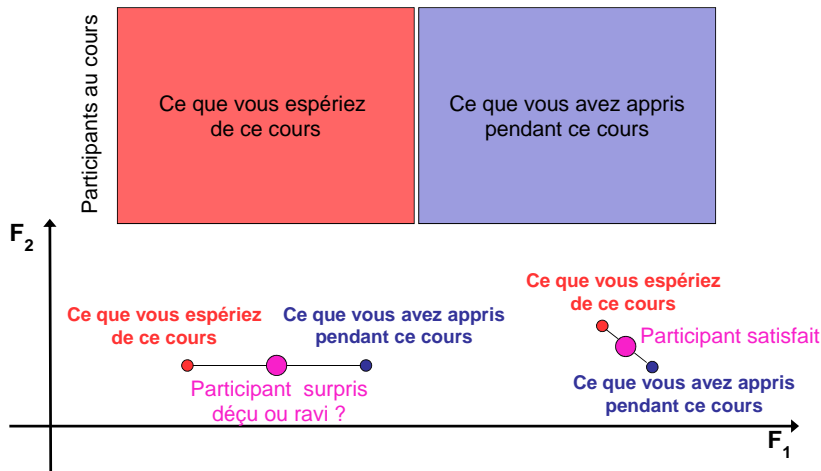
## Points partiels



## Points partiels

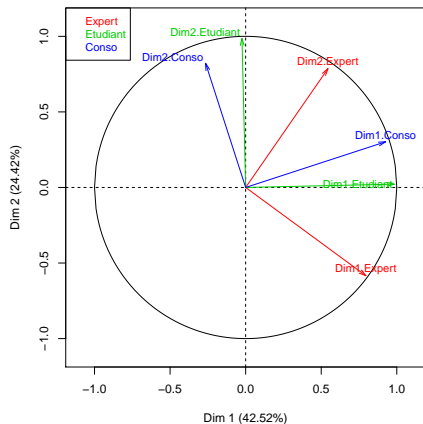


## Points partiels



## Relation avec les facteurs des analyses séparées

⇒ Les composantes principales des analyses séparées sont projetées en supplémentaires



- Les dimensions de l'ACP sur les données étudiant coïncident avec les dimensions de l'AFM
- Les deux premières dimensions de chaque groupe sont bien projetées

## Données mixtes

⇒ Groupes composés de variables quantitatives et groupes composés de variables qualitatives

L'AFM fonctionne "localement" comme :

- une ACP pour les variables quantitatives
- une ACM pour les variables qualitatives
- une AFC pour les tableaux de contingence

La pondération de l'AFM permet d'analyser tous ces types de données simultanément



## Données mixtes

⇒ Groupes composés de variables quantitatives et groupes composés de variables qualitatives

L'AFM fonctionne "localement" comme :

- une ACP pour les variables quantitatives
- une ACM pour les variables qualitatives
- une AFC pour les tableaux de contingence

La pondération de l'AFM permet d'analyser tous ces types de données simultanément

Cas particulier : si chaque groupe est composé d'une seule variable  
⇒ **Analyse Factorielle de Données Mixtes (AFDM)**

## Mise en œuvre d'une AFM

1. Définir la composition des groupes (la structure du tableau)
2. Définir les groupes actifs et les éléments supplémentaires
3. Réduire ou non les variables quantitatives ?
4. Réaliser l'AFM
5. Choisir le nombre de dimensions à interpréter
6. Interpréter simultanément le graphe des individus et des variables
7. Etude des groupes
8. Analyses partielles
9. Utilisation d'indicateurs pour enrichir l'interprétation

Fonction MFA du package FactoMineR

# Plan

Introduction

Analyse en Composantes Principales



Analyse des correspondances

Analyse des correspondances multiples

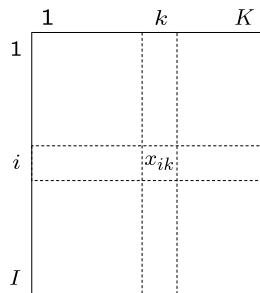
Analyse Factorielle Multiple

**Classification**

Conclusion

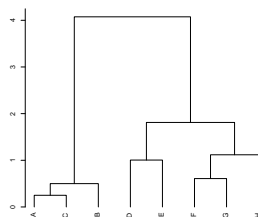
## Quelles données pour quels objectifs ?

La classification s'intéresse à des tableaux de données individus  $\times$  variables quantitatives



Objectifs : production d'une structure (arborescence) permettant :

- la mise en évidence de liens hiérarchiques entre individus ou groupes d'individus
- la détection d'un nb de classes « naturel » au sein de la population



# Critères

Ressemblance entre individus :

- distance euclidienne
- indice de similarité
- ...

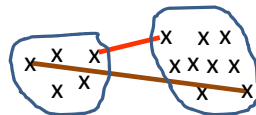
## Critères

Ressemblance entre individus :

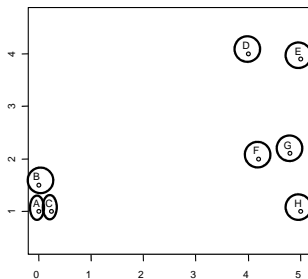
- distance euclidienne
- indice de similarité
- ...

Ressemblance entre groupes d'individus :

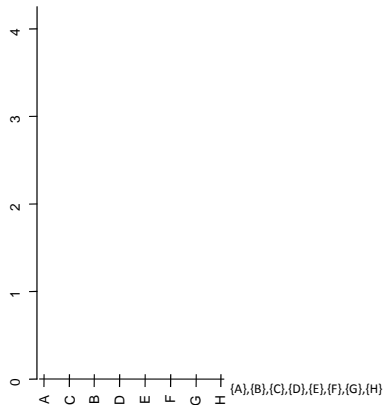
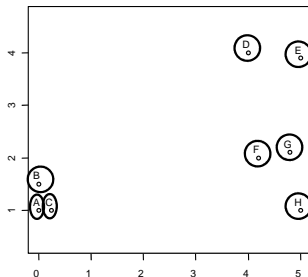
- saut minimum ou lien simple (**plus petite distance**)
- lien complet (**plus grande distance**)
- critère de Ward



# Algorithme



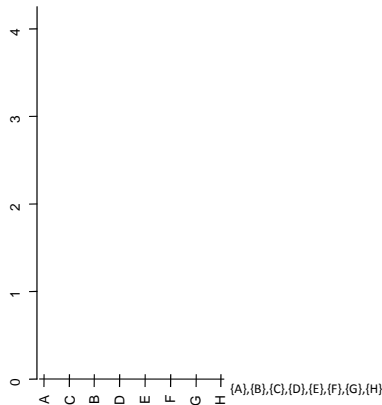
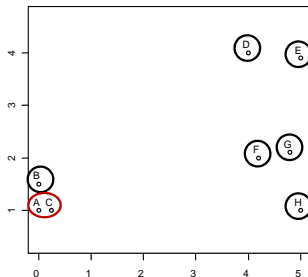
# Algorithme



	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

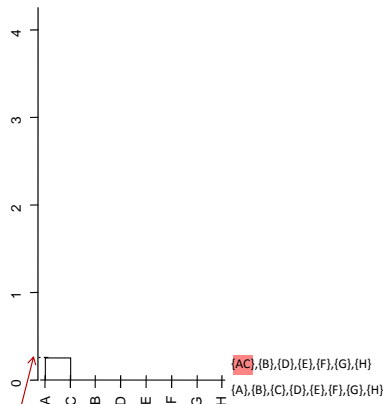
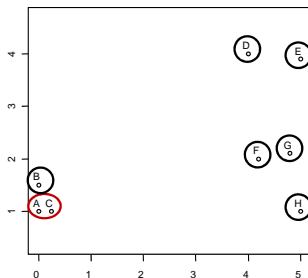


# Algorithme



	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

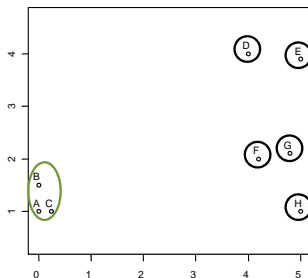
# Algorithme



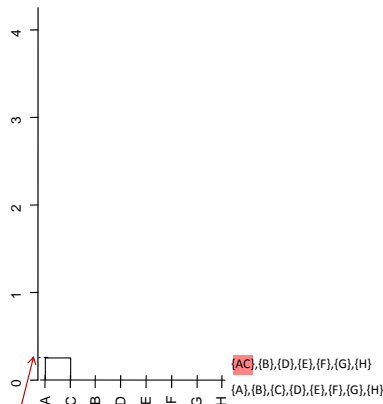
	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

1<sup>er</sup> regroupement

# Algorithme



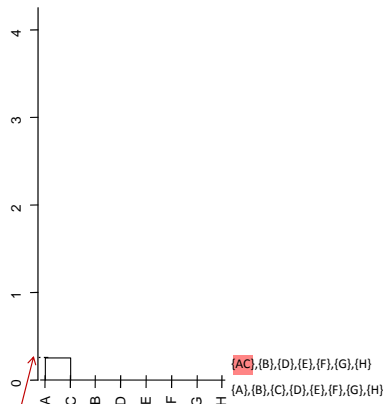
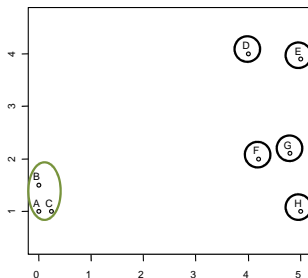
	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12



	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

1<sup>er</sup> regroupement

# Algorithme



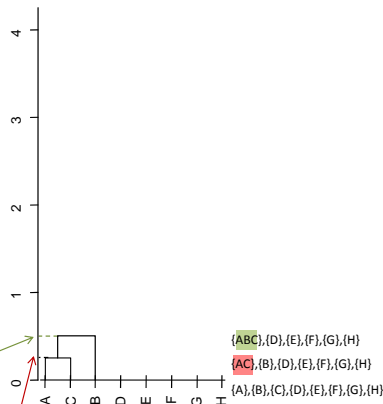
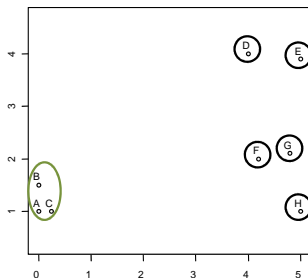
2<sup>e</sup> regroupement

	A	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

1<sup>er</sup> regroupement

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

# Algorithme



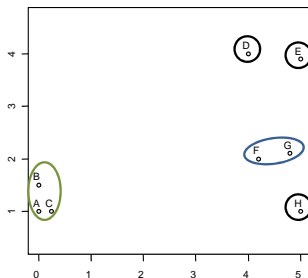
2<sup>e</sup> regroupement

	A	B	C	D	E	F	G
B	0.50						
D	4.80	4.72					
E	5.57	5.55	1.00				
F	4.07	4.23	2.01	2.06			
G	4.68	4.84	2.06	1.81	0.61		
H	4.75	5.02	3.16	2.90	1.28	1.12	

1<sup>er</sup> regroupement

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

# Algorithme

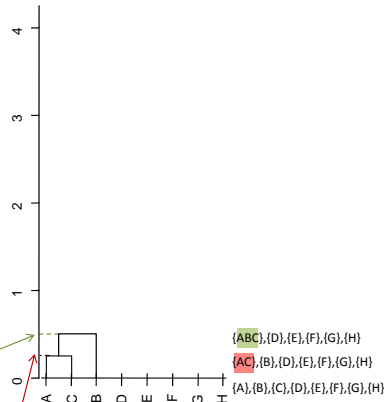


3<sup>e</sup> regroupement

	ABC	D	E	F	G
D	4.72				
E	5.55	1.00			
F	4.07	2.01	2.06		
G	4.68	2.06	1.81	0.61	
H	4.75	3.16	2.90	1.28	1.12

2<sup>e</sup> regroupement

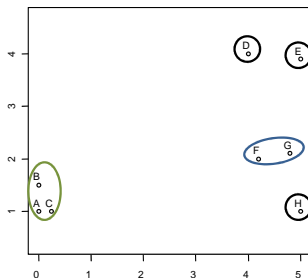
	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12



1<sup>er</sup> regroupement

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

# Algorithme

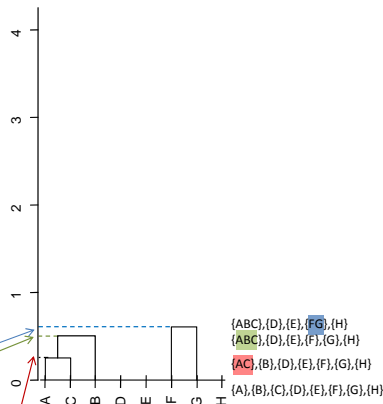


3<sup>e</sup> regroupement

	ABC	D	E	F	G
D	4.72				
E	5.55	1.00			
F	4.07	2.01	2.06		
G	4.68	2.06	1.81	0.61	
H	4.75	3.16	2.90	1.28	1.12

2<sup>e</sup> regroupement

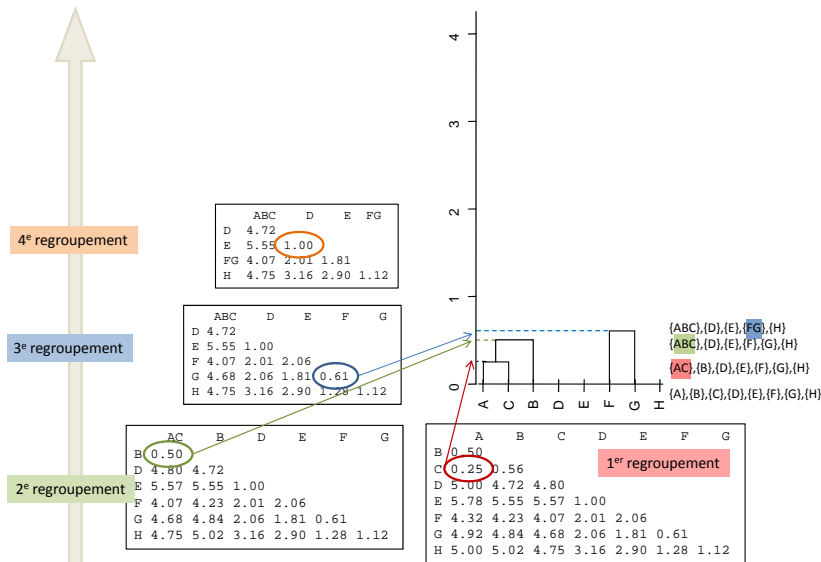
	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12



1<sup>er</sup> regroupement

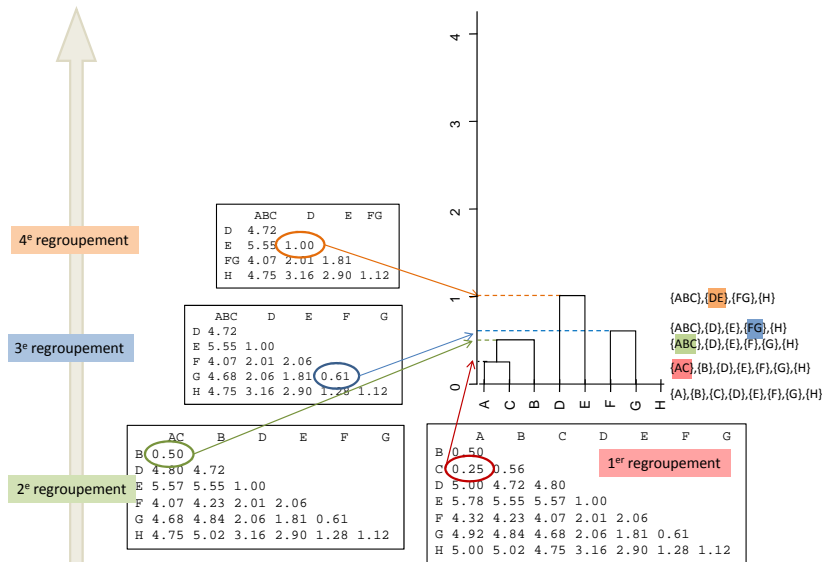
	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

# Algorithme

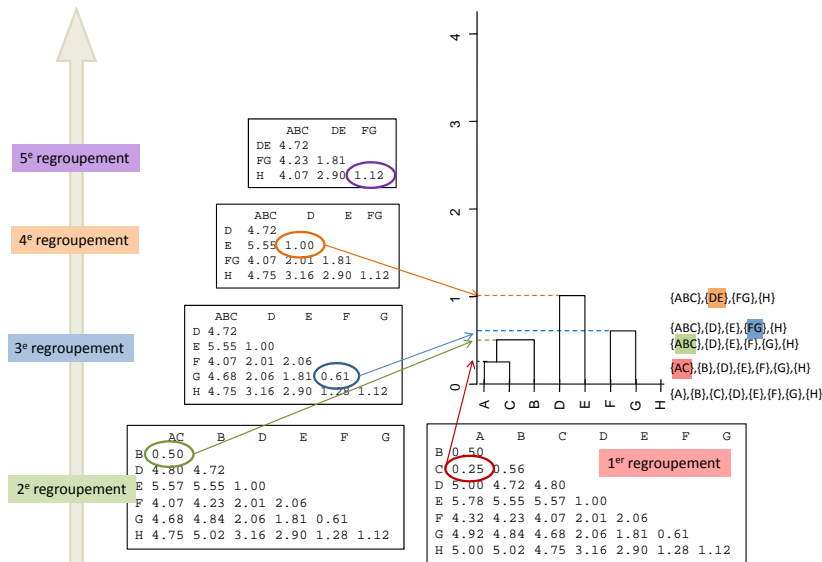




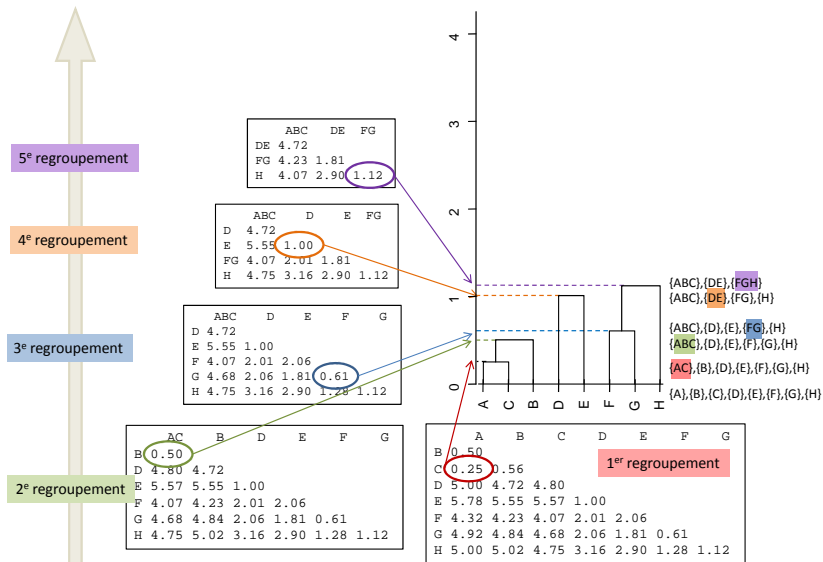
# Algorithme



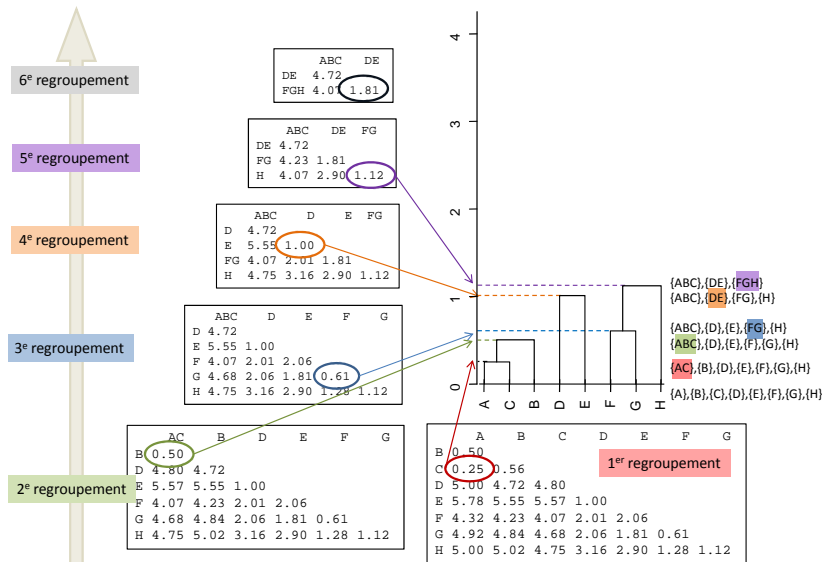
# Algorithme



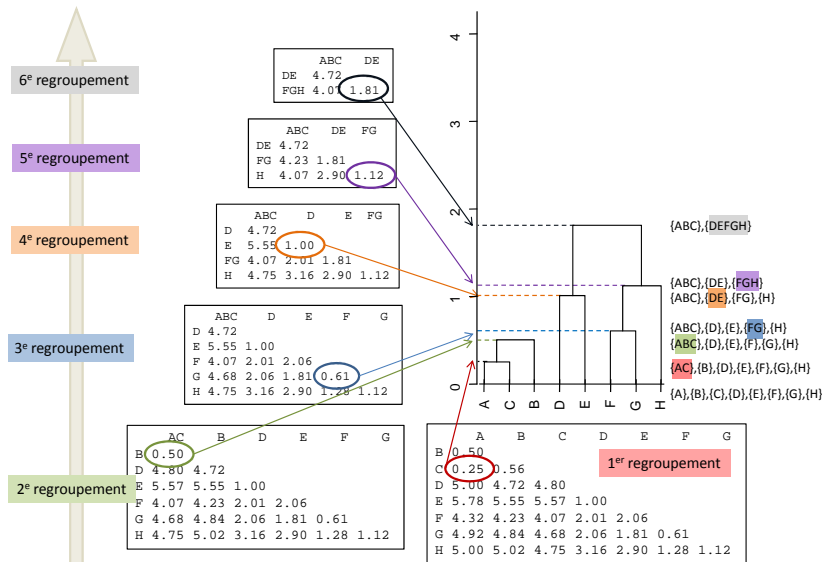
# Algorithme



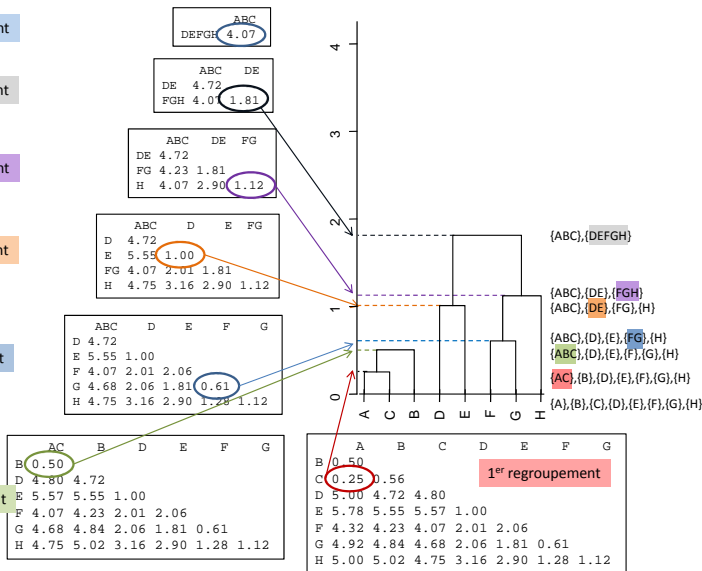
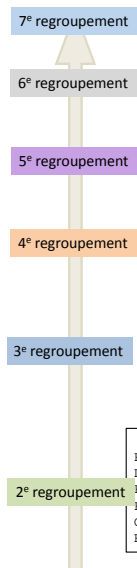
# Algorithme



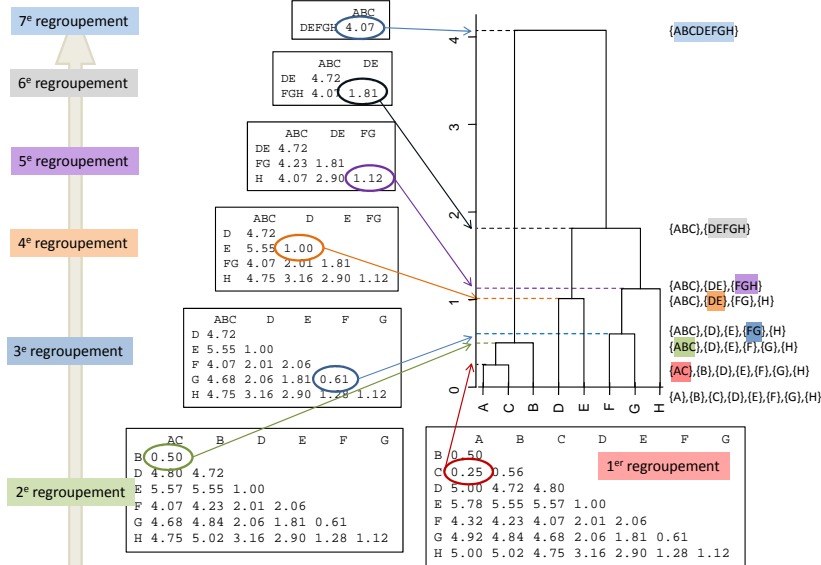
# Algorithme



# Algorithme



# Algorithme







## Qualité d'une partition

Quand une partition est-elle bonne ?

- Si les individus d'une même classe sont proches
- Si les individus de 2 classes différentes sont éloignés

## Qualité d'une partition

Quand une partition est-elle bonne ?

- Si les individus d'une même classe sont proches
- Si les individus de 2 classes différentes sont éloignés

Et mathématiquement ça se traduit par ?

- Variabilité intra-classe petite
- Variabilité inter-classes grande

## Qualité d'une partition

Quand une partition est-elle bonne ?

- Si les individus d'une même classe sont proches
- Si les individus de 2 classes différentes sont éloignés

Et mathématiquement ça se traduit par ?

- Variabilité intra-classe petite
- Variabilité inter-classes grande

⇒ En réalité, il s'agit d'1 seul critère

## Qualité d'une partition

Quand une partition est-elle bonne ?

- Si les individus d'une même classe sont proches
- Si les individus de 2 classes différentes sont éloignés

Et mathématiquement ça se traduit par ?

- Variabilité intra-classe petite
- Variabilité inter-classes grande

⇒ En réalité, il s'agit d'1 seul critère

**Attention** : ce critère ne peut être jugé en absolu car il dépend du nb d'individus et du nb de classes

## Les données température

- 15 individus : villes de France
- 12 variables : températures mensuelles moyennes (sur 30 ans)

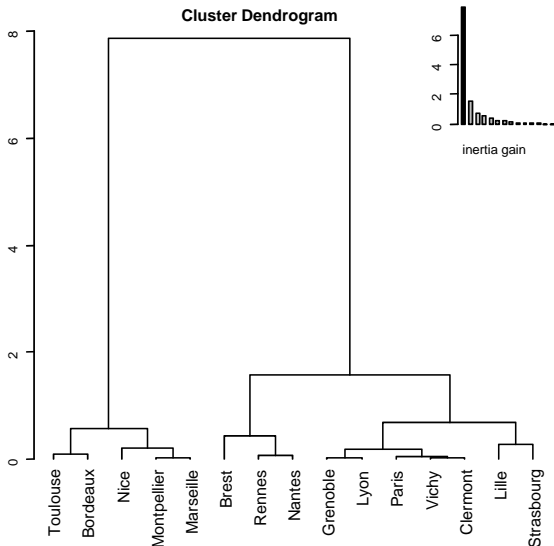
	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
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	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
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	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
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	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26

Quelles villes ont des profils météo similaires ?

Comment caractériser les groupes de villes ?

# Les données température : l'arbre hiérarchique

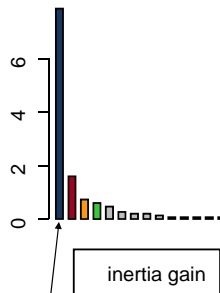
## Hierarchical clustering



## Les données température

### Pertes d'inertie inter lors du passage de

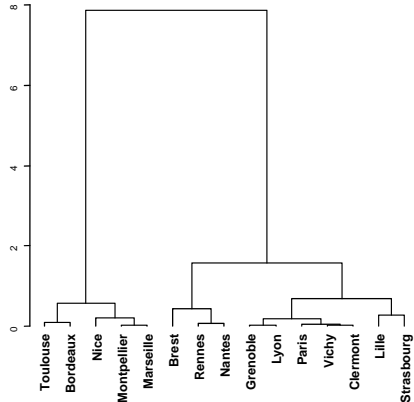
15 classes en 14 classes : 0.01  
14 classes en 13 classes : 0.02  
13 classes en 12 classes : 0.03  
12 classes en 11 classes : 0.05  
11 classes en 10 classes : 0.06  
10 classes en 9 classes : 0.09  
9 classes en 8 classes : 0.17  
8 classes en 7 classes : 0.19  
7 classes en 6 classes : 0.26  
6 classes en 5 classes : 0.42  
**5 classes en 4 classes : 0.56**  
**4 classes en 3 classes : 0.69**  
**3 classes en 2 classes : 1.56**  
**2 classes en 1 classe : 7.88**



Grosse perte si on passe de  
2 classes à 1 seule donc on  
préfère garder 2 classes

# Utilisation de l'arbre pour construire une partition

Doit-on faire 2 groupes ? 3 groupes ? 4 ?



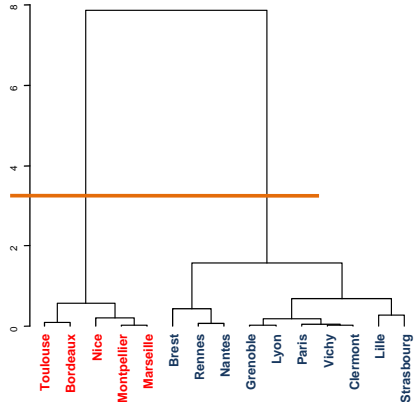


# Utilisation de l'arbre pour construire une partition

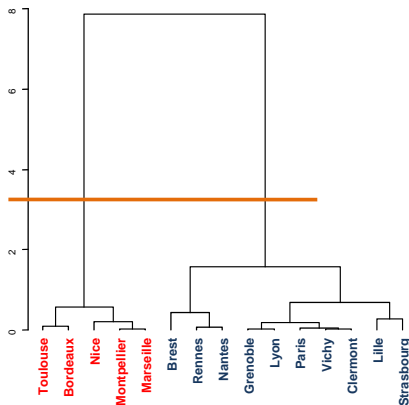
Doit-on faire 2 groupes ? 3 groupes ? 4 ?

Découpage en 2 groupes :

$$\frac{\text{Inertie inter}}{\text{Inertie totale}} = \frac{7.88}{12} = 66\%$$

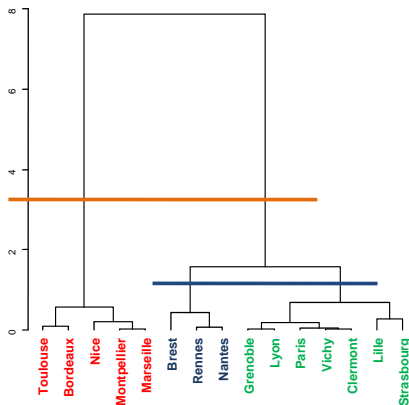


## Utilisation de l'arbre pour construire une partition



Séparer villes froides en 2 groupes :

# Utilisation de l'arbre pour construire une partition

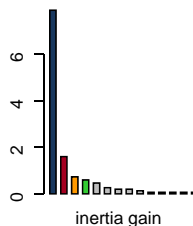
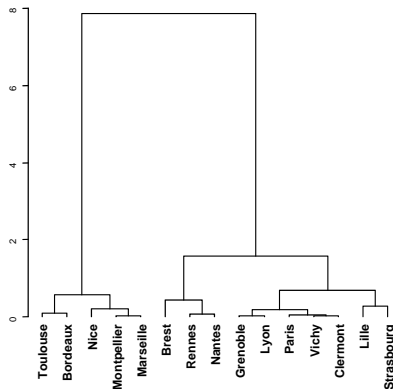


Séparer **villes froides** en 2 groupes :

$$\frac{\text{Inertie inter}}{\text{Inertie totale}} = \frac{1.56}{12} = 13\%$$

## Détermination d'un nombre de classes

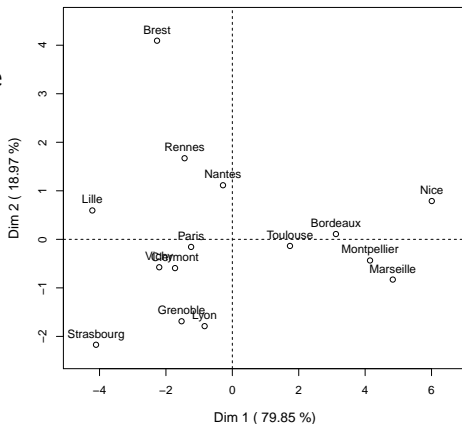
- A partir de l'arbre
- Dépend de l'usage (enquête, ...)
- A partir du diagramme des indices de niveau
- Critère ultime : interprétabilité des classes



# Algorithme de partitionnement : les K-means

## Algorithme d'agrégation autour des centres mobiles (K-means)

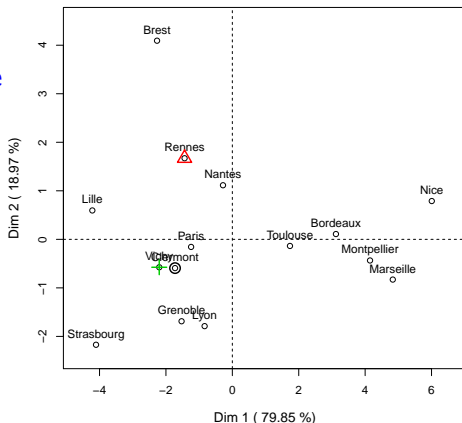
- Choisir  $Q$  centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les  $Q$  centres de gravité



# Algorithme de partitionnement : les K-means

## Algorithme d'agrégation autour des centres mobiles (K-means)

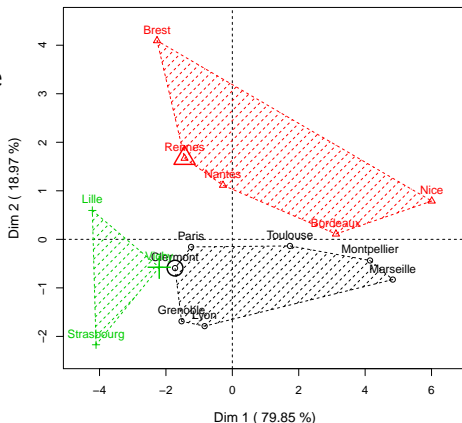
- Choisir  $Q$  centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les  $Q$  centres de gravité



# Algorithme de partitionnement : les K-means

## Algorithme d'agrégation autour des centres mobiles (K-means)

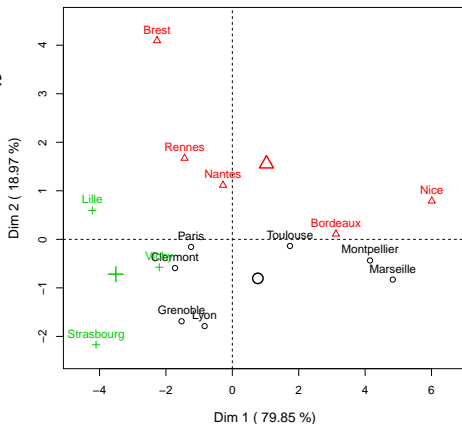
- Choisir  $Q$  centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les  $Q$  centres de gravité



# Algorithme de partitionnement : les K-means

## Algorithme d'agrégation autour des centres mobiles (K-means)

- Choisir  $Q$  centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les  $Q$  centres de gravité

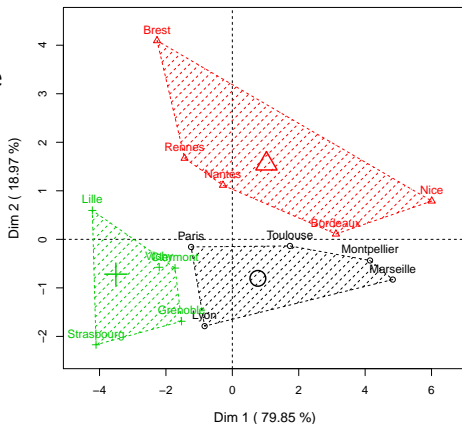




# Algorithme de partitionnement : les K-means

## Algorithme d'agrégation autour des centres mobiles (K-means)

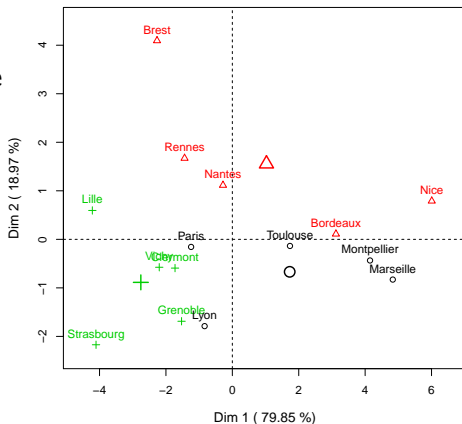
- Choisir  $Q$  centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les  $Q$  centres de gravité



# Algorithme de partitionnement : les K-means

## Algorithme d'agrégation autour des centres mobiles (K-means)

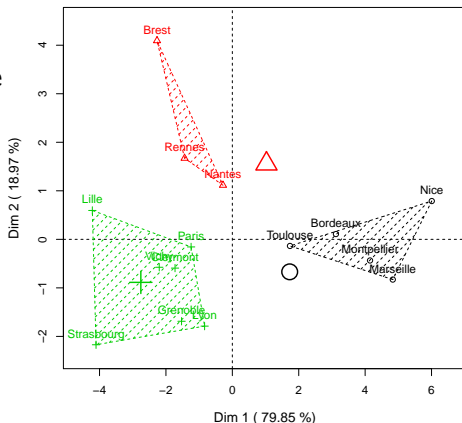
- Choisir  $Q$  centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les  $Q$  centres de gravité



# Algorithme de partitionnement : les K-means

## Algorithme d'agrégation autour des centres mobiles (K-means)

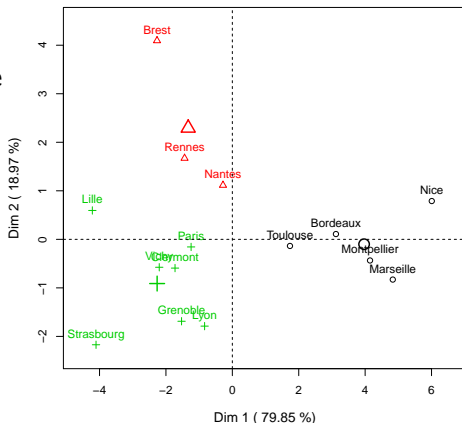
- Choisir  $Q$  centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les  $Q$  centres de gravité



# Algorithme de partitionnement : les K-means

## Algorithme d'agrégation autour des centres mobiles (K-means)

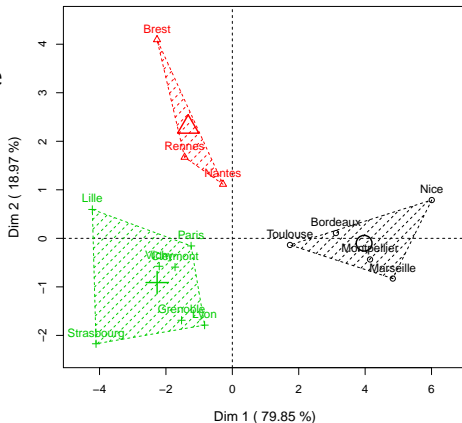
- Choisir  $Q$  centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les  $Q$  centres de gravité



# Algorithme de partitionnement : les K-means

## Algorithme d'agrégation autour des centres mobiles (K-means)

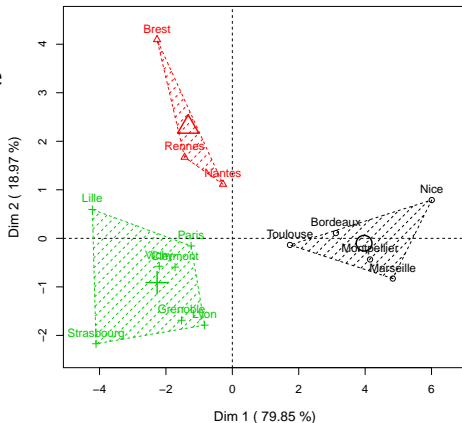
- Choisir  $Q$  centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les  $Q$  centres de gravité



# Algorithme de partitionnement : les K-means

## Algorithme d'agrégation autour des centres mobiles (K-means)

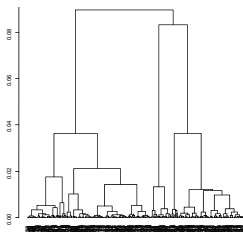
- Choisir  $Q$  centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les  $Q$  centres de gravité



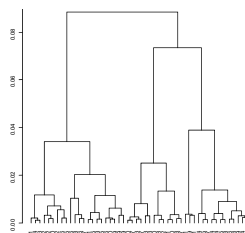
L'algorithme a convergé

## CAH en grandes dimensions

- Si beaucoup de variables : faire une ACP et ne conserver que les premières dimensions  $\Rightarrow$  on se ramène au cas classique
- Si beaucoup d'individus : algorithme de CAH trop long
  - Faire une partition (par K-means) en une centaine de classes
  - Construire la CAH à partir des classes (utiliser l'effectif des classes dans le calcul)
  - Obtention du « haut » de l'arbre de la CAH



Arbre sur données brutes



Arbre à partir de classes

## Enchaînement analyse factorielle - classification

- Données qualitatives : ACM renvoie des composantes principales qui sont quantitatives



## Enchaînement analyse factorielle - classification

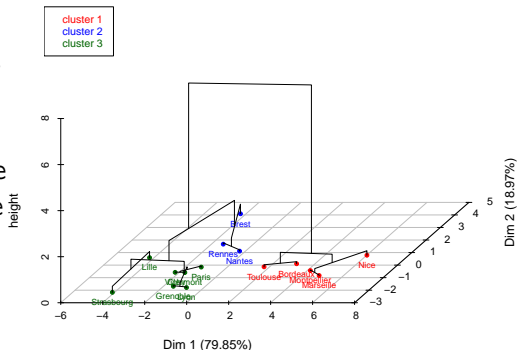
- Données qualitatives : ACM renvoie des composantes principales qui sont quantitatives
- L'analyse factorielle élimine les dernières composantes qui ne contiennent que du bruit  $\implies$  classification plus stable

## Enchaînement analyse factorielle - classification

- Données qualitatives : ACM renvoie des composantes principales qui sont quantitatives
- L'analyse factorielle élimine les dernières composantes qui ne contiennent que du bruit  $\Rightarrow$  classification plus stable

Hierarchical clustering on the factor map

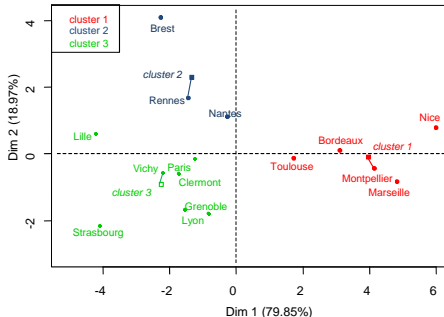
- Représentation de l'arbre et des classes sur un plan factoriel  $\Rightarrow$  vision continue avec AF, discontinue avec CAH ; vision de l'information sur d'autres axes avec CAH



# Constitution des classes - Édition des parangons

Parangon : individu le plus proche du centre d'une classe

classe 1 :	Montpellier	Bordeaux	Marseille	Nice	Toulouse
	0.419	1.141	1.193	2.242	2.256
classe 2 :	Rennes	Nantes	Brest		
	0.641	1.586	2.045		
classe 3 :	Vichy	Clermont	Grenoble	Paris	Lyon
	0.428	0.669	1.184	1.339	1.680



## Caractérisation des classes

Quelles variables caractérisent le mieux la partition ?

- Pour chaque variable quantitative :
  - construire le modèle d'analyse de variance entre la variable quantitative expliquée par la variable de classe
  - faire le test de Fisher de l'effet de la classe
- Trier les variables par probabilité critique croissante

	Eta2	P-value
Octo	0.8362	1.930e-05
Sept	0.8301	2.407e-05
Févr	0.8227	3.103e-05
Mars	0.8126	4.326e-05
Janv	0.8118	4.444e-05
Nove	0.8083	4.963e-05
Avri	0.7929	7.890e-05
Déce	0.7871	9.316e-05
Août	0.7864	9.503e-05
Juin	0.7241	4.409e-04
Mai	0.7164	5.205e-04
juil	0.7156	5.287e-04

# Caractérisation d'une classe par les variables quantitatives

$$\text{Valeur-test} = \frac{\bar{X}_q - \bar{X}}{\sqrt{\frac{s^2}{n_q} \left( \frac{N-n_q}{N-1} \right)}}$$

⇒ Si  $|\text{Valeur-test}| \geq 1.96$  alors variable  $X$  caractérise la classe  $q$

\$quantile\$'1'

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Sept	3.40	19.30	17.00	0.755	1.79	0.000678
Moye	3.39	13.80	11.80	0.742	1.55	0.000705
Avri	3.33	12.70	11.00	0.580	1.37	0.000871
Octo	3.32	14.50	12.30	0.941	1.77	0.000893
Mars	3.24	10.00	8.23	0.524	1.48	0.001210
Août	3.18	21.90	19.60	0.792	1.94	0.001490
Juin	3.00	19.80	17.80	0.727	1.73	0.002670
Mai	3.00	16.10	14.40	0.691	1.45	0.002720
Nove	2.97	9.88	7.93	0.999	1.74	0.003020
juil	2.92	22.10	19.80	1.000	2.06	0.003550
Févr	2.88	6.80	4.83	0.940	1.81	0.003940
Déce	2.54	6.66	4.85	0.896	1.89	0.011200
Janv	2.46	5.78	3.97	0.924	1.94	0.013700

## Caractérisation des classes par les variables qualitatives

Quelles variables caractérisent le mieux la partition ?

- Pour chaque variable qualitative, construire un test du  $\chi^2$  entre la variable et la variable de classe
- Trier les variables par probabilité critique croissante

```
$test.chi2
           p.value df
Région 0.001700272  6
```

Quelle modalité caractérise le mieux la classe 3 ?

- Pour chaque modalité, construire un test hypergéométrique pour voir si la modalité est sur-représentée

```
Classe 3
```

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Région=NE	100.00	42.86	20.00	0.077	1.769
Région=SE	57.14	57.14	46.67	0.505	0.667
Région=SO	0.00	0.00	13.33	0.267	-1.111
Région=NO	0.00	0.00	20.00	0.123	-1.542

# Plan

Introduction

Analyse en Composantes Principales



Analyse des correspondances

Analyse des correspondances multiples

Analyse Factorielle Multiple

Classification

Conclusion

# Démarche en analyse de données

## Démarche en analyse des données

FactoMineR

1. Y a-t-il des groupes de variables ?



oui → AFM

2. Quel est le type d'information ?

Tableau de contingence → AFC (ou AFMTC si plusieurs)

CA

MFA

Tableau individus – variables → ACP, ACM, AFDM ou AFM

PCA

MCA

FAMD

MFA

3. Éléments actifs ?

Quels éléments participent à la construction des axes factoriels ?



ind.sup, quanti.sup, quali.sup,  
row.sup, col.sup, group.sup

4. Nature des variables actives ?

Quantitative → ACP

Qualitative → si 2 variables, tableau croisé → AFC

→ sinon ACM

Mixte → AFDM

AFM si groupes

5. Doit-on réduire les variables quantitatives ?

oui → ACP réduite `scale.unit = TRUE`

6. Y a-t-il des données manquantes ?

oui → utiliser le package `missMDA` pour compléter le jeu de données

Lancer l'analyse factorielle

PCA, CA, MCA, FAMD ou MFA

Voir les résultats et construire les graphes

`summary`, `plot`

Décrire les axes factoriels par les variables initiales

`dimdesc`

(facultatif) Faire une classification des individus (et décrire les classes)

HCPC



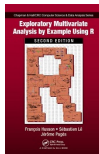
# Logiciels et aides à l'utilisateur

## Packages R :

- FactoMineR : pour mettre en oeuvre les méthodes
- Factoshiny : pour un menu déroulant et graphes interactifs
- missMDA : pour la gestion des données manquantes
- FactoInvestigate : pour les rapports automatisés

## Aides à l'utilisateur :

- site FactoMineR : <http://factominer.free.fr>
- site F. Husson : <https://husson.github.io>
- MOOC
- livre : Analyse de données avec R (2<sup>e</sup> ed)



- chaîne Youtube : <https://www.youtube.com/HussonFrancois>