Panorama sur les méthodes d'analyse exploratoire des données

François Husson

Unité de mathématiques appliquées, Agrocampus Ouest, Rennes

17-19 février 2020

husson@agrocampus-ouest.fr

Présentation

- Recherche : analyse de données, tableaux multiples, données manquantes
- Enseignement : cursus d'ingénieur, master *science des données*
- MOOC en analyse de données et MOOC en Sensométrie
- Formation continue : statistique avec R, analyse de données


2018

2nd ed: 2017
1st ed: 2011

2nd ed: 2016
1st ed: 2009

2nd ed: 2013
1st ed: 2005


2013

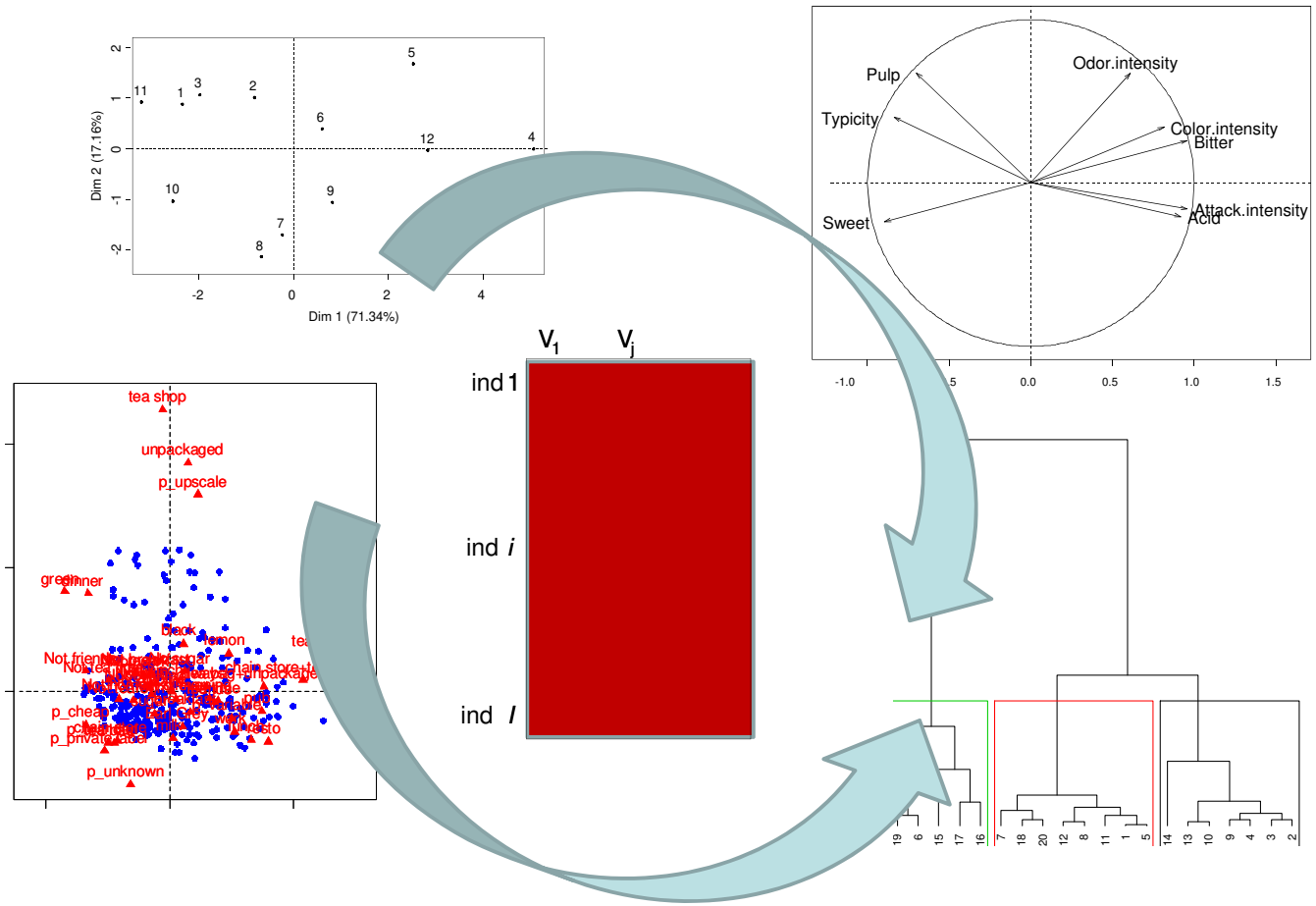
3rd ed: 2012
2nd ed: 2010
1st ed: 2008

2012

Packages: FactoMineR - missMDA - SensoMineR - Factoshiny - FactoInvestigate - RcmdrPlugin.FactoMineR

Plan

Les méthodes d'analyse de données



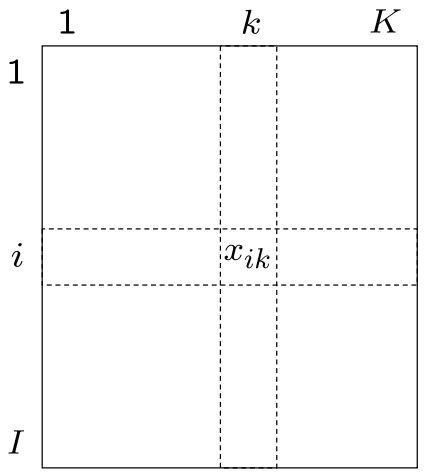
Objectifs :

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

Plan

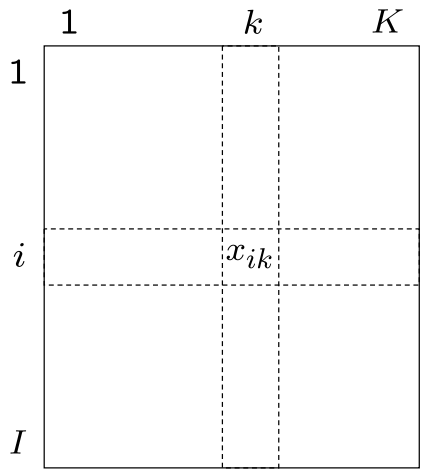
Quelles données ?

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



Quelles 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.preferene	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

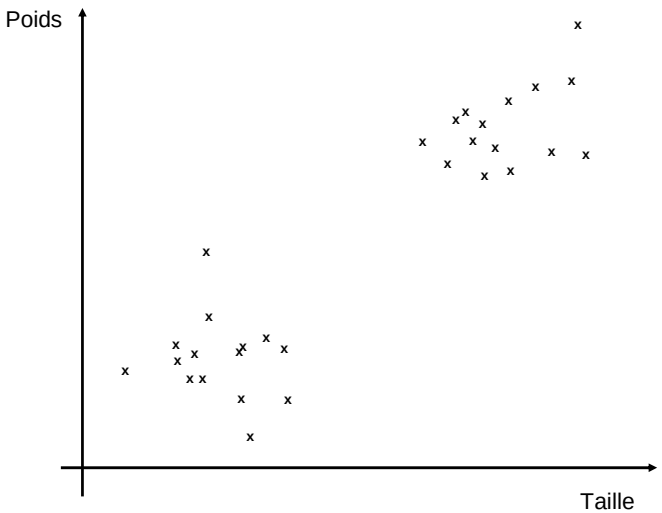
Lien entre les deux études

- 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 $\mathbb{R}^K$
- 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

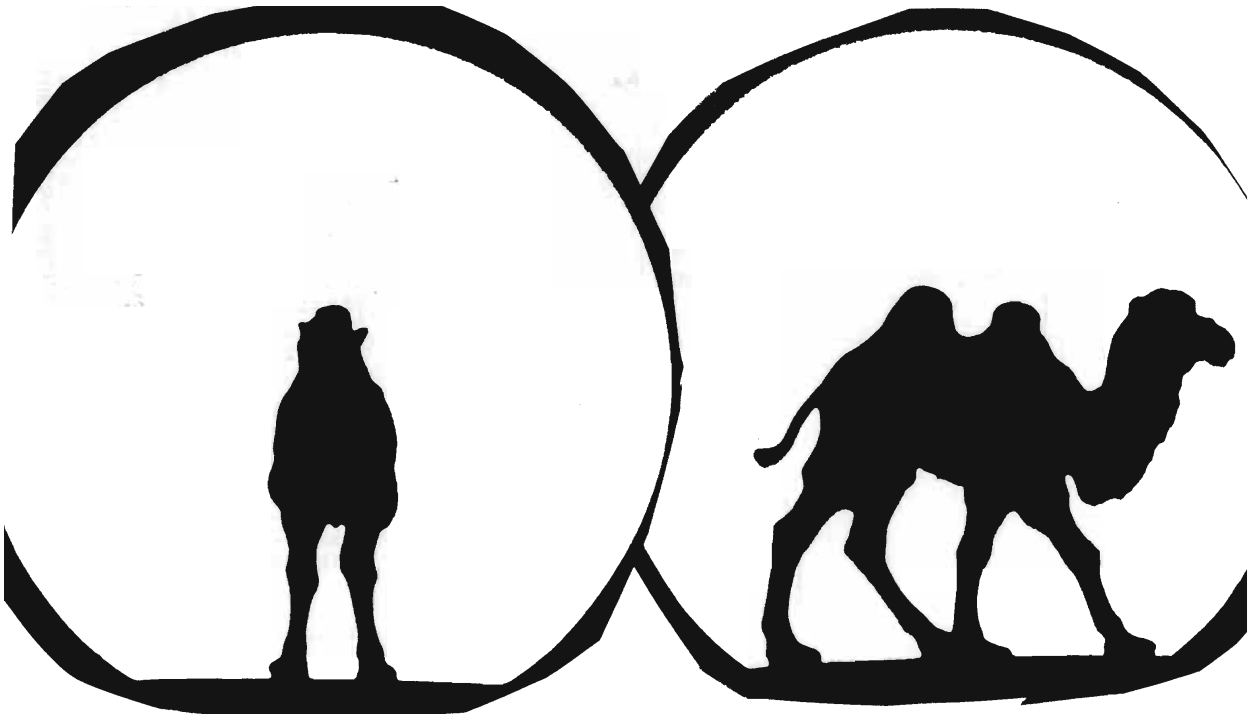
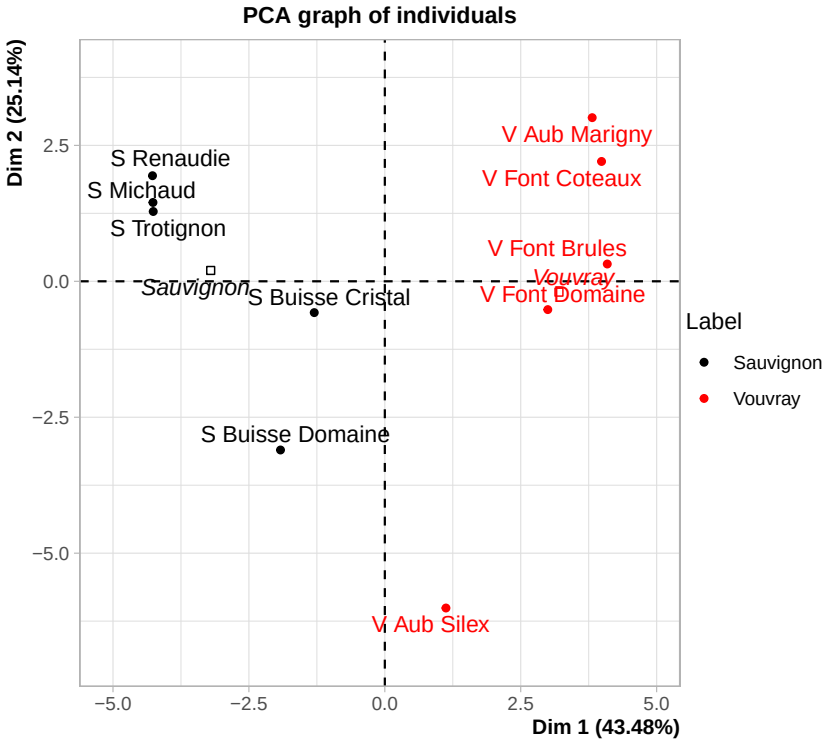


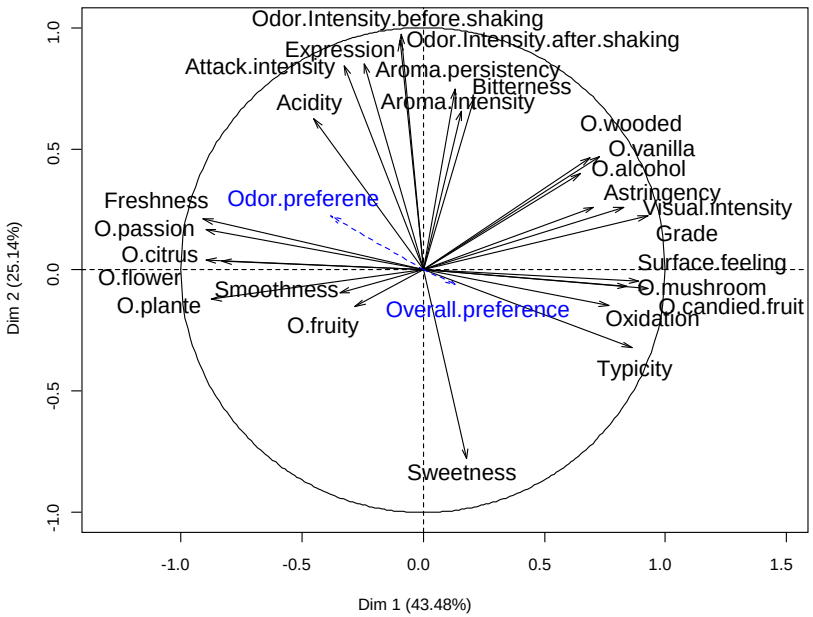
FIGURE – Quel animal ? source J.P. Fenelon

⇒ Meilleure approximation par projection

⇒ Meilleure représentation de la diversité, de la variabilité

Représentation des individus et des variables

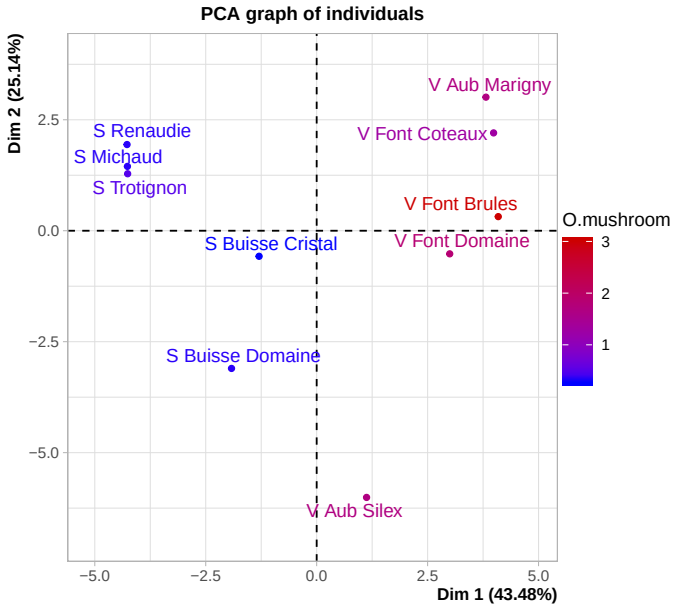


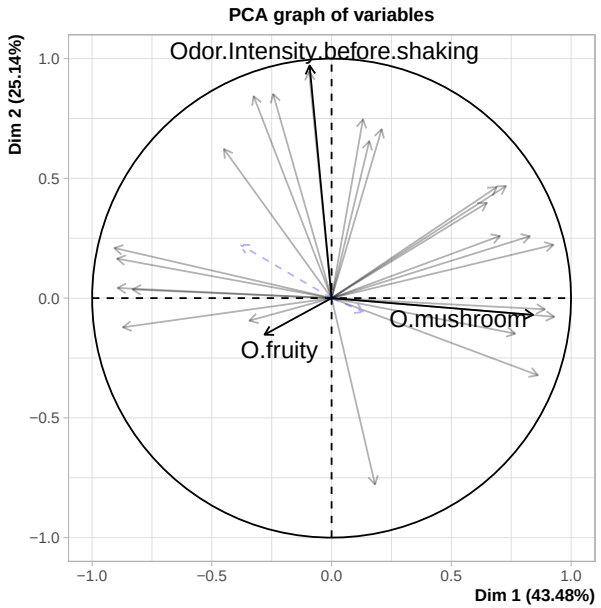


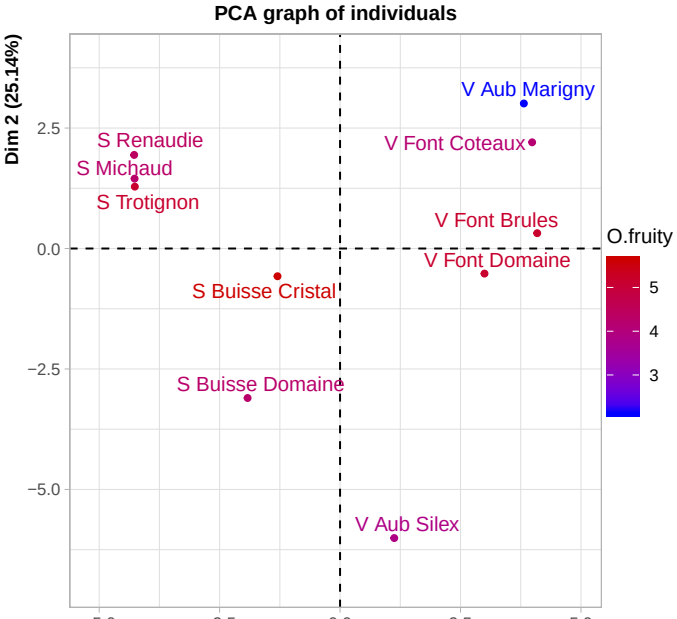
$\Longrightarrow$ Utiliser de l'information supplémentaire

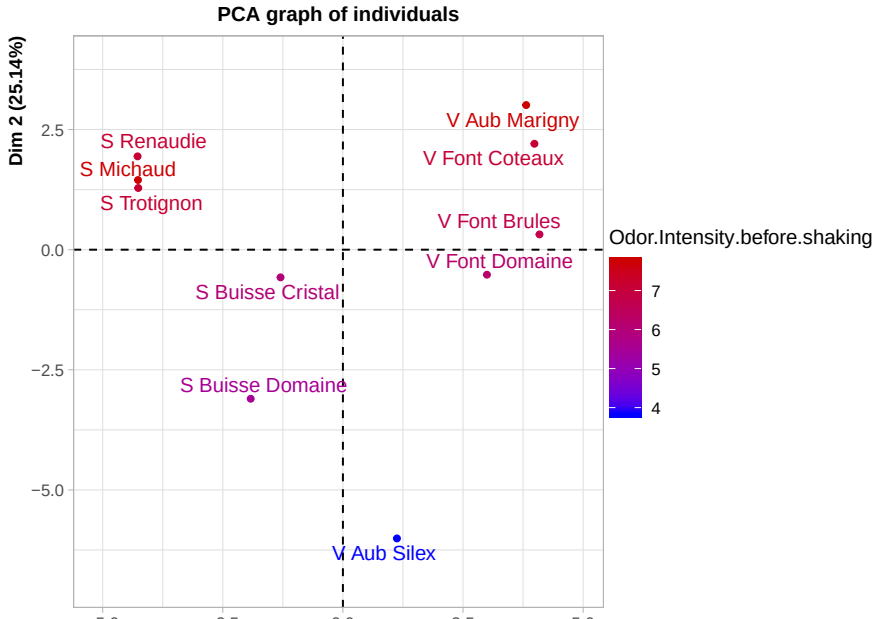
- Variables qualitatives : modalités au barycentre des individus qui prennent cette modalité
- Variables quantitatives : projection des variables

Représentation des individus et des variables



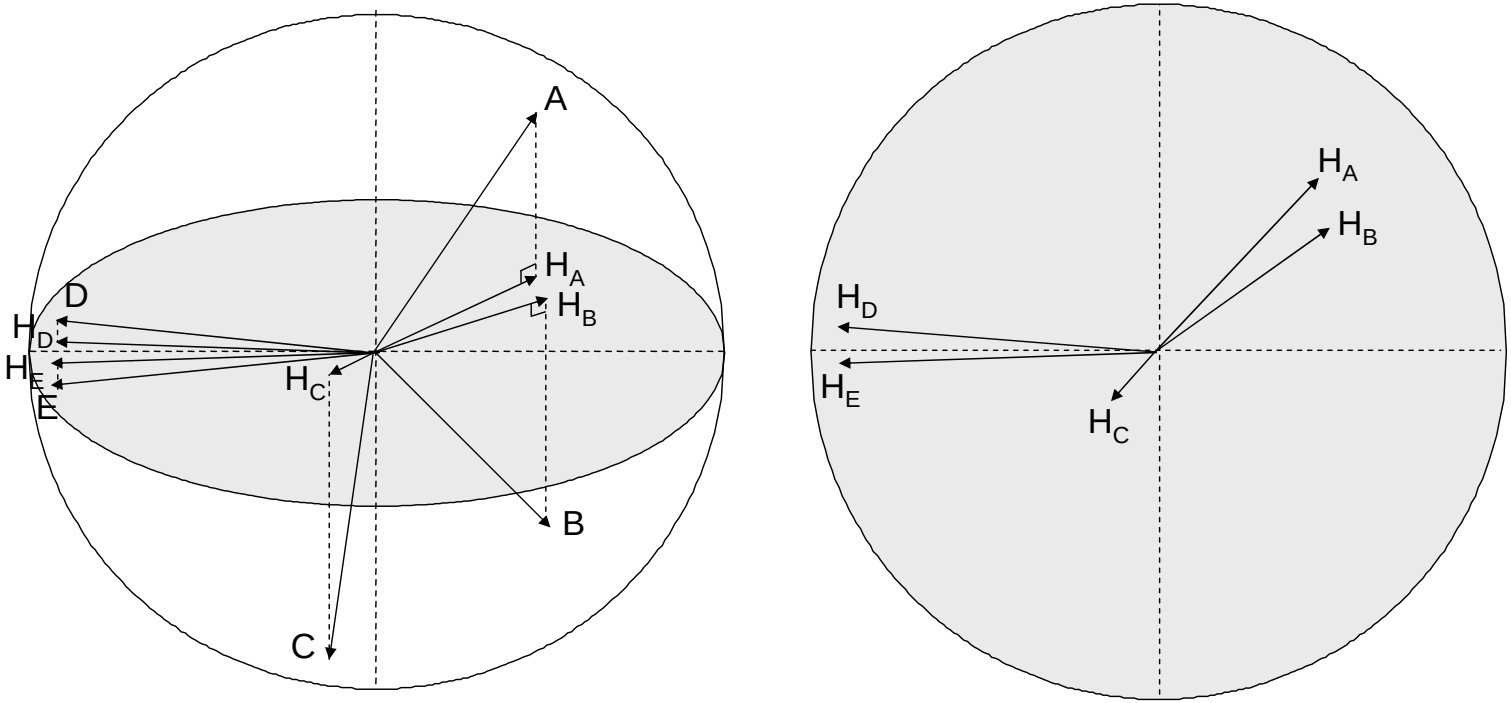






Projections...

$r(A, B) = cos(\theta_{A,B})$
$cos(\theta_{A,B}) \approx cos(\theta_{H_A,H_B})$ 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
O.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
O.mushroom           0.84 2.3e-03  Aroma.persistency             0.75 1.3e-02
  ...                 ...     ...  Bitterness                    0.71 2.3e-02
O.plante            -0.87 1.0e-03
O.flower            -0.89 4.9e-04
O.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 :

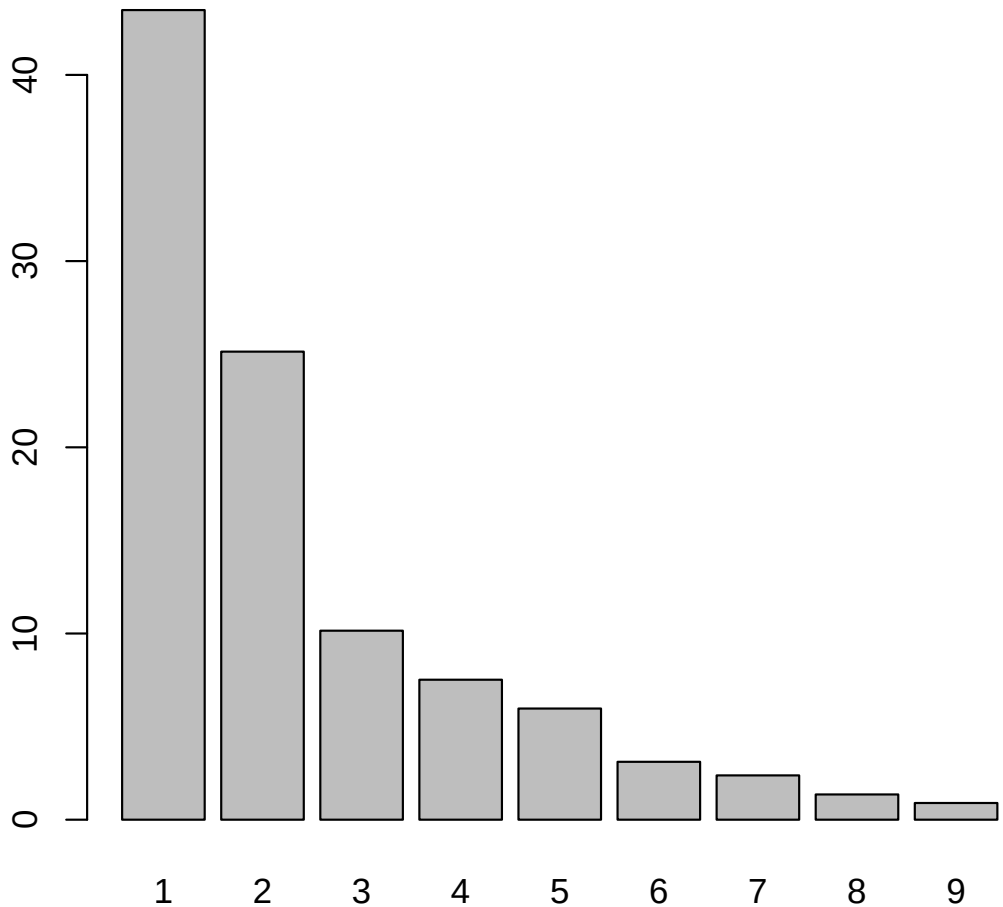
- réalisation d'une analyse de variance avec les coordonnées des individus en fonction de la variable qualitative
  - un F-test par variable
  - un test t de Student par modalité pour comparer la moyenne de la modalité à la moyenne générale

```
> 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
```

Pourcentage d'inertie

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



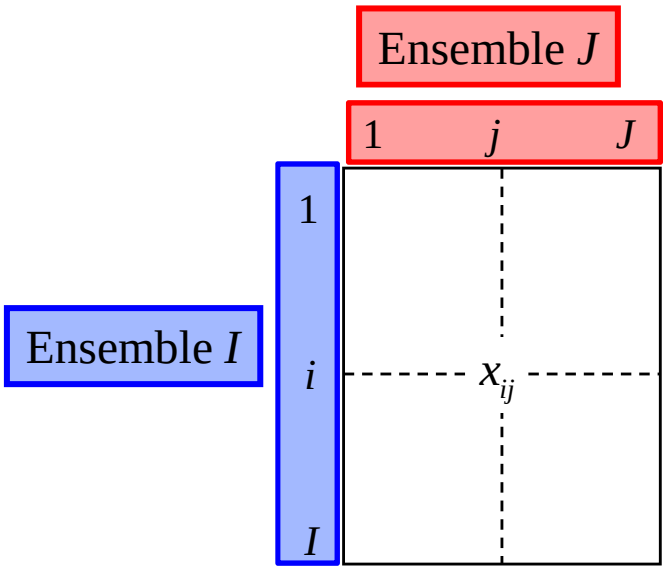
$\Rightarrow$ Choix d'un nombre de dimensions à interpréter

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

Tableau de correspondances



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

Personnages de Phèdre (Racine)	Mots	Nombre de fois que le personnage i a utilisé le mot j
Milieux	Espèces	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

$\Longrightarrow$ Exemples où le test d'indépendance du χ^2 peut être appliqué

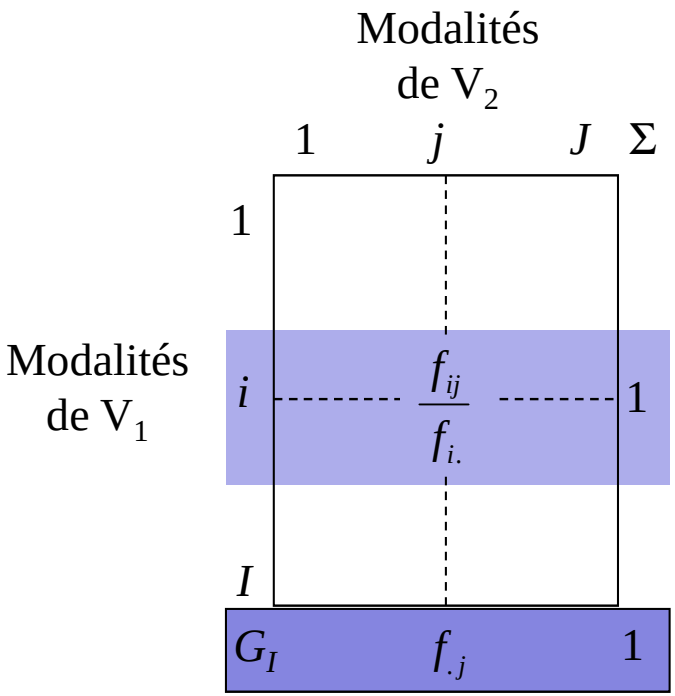
Données sur les prix Nobel

	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 ? Certains pays ont-ils le même profil ? Certaines disciplines ont-elles le même profil ?

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 V2 sachant que l'on possède la modalité *i* de V1

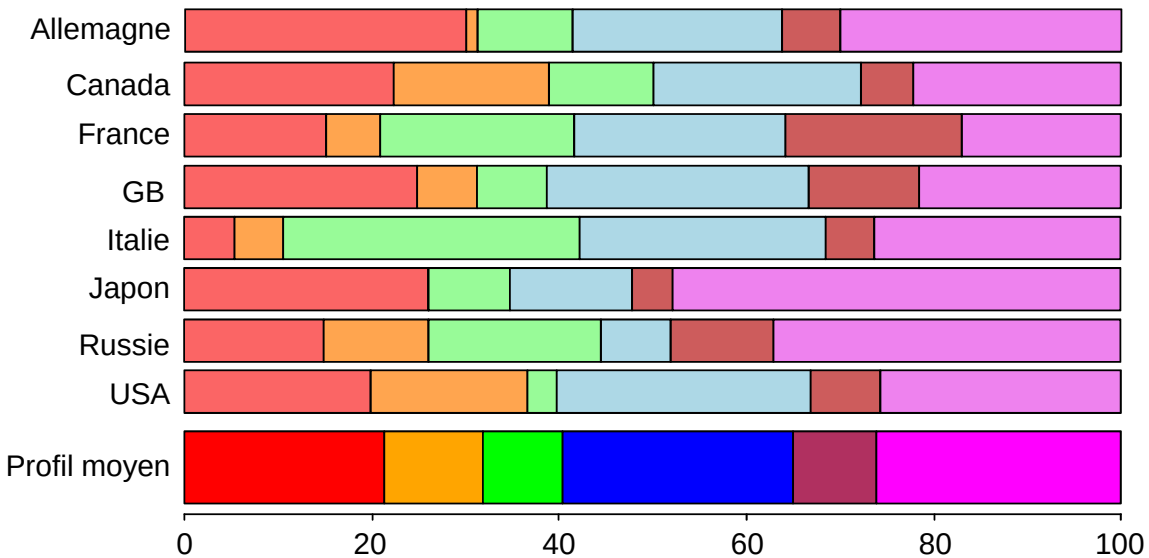
L'AFC compare les profils lignes au profil moyen

Profil ligne moyen = distribution marginale de V2
Profil de l'ensemble des individus étudiés

Approche multidimensionnelle de l'écart à l'indépendance

Comparaison du profil ligne au profil moyen

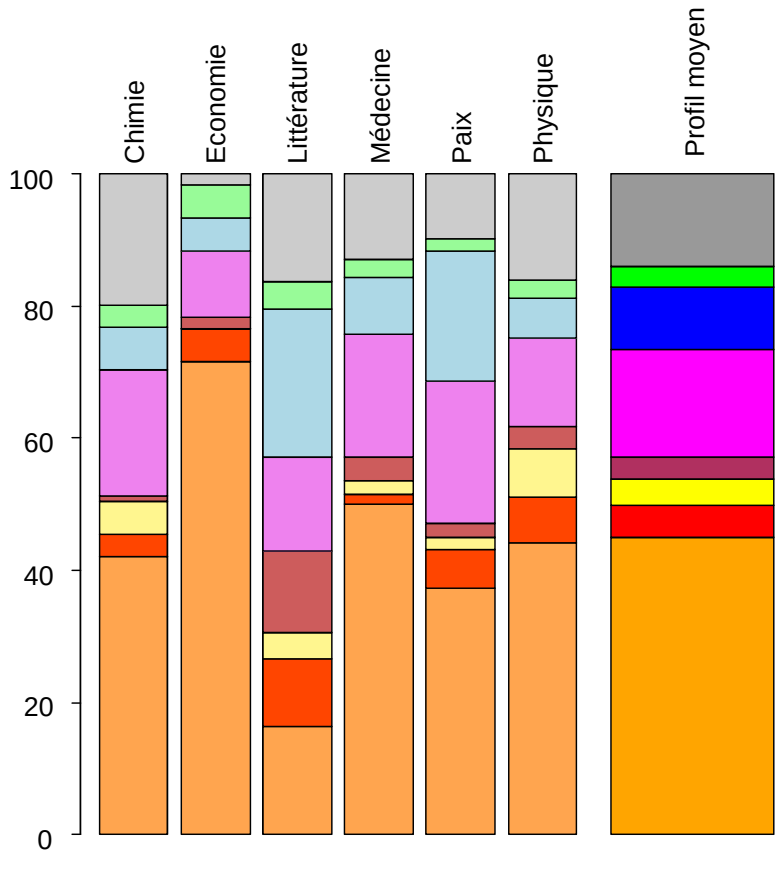
	Chimie	Eco	Lit.	Médecine	Paix	Physique	Somme
Allemagne	30.0	1.2	10.0	22.5	6.2	30.0	100
Canada	22.2	16.7	11.1	22.2	5.6	22.2	100
France	15.1	5.7	20.8	22.6	18.9	17.0	100
GB	24.7	6.5	7.5	28.0	11.8	21.5	100
Italie	5.3	5.3	31.6	26.3	5.3	26.3	100
Japon	26.1	0.0	8.7	13.0	4.3	47.8	100
Russie	14.8	11.1	18.5	7.4	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



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

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \underbrace{\sum_{j=1}^{J} \frac{f_{ij}}{f_{i.}} G_s(j)}$$

$F_s(i)$: 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
λ_s : inertie associée à l'axe s (en AFC $\lambda_s \leq 1$)

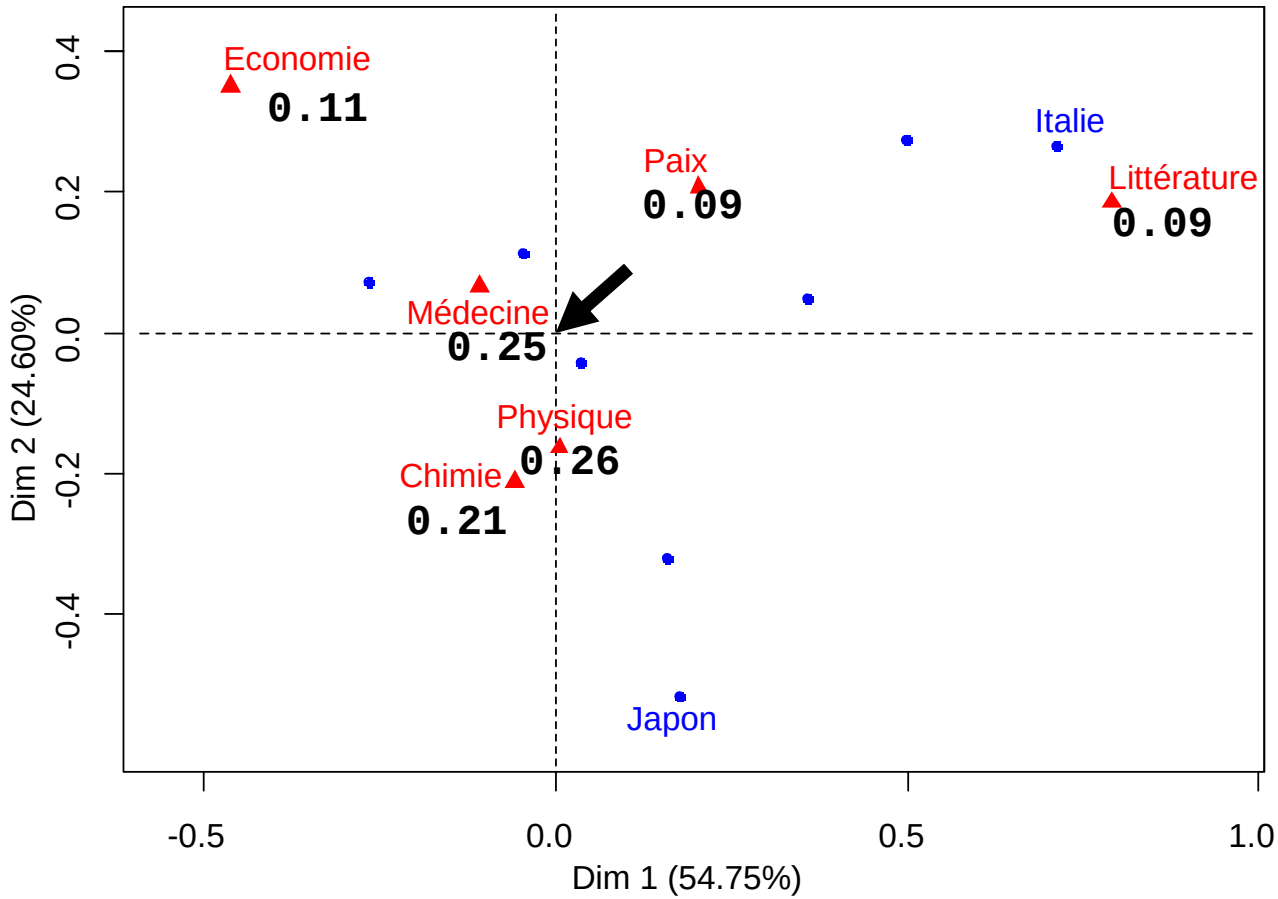
Le long de l'axe de rang s, on calcule le barycentre de toutes les colonnes, chaque colonne j étant affectée du poids $f_{ij}/f_{i.}$

Le barycentre est ensuite d'autant plus écarté de l'origine que λ_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)

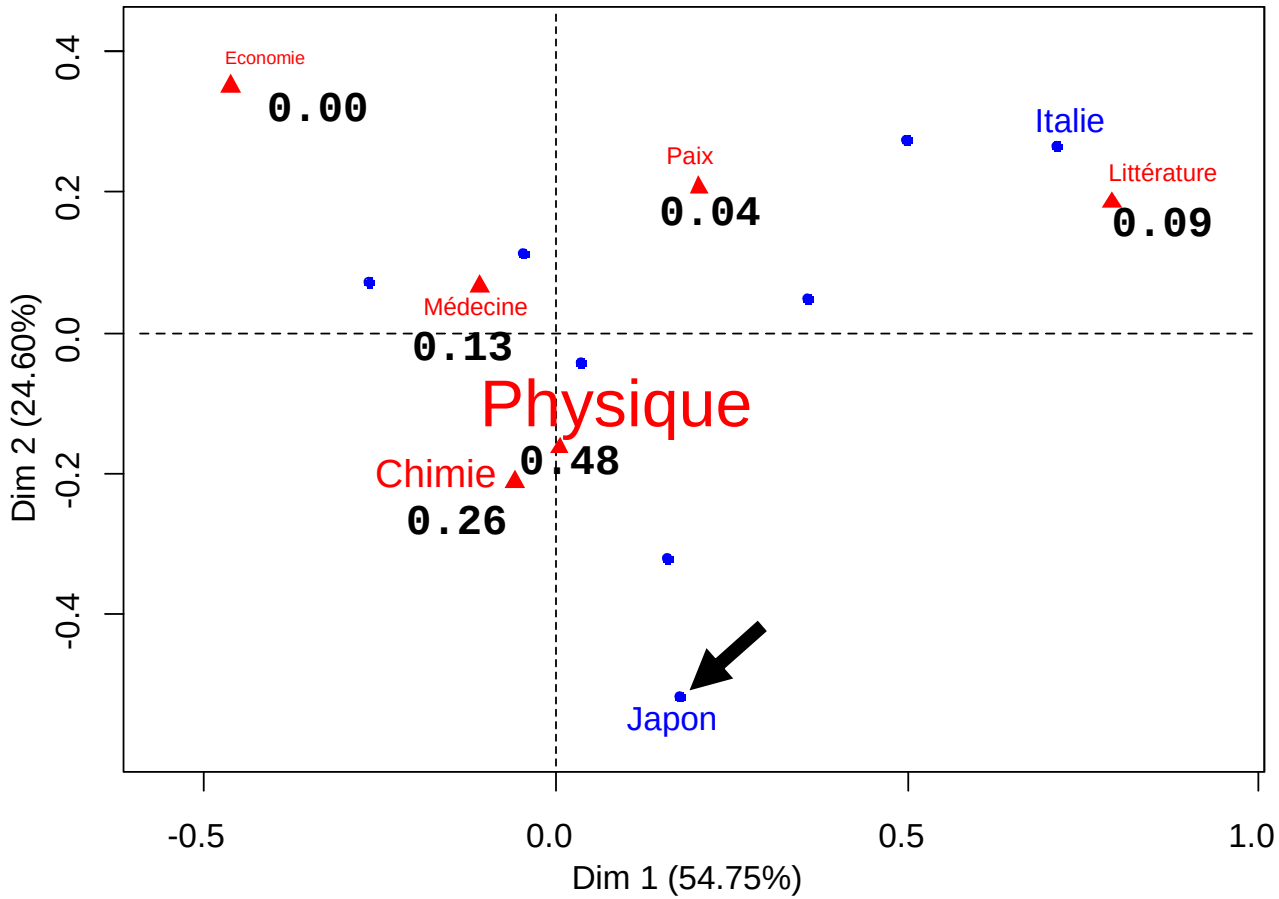
Et symétriquement : $G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} F_s(i)$

Propriété barycentrique



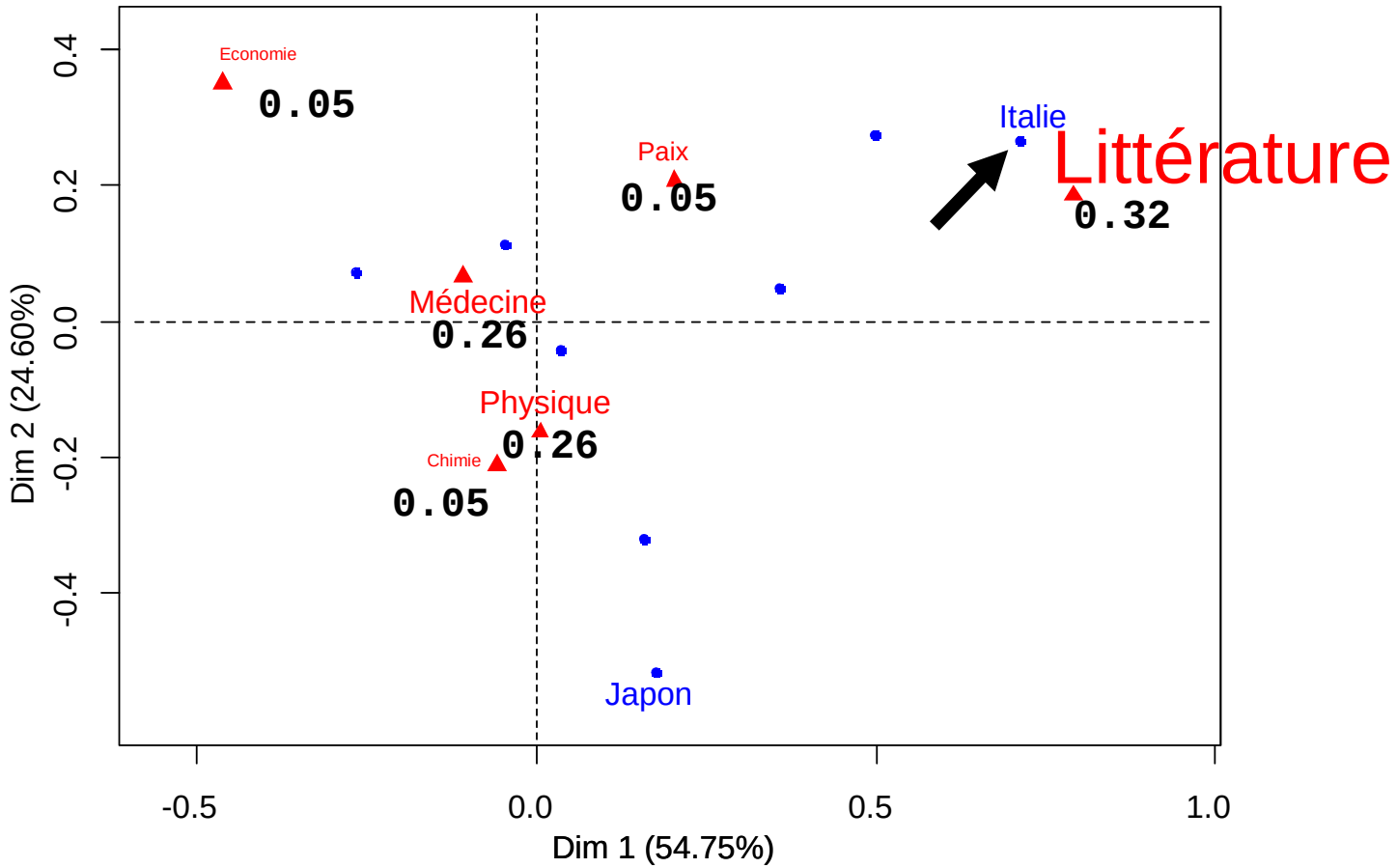
	Chimie	Economie	Littérature	Médecine	Paix	Physique
Italie	5.26	5.26	31.58	26.32	5.26	26.32
Japon	26.09	0.00	8.70	13.04	4.35	47.83
Profil moyen	21.23	10.53	8.60	24.56	8.95	26.14

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

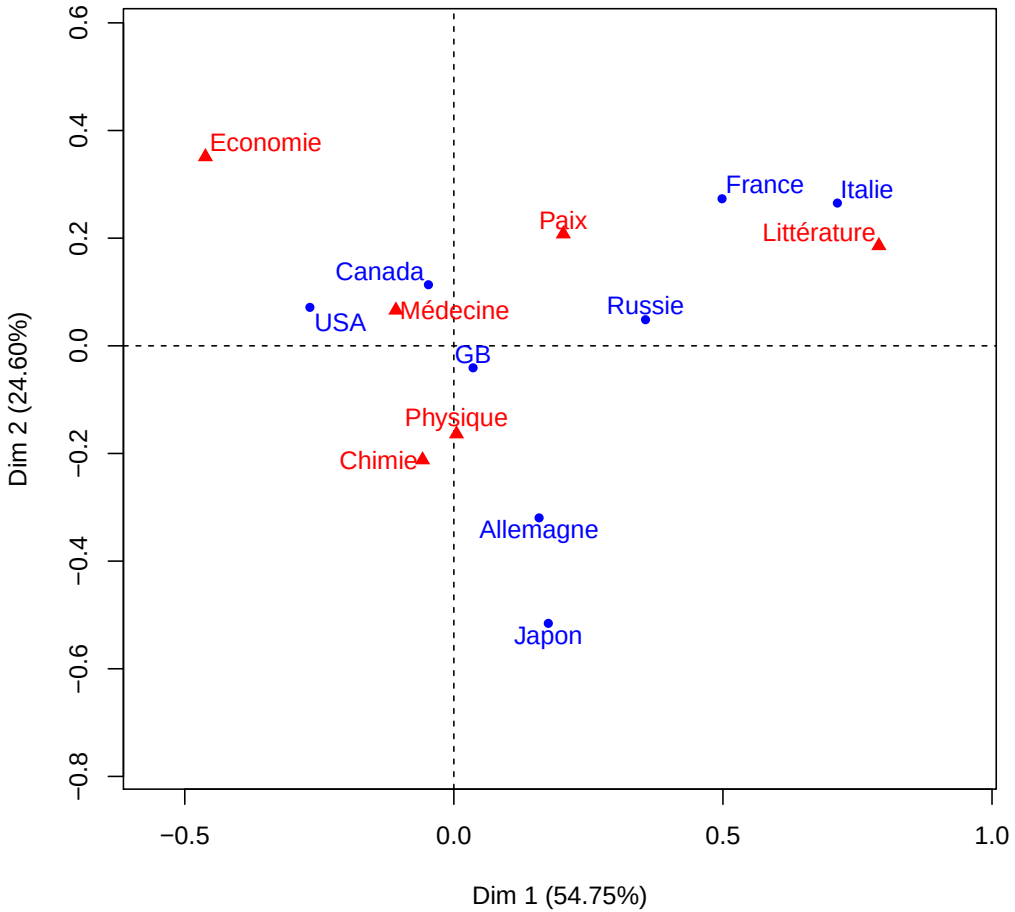


	Chimie	Economie	Littérature	Médecine	Paix	Physique
Italie	5.26	5.26	31.58	26.32	5.26	26.32
Japon	26.09	0.00	8.70	13.04	4.35	47.83
Profil moyen	21.23	10.53	8.60	24.56	8.95	26.14

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
  - Ligne i du côté des colonnes avec lesquelles elle s'associe le plus (et opposé aux colonnes avec lesquelles elle s'associe le moins)
  - Colonne j du côté des lignes avec lesquelles elle s'associe le plus (et à l'opposé des lignes avec lesquelles elle s'associe le moins)

Interprétation sur l'exemple

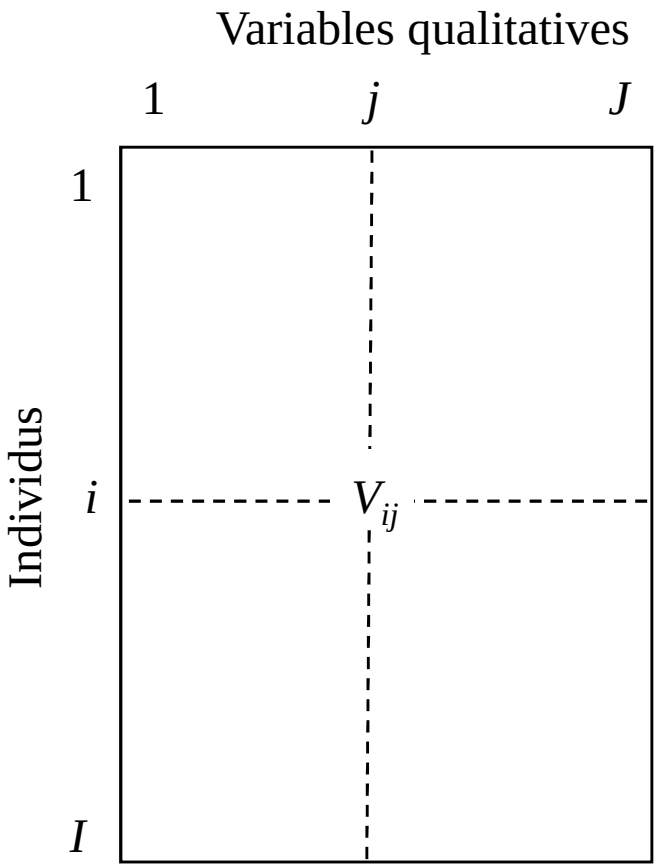


- opposition sciences - autres
  dans une moindre mesure, opposition physique/chimie - science économique
- positions des pays illustrent leur spécificité dans l'obtention des prix Nobel

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

Plan

Les données

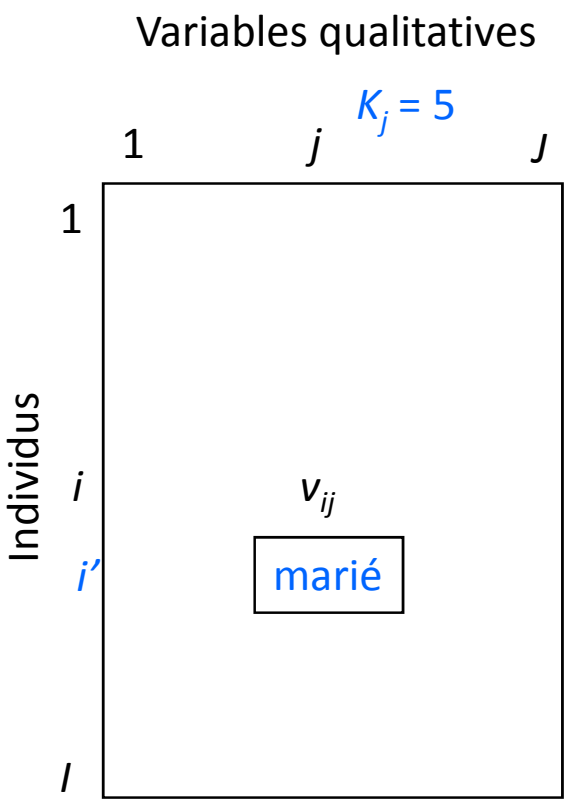


I individus
J variables qualitatives

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



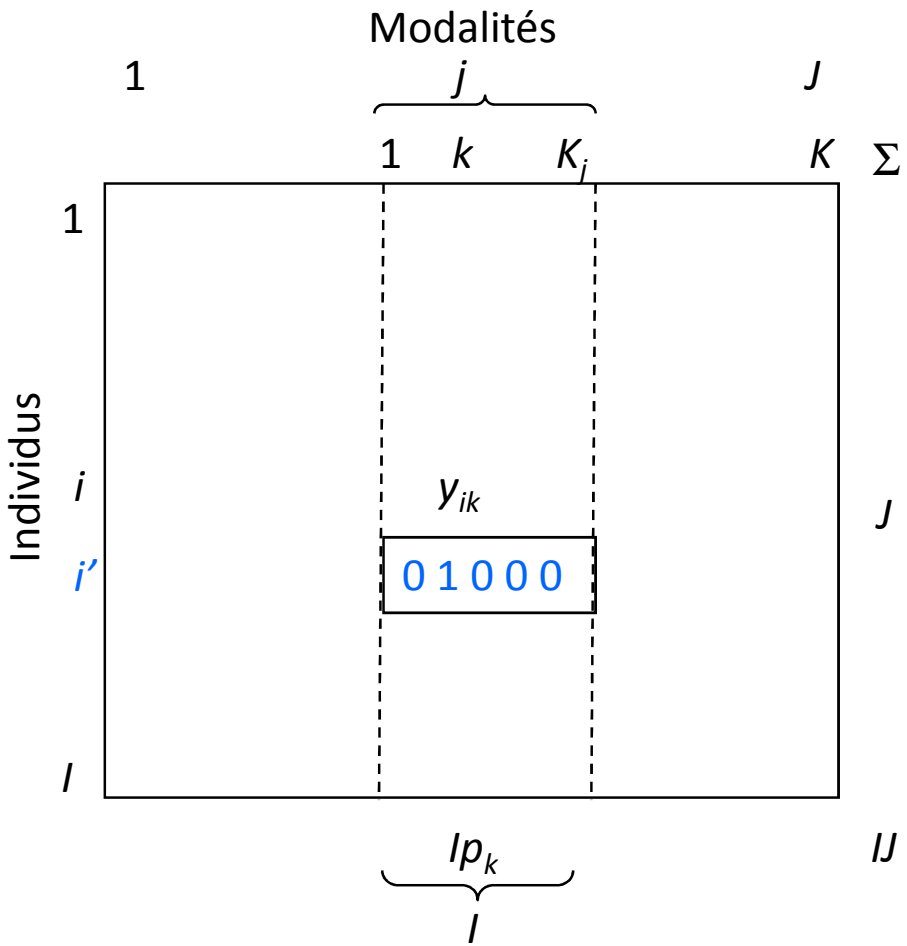


Tableau disjonctif complet (TDC)

Objectifs – problématique

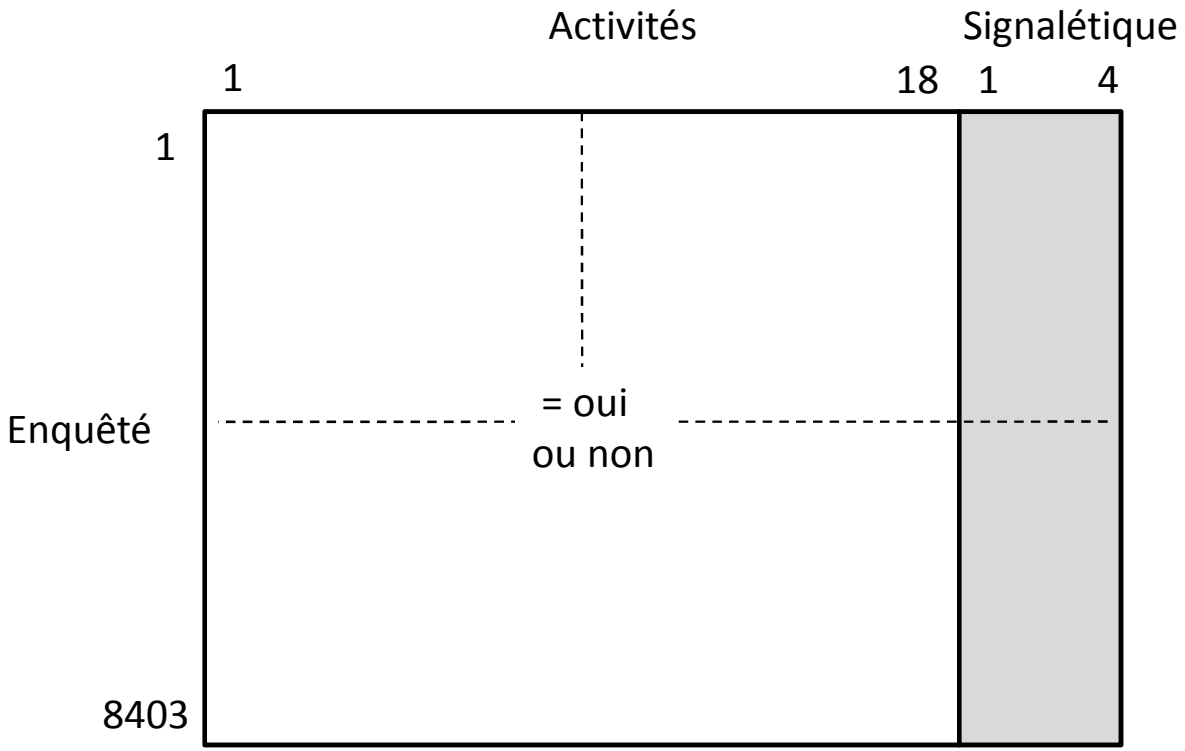
1. Etude des individus
   - Un individu = une ligne du TDC = ensemble de ses modalités
   - Ressemblance des individus Variabilité des individus
   - Principales dimensions de la variabilité des individus
     (en relation avec les modalités)
2. Etude des variables
   - Liaisons entre variables qualitatives
     (en relation avec les modalités)
   - Visualisation d'ensemble des associations entre modalités
   - Variable synthétique
     (Indicateur quantitatif fondé sur des variables qualitatives)

$\Rightarrow$ 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

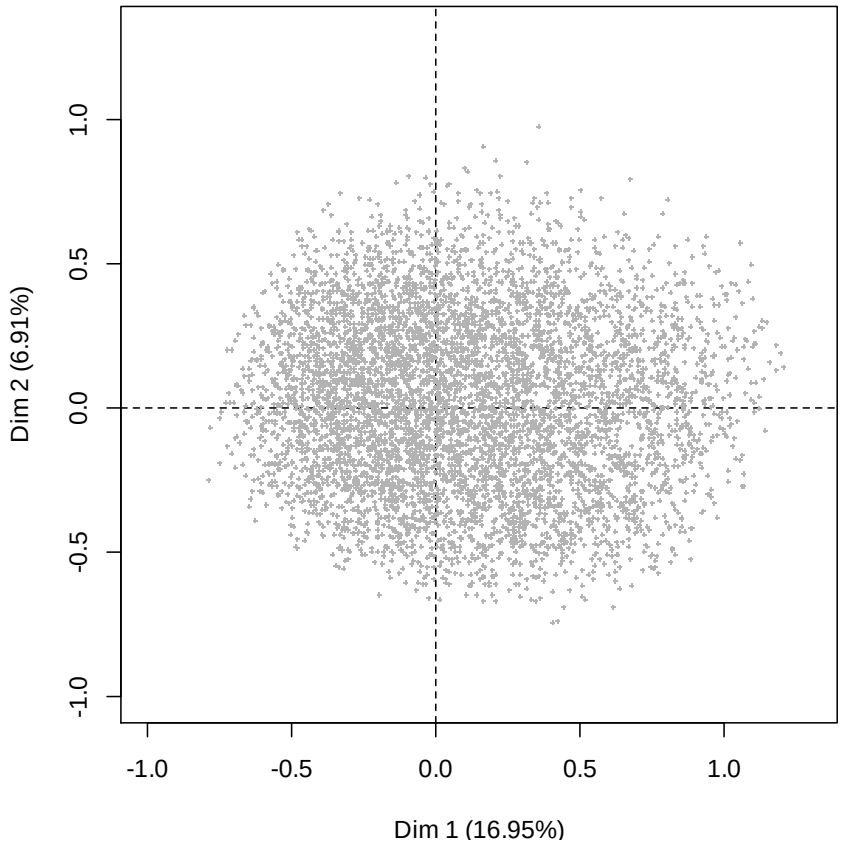
Les données loisirs



ACM : loisirs en actif, signalétique en supplémentaire

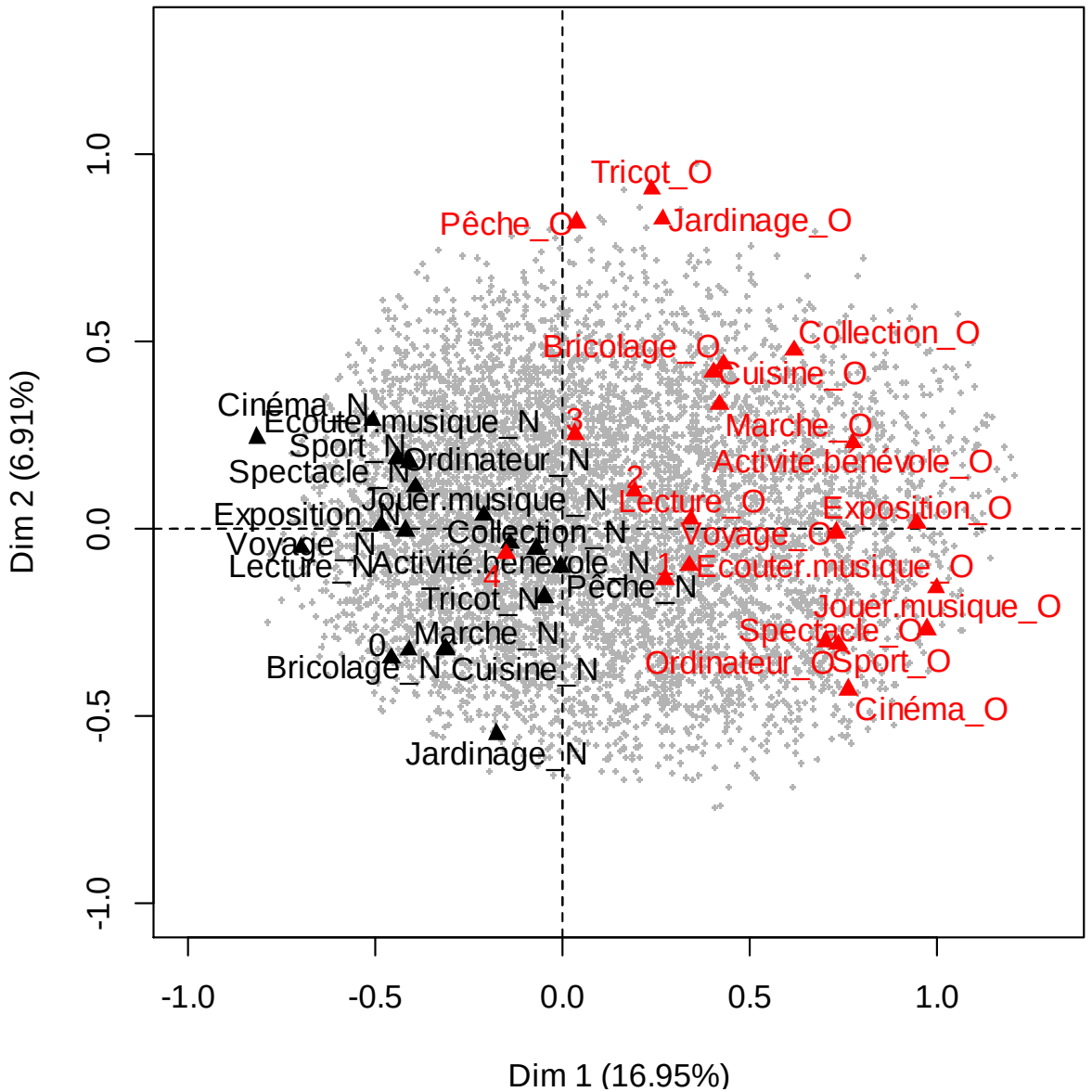
- 1 individu = profil d'activités
- Principales dimensions de variabilité des profils d'activités
- Liaisons entre ces dimensions et le signalétique

Représentation du nuage des individus

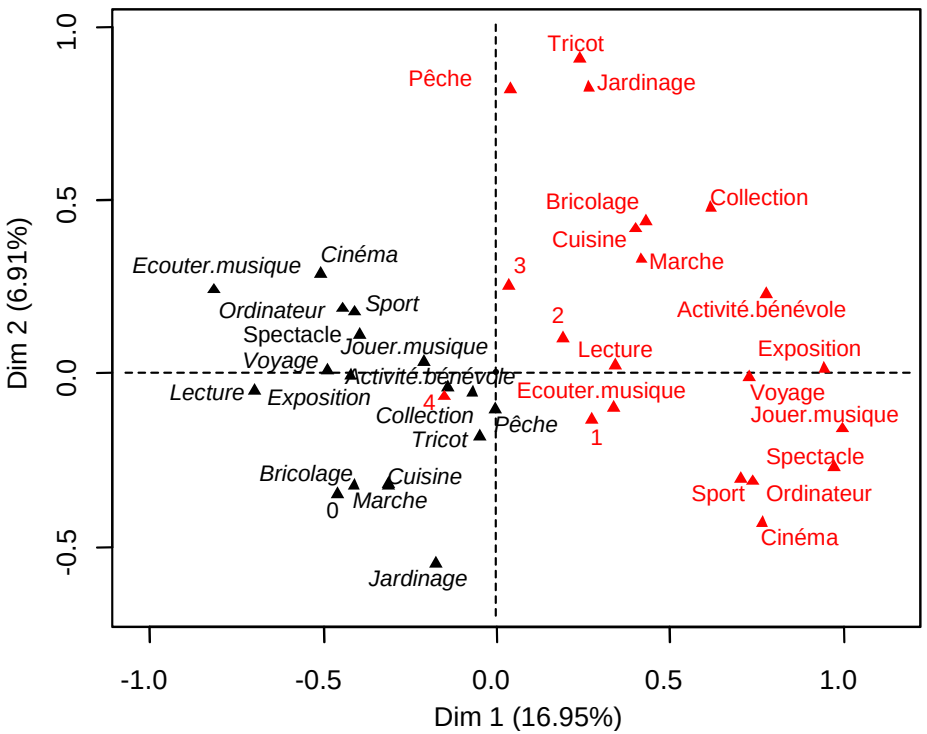


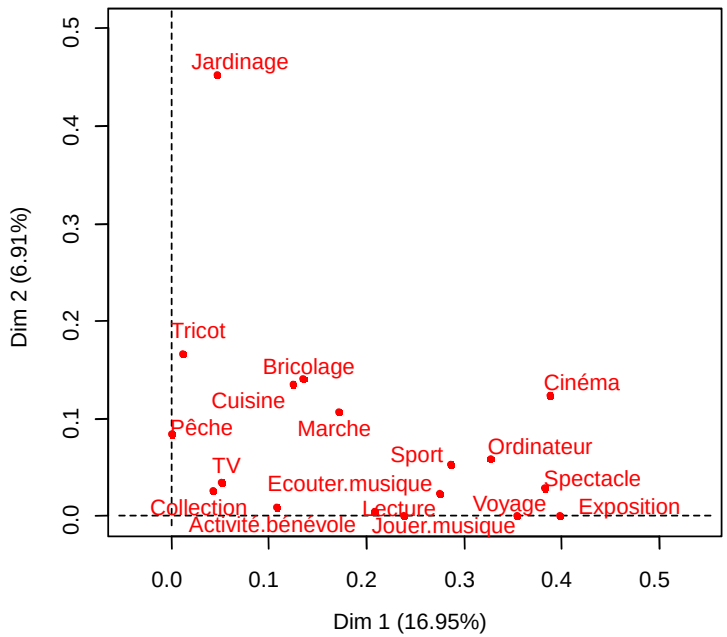
- 2 individus sont superposés s'ils prennent les mêmes modalités
- 2 individus ont en commun beaucoup de modalités : distance petite
- 2 individus dont l'un des 2 possède une modalité rare : distance grande pour prendre en compte la spécificité d'un des 2
- 2 individus ont en commun une modalité rare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

Représentations barycentriques – représentation simultanée



Graphe des modalités et graphe des variables

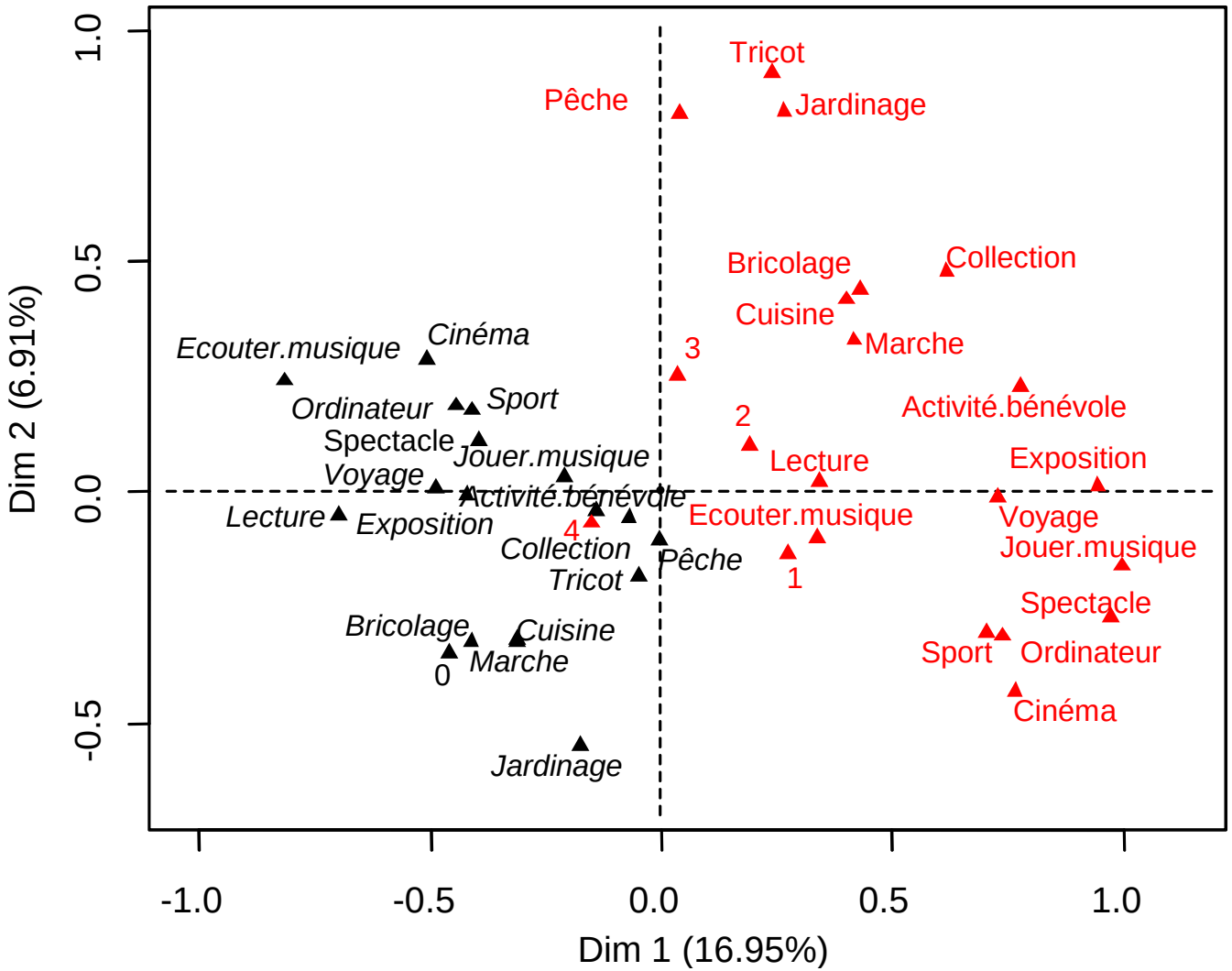




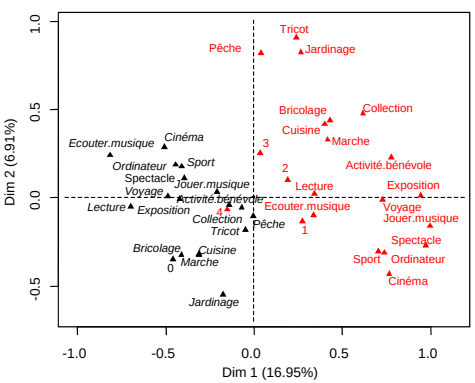
Conclusion

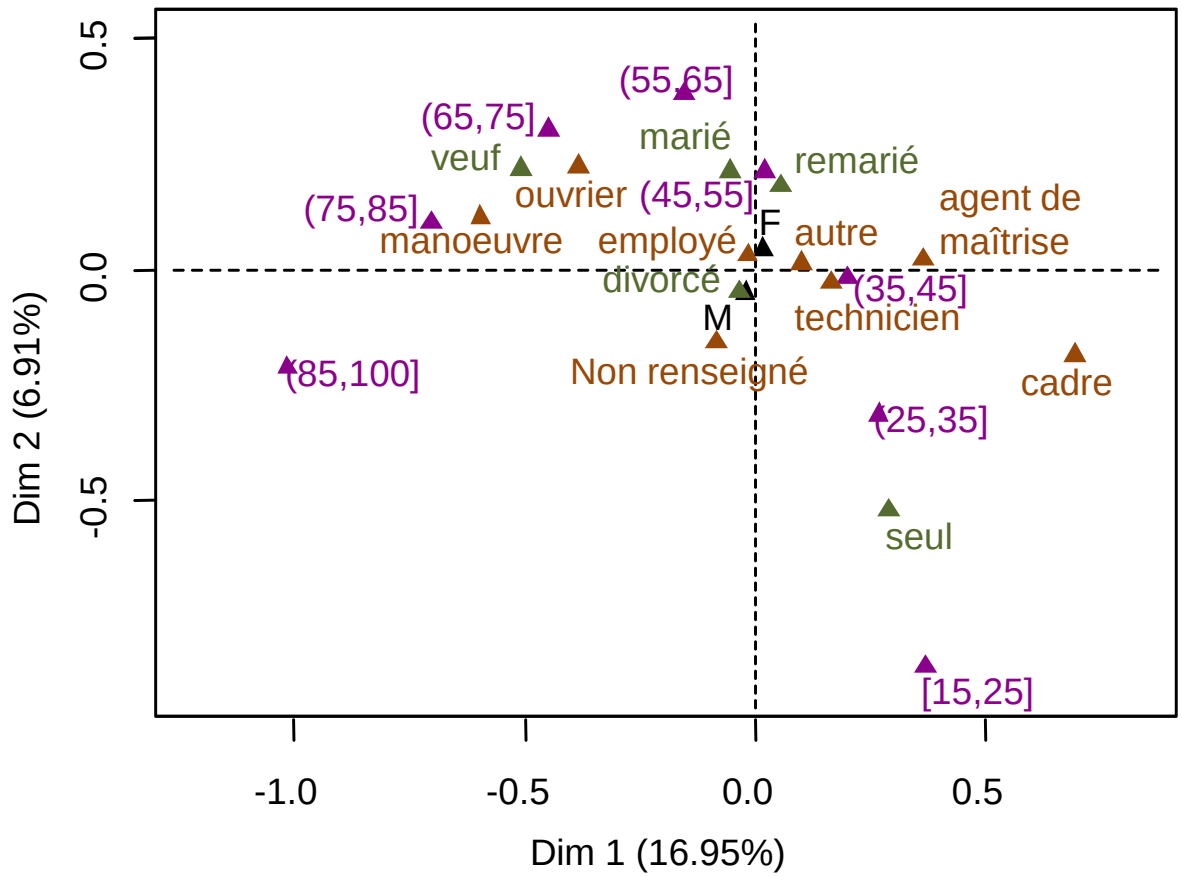
- L'ACM est la méthode factorielle adaptée aux tableaux individus $\times$ variables qualitatives
- Pourcentages d'inertie et qualités de représentation souvent faibles
- Deux dimensions ne suffisent pas à expliquer les fortes variabilité entre individus : il faut souvent interpréter plus de 2 dimensions
- Revenir aux données en analysant des tableaux de contingence par AFC
- L'ACM comme pré-traitement d'une classification

Représentation des modalités



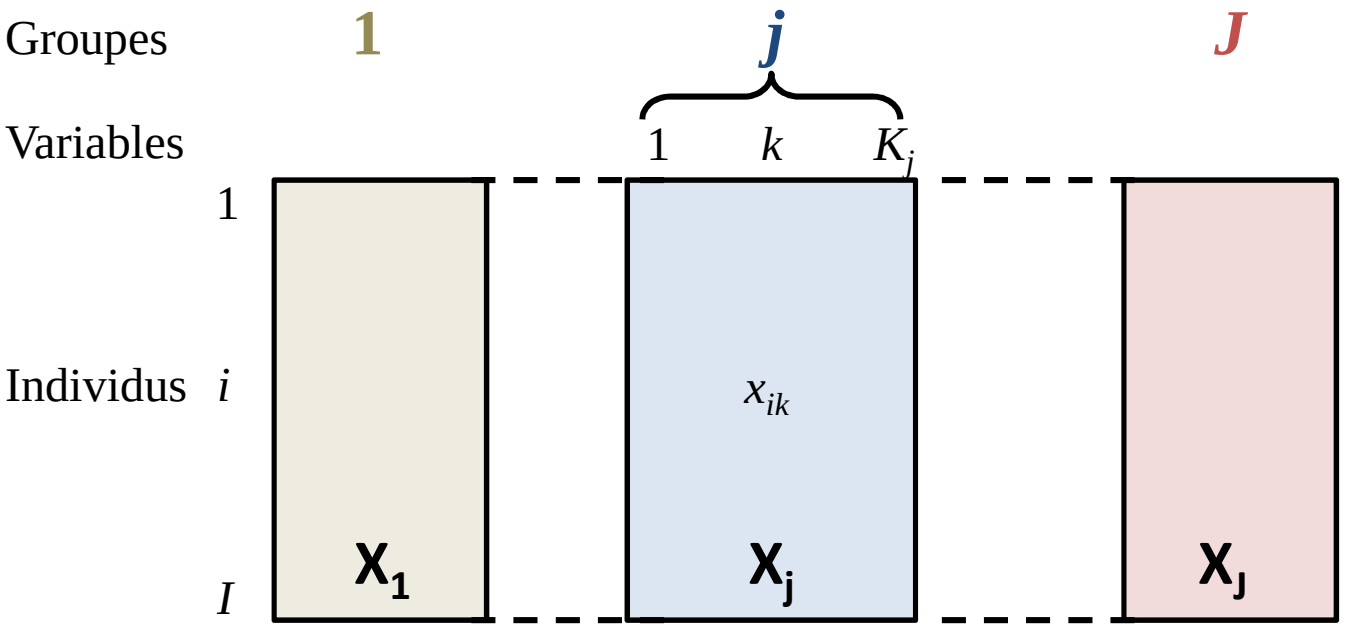
Représentation des modalités





Plan

L'Analyse Factorielle Multiple (AFM)



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

- enquête *mieux vivre* par pays (22 indicateurs de 5 domaines)
- tableau pays $\times$ indicateurs économique, sur plusieurs années
- questionnaire avec échelles de likert et questions qualitatives
- analyse textuelle d'un mouvement social par les journaux, à plusieurs dates

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)	Appréciation (60)	Cépage (1)
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 ?
- Peut-on compare les typologies des vins d'un jury à l'autre ?

Objectifs de l'AFM

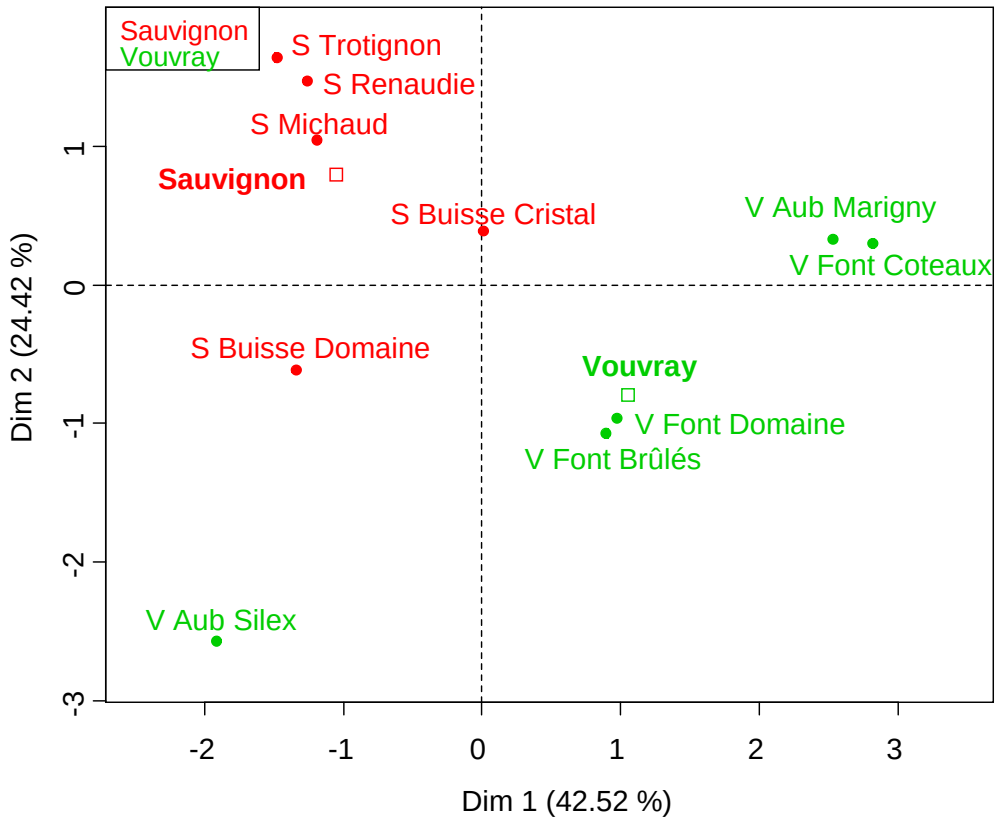
- Etudier les ressemblances entre individus du point de vue de l'ensemble des variables ET les relations entre variables

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

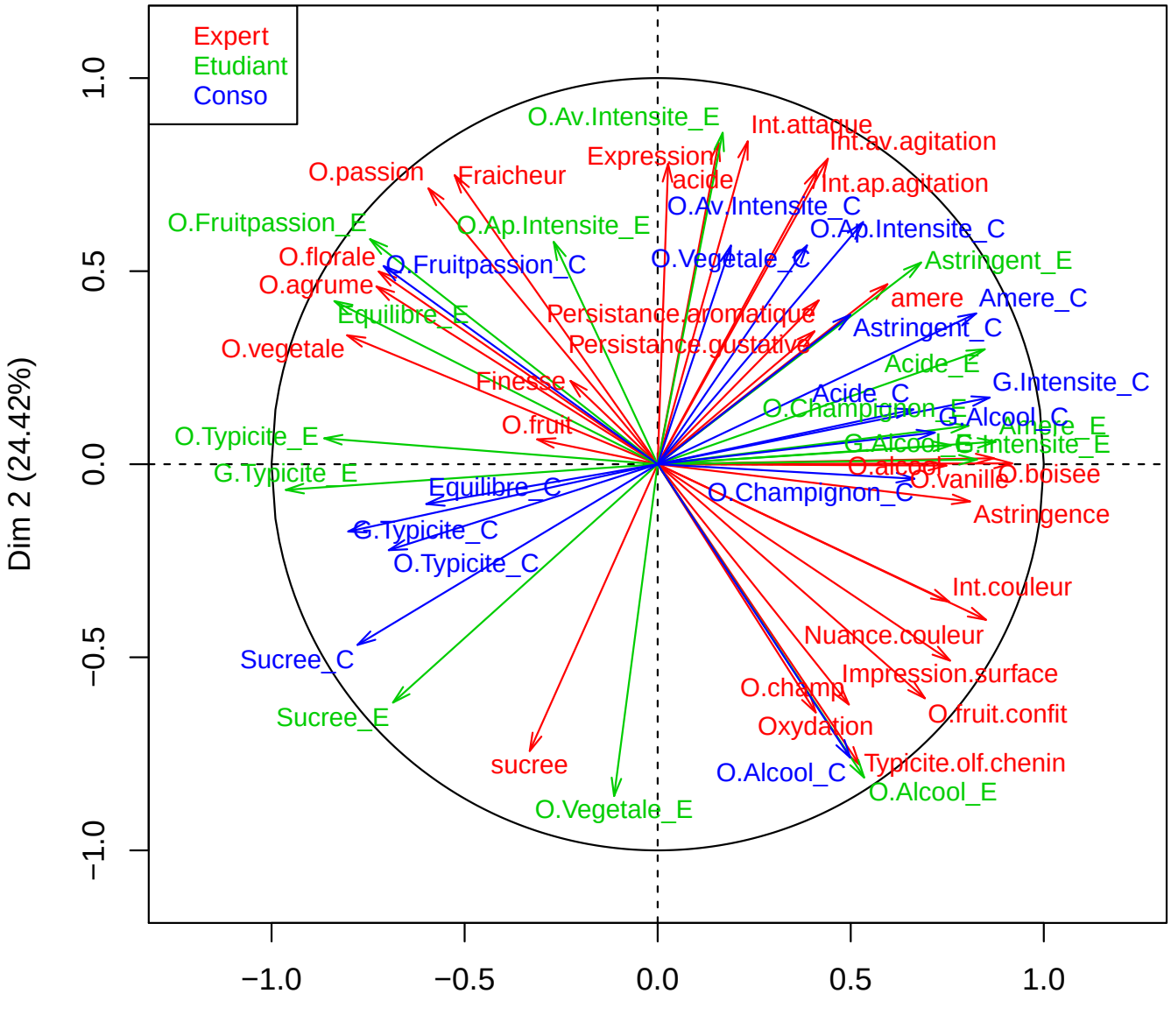
$\Rightarrow$ Equilibrer l'influence de chaque groupe dans l'analyse

Représentation des individus

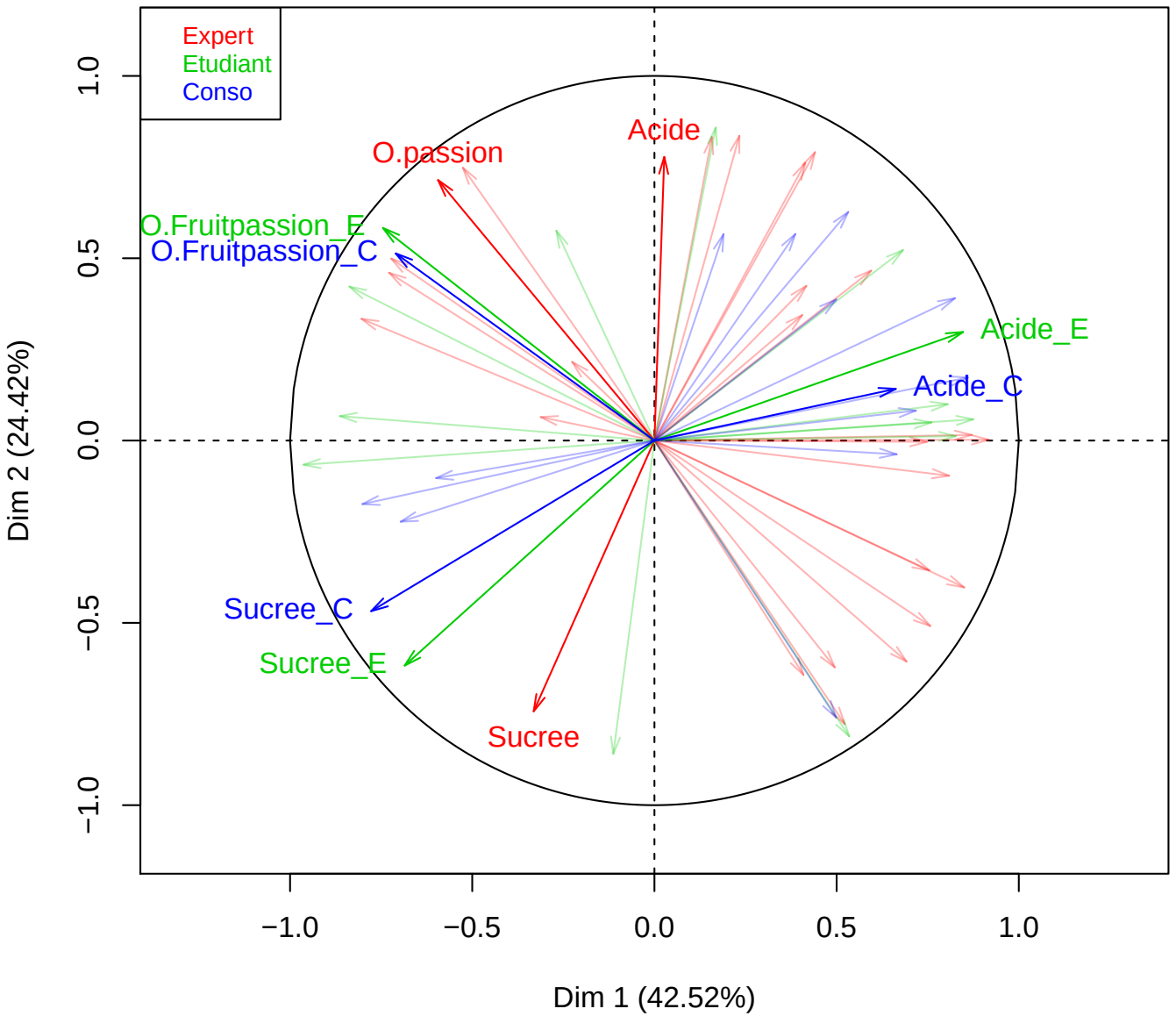


- Les deux cépages sont bien séparés
- Les Vouvray sont plus différents du point de vue sensoriel
- Plusieurs groupes de vins, ...

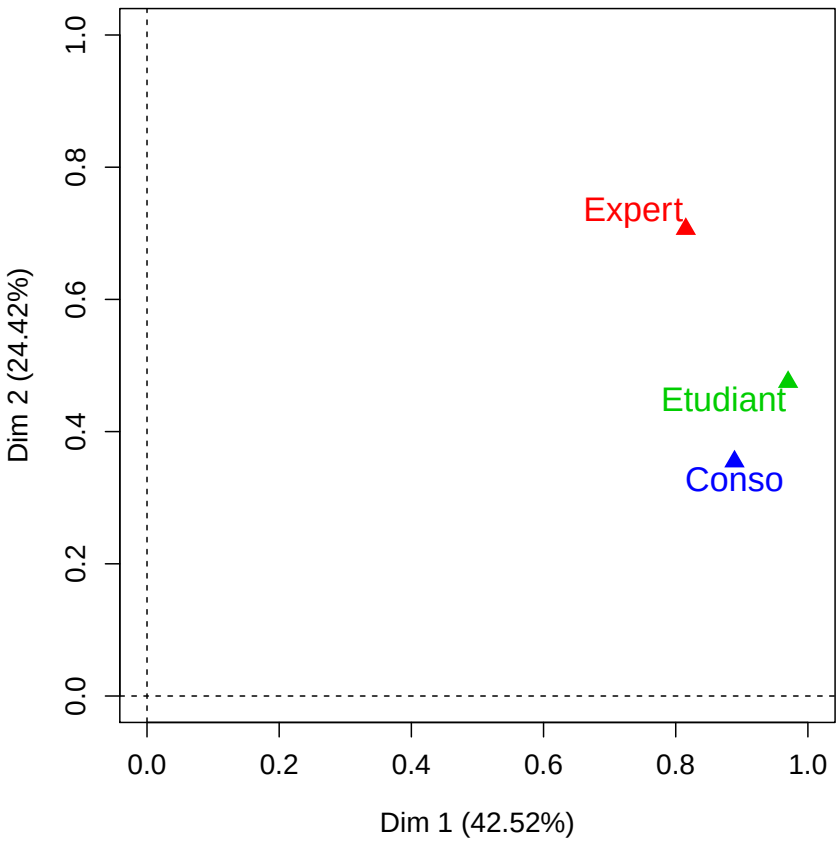
Représentation des variables



Représentation des variables



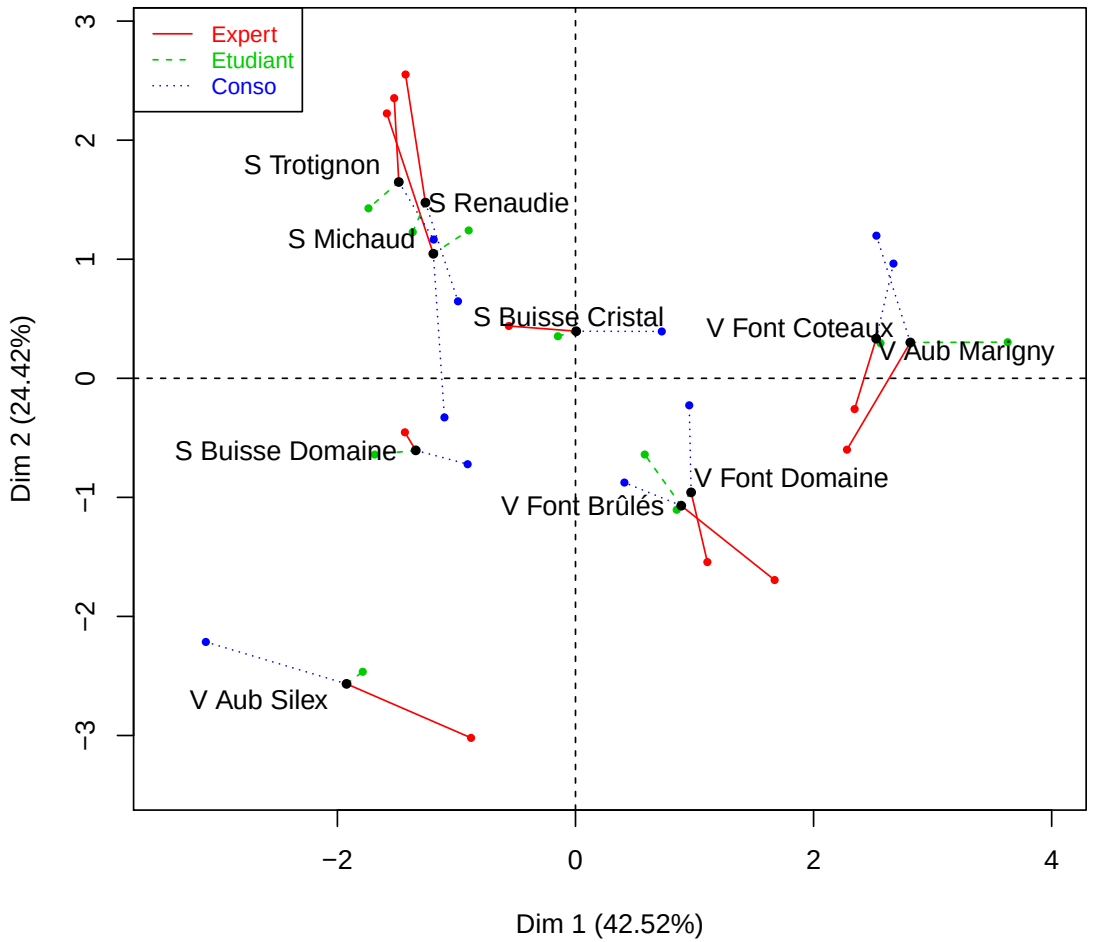
Représentation des groupes



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

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

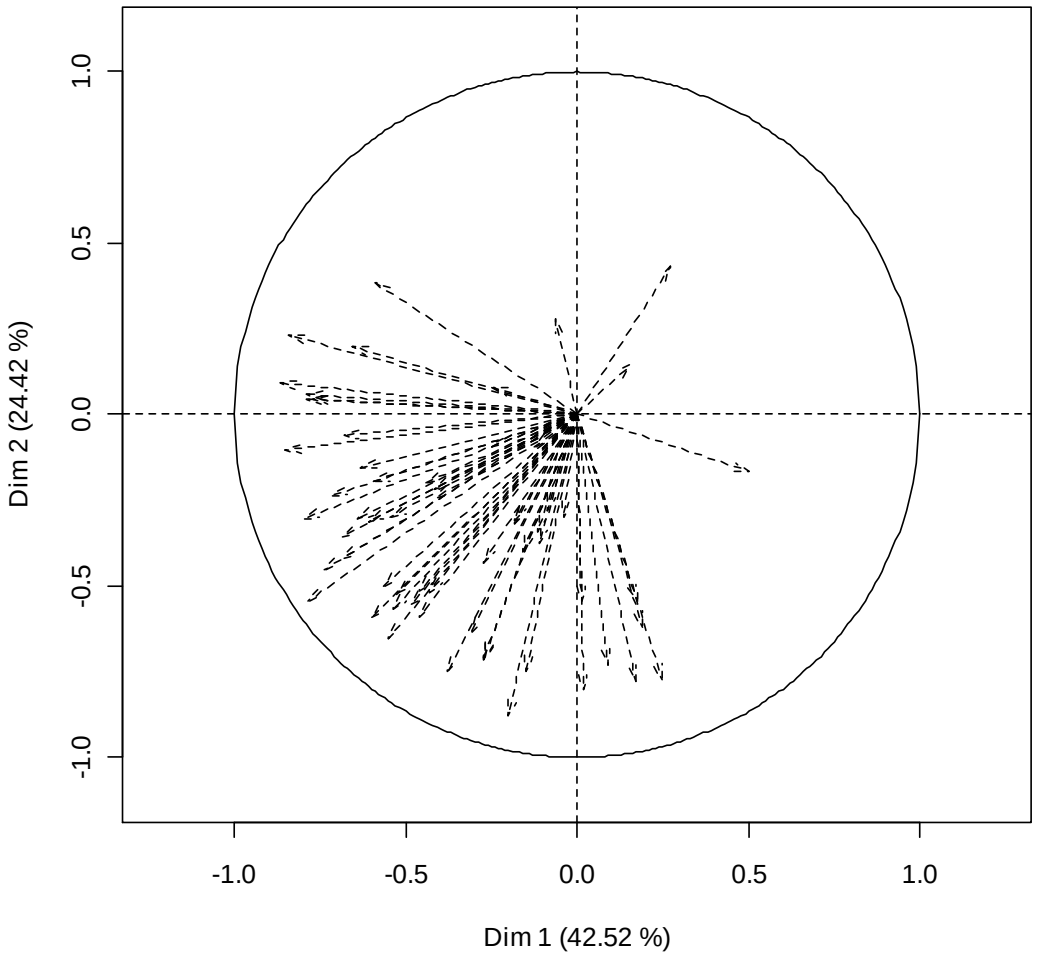
Représentation des points partiels



- 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

Représentation de variables supplémentaires

Le vin préféré est *Vouvray Aubussière Silex*



Les préférences sont liées à la description sensorielle

Mise en œuvre d'une AFM

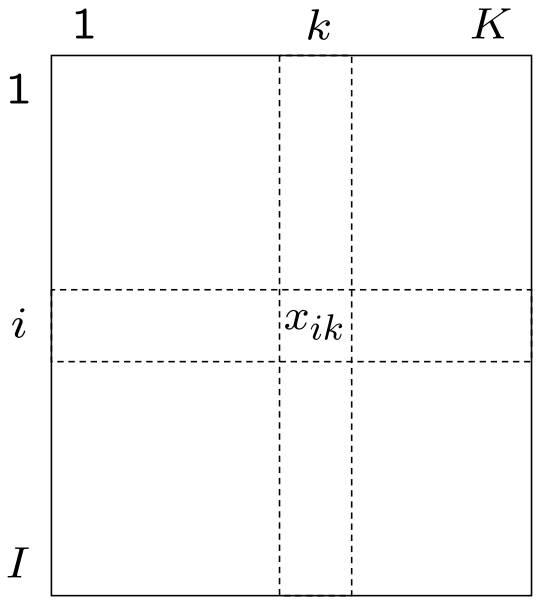
1. Définir la composition des groupes (la structure du tableau)
2. Définir les groupes actifs et les éléments supplémentaires
3. Réduire ou non les variables quantitatives ?
4. Réaliser l'AFM
5. Choisir le nombre de dimensions à interpréter
6. Interpréter simultanément le graphe des individus et des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

Fonction `MFA` du package `FactoMineR`

Plan

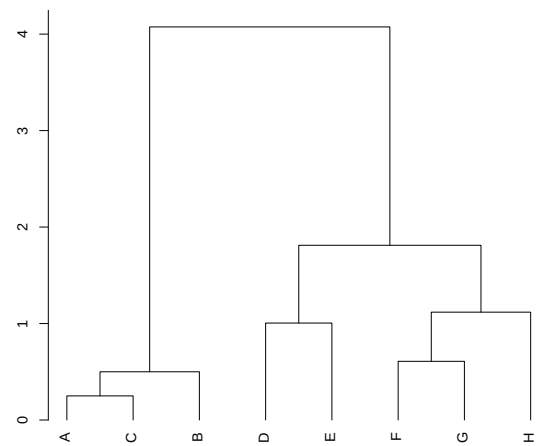
Quelles données pour quels objectifs ?

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



Algorithme

7e regroupement

6e regroupement

5e regroupement

4e regroupement

3e regroupement

2e regroupement

```
        ABC
DEFGH 4.07
```

```
        ABC    DE
DE     4.72
FGH    4.07  1.81
```

```
        ABC    DE    FG
DE     4.72
FG     4.23  1.81
H      4.07  2.90  1.12
```

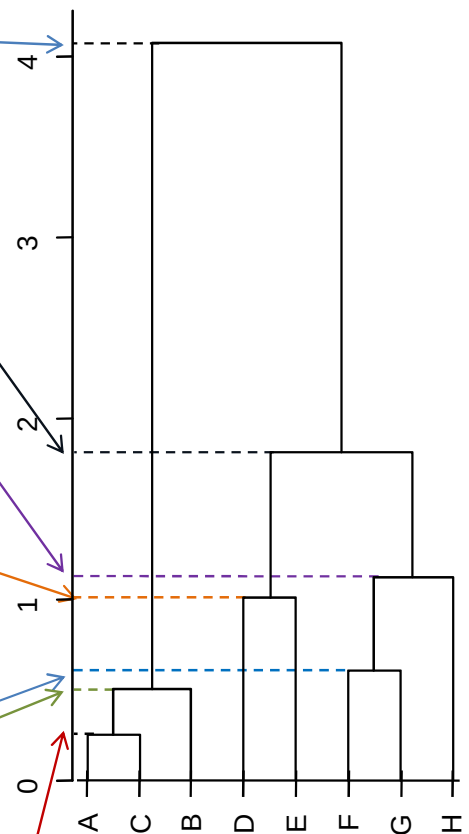
```
        ABC     D     E    FG
D      4.72
E      5.55  1.00
FG     4.07  2.01  1.81
H      4.75  3.16  2.90  1.12
```

```
        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
```

```
         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
```

1er regroupement



{ABCDEFGH}

{ABC},{DEFGH}

{ABC},{DE},{FGH}

{ABC},{DE},{FG},{H}

{ABC},{D},{E},{FG},{H}

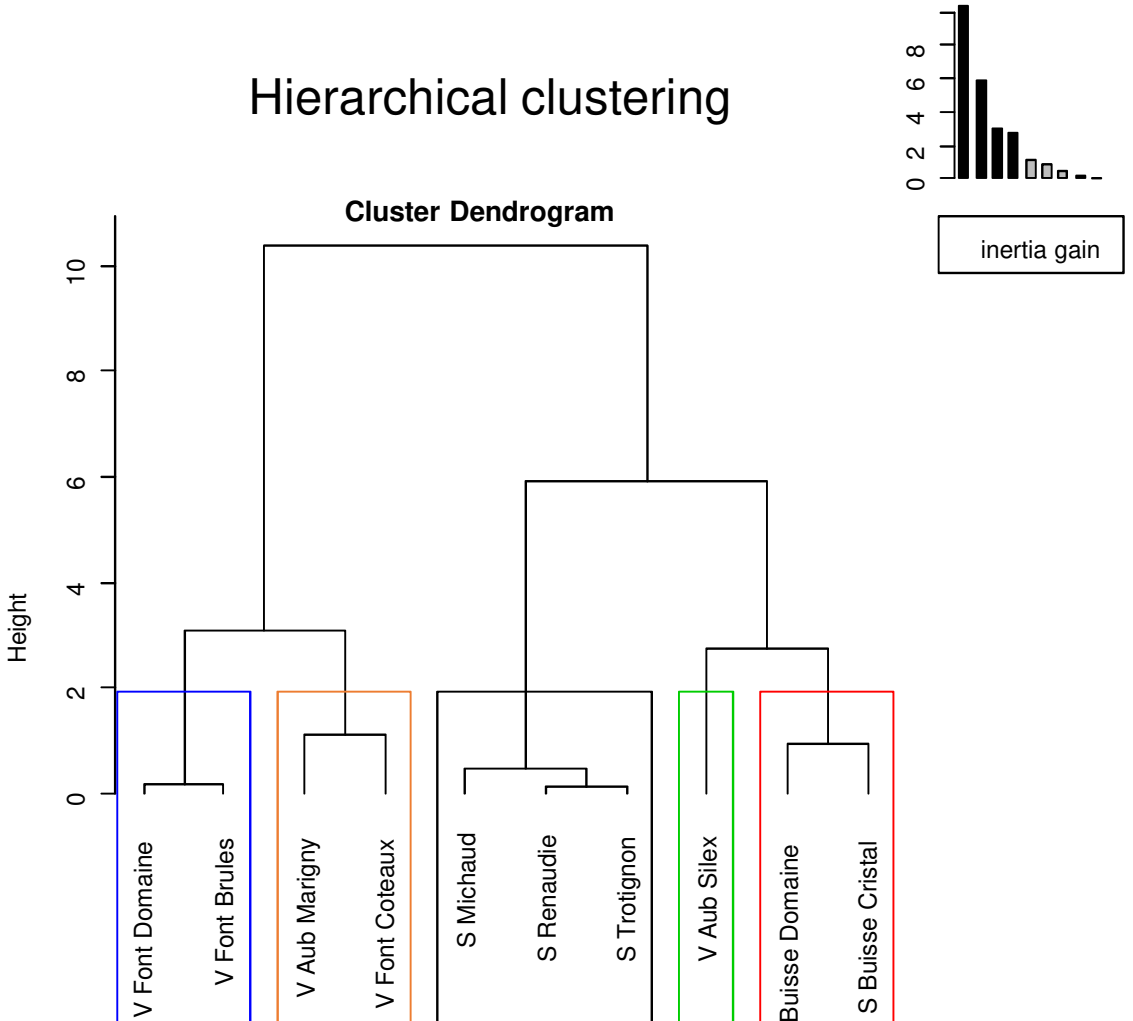
{ABC},{D},{E},{F},{G},{H}

{AC},{B},{D},{E},{F},{G},{H}

{A},{B},{C},{D},{E},{F},{G},{H}

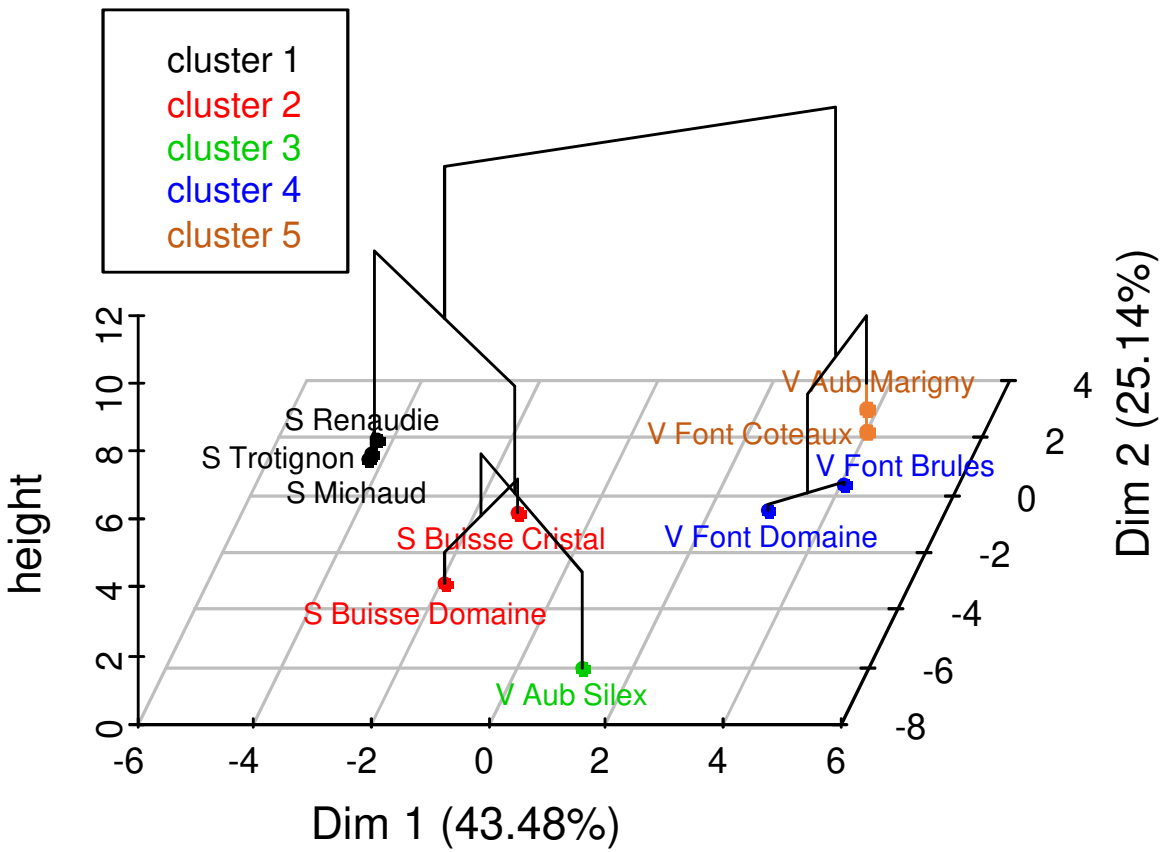
Classification Ascendante Hiérarchique (CAH)

- peut-on faire des classes d'individus qui se ressemblent ?
- comment décrire ces classes ?



Classification et plan factoriel

Représentation de l'arbre et des classes sur un plan factoriel



Plan

Démarche en analyse de données

Démarche en analyse des données `FactoMineR`

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

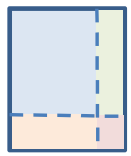
oui ➔ AFM

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. Eléments 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`

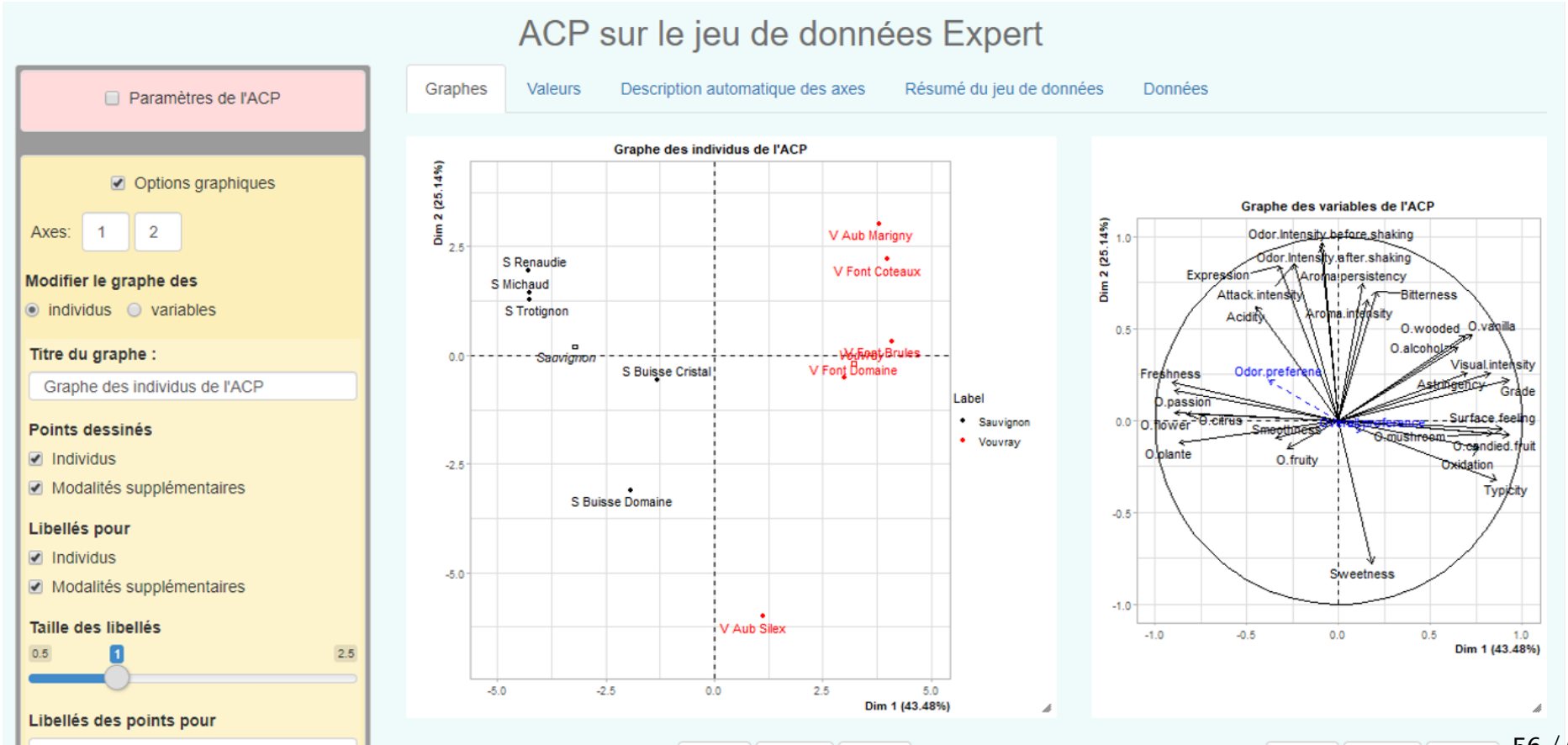
Décrire les axes factoriels par les variables initiales `dimdesc`

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

Graphiques interactifs avec le package Factoshiny

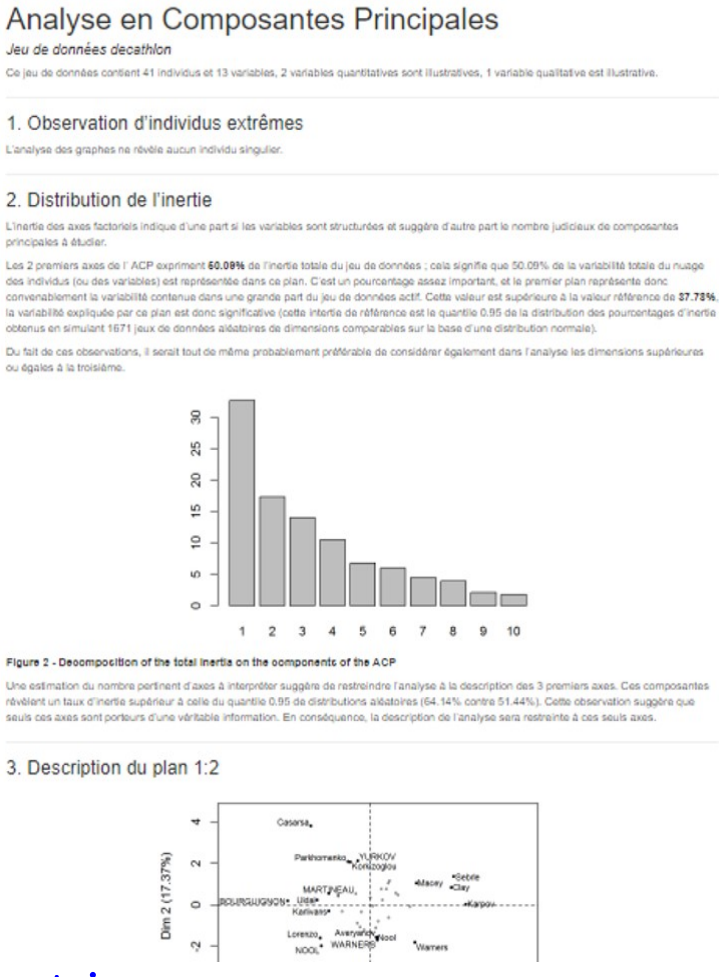
- Réaliser des analyses sans besoin de maîtriser le code
- Visualisation en temps réel des modifications apportées

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



Rapport automatisé avec le package FactoInvestigate

Propose une interprétation des résultats basée sur l'objet résultat



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 axes factoriels 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 expriment **50.09%** de l'inertie totale du jeu de données ; cela signifie que 50.09% 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 actif. Cette valeur est supérieure à la valeur référence de **37.78%**, la variabilité expliquée par ce plan est donc significative (cette intertie de référence est le quantile 0.95 de la distribution des pourcentages d'inertie obtenus en simulant 1671 jeux de données aléatoires de dimensions comparables sur 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 - Decomposition of the total inertia 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 quantile 0.95 de distributions aléatoires (64.14% 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

http://factominer.free.fr/reporting

Matériel sur FactoMineR

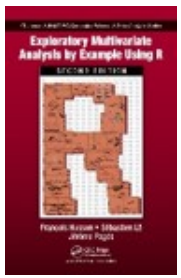
- FactoMineR : pour mettre en œuvre les méthodes
- Factoshiny : pour un menu déroulant et graphes interactifs
- missMDA : pour la gestion des données manquantes
- FactoInvestigate : pour les rapports automatisés

- site FactoMineR : `http://factominer.free.fr`
- site F. Husson : https://husson.github.io

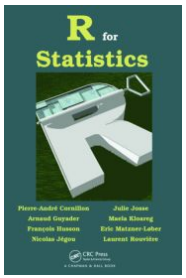
- 2 articles dans J. of stat. software (FactoMineR, missMDA)
- 2 articles dans R journal (CA-galt, MFACT)

Aides à l'utilisateur

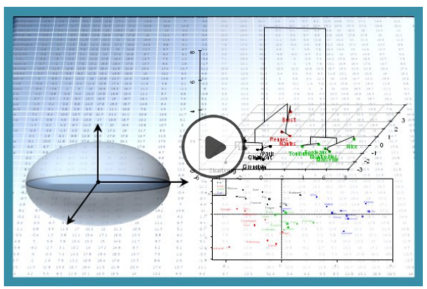
Analyse de données avec R (2e ed)



R pour la stat. et sc. des données



MOOC analyse de données multidimensionnelles



Playlists en analyse de données :

- sur l'ACP, on PCA
- sur l'AFC, on correspondence analysis,
- sur l'ACM, on multiple correspondence analysis (MCA),
- sur la classification, on clustering,
- sur l'AFM, on multiple factor analysis (MFA),
- sur la gestion de données manquantes, on handling missing values

Un exemple en linguistique

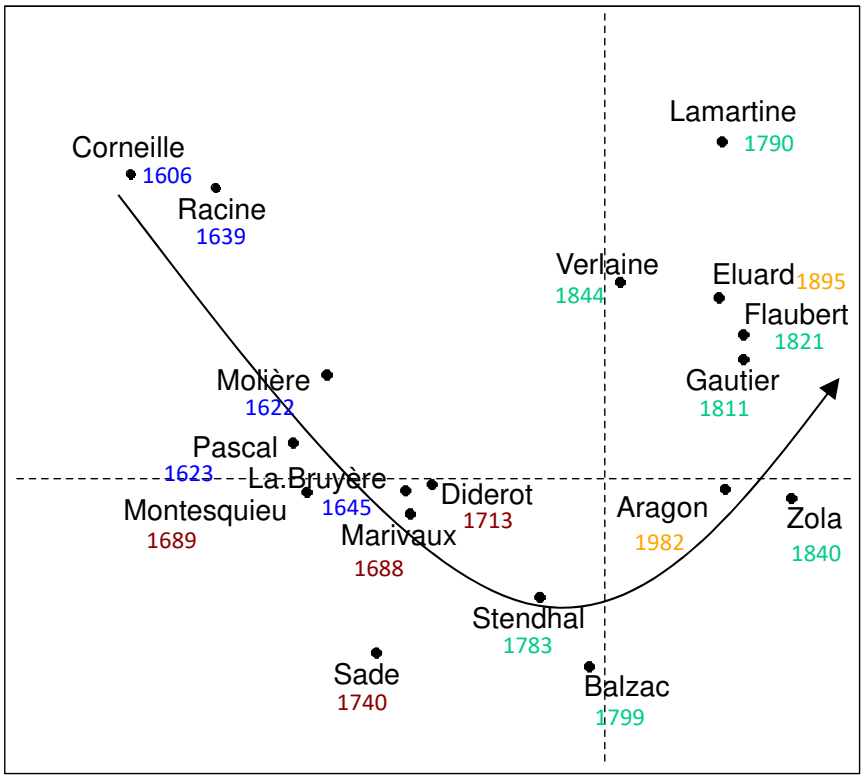
- Aragon (23 textes) : FeuJoie, Perpétuel, Destinées, Snark, Peinture, ...
- Balzac (49 textes) : *Chouans*, *Physiologie*, *Vendetta*, *Gobseck*, ...
- Corneille (34 textes) : *Mélite*, *Clitandre*, *Veuve*, *Gelerie*,*Suivante*, ...
- ...

On conserve les mots cités au moins 100 fois

accord	264	0	88	44	...
affaire	1029	2040	74	154	...
âge	545	629	92	108	
ah	219	0	0	0	
air	2093	2009	95	191	
allemagne	366	0	0	0	
allemand	476	0	0	0	
amant	303	760	566	0	
âme	478	2190	1101	240	
ami	1090	2583	307	407	
amour	1374	3286	1791	167	
an	1812	3009	112	182	
anglais	315	0	0	0	
. . .					

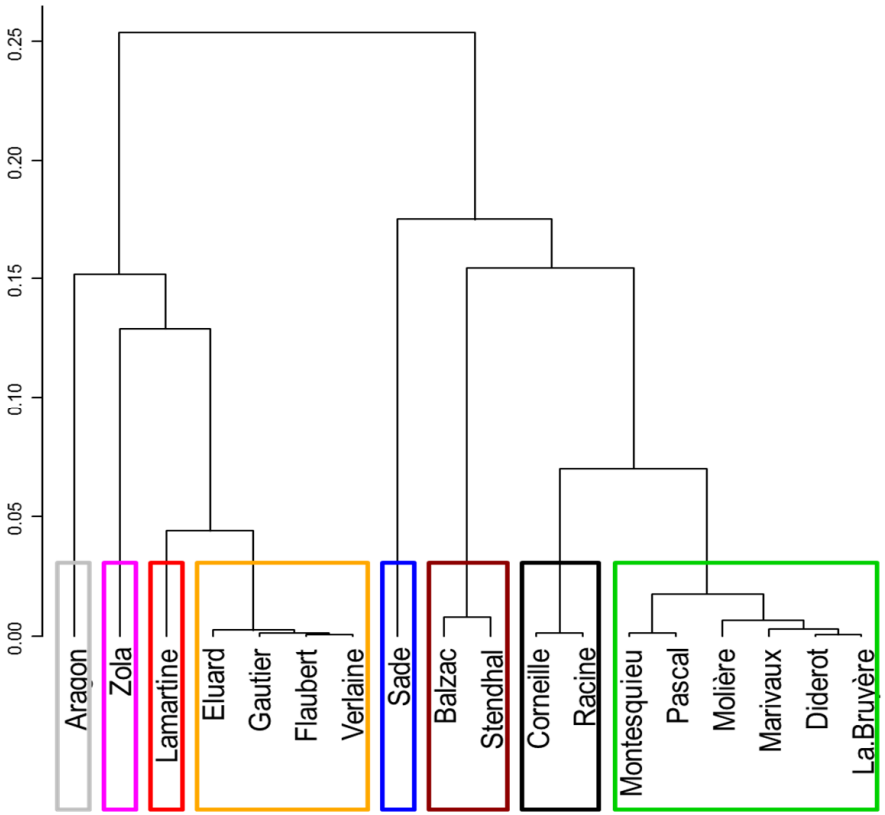
978 mots

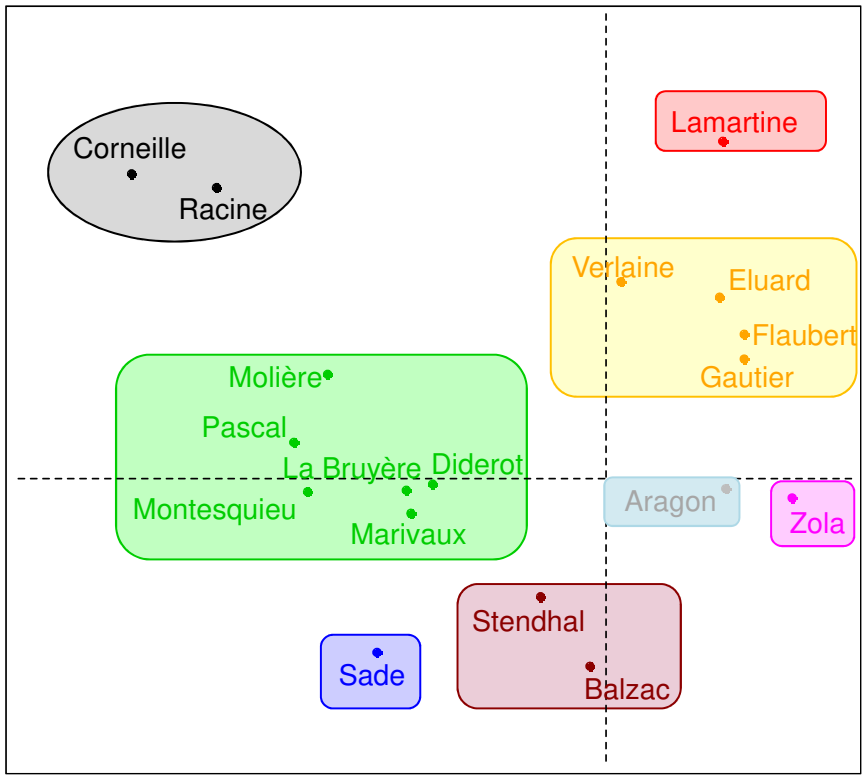
Un exemple en linguistique : l'analyse des correspondances



- Corneille et Racine sont proches et très éloignés de Zola.
  Ce sont 2 auteurs classiques du 17ème tandis que Zola est un naturaliste du 19ème
- Évolution du vocabulaire selon les siècles

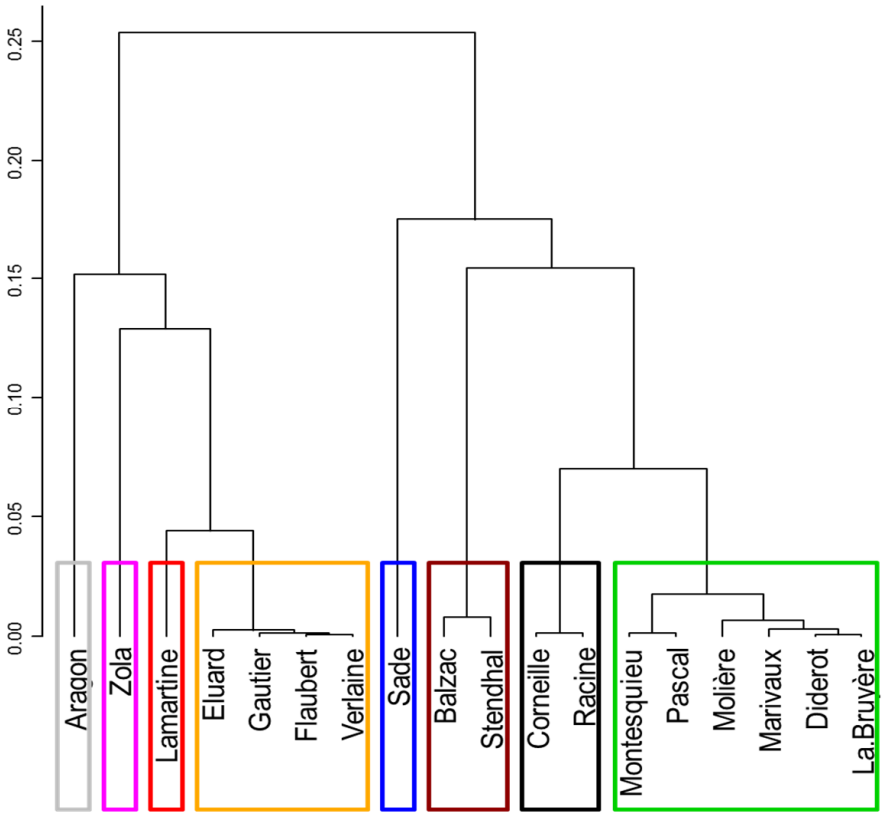
Un exemple en linguistique : caractérisation des classes

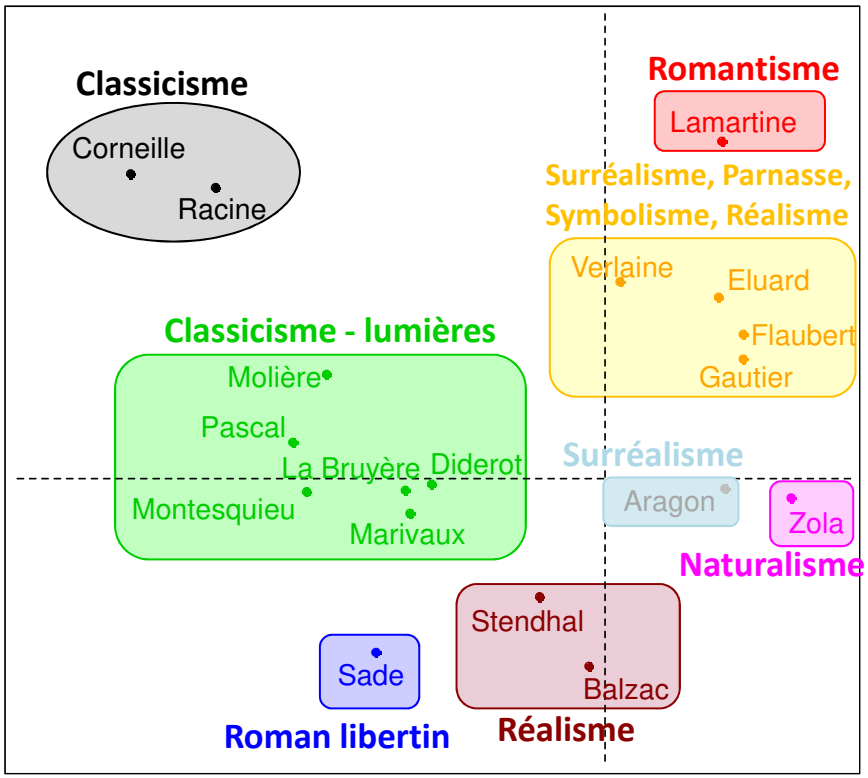




- La classification retrouve des classes d'auteurs connues

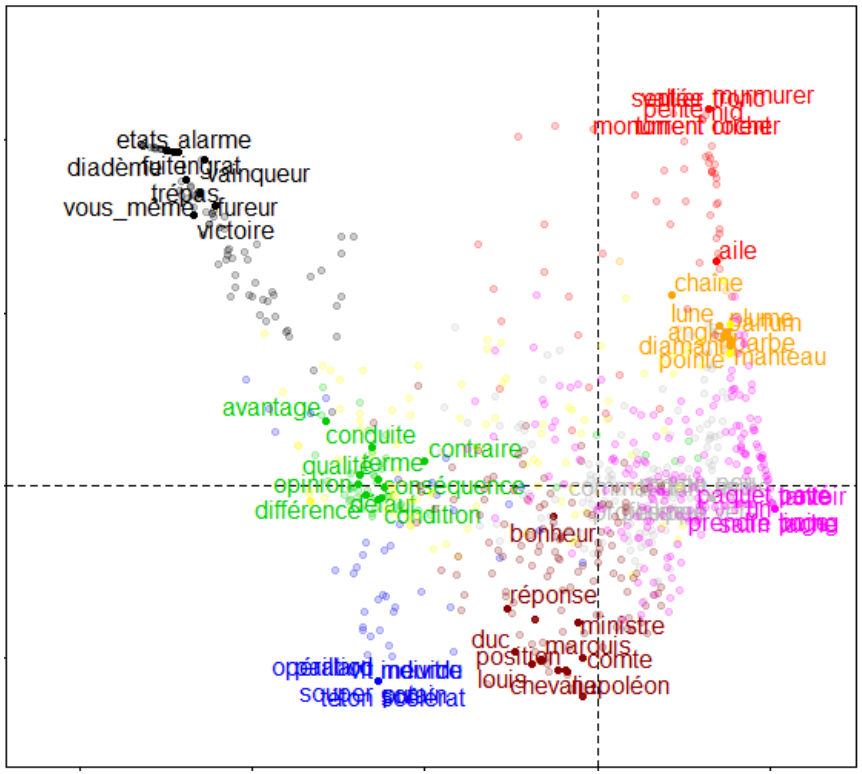
Un exemple en linguistique : caractérisation des classes

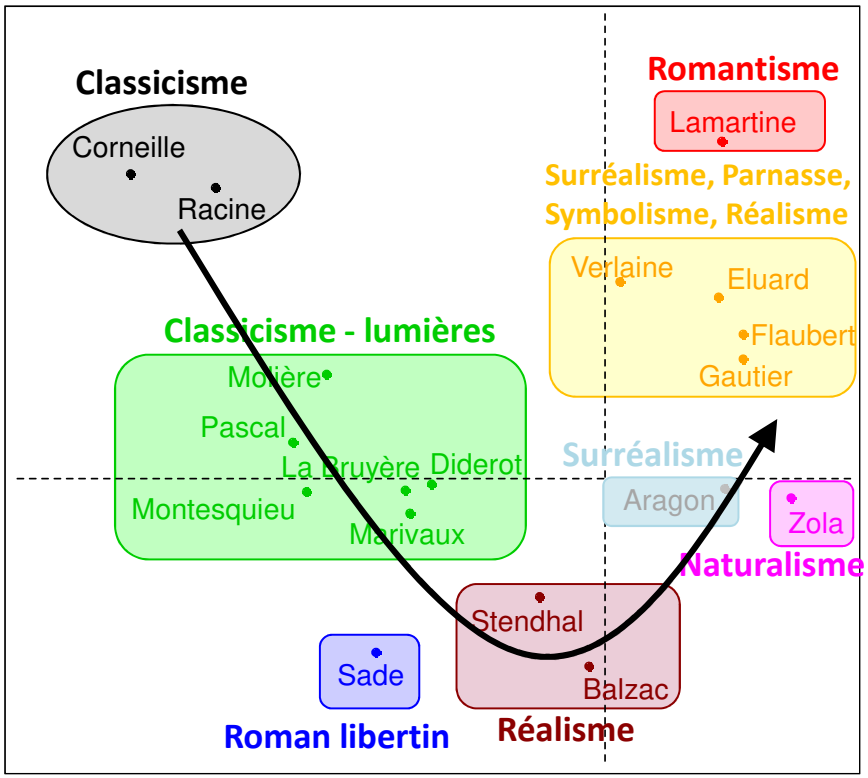




- Stendhal et Balzac (réalistes) sont très éloignés de Lamartine (romantique). On retrouve ici que les auteurs réalistes ont un point commun : s'éloigner des excès romantiques !
- Points communs naturalistes / réalistes : montrer la société telle qu'elle est, le roman devient le miroir de la société

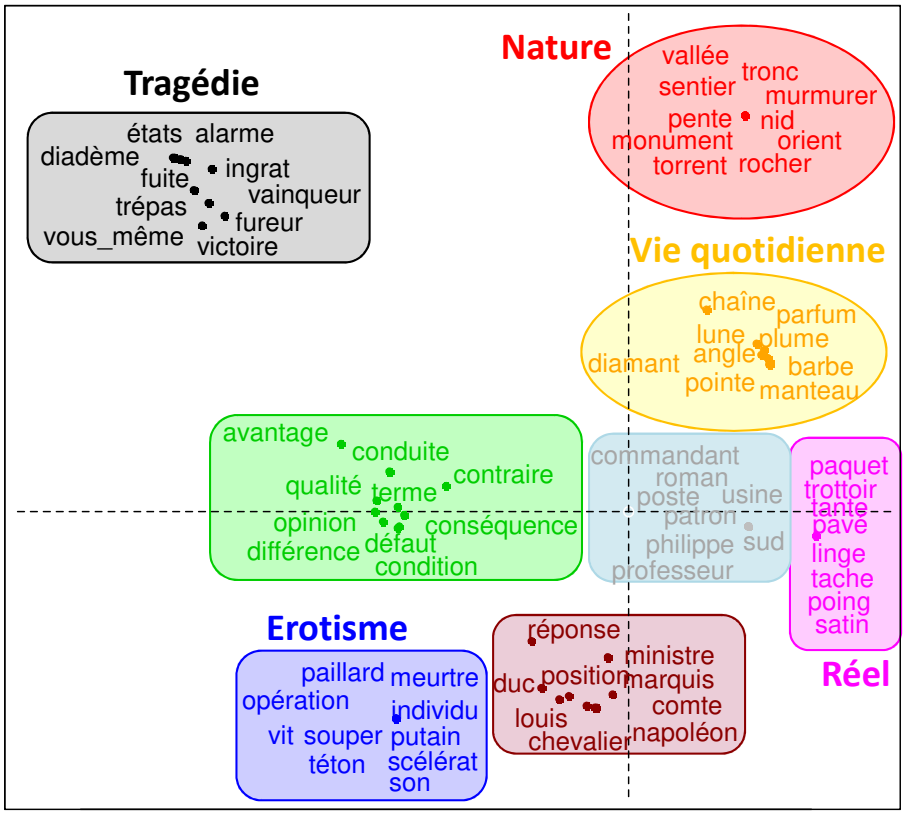
Un exemple en linguistique : interprétation ...

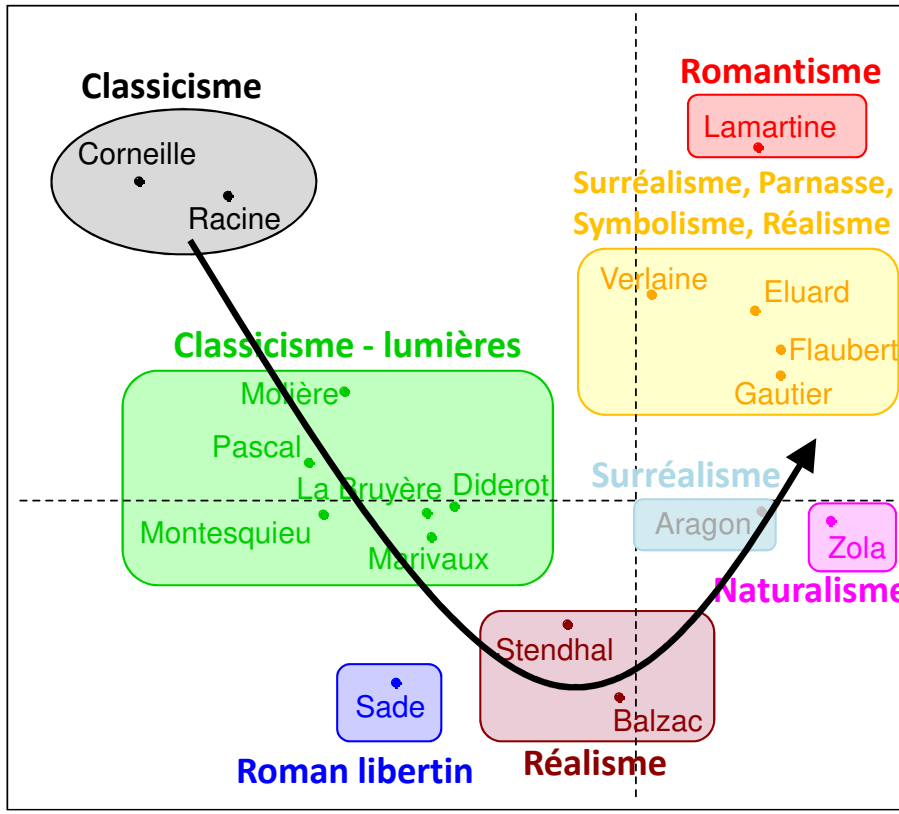




Les mots permettent de caractériser les sujets de prédilection des auteurs et les courants littéraires

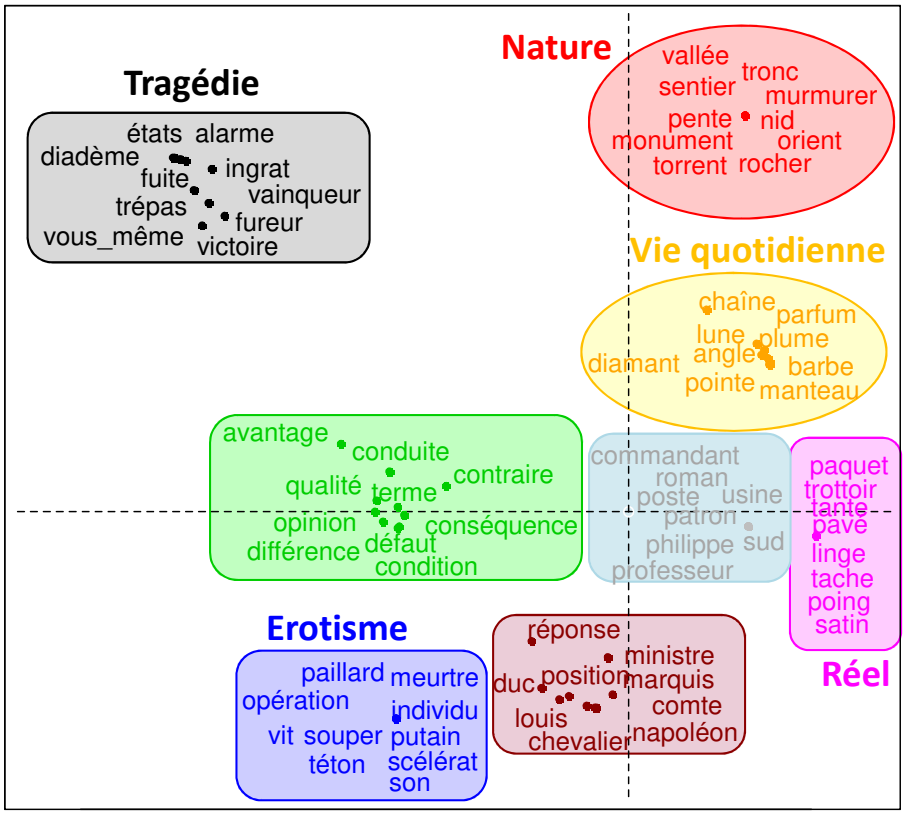
Un exemple en linguistique : interprétation ...

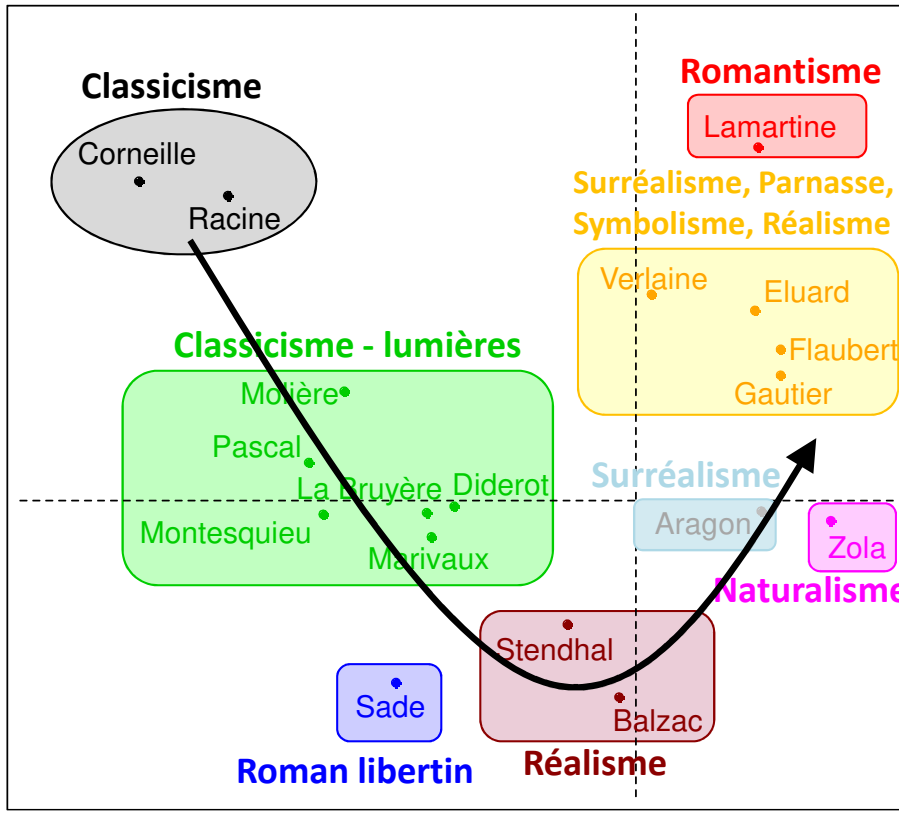




- Le naturalisme est la suite logique du réalisme : le naturalisme montre le milieu où vit le protagoniste pour expliquer son comportement de façon "scientifique"
- Évolution du vocabulaire selon les courants littéraires

Un exemple en linguistique : interprétation ...





- Le naturalisme est la suite logique du réalisme : le naturalisme montre le milieu où vit le protagoniste pour expliquer son comportement de façon "scientifique"
- Évolution du vocabulaire selon les courants littéraires

Une vidéo pour en savoir plus sur cet exemple