Panorama sur l'analyse de données

François Husson

https://husson.github.io

Unité de mathématiques appliquées, Institut Agro, Rennes

École doctorale – 20 janvier 2020

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

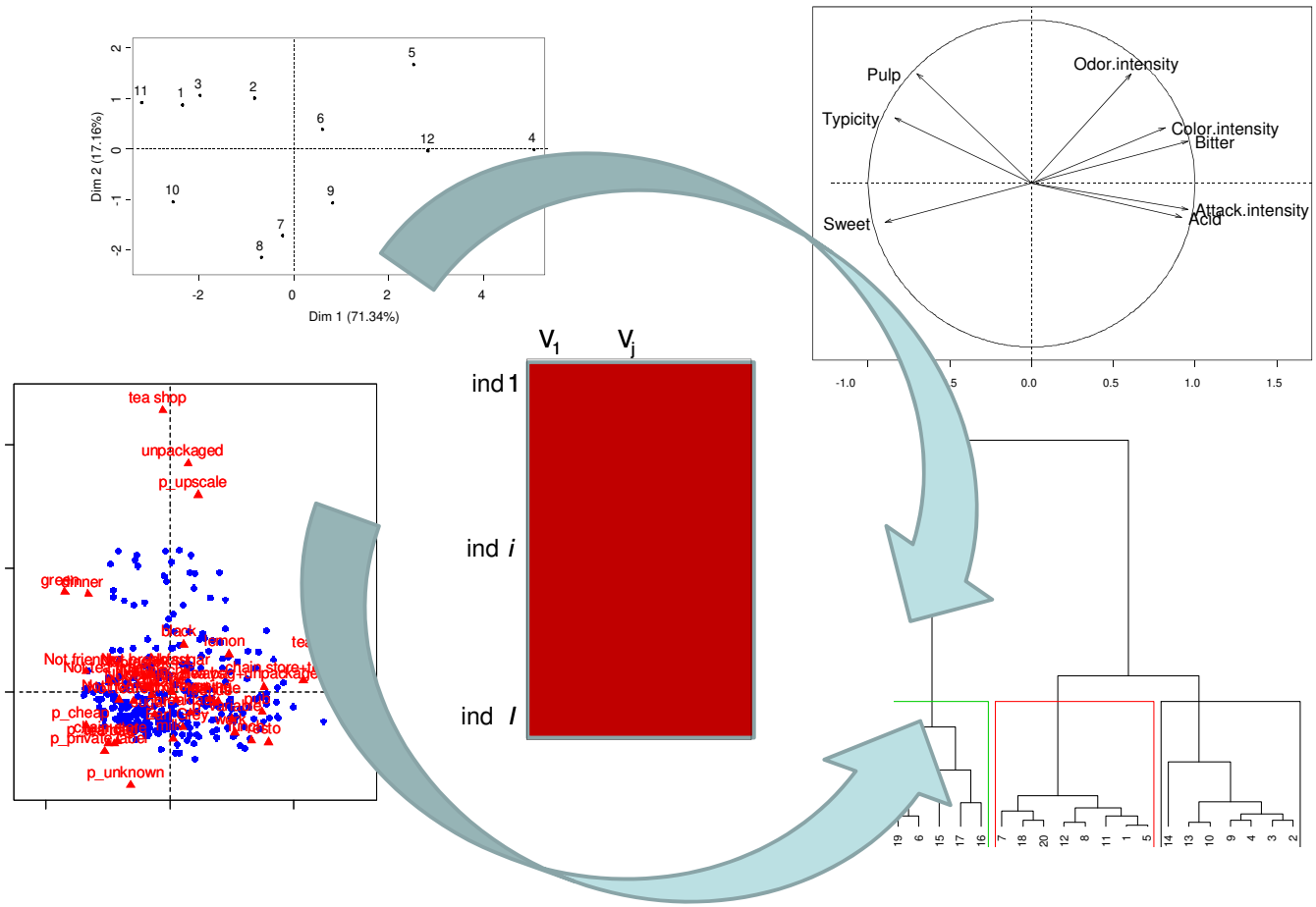

2012

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

Plan

Panorama des méthodes

Les méthodes d'analyse de données

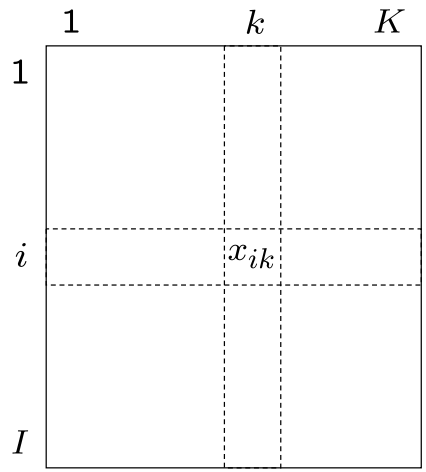


Objectifs :

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

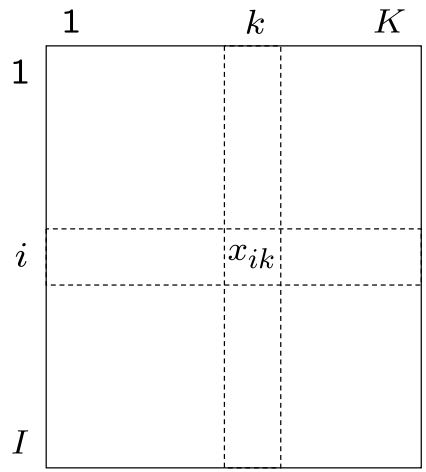
L'analyse en Composantes Principales (ACP)

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



L'analyse en Composantes Principales (ACP)

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



- Économie : valeur de l'indicateur *k* dans la région *i*
- Psychologie : degré d'accord de l'individu *i* avec l'affirmation *k*
- Sociologie : tps passé à l'activité *k* par les individus de la CSP *i*
- Enquête PISA : note de l'élève *i* dans la discipline *k*

Les données vins

- 10 individus : vins blancs du Val de Loire

- Quels vins se ressemblent ? Peut-on faire des groupes de vins ?
- Comment caractériser un vin ?
- Quels descripteurs se ressemblent ?

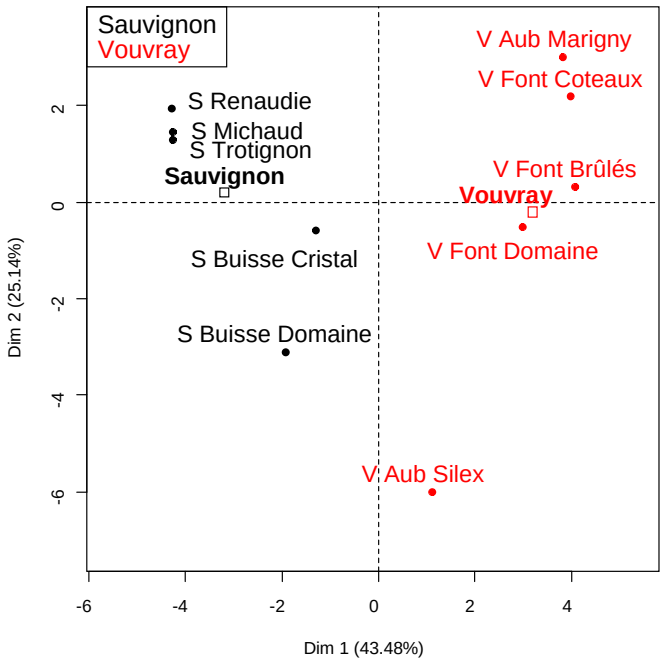
Les données vins

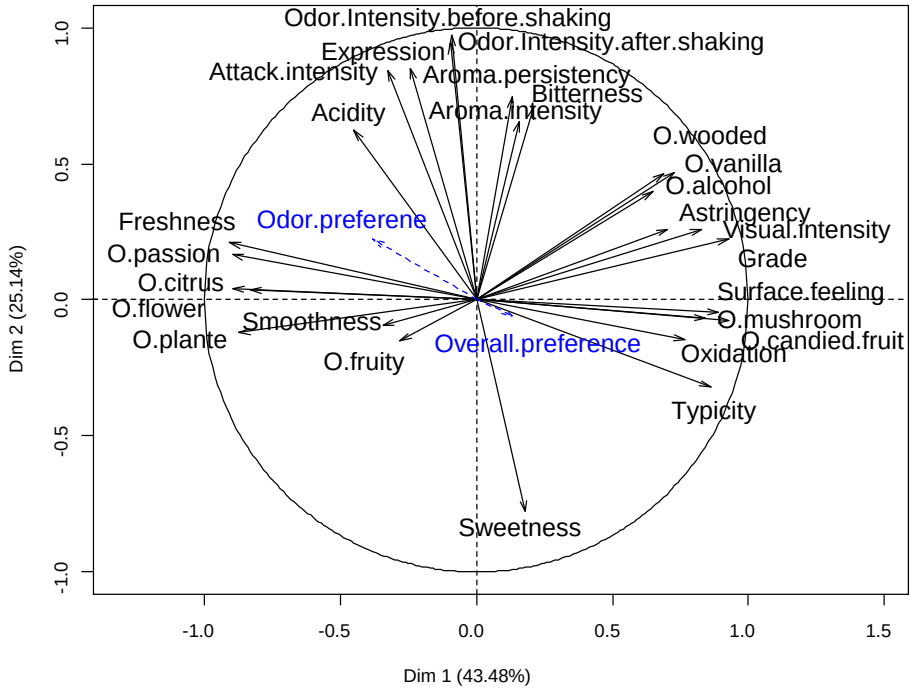
- 10 individus : vins blancs du Val de Loire
- 27 variables quantitatives : descripteurs sensoriels
  - mais aussi 2 variables d'appréciation
  - et 1 variable qualitative : label des vins (Vouvray - Sauvignon)

	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

- Quels vins se ressemblent ? Peut-on faire des groupes de vins ?
- Comment caractériser un vin ?
- Quels descripteurs se ressemblent ?

Représentation des individus et des variables

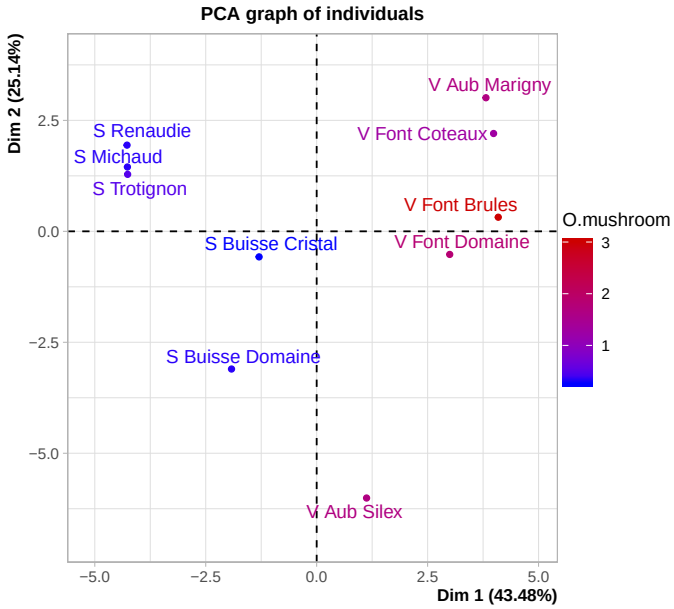


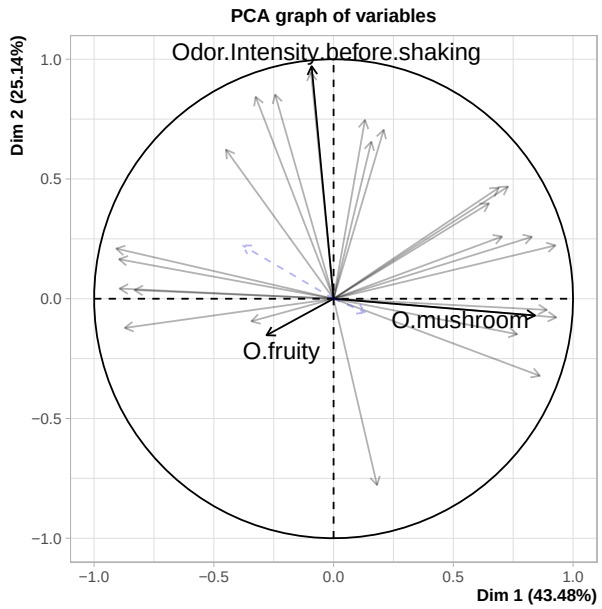


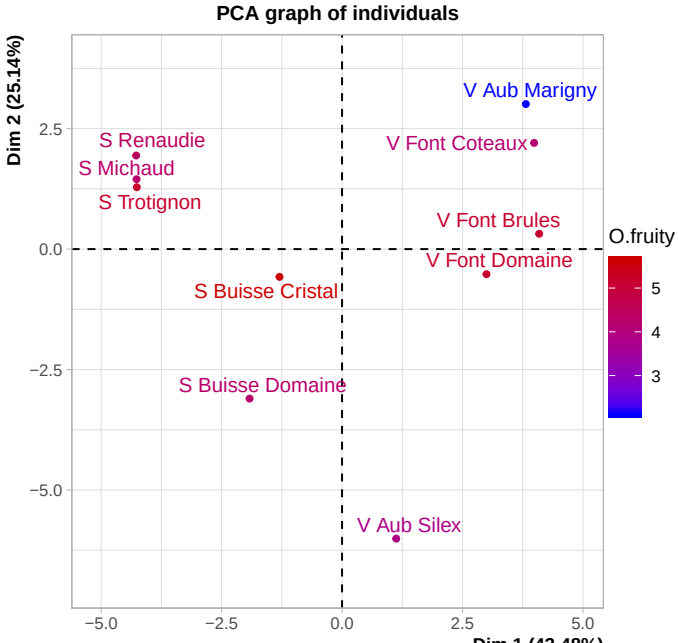
$\Longrightarrow$ Utilisation d'information supplémentaire

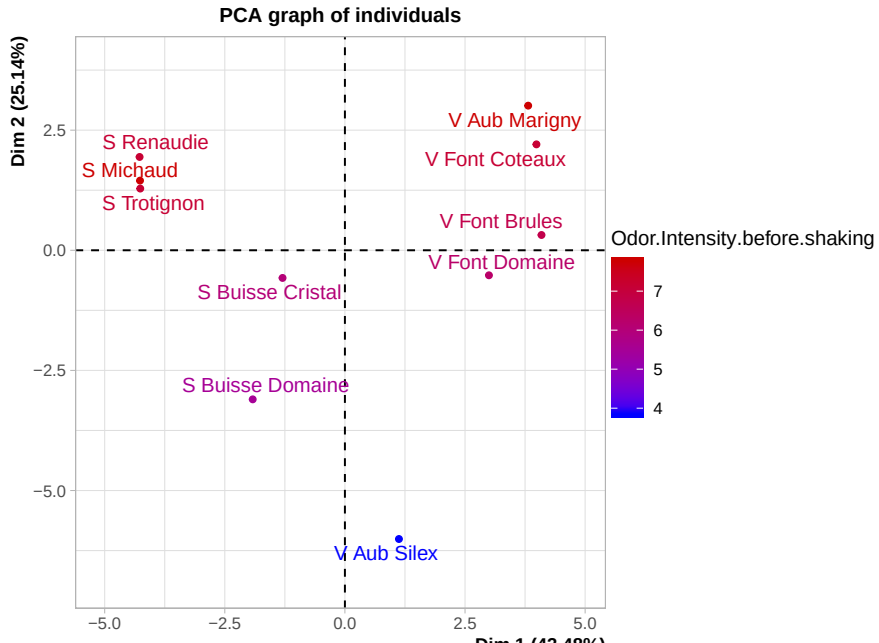
- la variable qualitative *cépage*
- les variables quantitatives d'*appréciation*

Représentation des individus et des variables

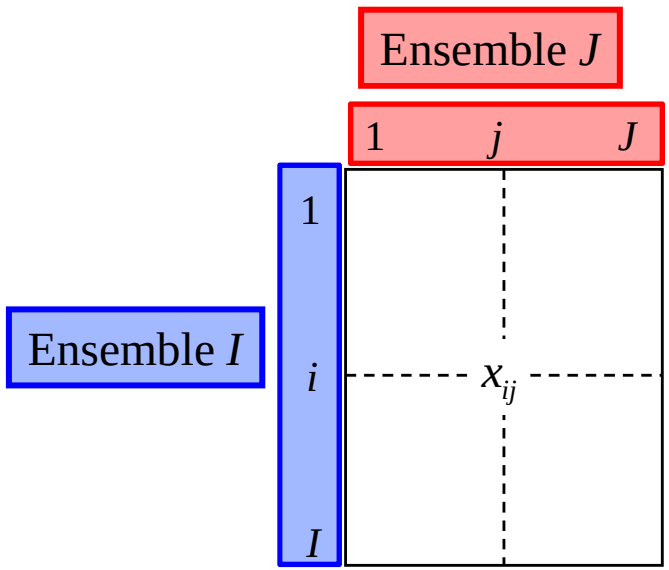








L'analyse des correspondances (AFC)



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

- Nombre de votes pour le candidat i dans le département j
- Nombre d'individus de la CSP i et de la classe d'âge j
- Analyse textuelle : nb de fois où le candidat i utilise le mot j

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

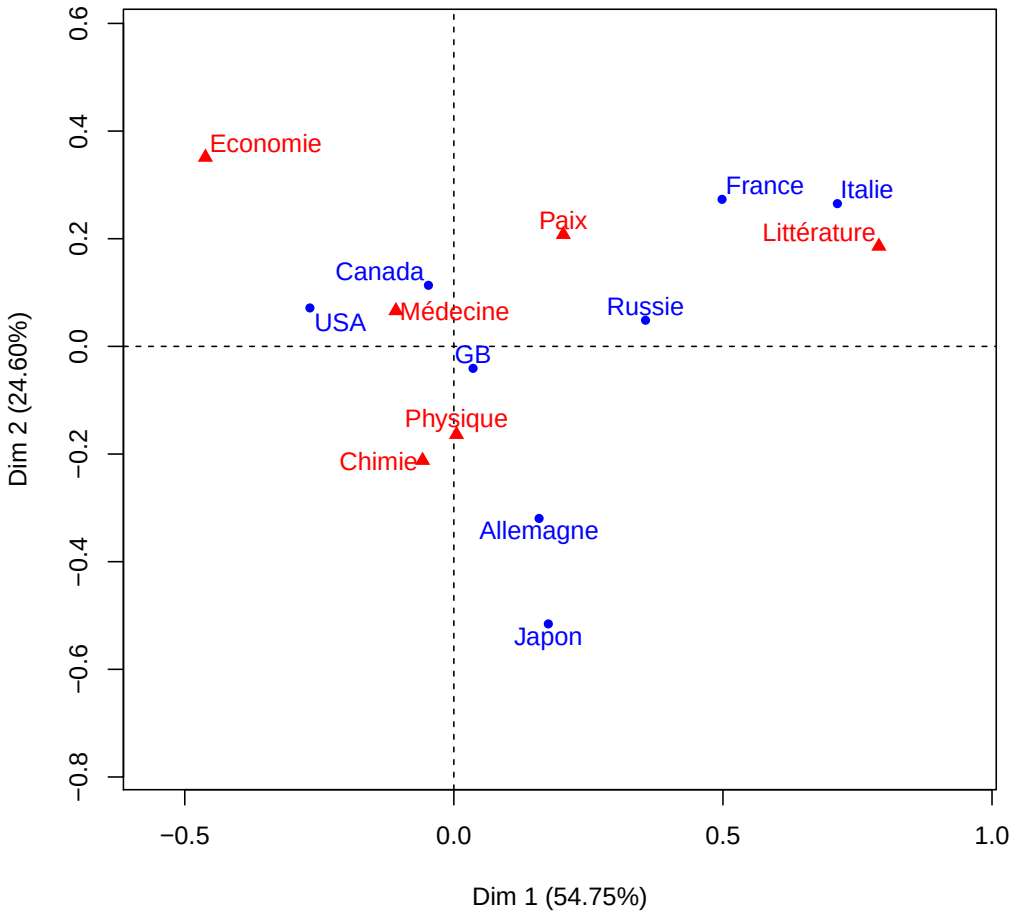
Données sur les prix Nobel

	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 ?

 On s'intéresse aux données relatives (on ne veut pas différencier petits et gros pays)

Exemple des prix Nobel



- 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 qui aide la compréhension du tableau (a fortiori avec de grands tableaux)

L'Analyse des correspondances multiples (ACM)

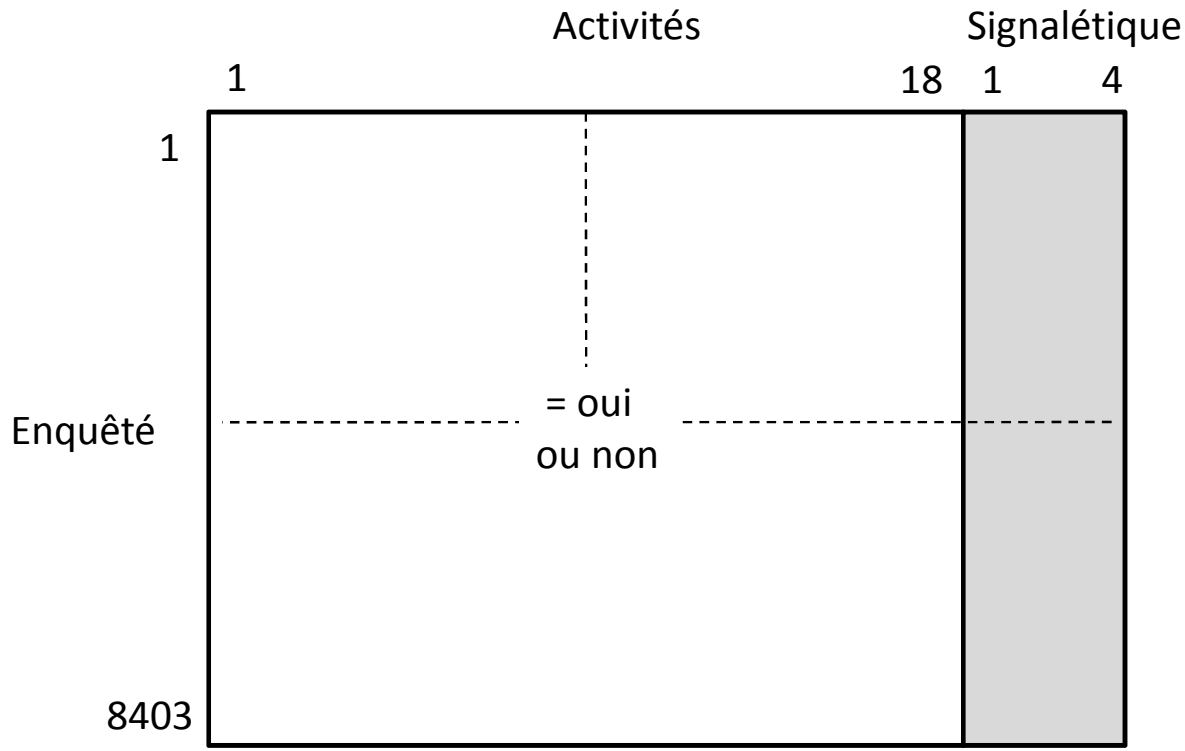
Pour analyser des questionnaires (tableau individus - variables qualitatives)

L'Analyse des correspondances multiples (ACM)

Pour analyser des questionnaires (tableau individus - variables qualitatives)

- 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

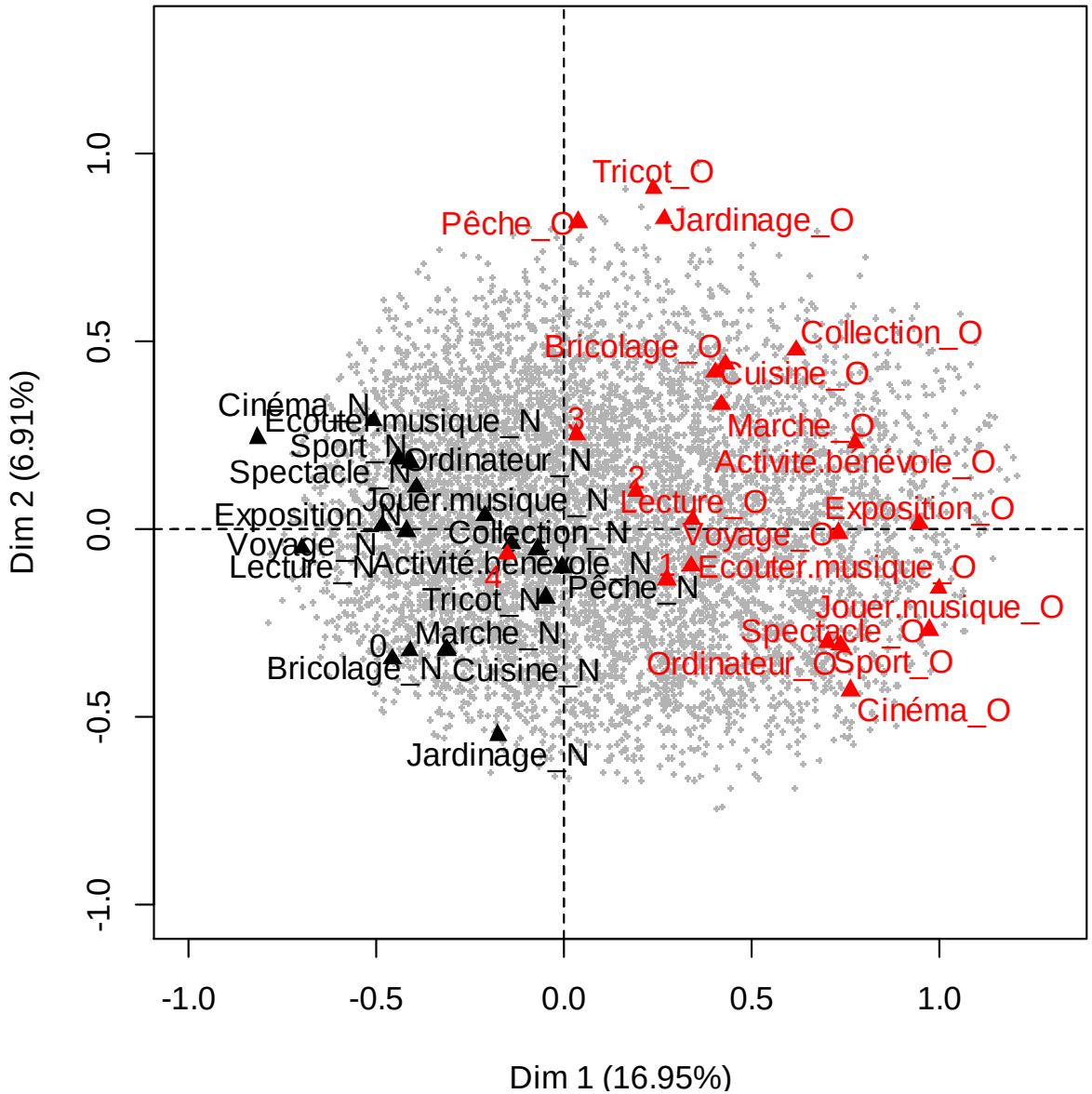
Exemple : les données loisirs



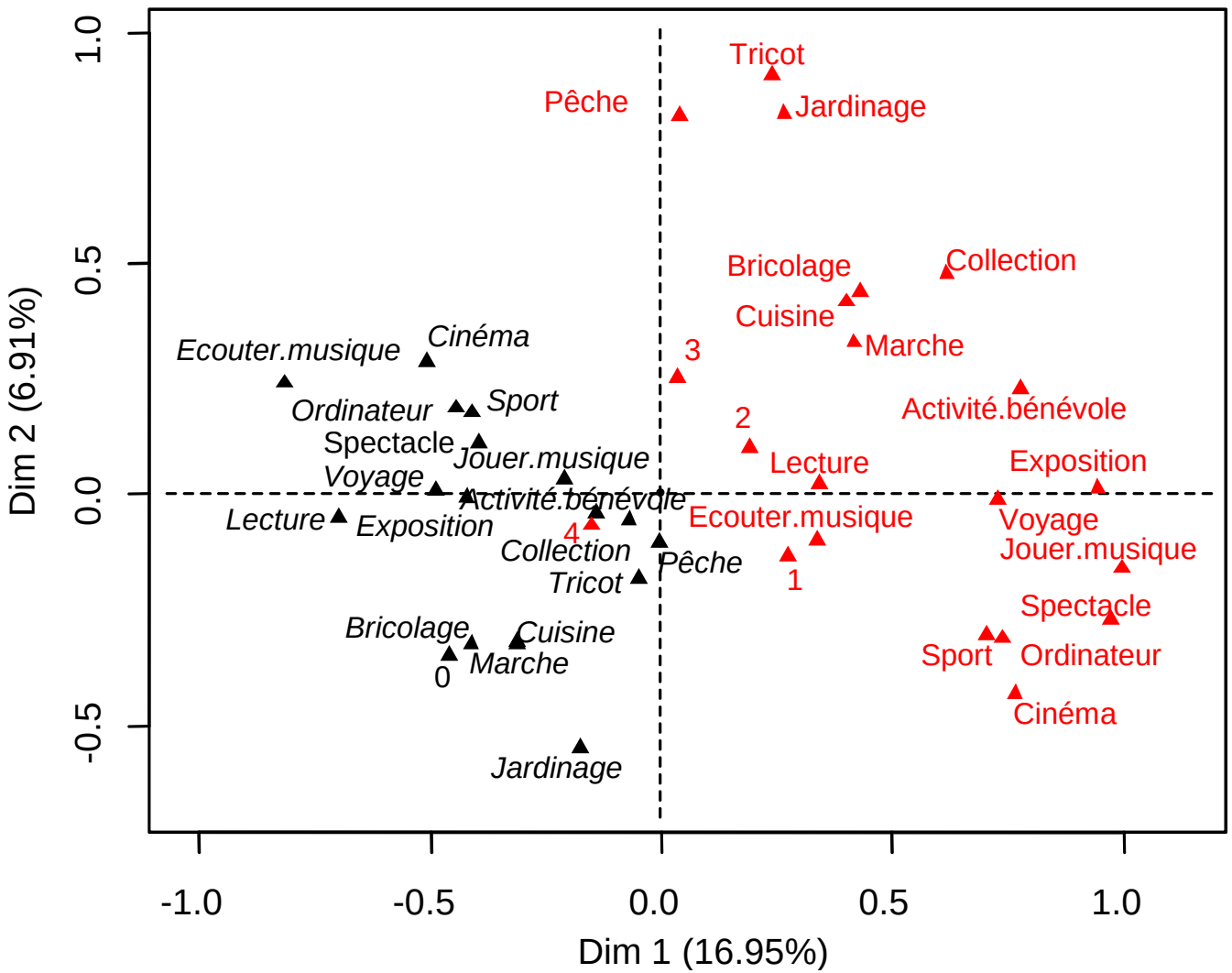
ACM : loisirs en actif, signalétique en supplémentaire

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

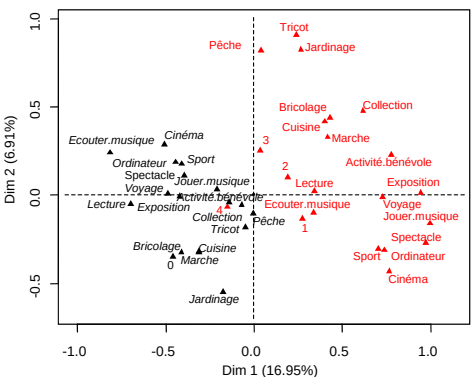
Représentation simultanée

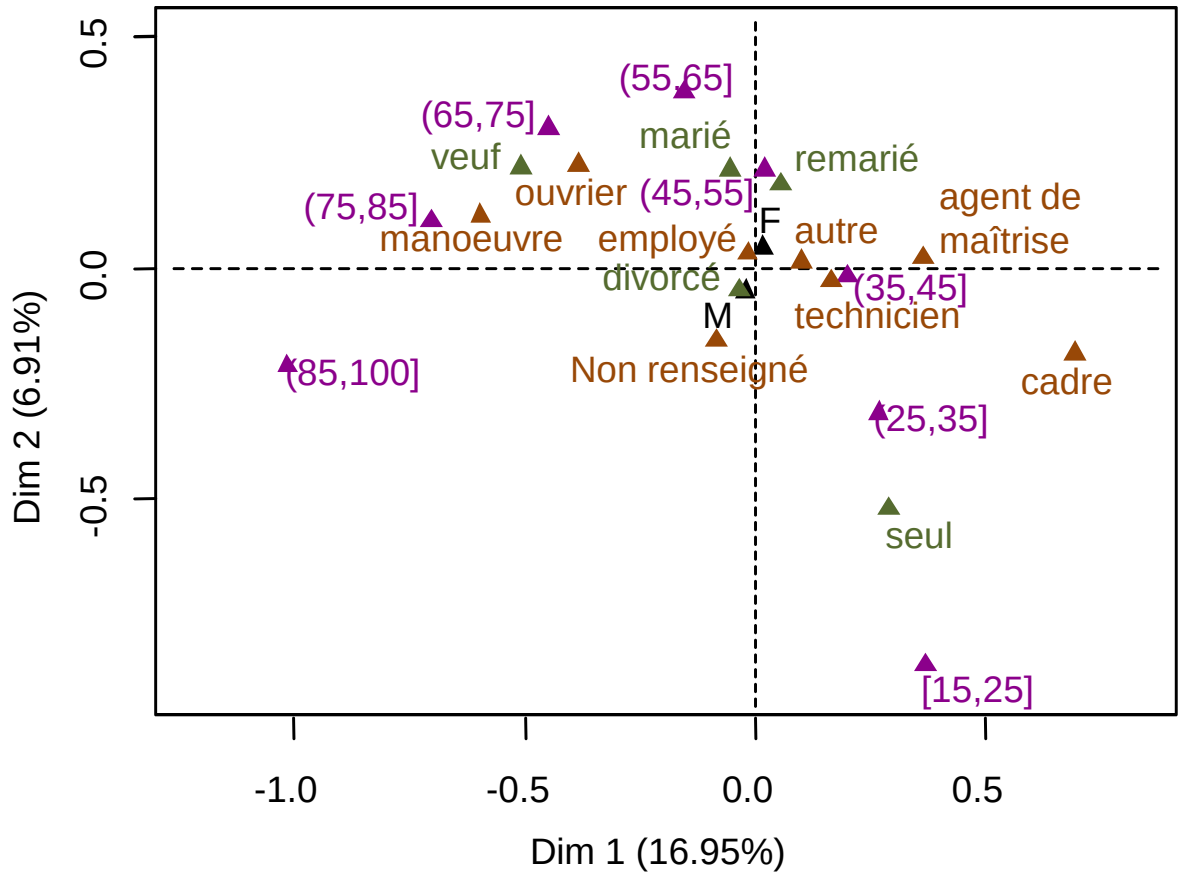


Représentation des modalités

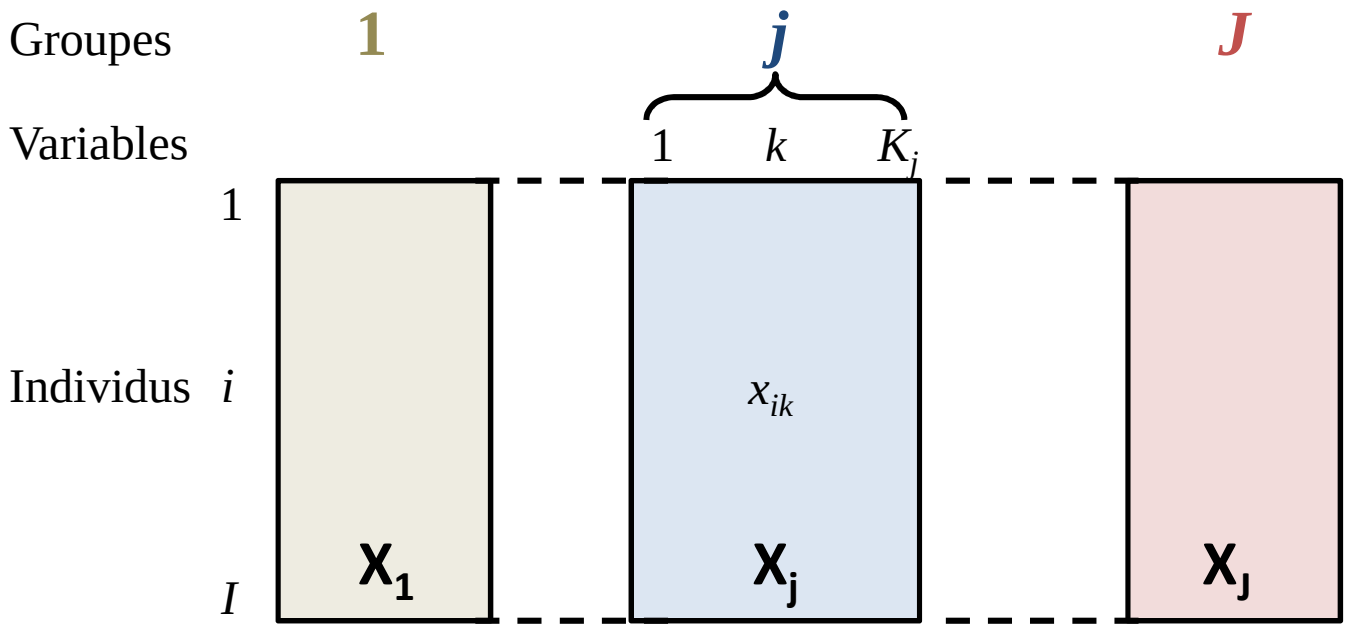


Représentation des modalités



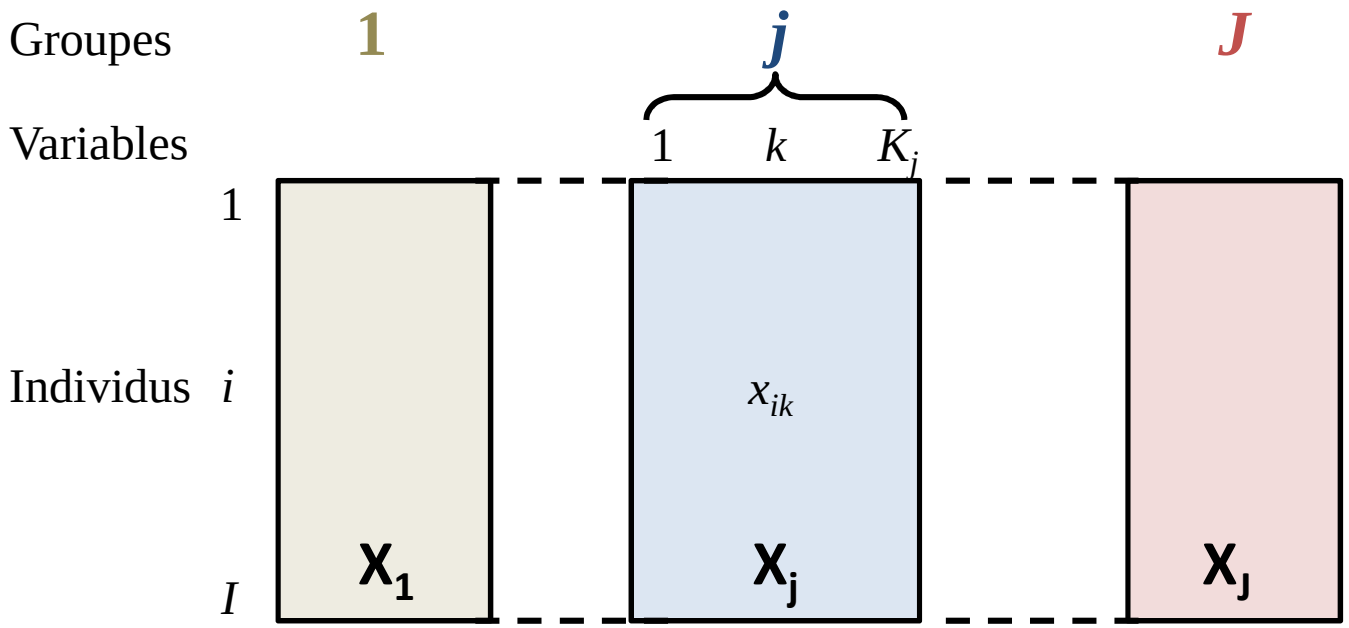


L'Analyse Factorielle Multiple (AFM)



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

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)
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 ?

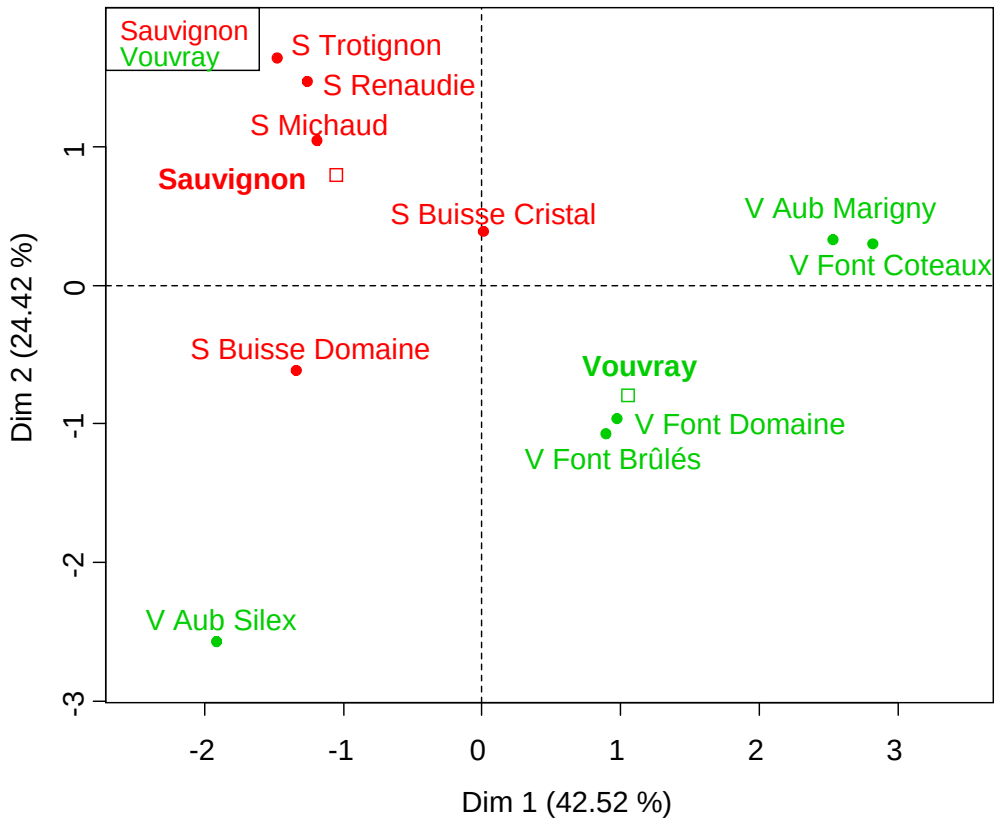
Description sensorielle de vins : comparaison de jurys

- 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					

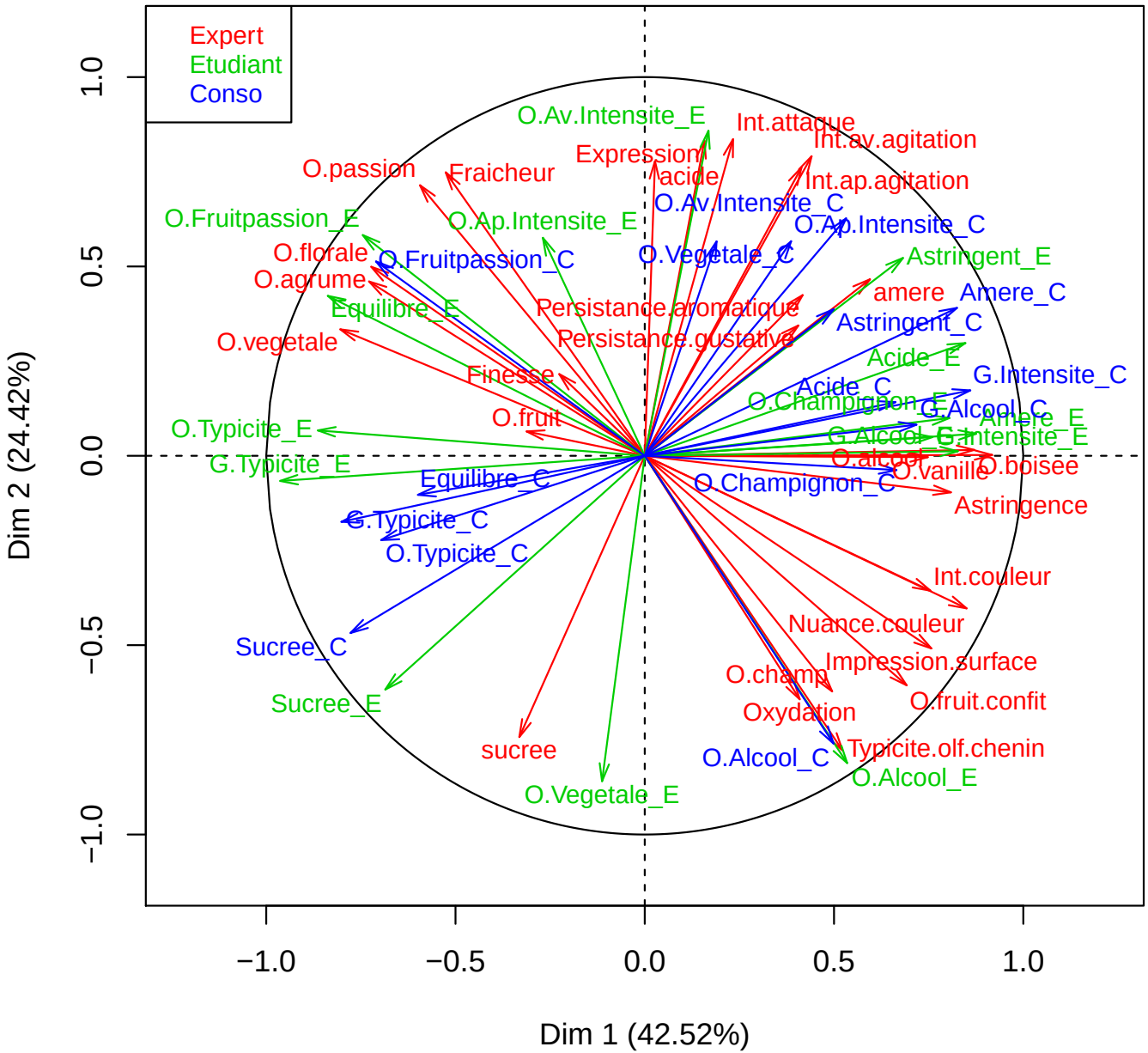
- Comment caractériser les vins ?
- Les vins sont-ils décrits de la même façon par les différents jurys ? Y-a t'il des spécificités par jury ?
- Peut-on compare les typologies des vins d'un jury à l'autre ?

Représentation des individus

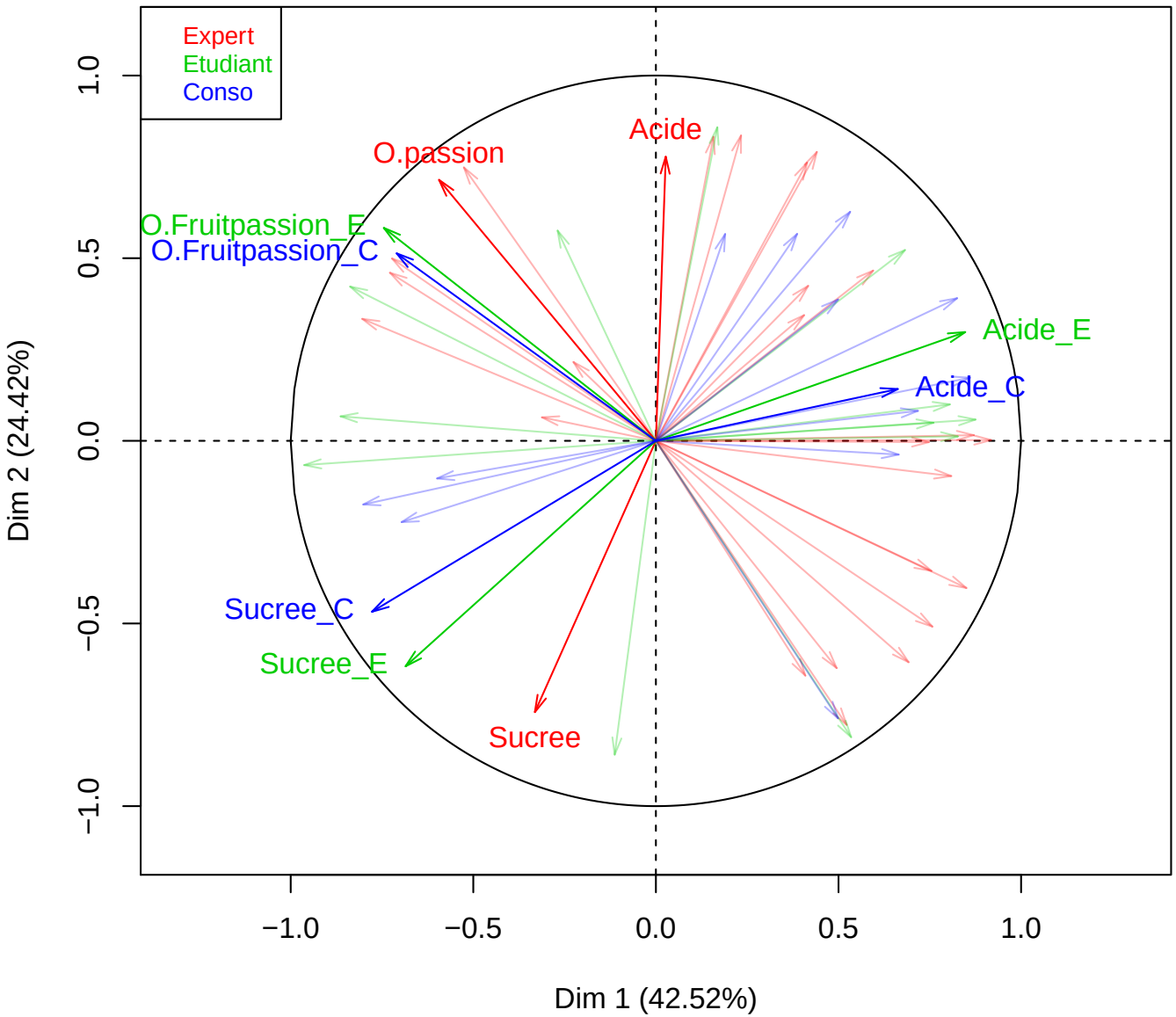


- Les deux cépages sont bien séparés
- Les Vouvray sont plus différents du point de vue sensoriel
- Plusieurs groupes de vins, ...

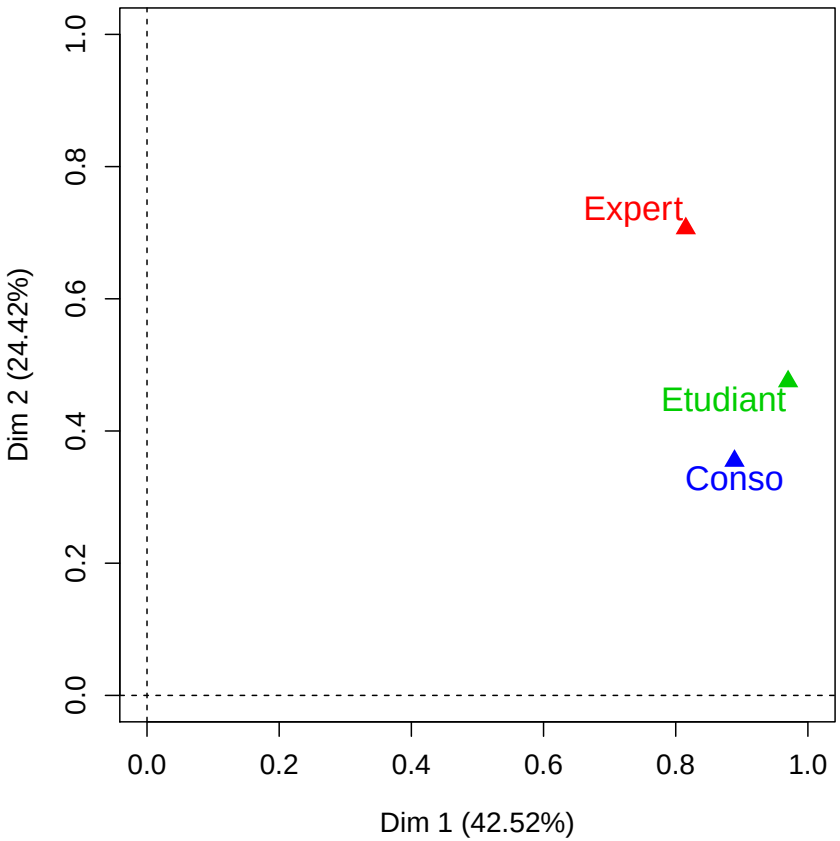
Représentation des variables



Représentation des variables



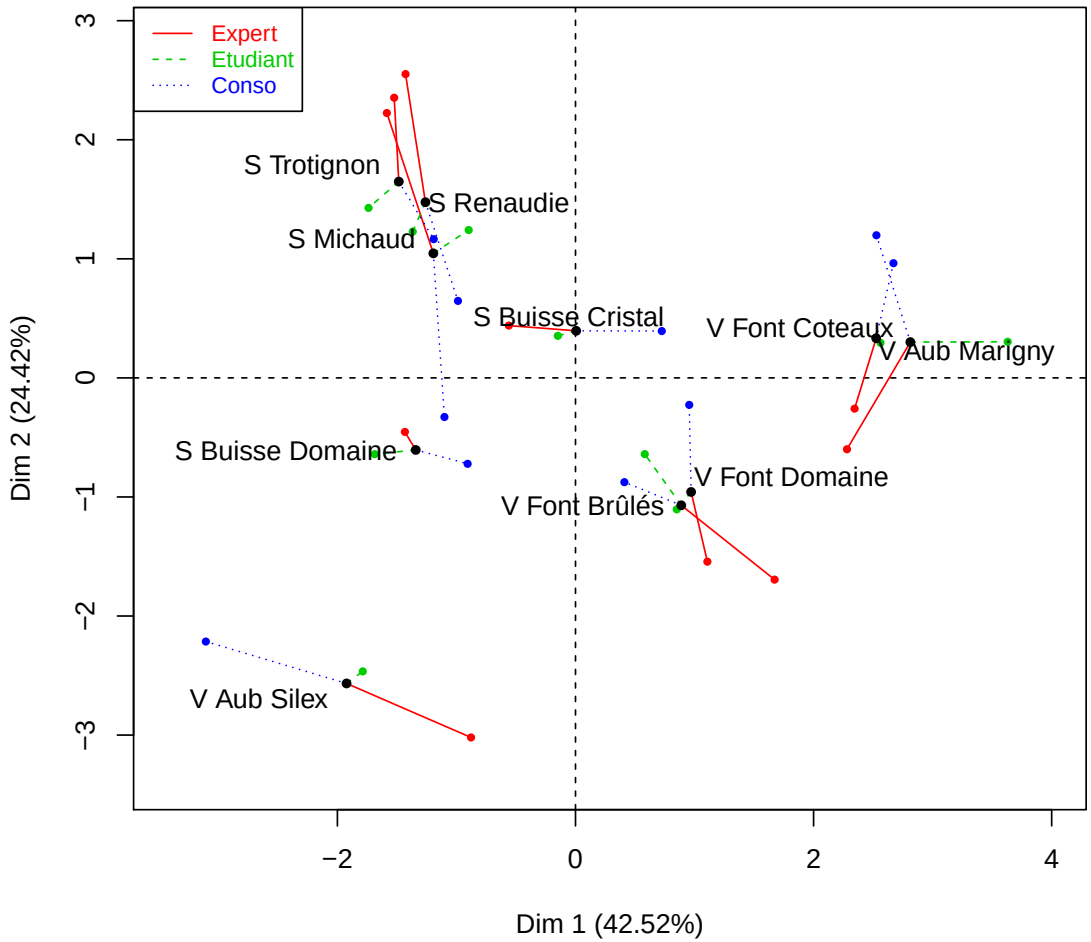
Représentation des groupes



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

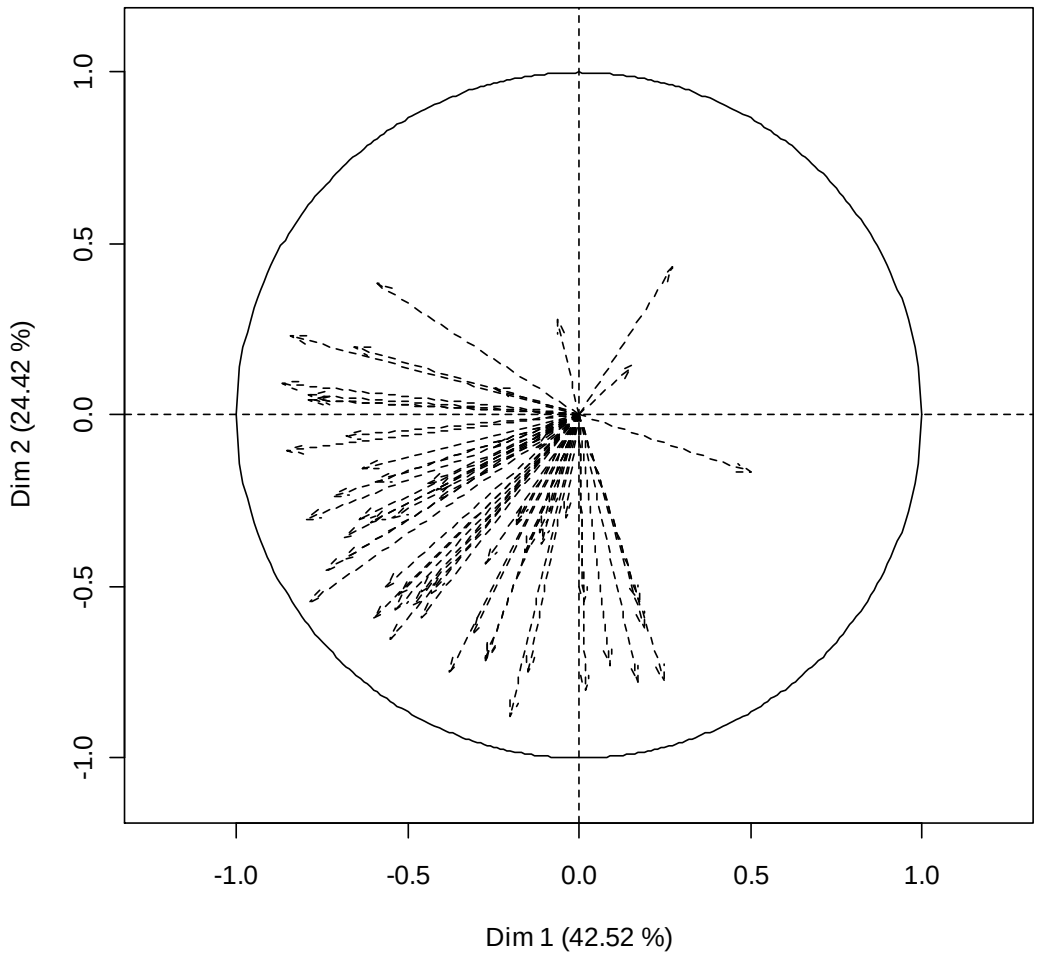
$\Rightarrow$ Ce graphe fournit une comparaison synthétique des groupes
$\Rightarrow$ Les positions relatives des individus sont-elles similaires d'un groupe à l'autre ?

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

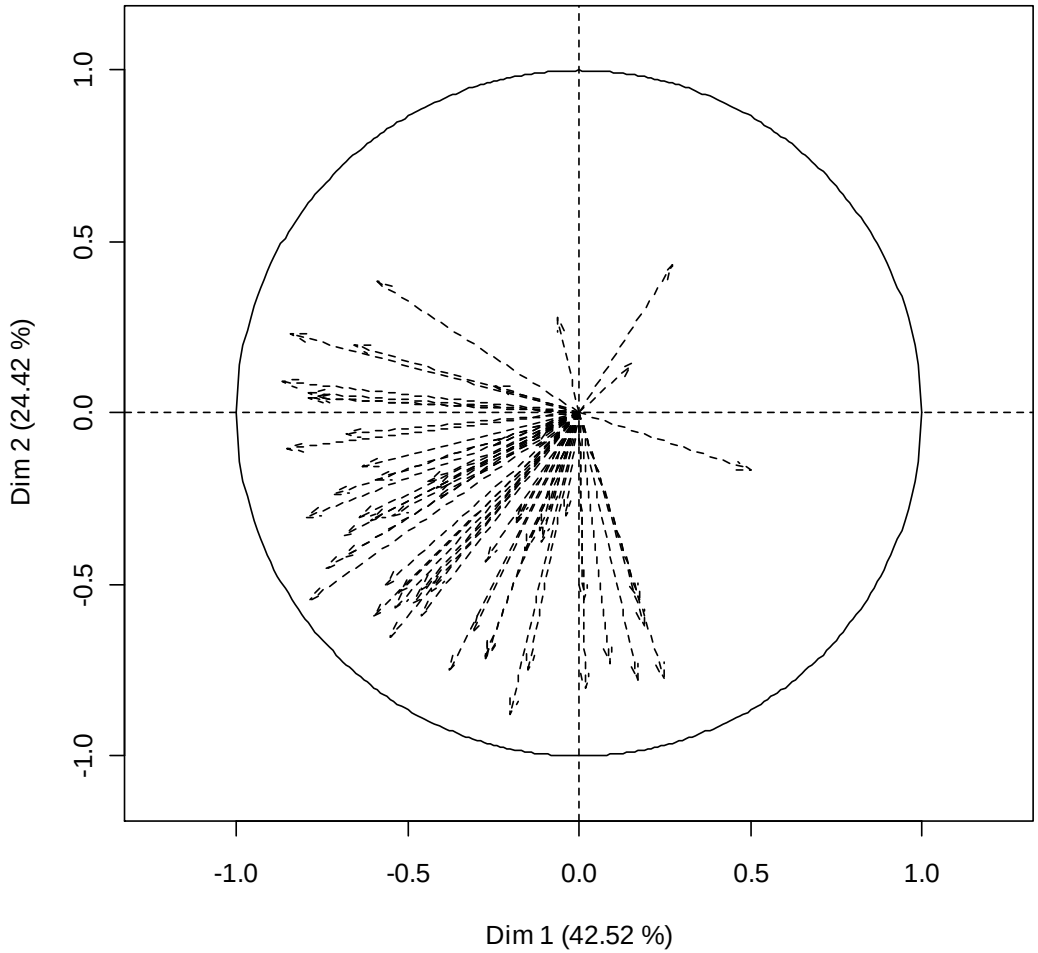
Représentation de variables supplémentaires



Les préférences sont liées à la description sensorielle

Représentation de variables supplémentaires

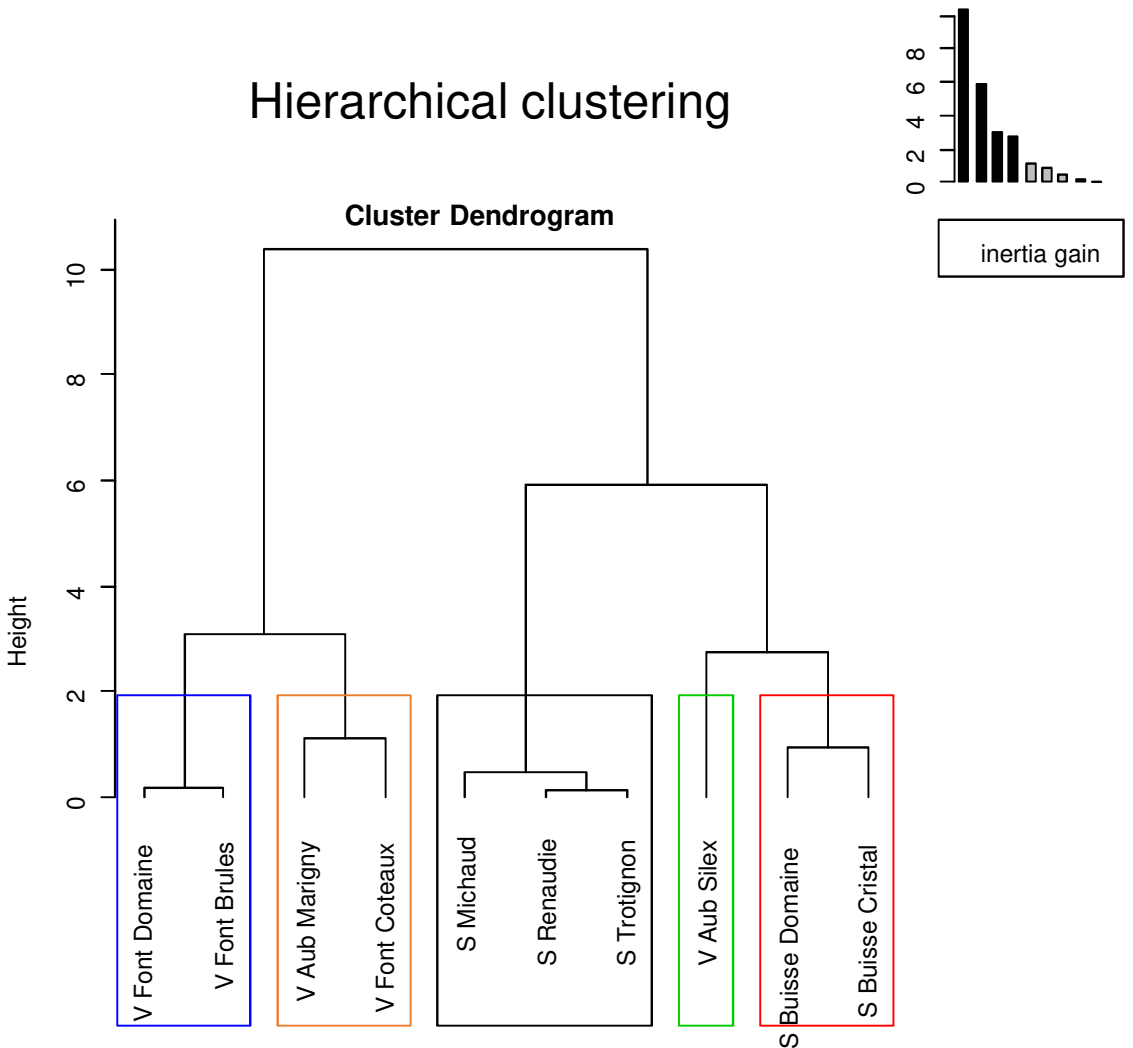
Le vin préféré est *Vouvray Aubussière Silex*



Les préférences sont liées à la description sensorielle

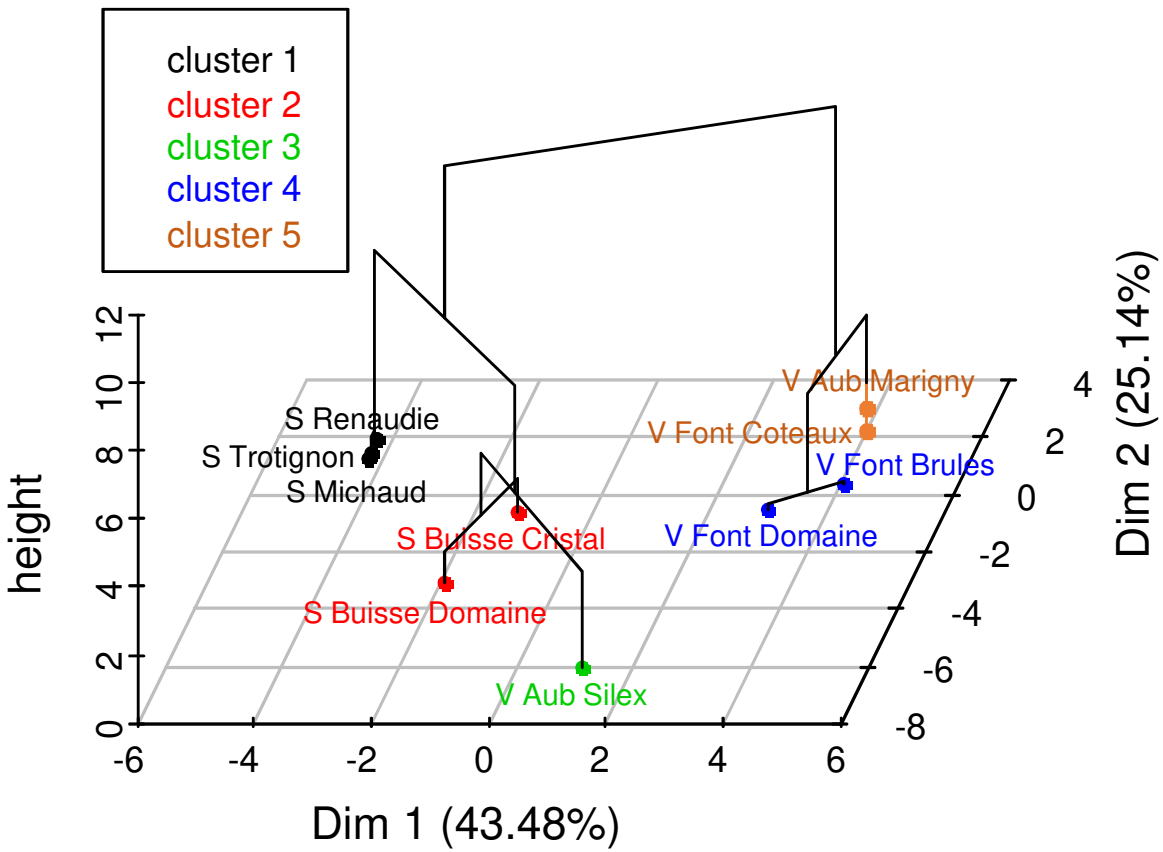
Classification Ascendante Hiérarchique (CAH)

- 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

Panorama des méthodes

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 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`
Hiérarchie 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`

Gestion de données manquantes avec le package missMDA

- Impute les données de façon optimale pour une analyse factorielle
- http://factominer.free.fr/missMDA

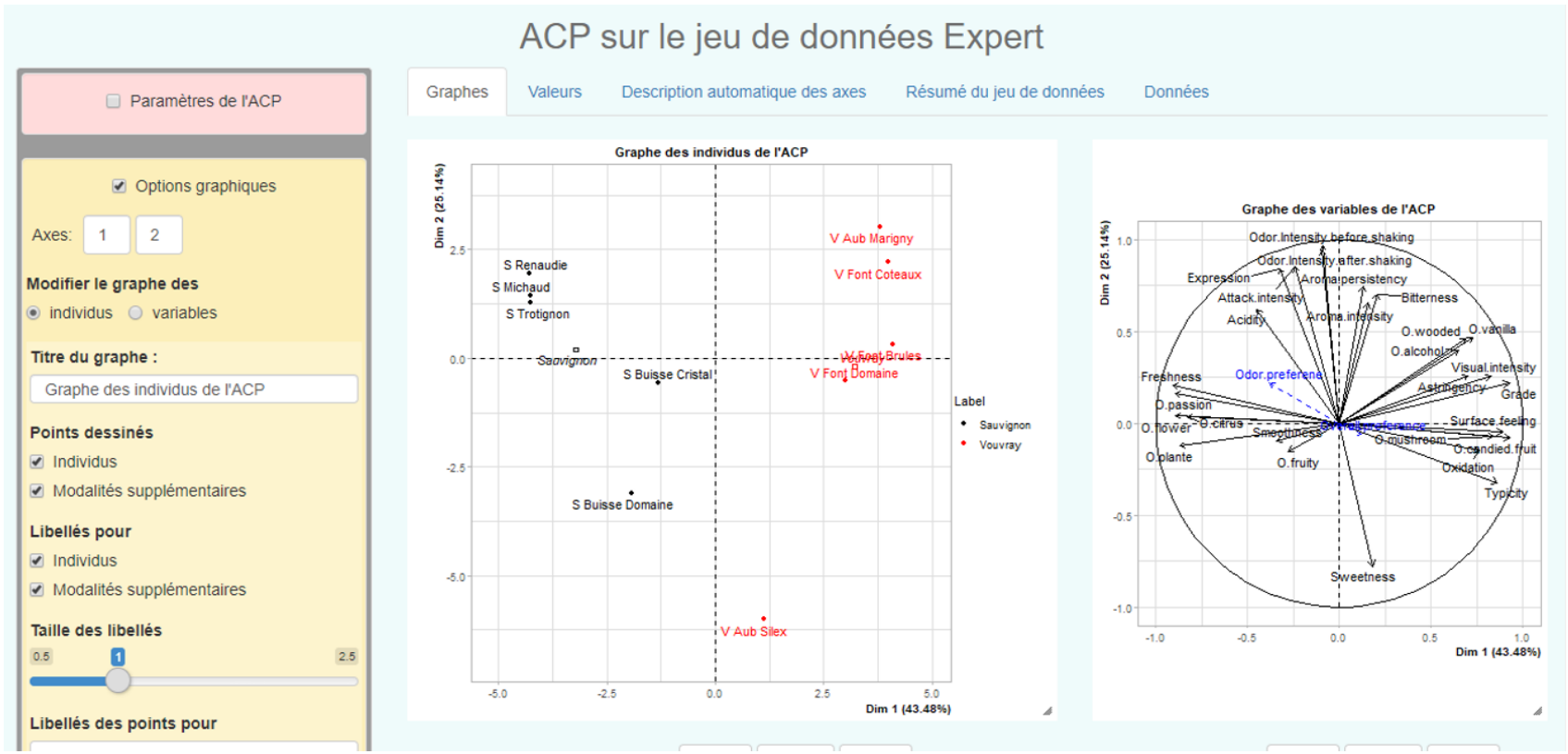
```
library(missMDA)
data(orange)
nb <- estim_ncpPCA(orange, scale=TRUE)          ## Estime le nb de dimensions
comp <- imputePCA(orange, ncp=2, scale=TRUE)  ## Complète le tableau
res.pca <- PCA(comp$completeObs)                ## Effectue l'ACP

mi <- MIPCA(orange, scale = TRUE, ncp=2)        ## Imputation multiple
plot(mi)
```

Graphiques interactifs avec le package Factoshiny

- Interface - graphes interactifs - gestion de données manquantes
- Vidéo de démonstration

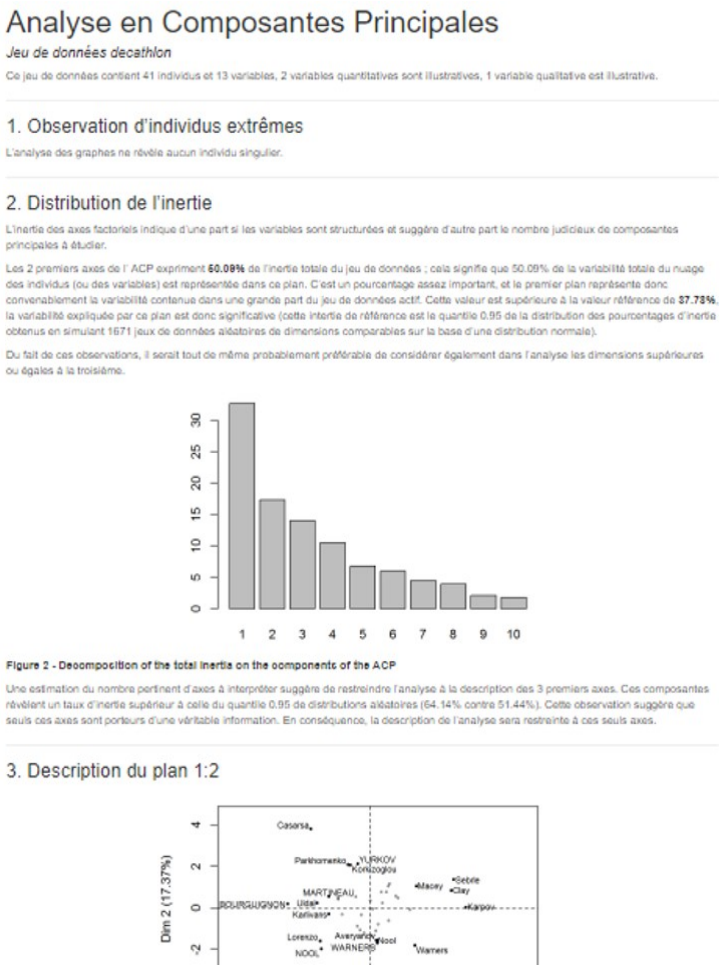
```
library(Factoshiny)
vins <- read.table("https://husson.github.io/img/vins_expert.csv", header=TRUE,
     sep=";", row.names=1)
res <- Factoshiny(vins)
```



Rapport automatisé avec le package FactoInvestigate

Propose une interprétation des résultats basée sur l'objet résultat

```
library(FactoInvestigate)
vins <- read.table("https://husson.githu
                   header=TRUE, sep=";"
res.pca <- PCA(vins,quanti.sup=29:30,qu
Investigate(res.pca)
```



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 inertie 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

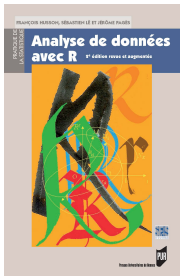
- FactoMineR : pour mettre en œuvre les méthodes
- Factoshiny : pour un menu déroulant et graphes interactifs
- missMDA : pour la gestion des données manquantes
- FactoInvestigate : pour les rapports automatisés

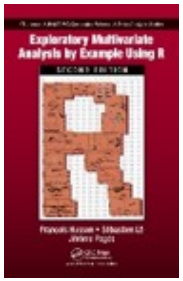
- site FactoMineR : http://factominer.free.fr
- site F. Husson : https://husson.github.io
- Google group
  https://groups.google.com/group/factominer-users/

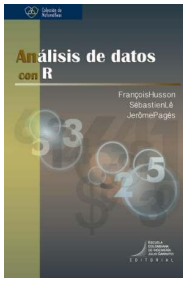
- 2 articles dans J. of stat. software (FactoMineR, missMDA)
- 2 articles dans R journal (CA-galt, MFACT)

Aides à l'utilisateur

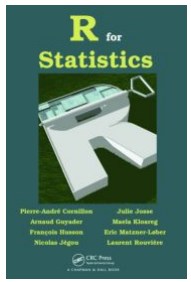
Analyse de données avec R (2e ed)



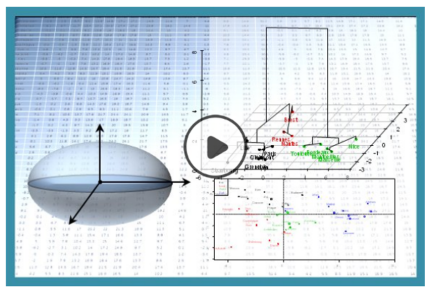




R pour la stat. et sc. des données



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*, …
- …

Un exemple en linguistique

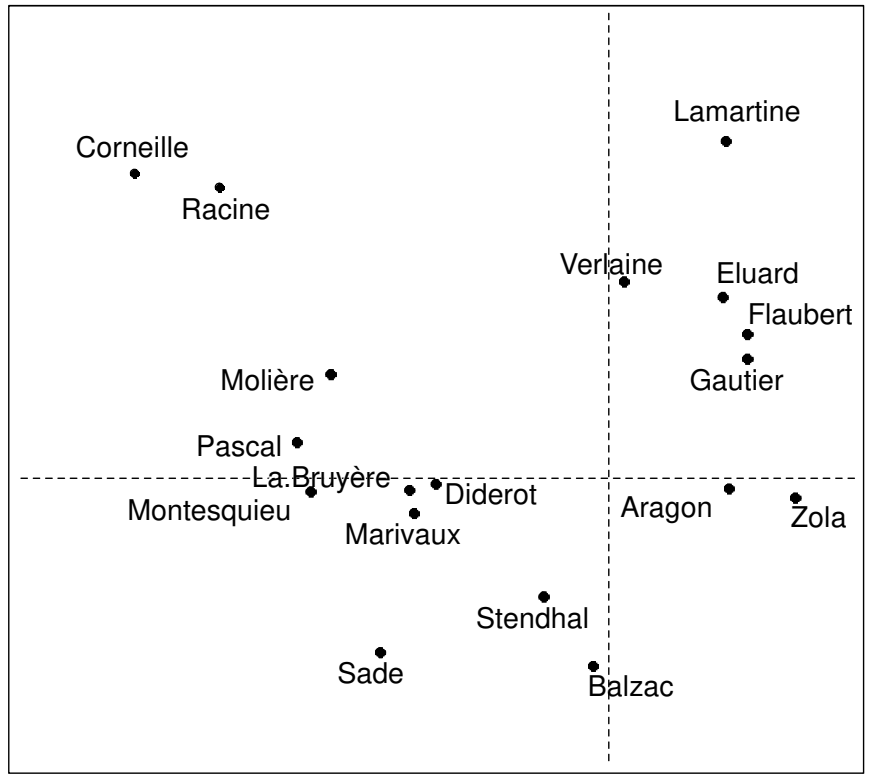
- Aragon (23 textes) : FeuJoie, Perpétuel, Destinées, Snark, Peinture, ...
- Balzac (49 textes) : *Chouans*, *Physiologie*, *Vendetta*, *Gobseck*, ...
- Corneille (34 textes) : *Mélite*, *Clitandre*, *Veuve*, *Gelerie*,*Suivante*, ...
- ...

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	
. . .					

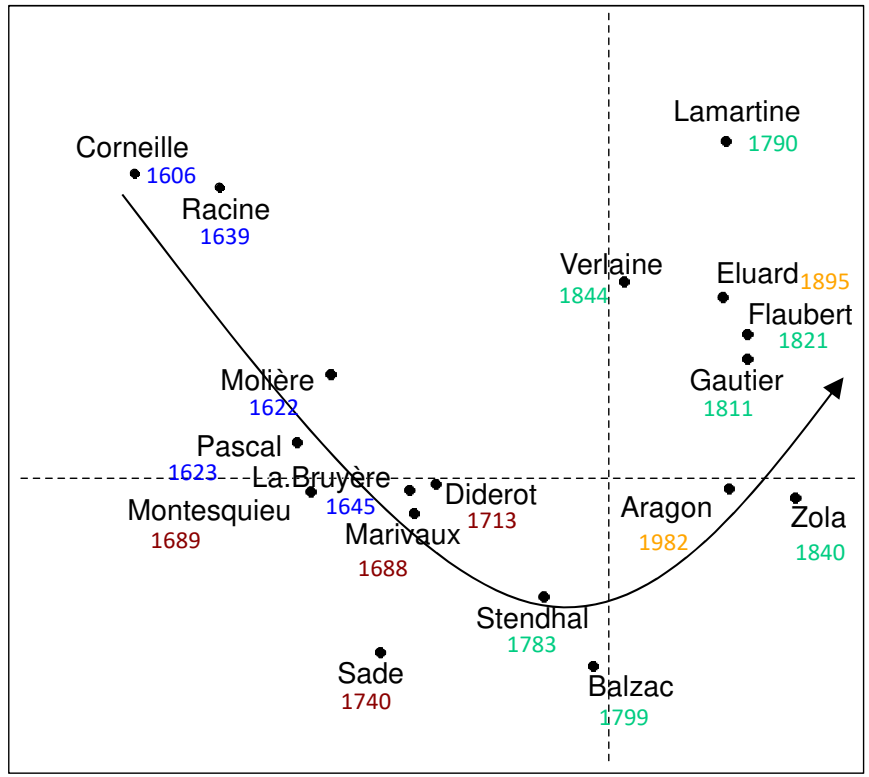
978 mots

Un exemple en linguistique



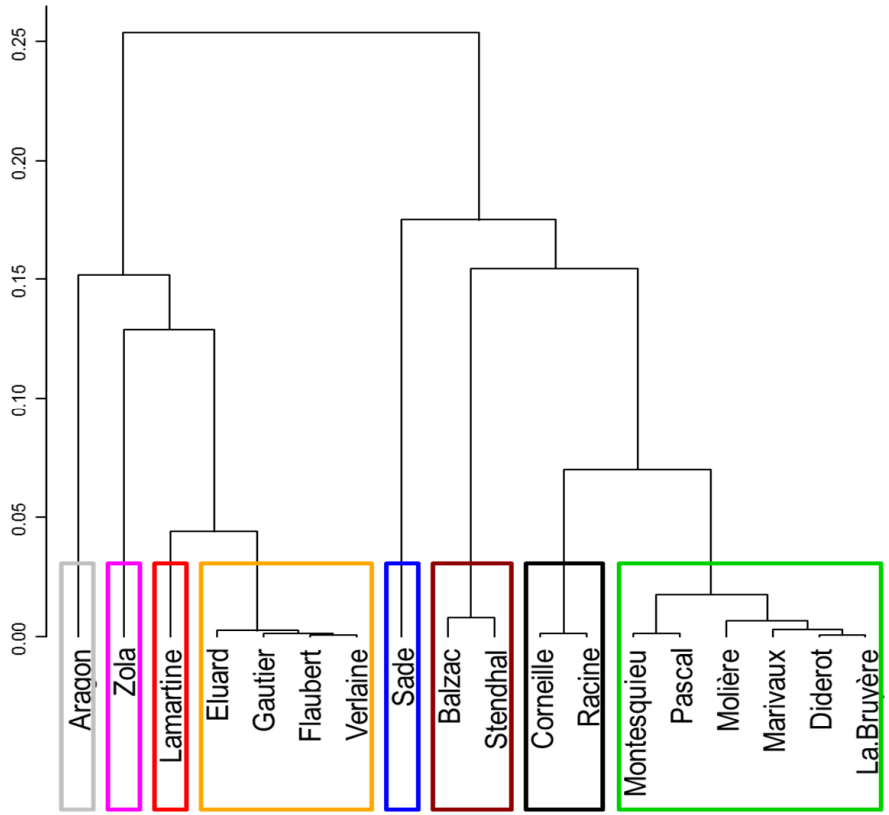
Avec l'AFC, les auteurs sont d'autant plus proches qu'ils emploient les mots dans les mêmes proportions, i.e. qu'ils s'intéressent aux mêmes sujets et ont les mêmes préoccupations

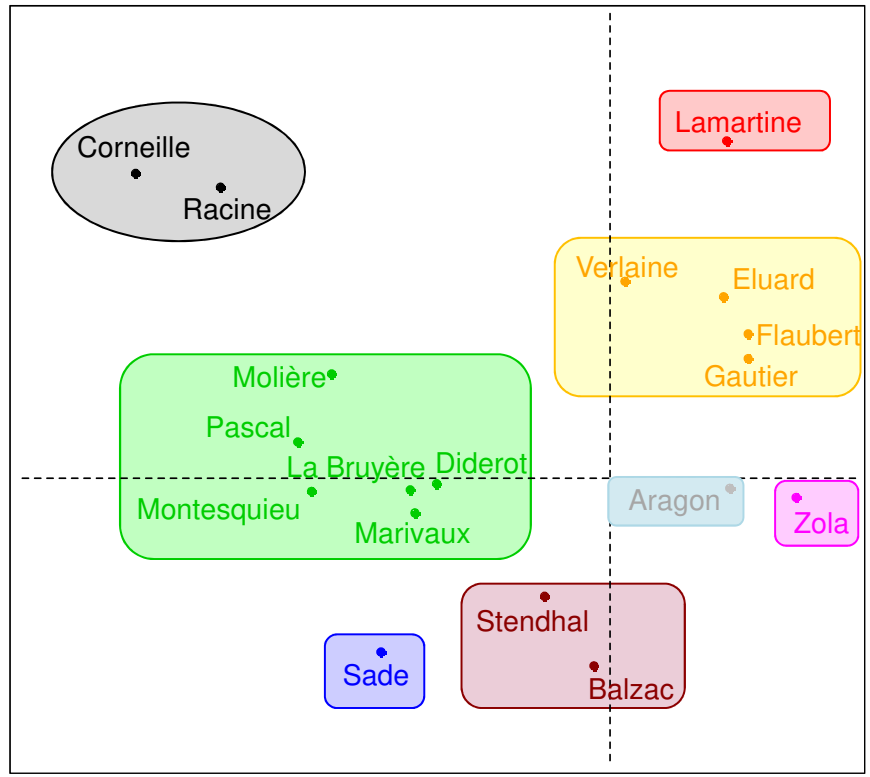
Un exemple en linguistique



- 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

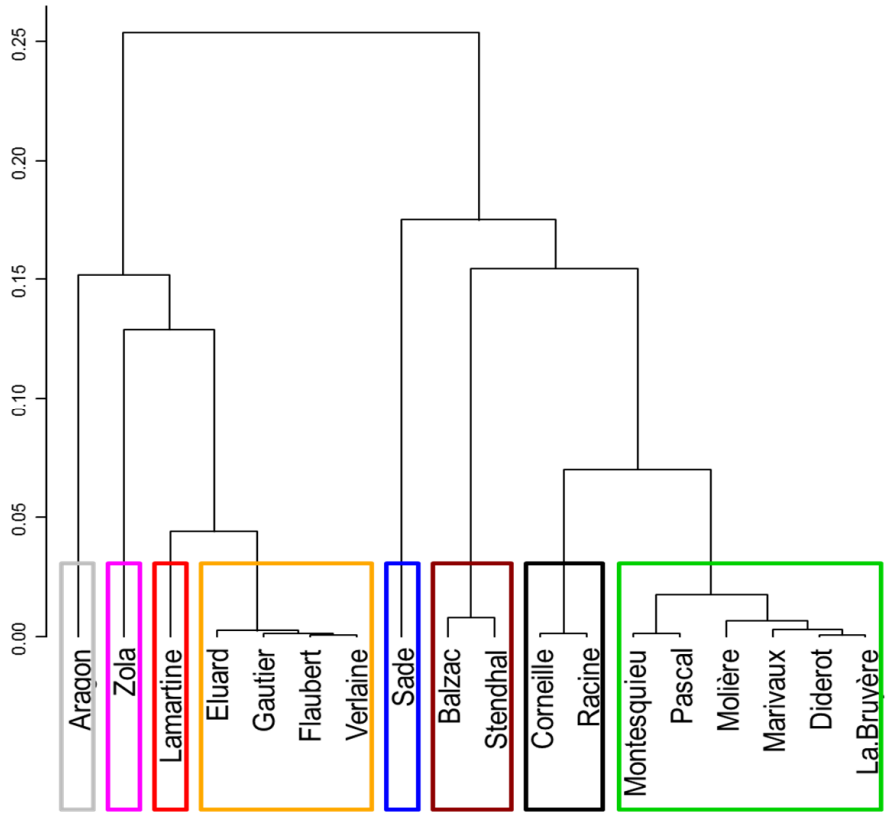
Un exemple en linguistique

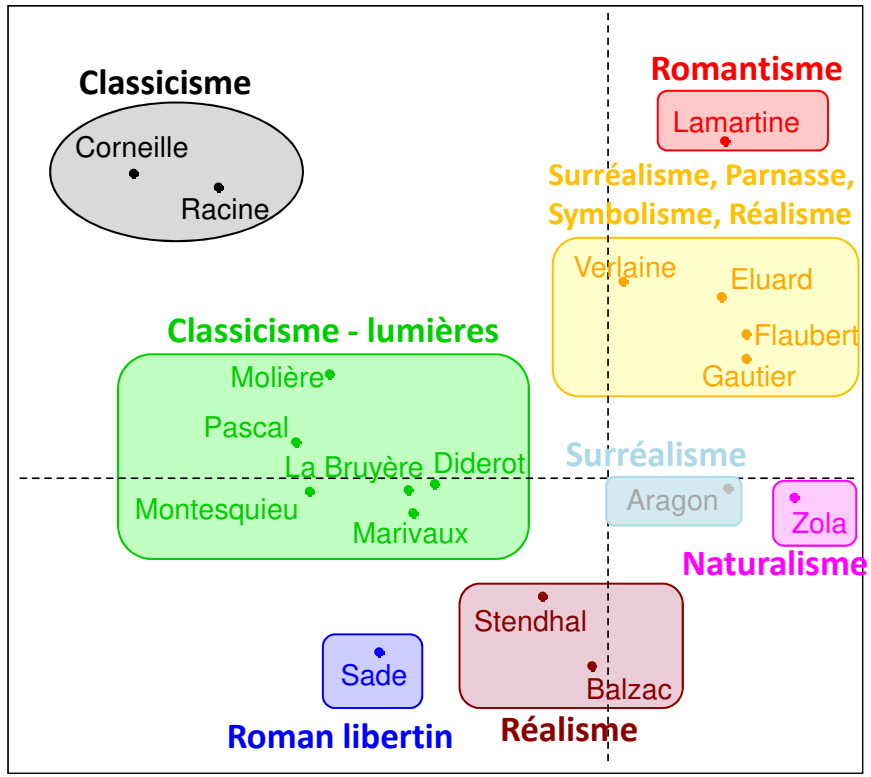




- La classification retrouve des classes d'auteurs connues

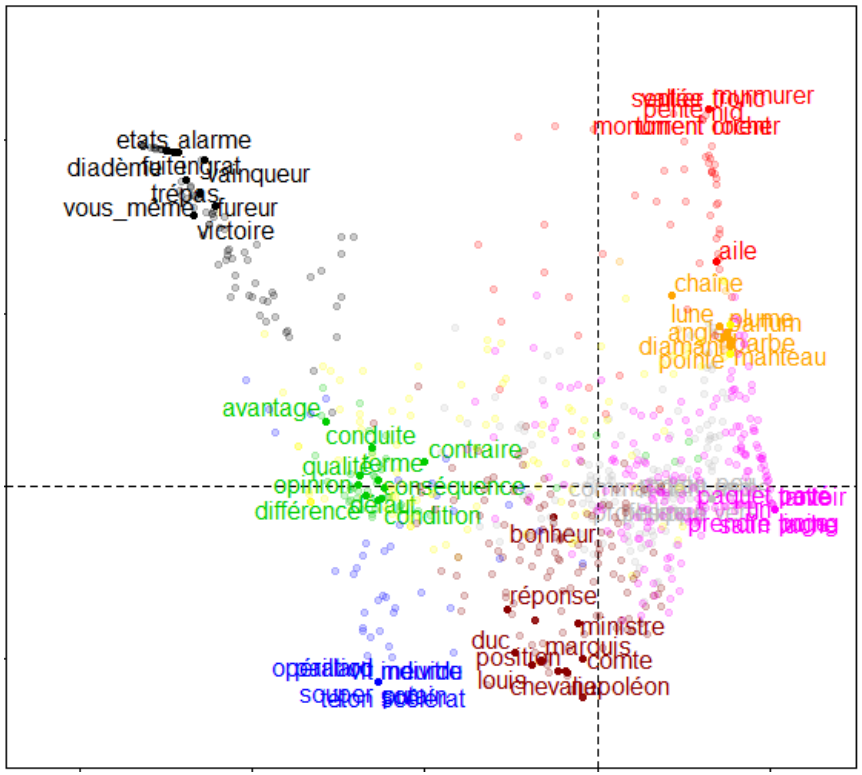
Un exemple en linguistique

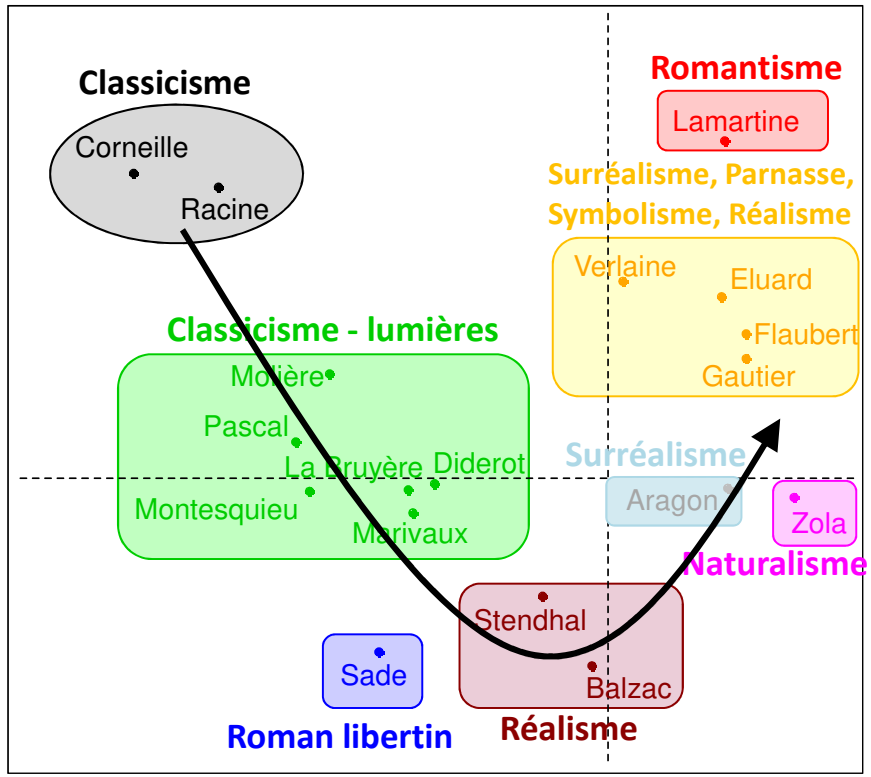




- Stendhal et Balzac (réalistes) sont très éloignés de Lamartine (romantique). On retrouve ici que les auteurs réalistes ont un point commun : s'éloigner des excès romantiques !
- Points communs naturalistes / réalistes : montrer la société telle qu'elle est, le roman devient le miroir de la société

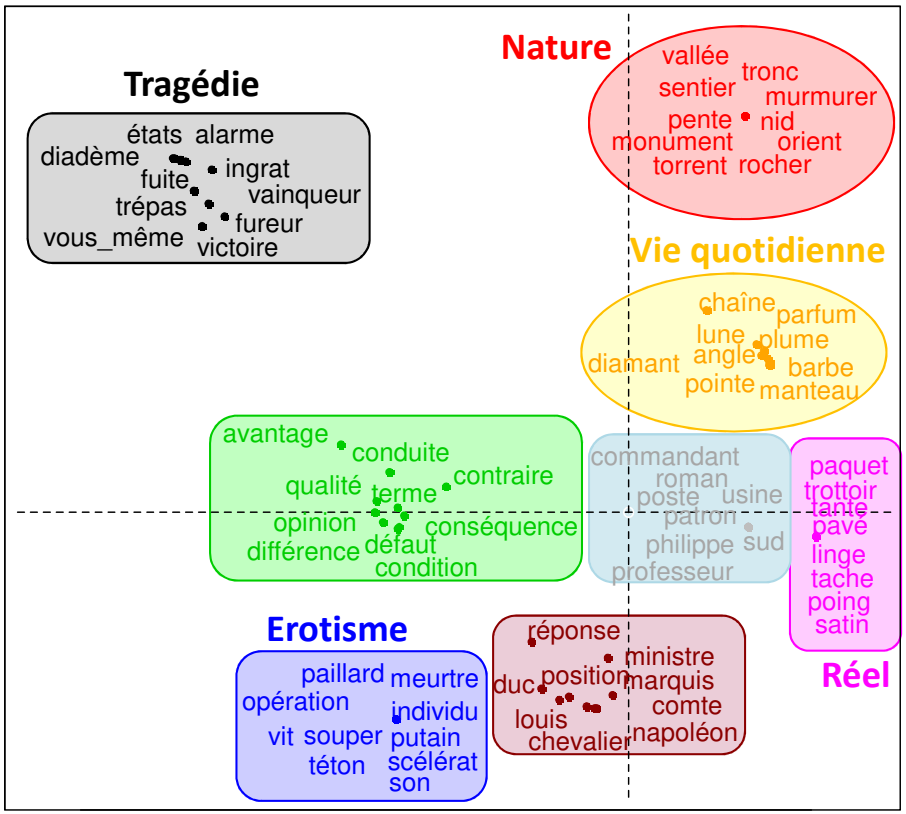
Un exemple en linguistique

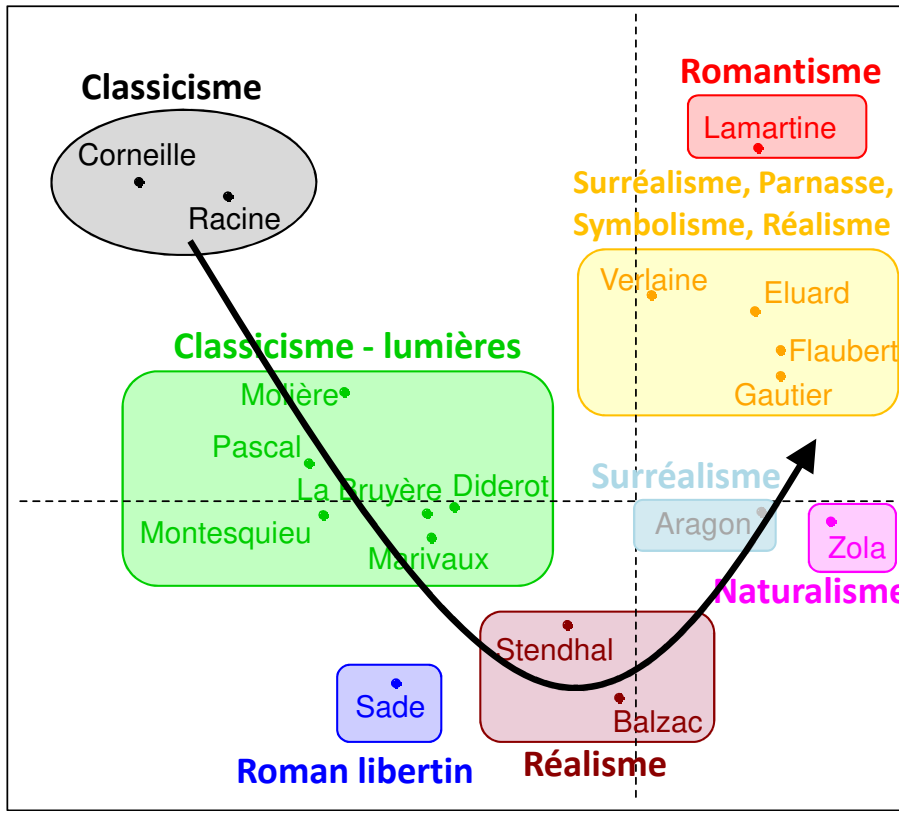




Les mots permettent de caractériser les sujets de prédilection des auteurs et les courants littéraires

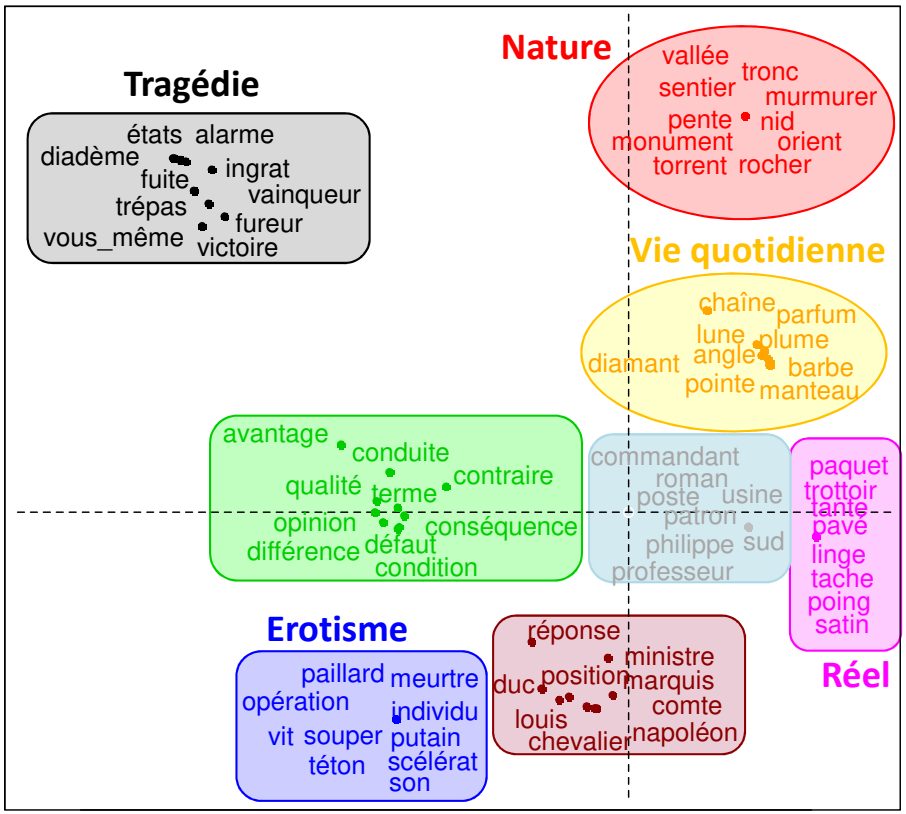
Un exemple en linguistique

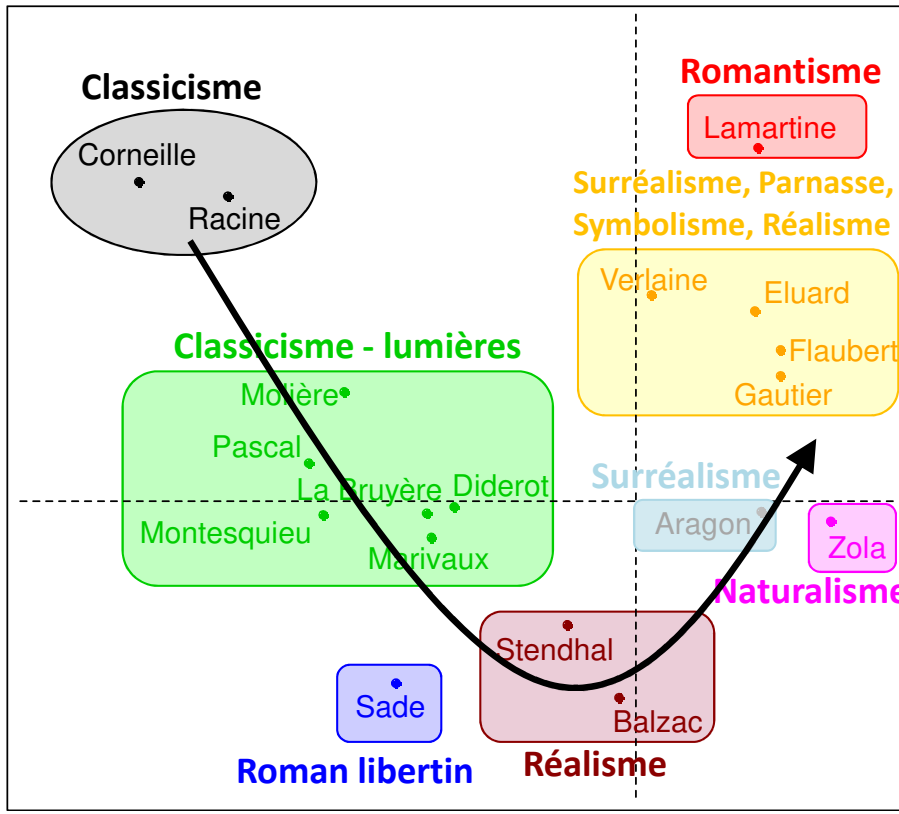




- Le naturalisme est la suite logique du réalisme : le naturalisme montre le milieu où vit le protagoniste pour expliquer son comportement de façon "scientifique"
- Évolution du vocabulaire selon les courants littéraires

Un exemple en linguistique





- 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