Classification non supervisée - Introduction

Cathy Maugis-Rabusseau

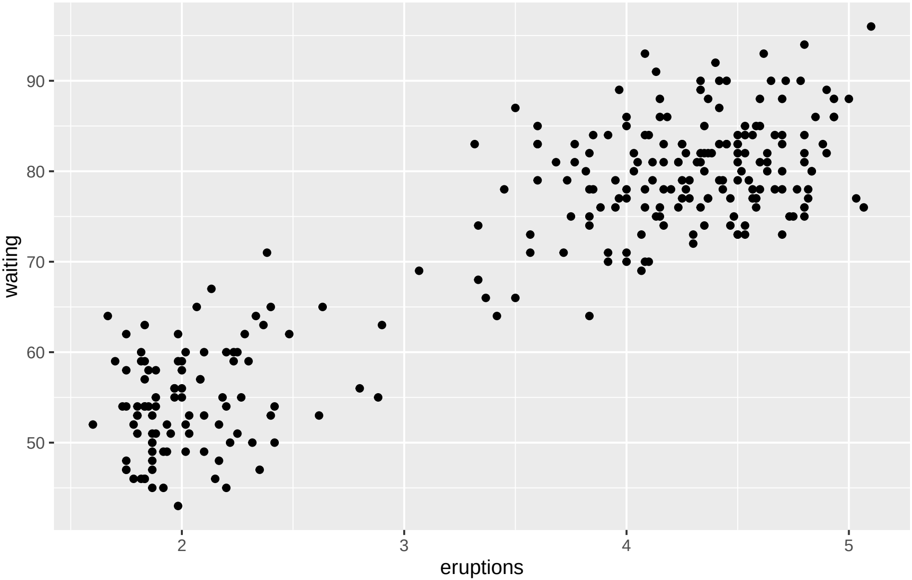
4modIA / INSA Toulouse & ENSEEIHT

2022-2023

Plan

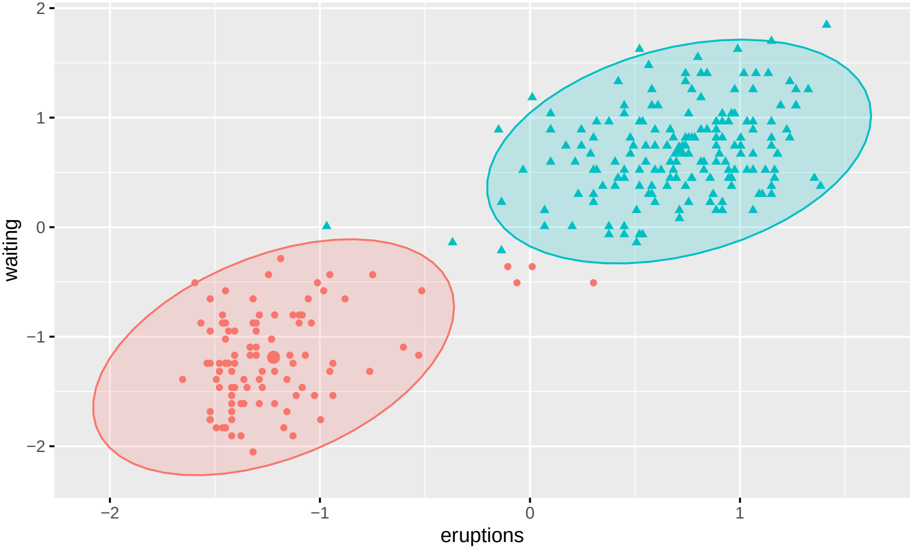
Old Faithful Geyser Data

Azzalini and Bowman [2]



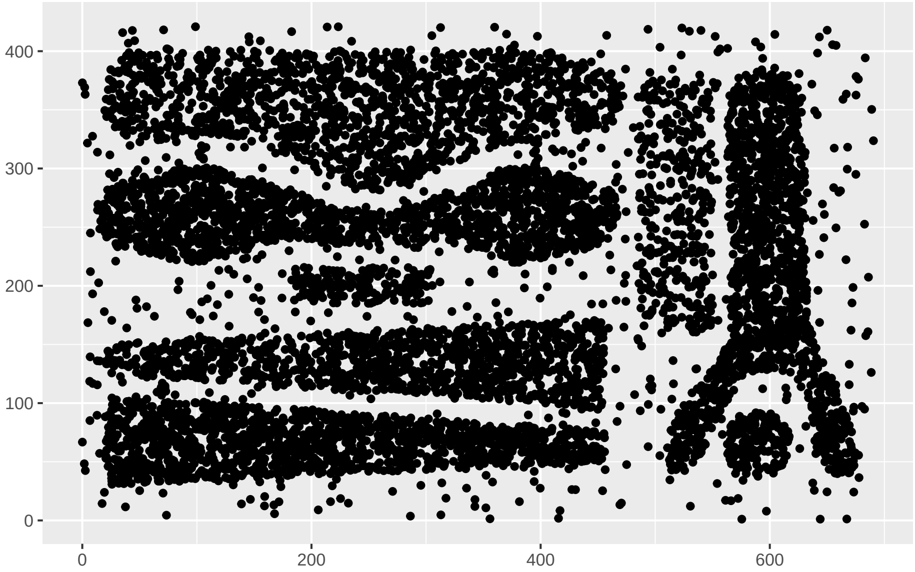
Old Faithful Geyser Data

Azzalini and Bowman [2]

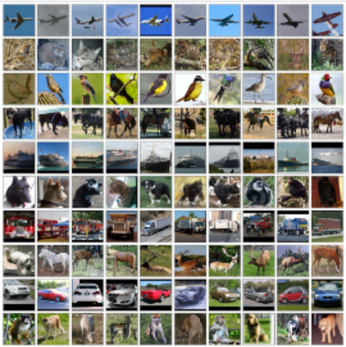


Exemple de classification non supervisée de formes

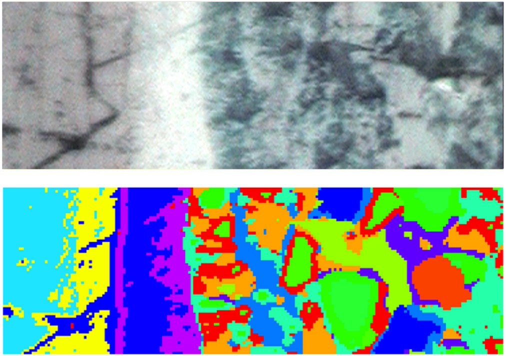
George and Eui-Hong [5]



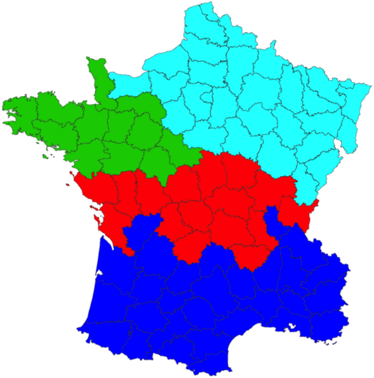
Exemple de classification non supervisée d'images



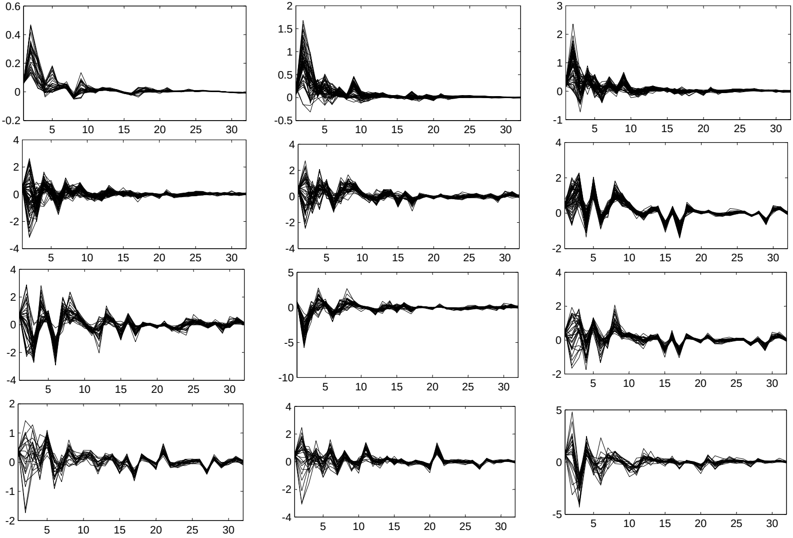
Exemple avec contraintes spatiales



Le Pennec and Cohen (2011)



Exemple de clustering de courbes



Plan

Les données

- On observe n individus décrits par p variables

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \text{ avec } x_i = (x_{i1}, \dots, x_{ip}) \in \mathcal{X}$$

- L'ensemble $\mathcal{X}$ peut-être très variable : $\mathcal{X} = \mathbb{R}^p$, $\{0,1\}^p$, $]-\pi,\pi]^p$, $\mathbb{R}^q \times \{0,1\}^{p-q}$, ...
- On peut partir du
  - Tableau initial des mesures
  - Tableau des mesures transformées
  - Tableau des coordonnées après une réduction de dimension

Objectif du clustering

- Soit $\mathbf{X}$ la matrice de données décrivant n individus

- **Classification** : organisation d'un ensemble d'individus hétérogènes en un ensemble de classes homogènes

- **Non supervisée** : on ne dispose d'aucune partition a priori des n individus et on ne connait pas le nombre de classes K.

$$\Longleftrightarrow$$

Déterminer K classes $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ des n individus à partir de $\mathbf{X}$ telles qu'une classe est une collection d'individus **similaires** entre eux et **dissimilaires** aux individus des autres classes (classes bien séparées).

Impossibilité d'une recherche exhaustive

- On n'abordera ici que des méthodes de "classification dure" : un individu n'appartient qu'à une seule classe

$$\forall i \in \{1, \ldots, n\},\ \exists! k \in \{1, \ldots, K\};\ i \in \mathcal{C}_k.$$

- Recherche exhaustive:

Le nombre de partitions d'un ensemble de n individus en K classes (nombre de Stirling de 2ème espèce)

$$\frac{1}{K!} \sum_{j=0}^{K} (-1)^j (K-j)^n C_K^j$$

- $\simeq 10^{47}$ partitions de $n = 100$ individus en $K = 3$ classes
- $\simeq 10^{68}$ partitions de $n = 100$ individus en $K = 5$ classes

$\Longrightarrow$ recherche exhaustive impossible.

Vocabulaire

Attention à la confusion de terminologie entre le français et l'anglais!

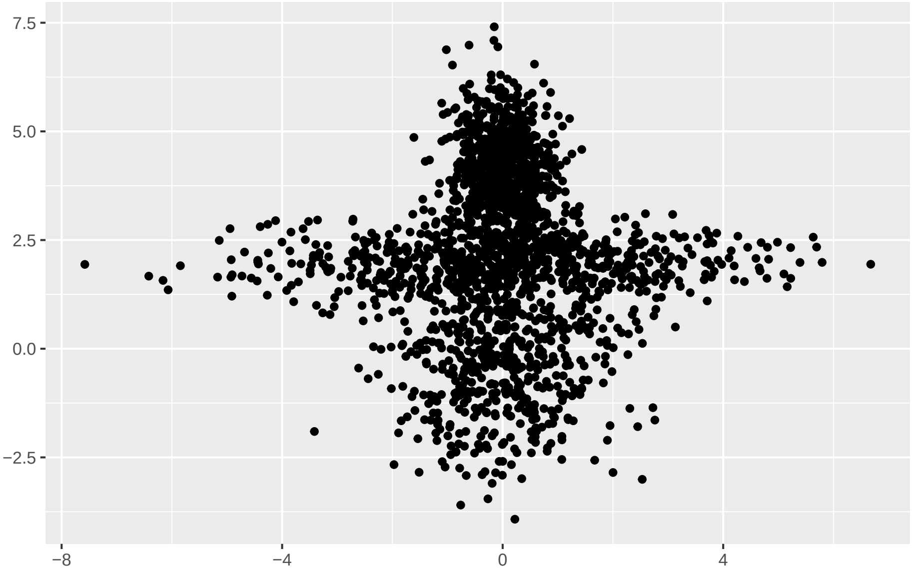
- Classification non supervisée : On ne connaît rien a priori sur les classes

En anglais : Clustering (unsupervised classification)

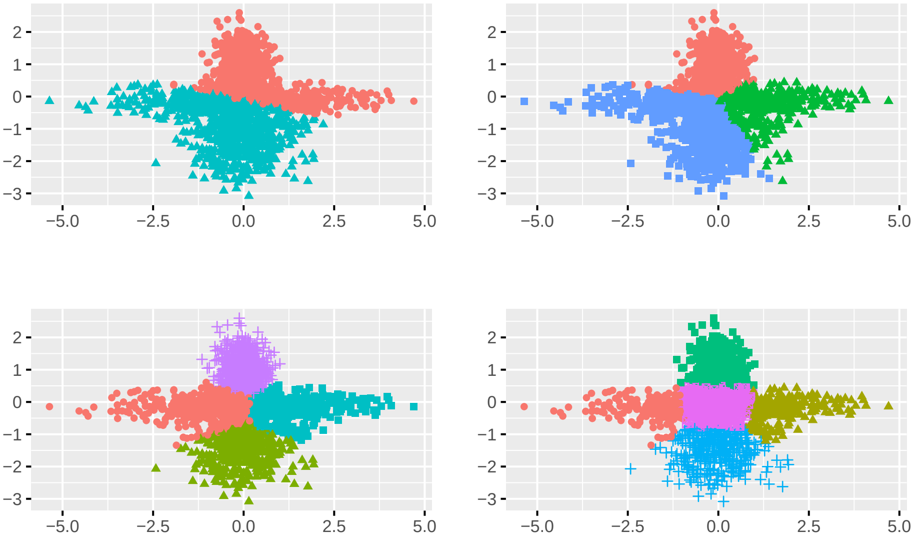
- Classification supervisée : On veut classer un nouvel individu à partir de la connaissance de classes définies a priori.

En anglais : Classification, discriminant analysis

Combien de classes ?



Combien de classes ?



Catégories de méthodes

- Les méthodes de clustering peuvent se différencier par
  - Type de “ressemblance” entre individus en terme de distance, de distribution de probabilité . . .
  - Type de “partitionnement” : hard ou fuzzy clustering
- Grandes catégories de méthodes :
  - Méthodes fondées sur une distance : méthodes hiérarchiques, méthodes par partitionnement, . . .
  - Méthodes basées sur la distribution probabiliste des données
  - Méthodes basées sur les réseaux de neurones
  - . . .

Plan

Comment comparer deux clusterings ?

- On suppose que l'on a obtenu deux partitions à partir des mêmes données $\mathbf{X}$

$$\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\} \text{ et } \tilde{\mathcal{P}}_{\tilde{K}} = \{\tilde{\mathcal{C}}_1, \ldots, \tilde{\mathcal{C}}_{\tilde{K}}\}$$

- Les nombres de classes K et $\tilde{K}$ peuvent être différents !

- Question : comment comparer ces deux classifications ?

Table de contingence

- On peut utiliser une **table de contingence** pour observer si des classes sont communes, des classes sont splittées, ...

	$\tilde{\mathcal{C}}_1$	$\tilde{\mathcal{C}}_2$	...	$\tilde{\mathcal{C}}_{\tilde{K}}$	Sums
$\mathcal{C}_1$	n_{11}	n_{12}	...	$n_{1\tilde{K}}$	a_1
$\mathcal{C}_2$	n_{21}	n_{22}	...	$n_{2\tilde{K}}$	a_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{C}_K$	n_{K1}	n_{K2}	...	$n_{K\tilde{K}}$	a_K
Sums	b_1	b_2	...	$b_{\tilde{K}}$	n

avec $n_{k\ell} = \sharp \left\{ i \in \{1, \dots, n\};\ i \in \mathcal{C}_k \cap \tilde{\mathcal{C}}_\ell \right\}$, $a_k = \sum_{\ell=1}^{\tilde{K}} n_{k\ell}$ et $b_\ell = \sum_{k=1}^{K} n_{k\ell}$.

- Exemple : Classification en 2 et 4 classes

```
      clust2
clust1    1   2   3   4 Sum
   1     25  45   0  27  97
   2      3   0  50   0  53
   Sum   28  45  50  27 150
```

Rand Index (RI)

$$RI(\mathcal{P}_K, \tilde{\mathcal{P}}_{\tilde{K}}) = \frac{A+D}{A+B+C+D}$$

avec

$A =$	Nb de paires d'indiv.	groupés	dans $\mathcal{P}_K$ et	groupés	dans $\tilde{\mathcal{P}}_{\tilde{K}}$
$B =$	" "	groupés	" "	séparés	" "
$C =$	" "	séparés	" "	groupés	" "
$D =$	" "	séparés	" "	séparés	" "

- RI = proportion de paires de points qui sont groupées de la même façon dans les deux partitions.
- Quand $a_k = \sum_{\ell=1}^{\tilde{K}} n_{k\ell}$, on utilise plutôt la version normalisée ARI (Adjusted Rand Index).

Adjusted Rand Index (ARI)

$$ARI(\mathcal{P}_K, \tilde{\mathcal{P}}_{\tilde{K}}) = \frac{RI - \mathbb{E}[RI]}{max(RI) - \mathbb{E}[RI]}$$

avec

- $$\begin{aligned} \mathbb{E}(RI) &= \text{indice obtenu en partitionnant les données au hasard} \\ &= \left[\sum_k \begin{pmatrix} a_k \\ 2 \end{pmatrix} \sum_\ell \begin{pmatrix} b_\ell \\ 2 \end{pmatrix}\right] / \begin{pmatrix} n \\ 2 \end{pmatrix} \end{aligned}$$
- $\mathrm{RI} = \sum_{k\ell} \begin{pmatrix} n_{k\ell} \\ 2 \end{pmatrix}$
- $\max(\mathrm{RI}) = \frac{1}{2}\left[\sum_k \begin{pmatrix} a_k \\ 2 \end{pmatrix} + \sum_\ell \begin{pmatrix} b_\ell \\ 2 \end{pmatrix}\right]$

Plus le ARI est proche de 1, plus les deux partitions se ressemblent

Adjusted Rand Index (ARI)

- Exemple :

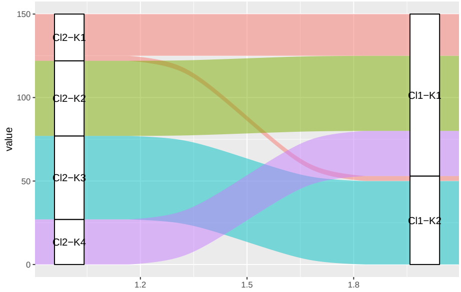
```
addmargins(table(clust1,clust2))
```

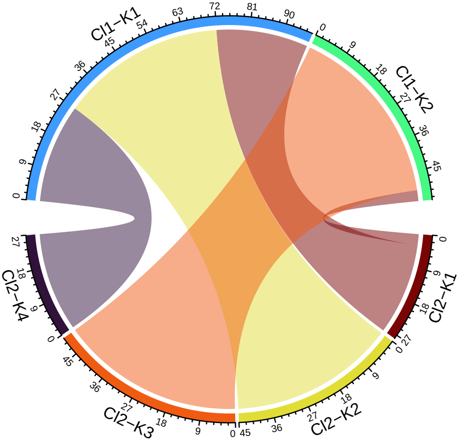
```
      clust2
clust1    1   2   3   4 Sum
   1     25  45   0  27  97
   2      3   0  50   0  53
   Sum   28  45  50  27 150
```

```
adjustedRandIndex(clust1,clust2)
```

```
[1] 0.4412583
```

Quelques outils de visualisation





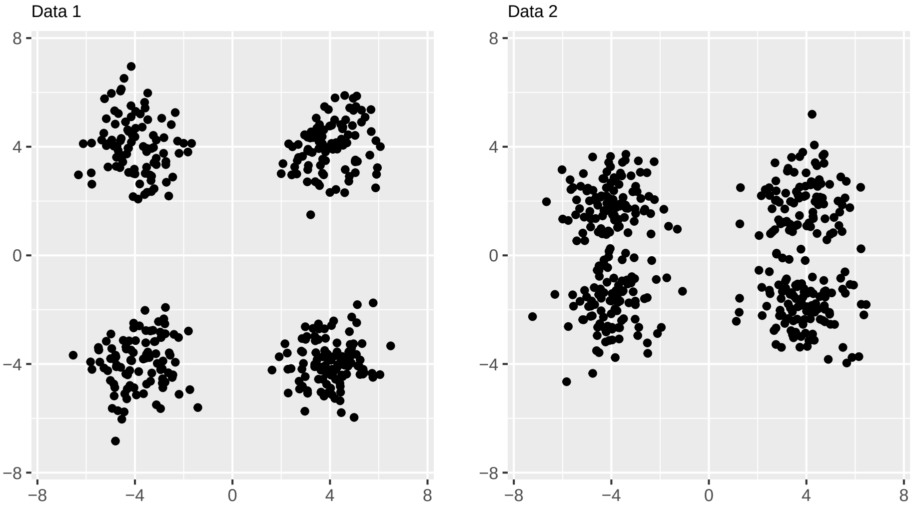
Plan

Plan du cours

- Chapitre 1 : (Dis)similarités, distances et inerties
- Chapitre 2 : Classification non supervisée par partitionnement et DBSCAN
- Chapitre 3 : Classification non supervisée hiérarchique
- Chapitre 4 : Classification par modèles de mélanges finis

Données simulées Data1 et Data2

- Jeux de données jouet ($n = 400$, $p = 2$)

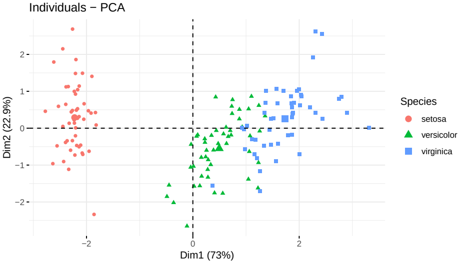


Données Iris [1]

- 3 espèces d'iris : setosa (50), versicolor (50) et virginica (50)
- Mesures en centimètres de longueur du sépale, largeur du sépale, longueur du pétale et largeur du pétale

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Fig. 2 – *I.setosa, I.versicolor, I.Virginica*



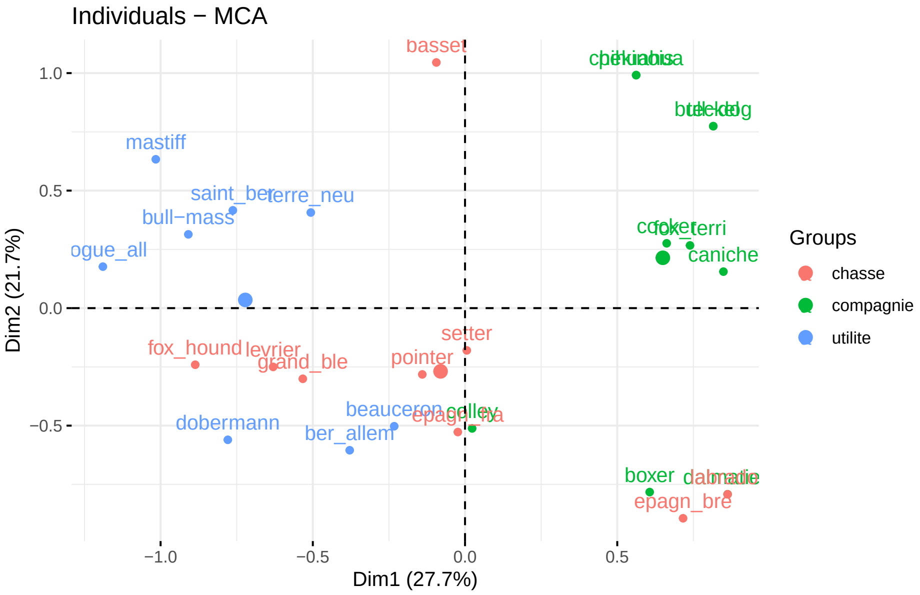
Données Races de chien [3]

- Données : 27 races de chiens décrites par 6 variables

- 6 variables descriptives qualitatives :
  - taille : petite (1), moyenne (2), grande (3)
  - poids : petite (1), moyenne (2), grande (3)
  - vélocité : petite (1), moyenne (2), grande (3)
  - intelligence : petite (1), moyenne (2), grande (3)
  - affectation : faible (1), forte (2)
  - agressivité : faible (1), forte (2)

- 1 autre variable "fonction" : compagnie (1), chasse (2), utilité (3)

Données Races de chien



Données de maladie du coeur [4]

- $n = 270$ individus
- Variables binaires: sexe, sucre dans le sang à jeun > 120mg/dl, angine induite par l'effort
- Variables nominales: douleurs à la poitrine (4 types),résultat électrocardiagraphique au repos (3 types), ...
- Variables réelles: Age ,pression artérielle au repos, taux de cholestérol, fréquence cardiaque maximale atteinte, ...
- Données disponibles sur le site de l'UCI (Lien)

	Age	Sex	ChestPainType	RestBloodPressure	SerumCholestoral	FastingBloodSugar
1	70	1	4	130	322	0
2	67	0	3	115	564	0
3	57	1	2	124	261	0
4	64	1	4	128	263	0
5	74	0	2	120	269	0
6	65	1	4	120	177	0

	ResElectrocardiographic	MaxHeartRate	ExerciseInduced	Slope	MajorVessels	Thal
1	2	109	0	2	3	3
2	2	160	0	2	0	7
3	0	141	0	1	0	7
4	0	105	1	2	1	7
5	2	121	1	1	1	3
6	0	140	0	1	0	7

References I

[1] Edgar Anderson. "The irises of the Gaspe Peninsula". In: *Bull. Am. Iris Soc.* 59 (1935), pp. 2–5.

[2] Adelchi Azzalini and Adrian W Bowman. "A look at some data on the Old Faithful geyser". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 39.3 (1990), pp. 357–365.

[3] A Bréfort. "L'étude des races canines à partir de leurs caractéristiques qualitatives". In: *Groupe HEC-Jouy en Josas* (1982).

[4] John Crowley and Marie Hu. "Covariance analysis of heart transplant survival data". In: *Journal of the American Statistical Association* 72.357 (1977), pp. 27–36.

[5] Karypis George and Han Eui-Hong. "Hierarchical clustering using dynamic modelling". In: *Computer* 4 (1999), pp. 68–75.