Chapitre 5-1) Clustering / Classification non supervisée - Introduction

Maxime El Masri

3 MIC / INSA Toulouse

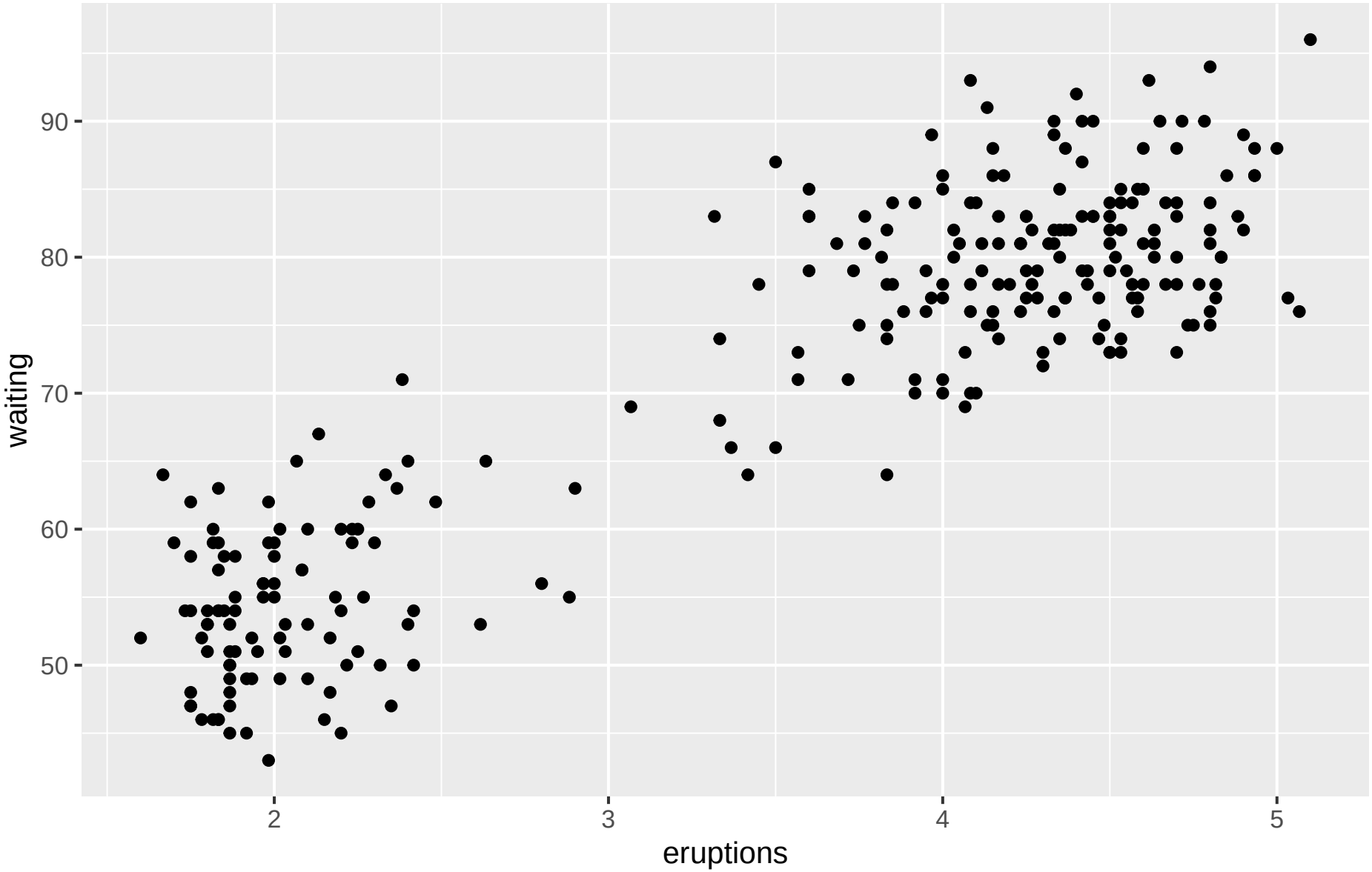
2023-2024

Plan du cours

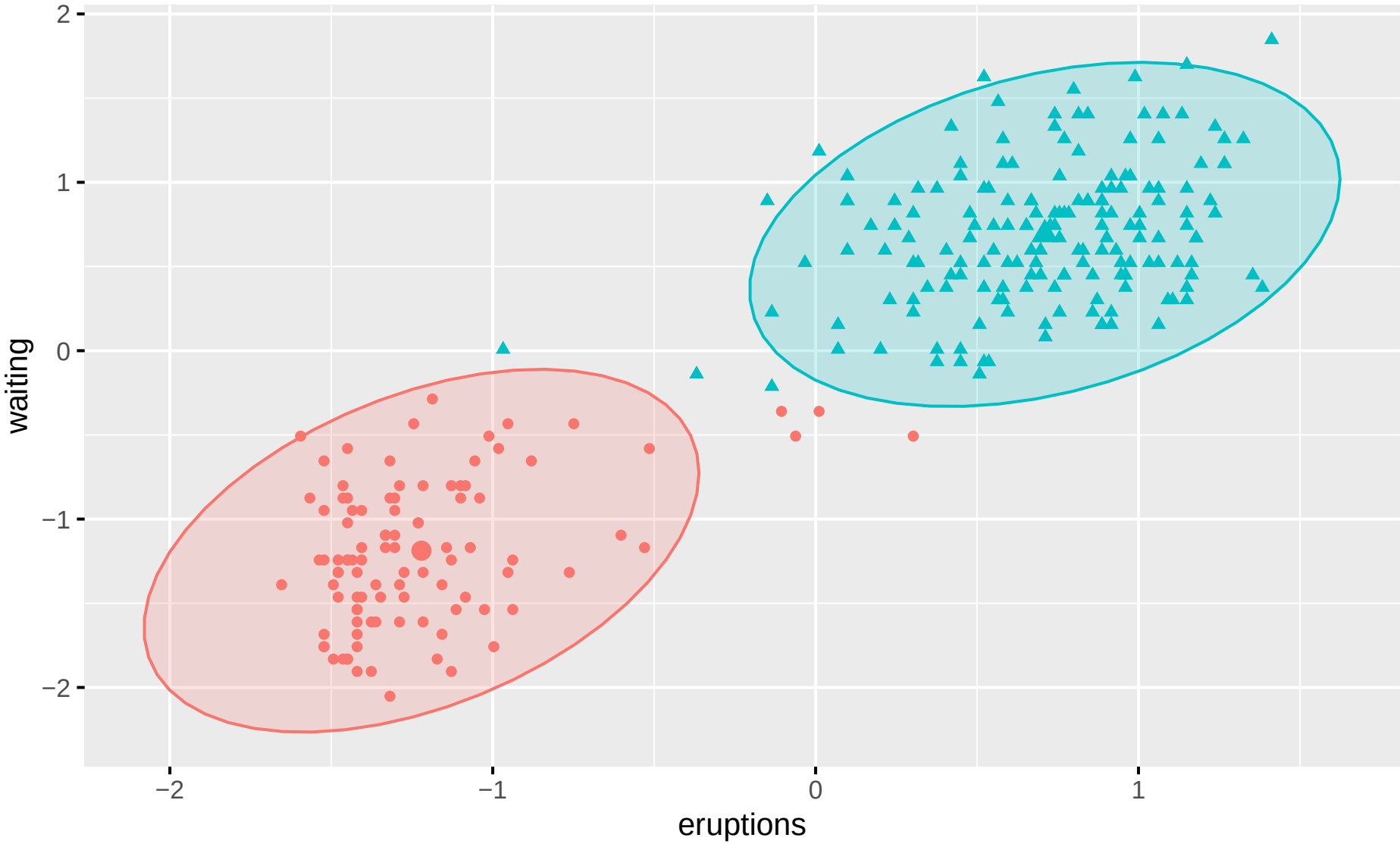
- Partie 1 : Introduction
- Partie 2 : (Dis)similarités, distances et inerties
- Partie 3 : Méthode des K-means et DBSCAN
- Partie 4 : Classification hiérarchique

Partie 1

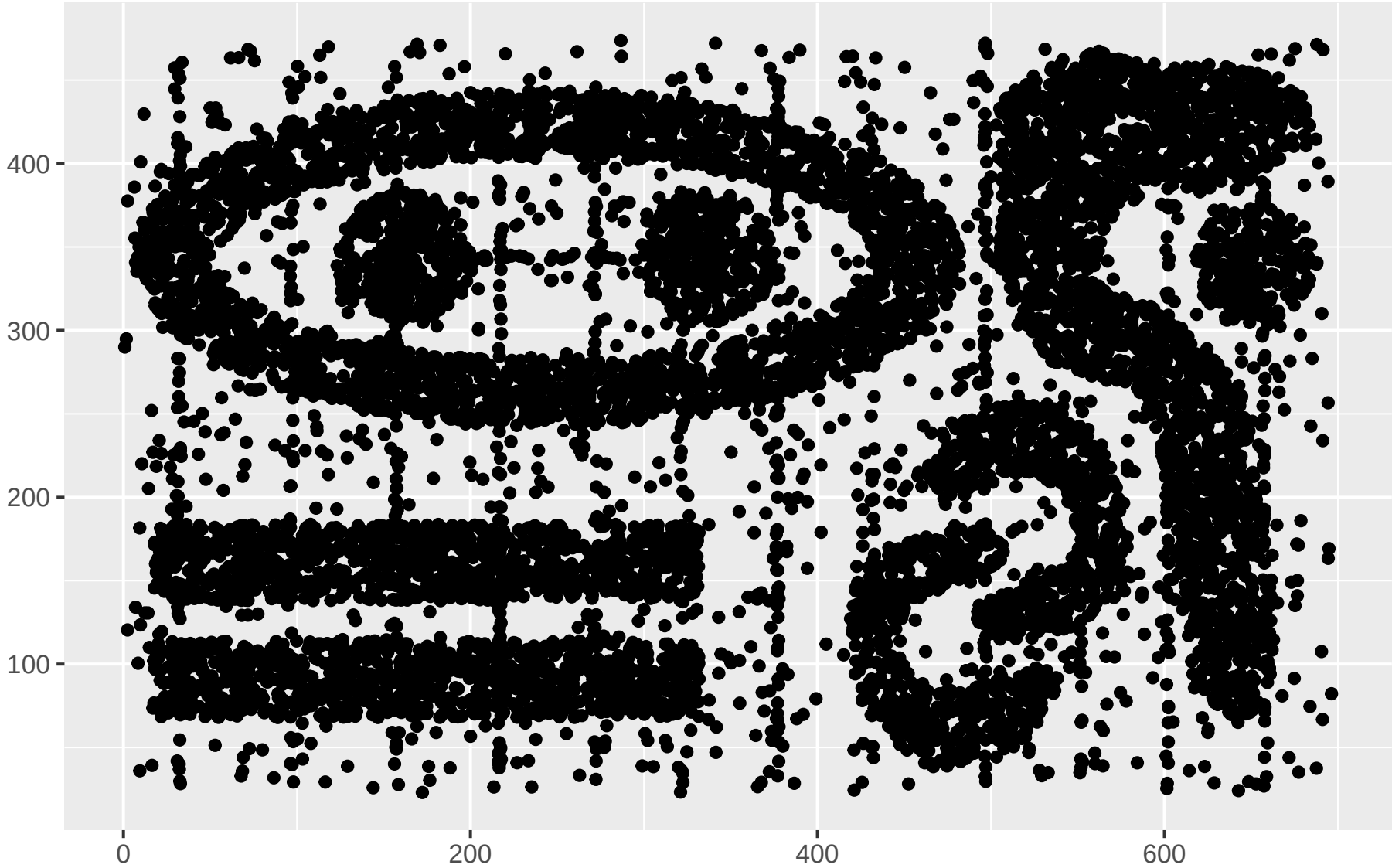
Données “Old Faithful Geyser”



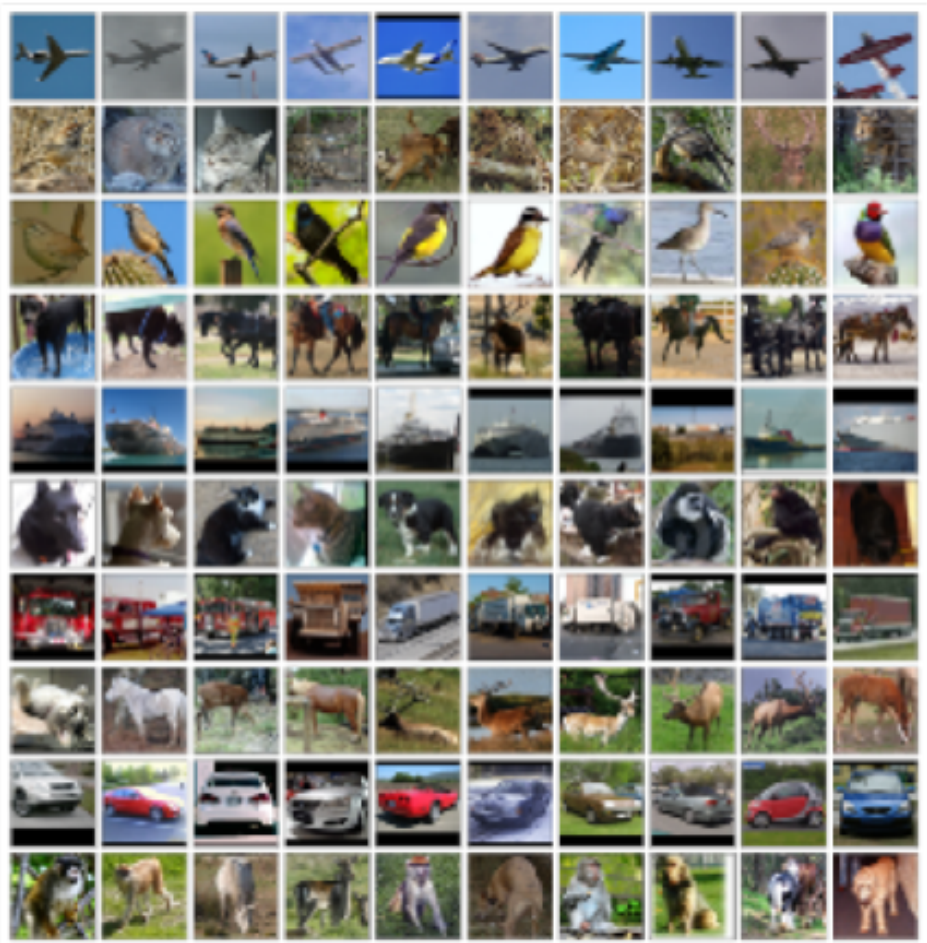
Données “Old Faithful Geyser”



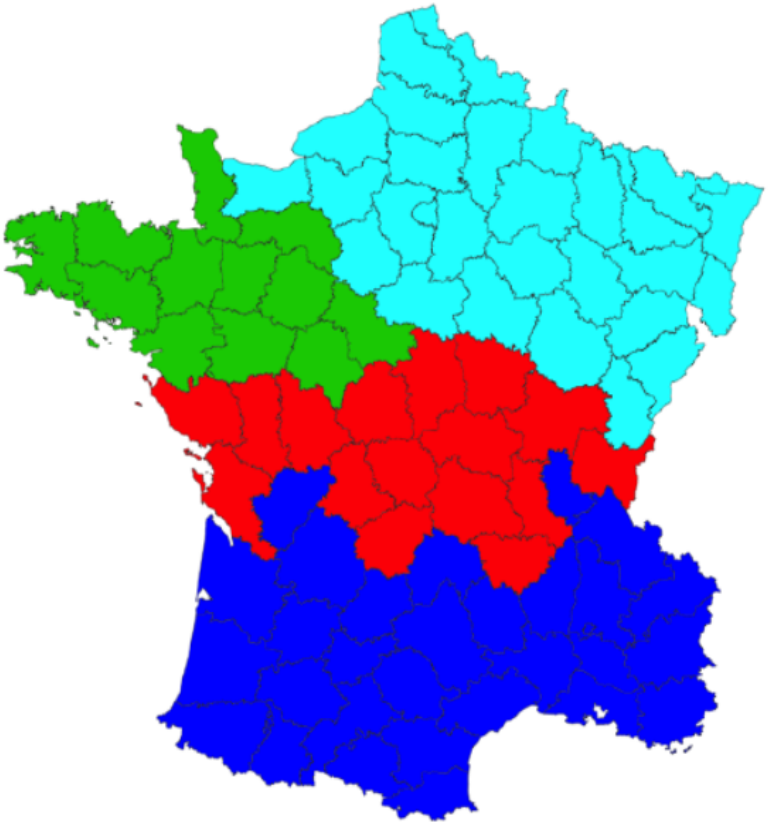
Exemple de classification non supervisée de formes



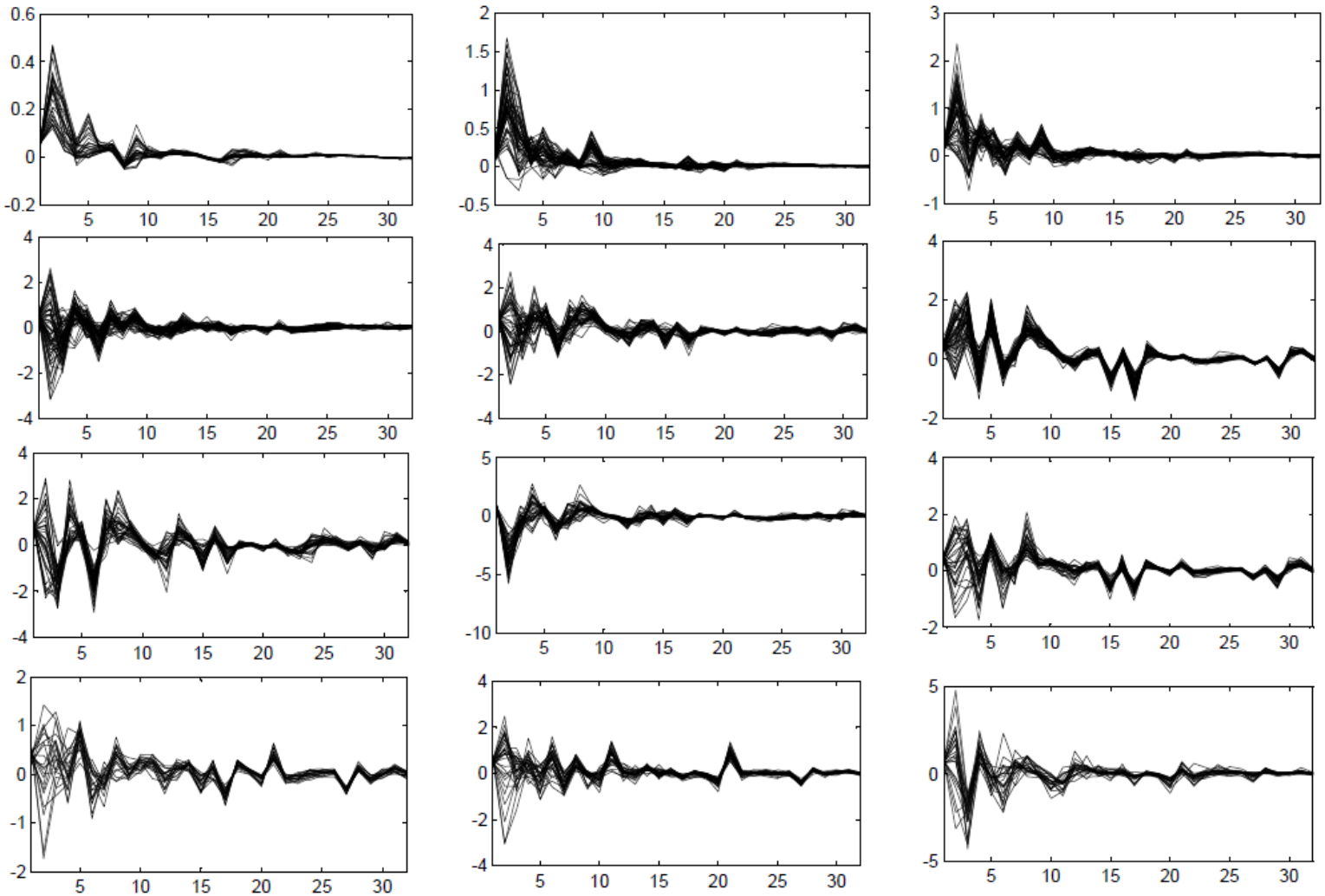
Exemple de classification non supervisée d'images



Exemple classification non supervisée avec contraintes spatiales



Exemple de classification non supervisée de courbes



Partie 1

Les données

- On observe n individus décrits par p variables

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

- 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.
- Objectif : 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).

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

Impossibilité d'une recherche exhaustive

- On n'abordera ici que des méthodes de "classification dure" :

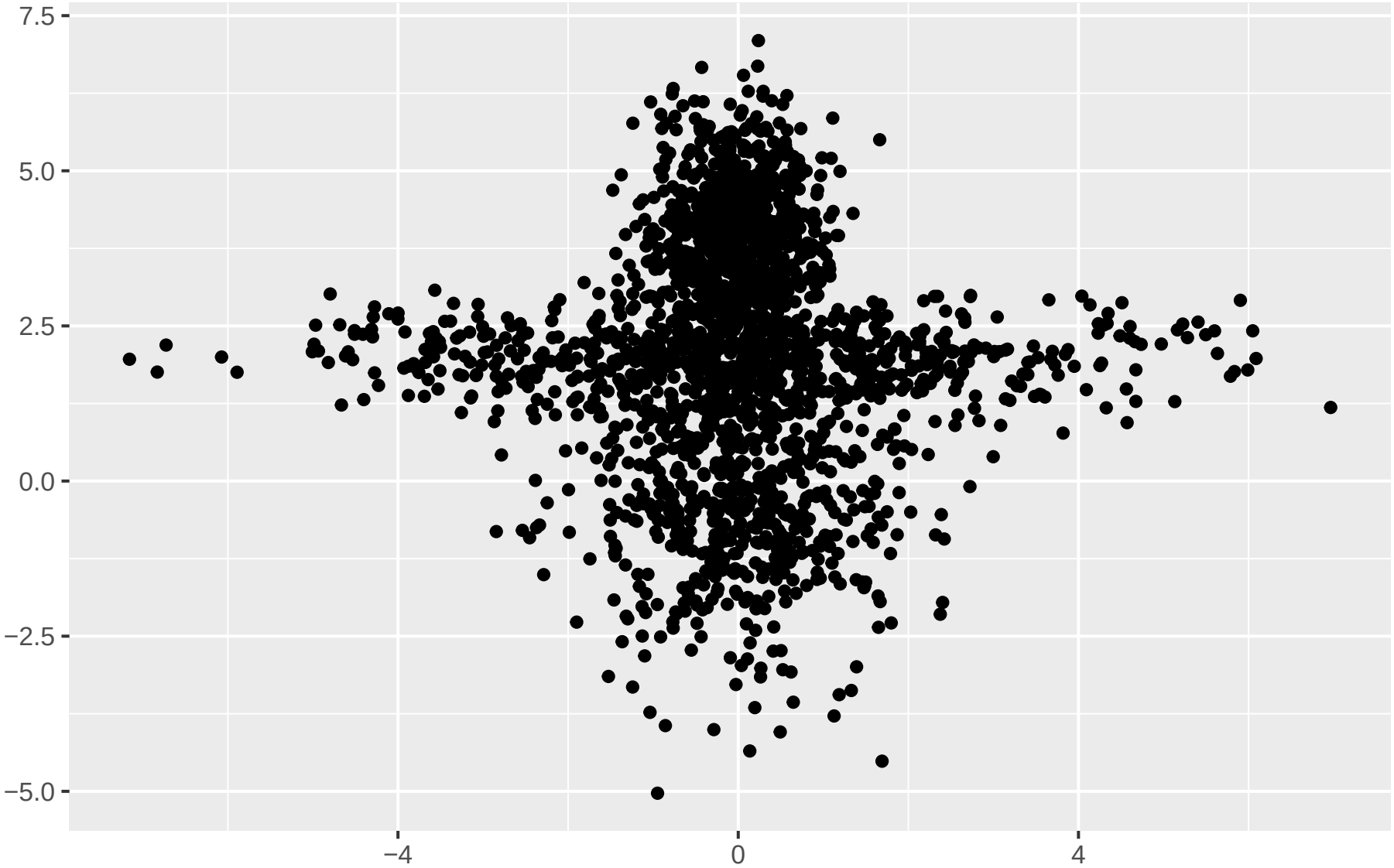
un individu n'appartient qu'à une seule classe

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

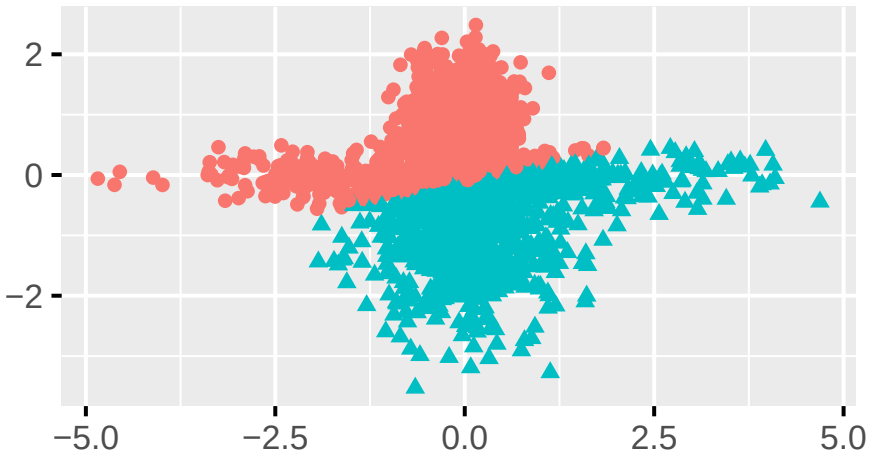
- Nombre de partitions d'un ensemble de n individus en K classes (donné par le nombre de Stirling) $\sim \frac{K^n}{K!}$. Par exemple :
  - $\simeq 10^{47}$ partitions de $n = 100$ individus en $K = 3$ classes
  - $\simeq 10^{68}$ partitions de $n = 100$ individus en $K = 5$ classes

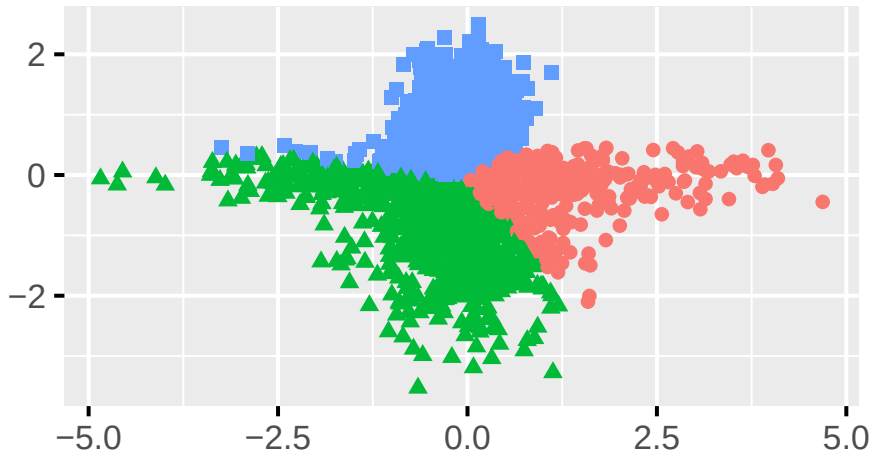
$\longrightarrow$ **recherche exhaustive impossible**.

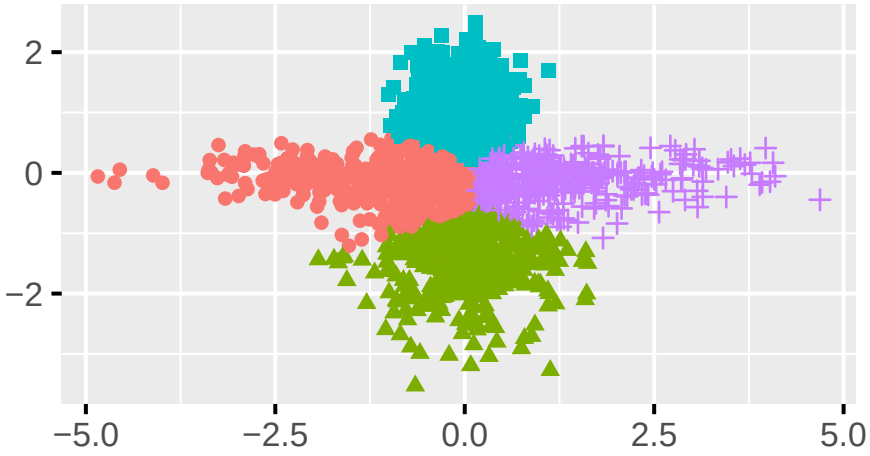
Combien de classes ?

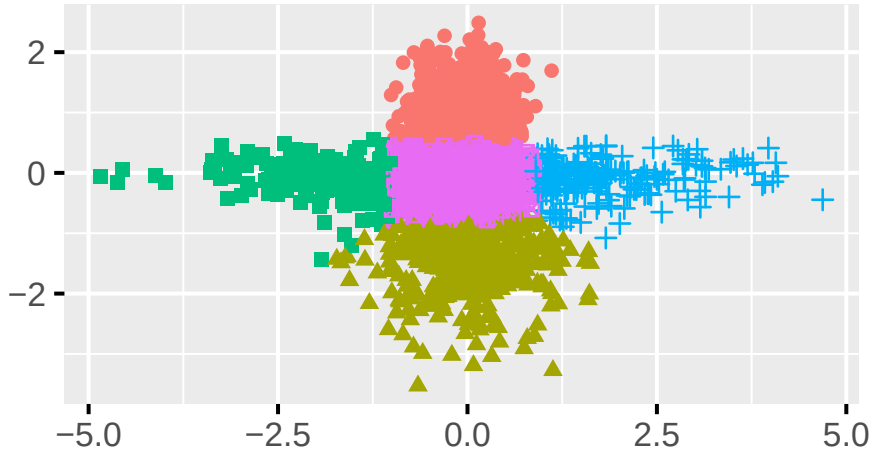


Combien de classes ?









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
  - ...

Dans ce cours

- 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

Partie 1

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 séparées, ...

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

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.

- Remarque : $A + B + C + D = \binom{n}{2}$

Adjusted Rand Index (ARI)

$$ARI(\mathcal{P}_K, \tilde{\mathcal{P}}_{\tilde{K}}) = \frac{U - W}{V - W}$$

avec

$$W = \left[\sum_k \binom{a_k}{2} \sum_\ell \binom{b_\ell}{2}\right] / \binom{n}{2}$$

$$U = \sum_{k\ell} \binom{n_{k\ell}}{2}$$

$$V = \frac{1}{2}\left[\sum_k \binom{a_k}{2} + \sum_\ell \binom{b_\ell}{2}\right]$$

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

Exemple

$n = 4$, $P_2 = (ab)(cd)$, $\tilde{P}_2 = (a)(bcd)$.

- Table de contingence :

	$\tilde{\mathcal{C}}_1 = (a)$	$\tilde{\mathcal{C}}_2 = (bcd)$	Sums
$\mathcal{C}_1 = (ab)$	$n_{11} = 1$	$n_{12} = 1$	$a_1 = 2$
$\mathcal{C}_2 = (cd)$	$n_{21} = 0$	$n_{22} = 2$	$a_2 = 2$
Sums	$b_1 = 1$	$b_2 = 3$	$n = 4$

- RI$=\frac{1+2}{6} = \frac{1}{2}$, en effet, $\binom{4}{2} = 6$, seuls c et d sont groupés dans P et $\tilde{P}$, (a, c) et (a, d) sont séparés dans les 2 partitions.
- ARI$=\frac{1-1}{5/2-1} = 0$. (exercice)

Quelques outils de visualisation