Université Paul Sabatier
TOULOUSE III

BECNM – Analyses de Données Multivariées

Méthodes de classification

Gaël Grenouillet
gael.grenouillet@univ-tlse3.fr

Quelques termes

- **Classification** : action de constituer des classes selon un ou plusieurs critères à partir d'éléments (personnes, des objets ou des notions)

- **Classement** : action de ranger dans la classe la plus appropriée

- **Clustering** : regroupement automatique

- **Apprentissage supervisé :** étant donné un ensemble de classes (connues), établir les « meilleures » règles de classement.

- **Apprentissage non supervisé** : aucune connaissance (hypothèse) de classes au départ.

Apprentissage supervisé

Les classes sont connues, on dispose d'exemples de chaque classe

- Approche probabiliste
- Notion d'apprentissage (*machine learning*)

Objectif :

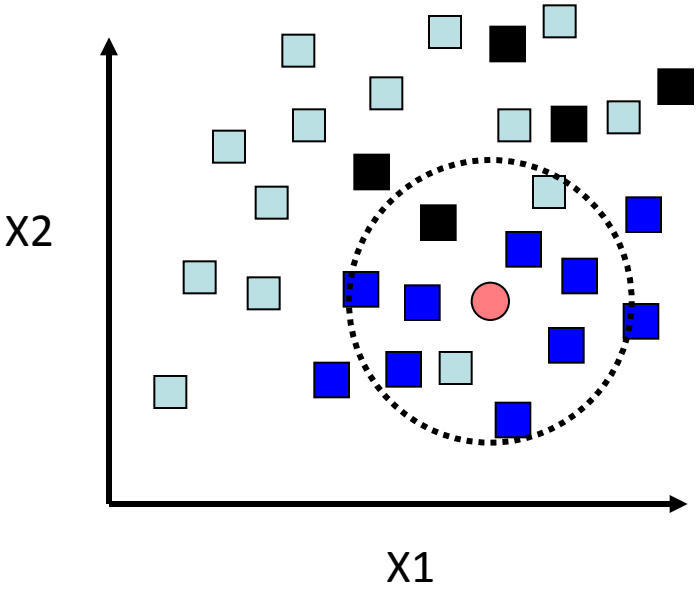
- modéliser la relation entre les observations et l'information cible (classe d'appartenance)
- identifier la classe d'appartenance d'un objet à partir d'un ensemble de descripteurs (caractéristiques)

Nombreux outils :

K plus proches voisins
Arbres de décision
Analyse Discriminante
Régression logistique
Réseaux bayésiens
Réseaux de neurones
Algorithmes génétiques
SVM (Support Vector Machines)

K plus proches voisins (K-nearest neighbor, K-NN)

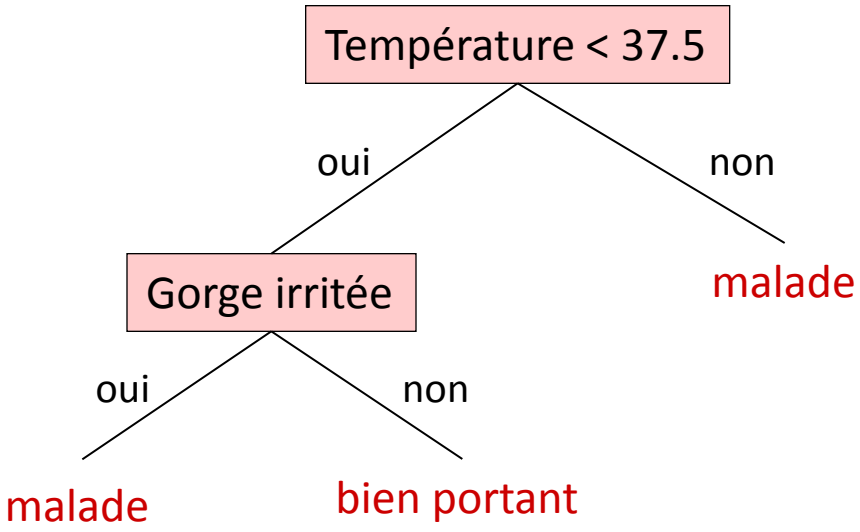
- Approche très simple
- Pas d'apprentissage (aucun modèle n'est induit à partir des données)
- Une donnée de classe inconnue est comparée à toutes les données stockée. On choisit pour la nouvelle donnée la classe majoritairement représentée par les K plus proches voisins.
- Procédure lourde (temps de calculs importants)



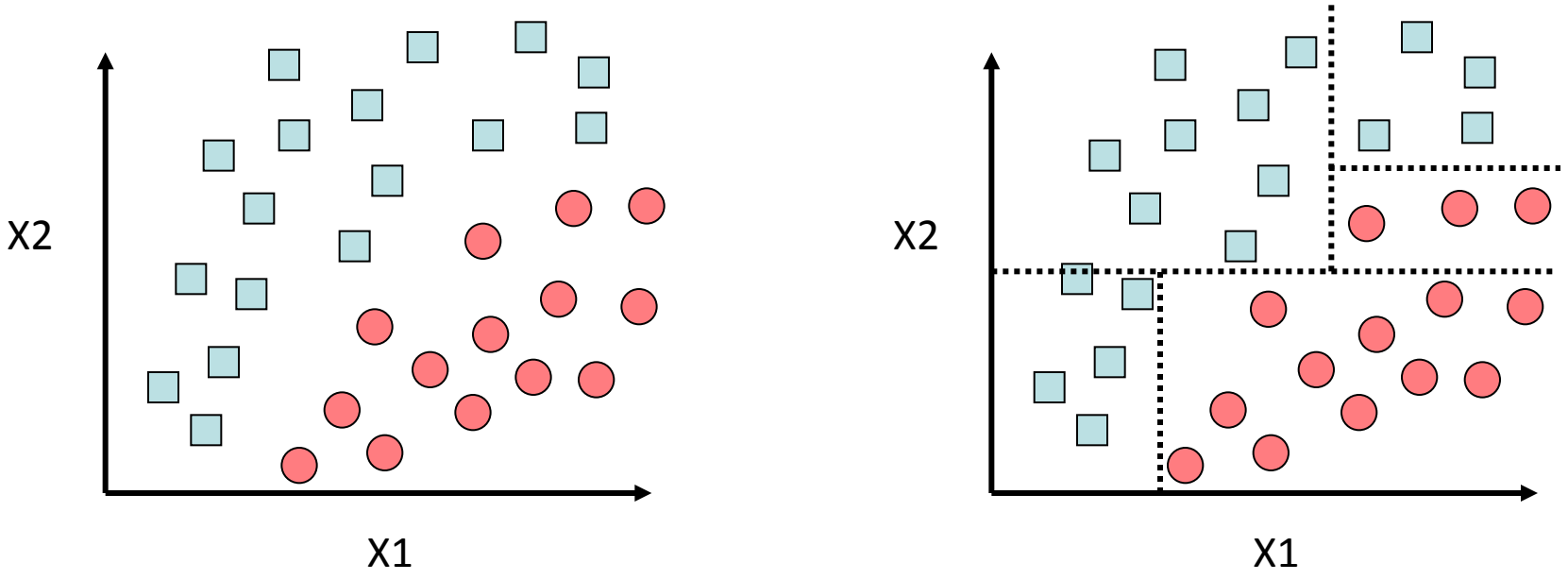
K=10 : ● → ■

Arbres de décisions

- Outil d'aide à la décision
- Vise à produire une procédure de classification interprétable (lisibilité du modèle de prédiction)
- Capacité à sélectionner automatiquement les variables discriminantes



Arbres de décisions



Décomposition du problème de classification en une suite de tests correspondant à une partition de l'espace des données en sous-régions homogènes en terme de classe

Apprentissage non supervisé

Les classes ne sont pas connues

Objectif :

A partir de n observations, constituer k groupes tels que :

- ces groupes soient constitués d'observations semblables
- ces groupes soient le plus différents possibles

- **Méthodes non hiérarchiques** (par partitionnement) :

  Construire k partitions et les corriger
  jusqu'à obtenir une similarité satisfaisante

K-means
K-medoids
Clarans
Self-Organizing Map

- **Méthodes hiérarchiques** :

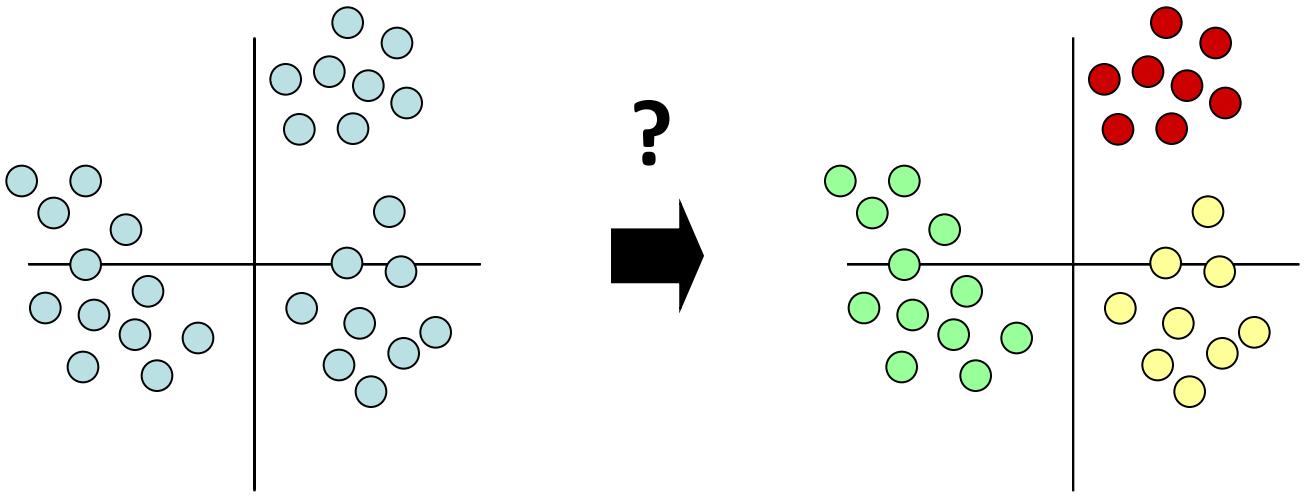
  Créer une décomposition hiérarchique par
  agglomération ou division de groupes similaires
  ou dissimilaires

Hierarchical clustering
Agnes
Diana
Birch
Cure
Rock

Méthodes non hiérarchiques

K-means

« Découper » un nuage de point en plusieurs sous-nuages
Chaque sous-nuage est caractérisé par son centre de gravité



Algorithme de partition → **centres mobiles** (Forgy 1965)

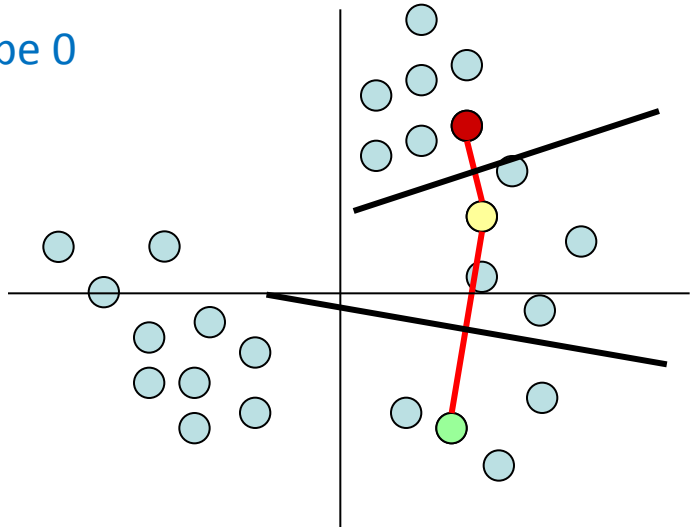
Ensemble I (n individus, p variables) à partitionner en **k classes**
Les **n points** sont munis d'une **distance notée d**
(distance Euclidienne ou du χ^2)

Méthodes non hiérarchiques

K-means

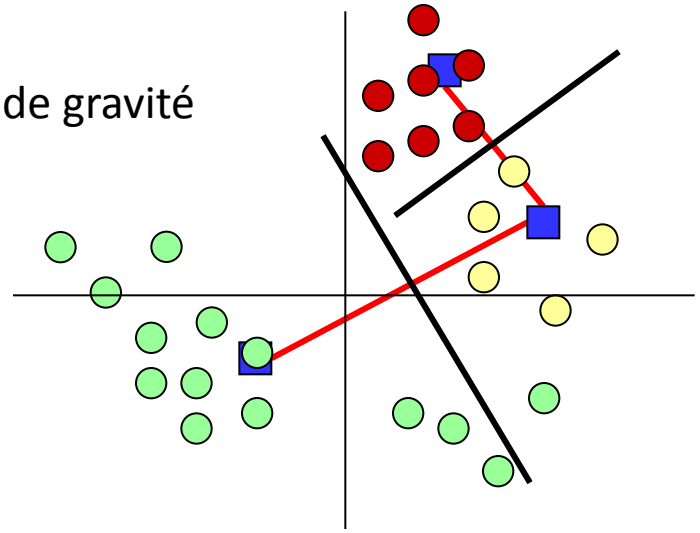
Etape 0

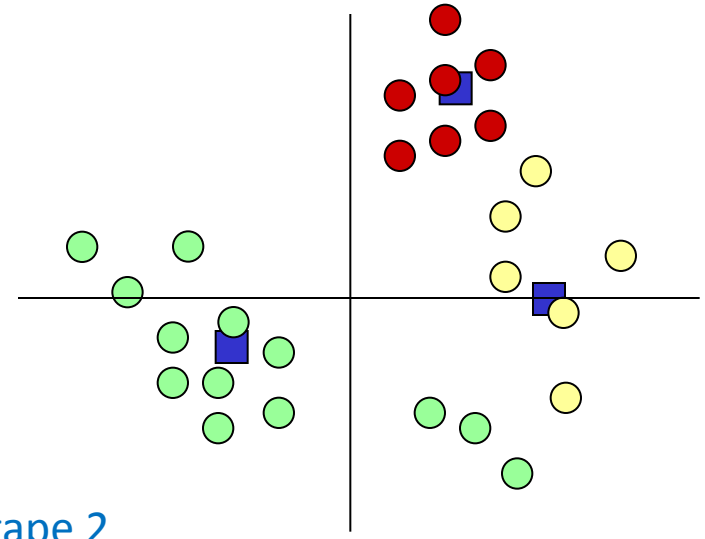
k=3



Etape 1

3 centres de gravité

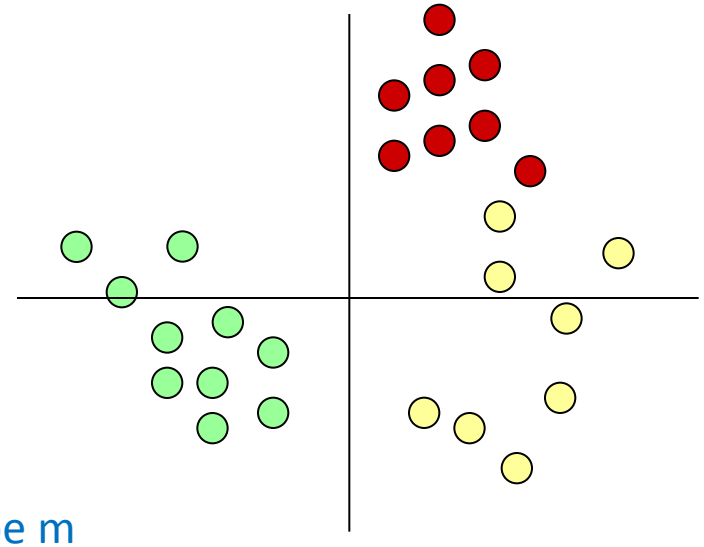




Etape 2

Modification des classes

Nouveaux centres de gravité

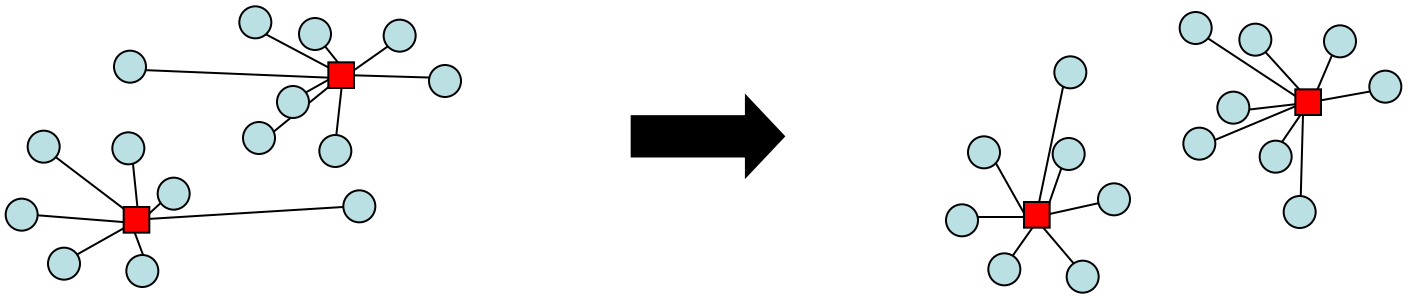


Etape m

3 groupes stables

Méthodes non hiérarchiques

K-means



(étape n) **(étape n+1)**

Centres C^n

A l'étape n, la dispersion intra-groupe correspond à la distance des objets i, de centre C^n

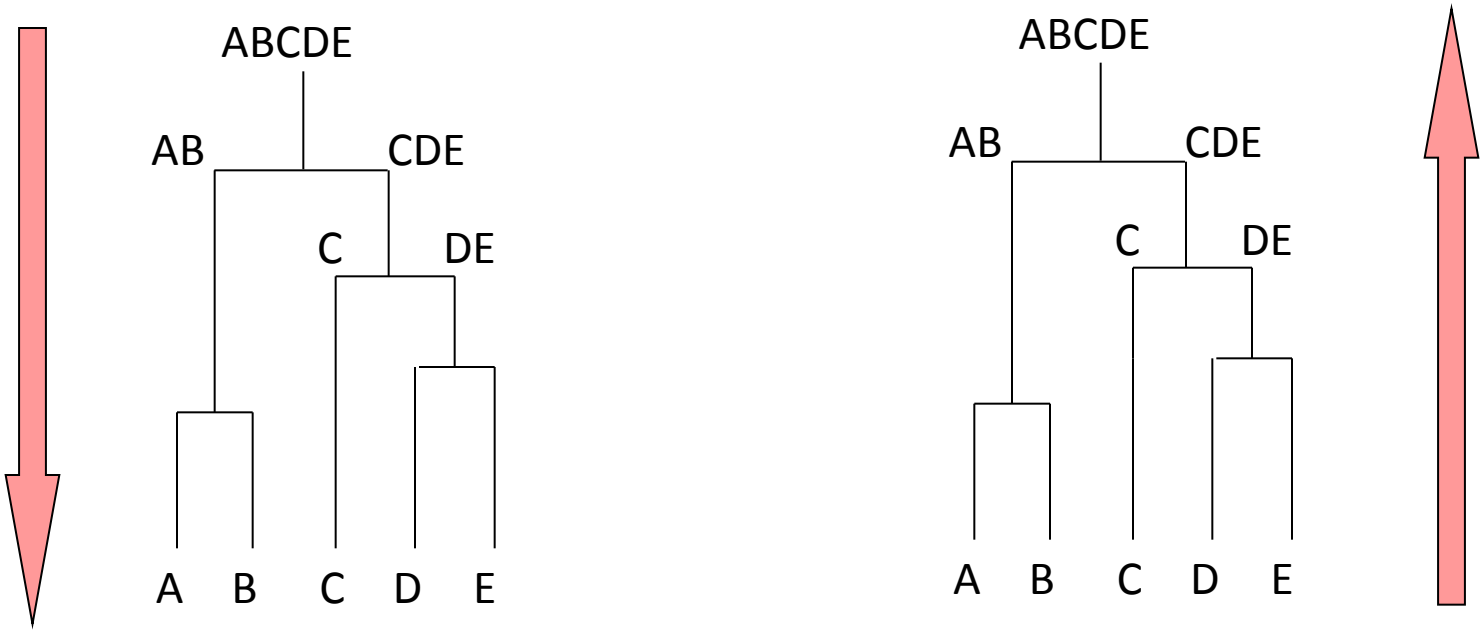
Les objets i sont ensuite ré-attribués aux groupes en fonction de la distance minimale les séparant des C^{n+1}

Entre (n) et (n+1), la dispersion intra-groupe diminue (ou reste stable)

Méthodes hiérarchiques

Le nombre de groupe attendu n'est pas précisé
(contrairement à la classification non-hiérarchique)

- Hiérarchie **descendante** ou **ascendante**



Algorithme divisif **Algorithme agrégatif**

Représentation graphique sous la forme d'un **dendrogramme**

Classif. hiérarchique ascendante

Algorithme de base

Algorithme agrégatif

fonction `hclust()`

Etape 1 : n individus

Etape 2 : Matrice de distance

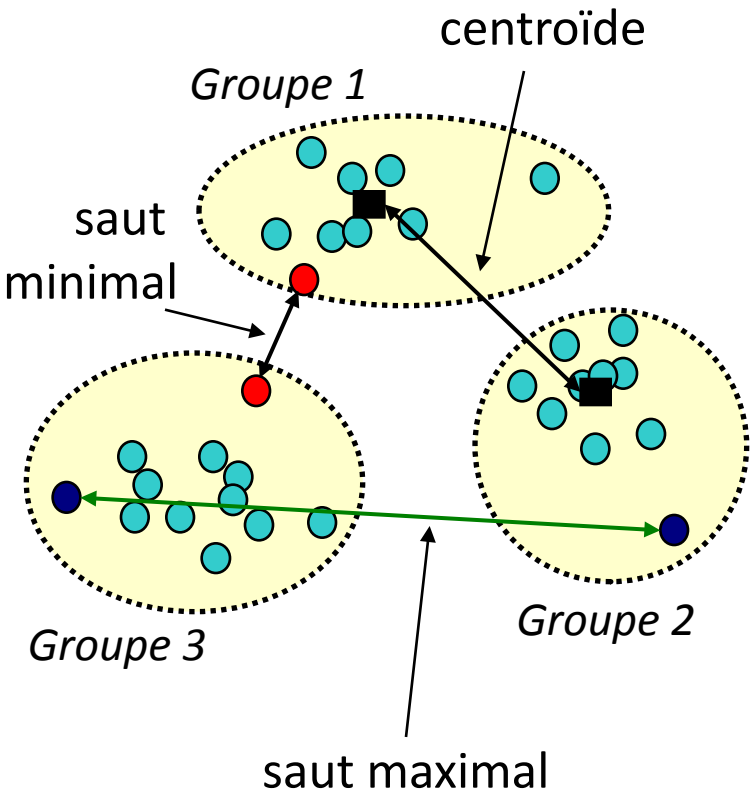
Etape 3 : Agrégation des 2 objets les plus proches (plus faible distance) → n-1 éléments

Etape 4 : Recalcul de la matrice des distance : nouvelles distances des n-2 éléments au groupe formé

A partir de ce stade, l'agrégation des n-1 éléments (1 groupe et n-2 objets isolés) dépendra du critère d'agrégation choisi (=mesure de dissimilarité entre deux groupes)

Classif. hiérarchique ascendante

Critère d'agrégation



- **saut minimal / lien minimum** (*single linkage*) : distance entre les 2 plus proches voisins de 2 groupes

- **saut maximal / lien maximum** (*complete linkage*) : distance entre les 2 membres les plus distants de 2 groupes

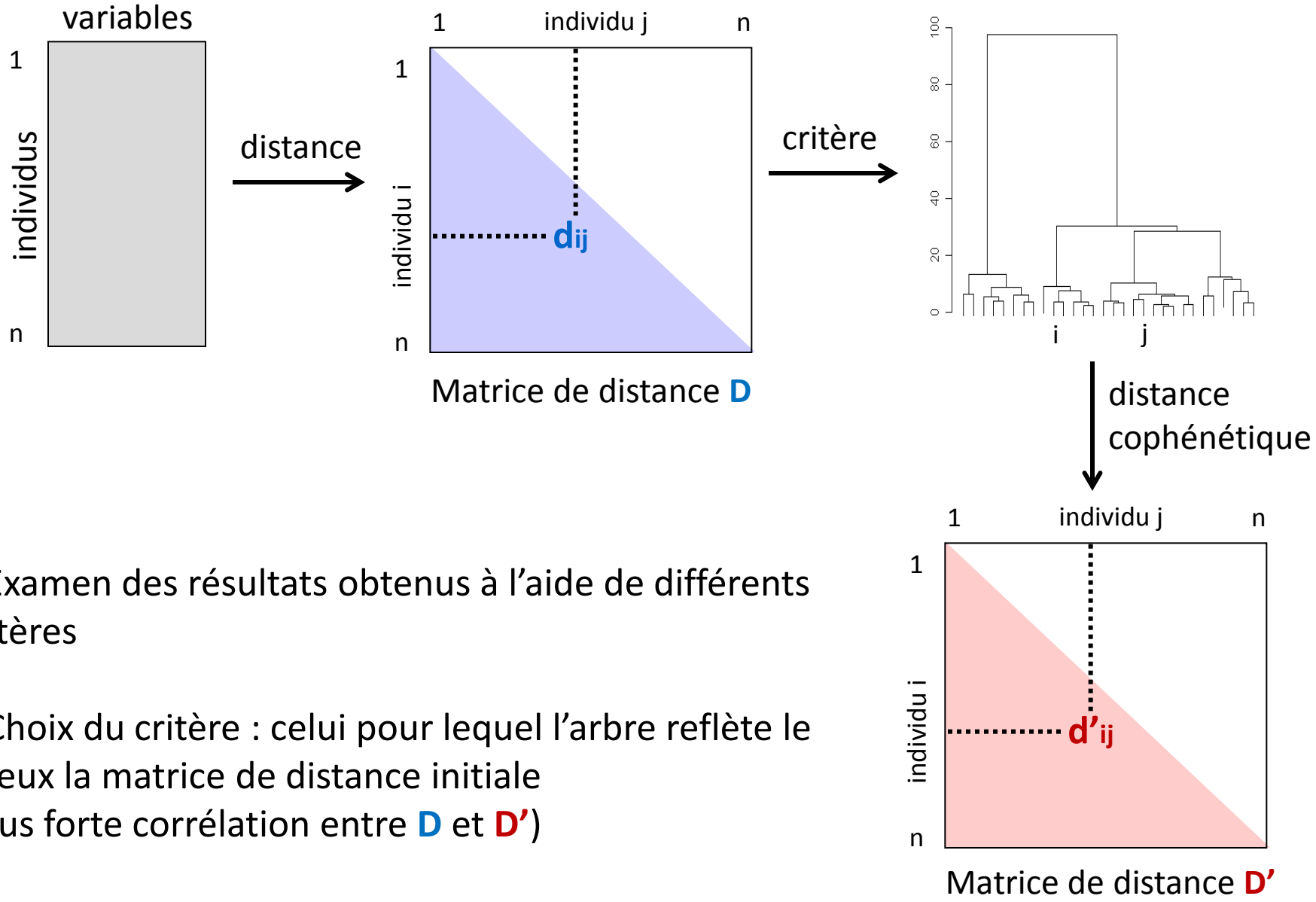
- **saut moyen** (*group average, UPGMA*) : distance moyenne entre tous les membres des 2 groupes

- **centroïdes** : distance entre les moyennes (centres de gravité des 2 groupes)

- **méthode de Ward** : proche du groupement moyen mais minimise l'augmentation de la variance intra-groupe à chaque regroupement

Classif. hiérarchique ascendante

Choix du critère d'agrégation



- Examen des résultats obtenus à l'aide de différents critères

- Choix du critère : celui pour lequel l'arbre reflète le mieux la matrice de distance initiale
(plus forte corrélation entre **D** et **D'**)

BECNM – Analyses de Données Multivariées

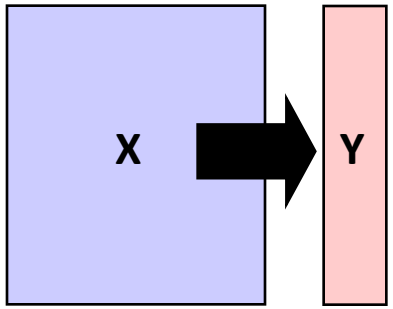
Analyse Factorielle Discriminante (AFD)

- *linear discriminant analysis*
- *canonical discriminant analysis*
- *factorial discriminant analysis*
- *discriminant function analysis*

Gaël Grenouillet
gael.grenouillet@univ-tlse3.fr

Introduction

- **Originalité** : Peut être considérée comme une extension de la régression multiple avec une variable dépendante qualitative



Y qualitative
Les groupes prédéfinis sont :
- connus *a priori*
- inconnus *a priori*, définis par une méthode de classification

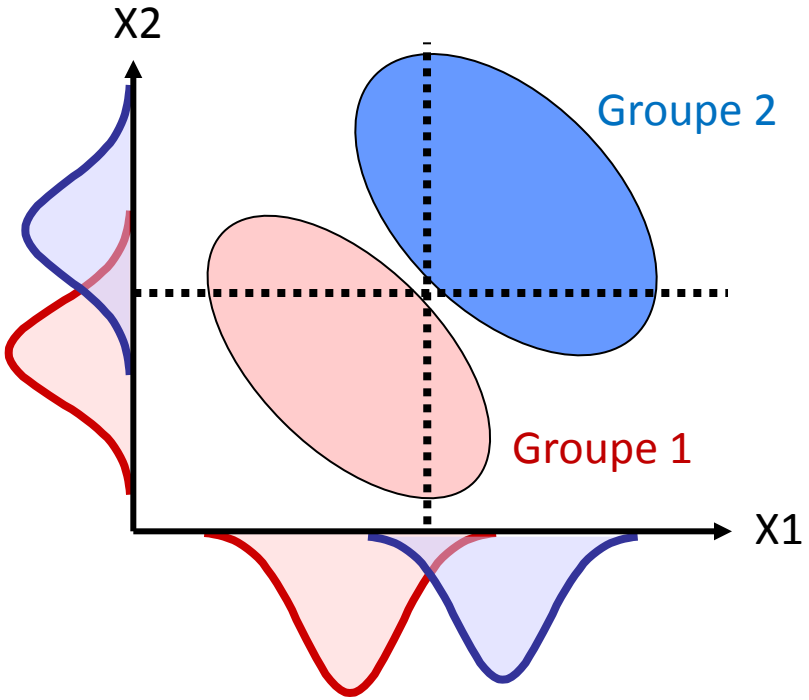
- **Objectifs** :
  - différencier des groupes existants
  - affecter un individu à un groupe en y associant une probabilité

- **Applications** :

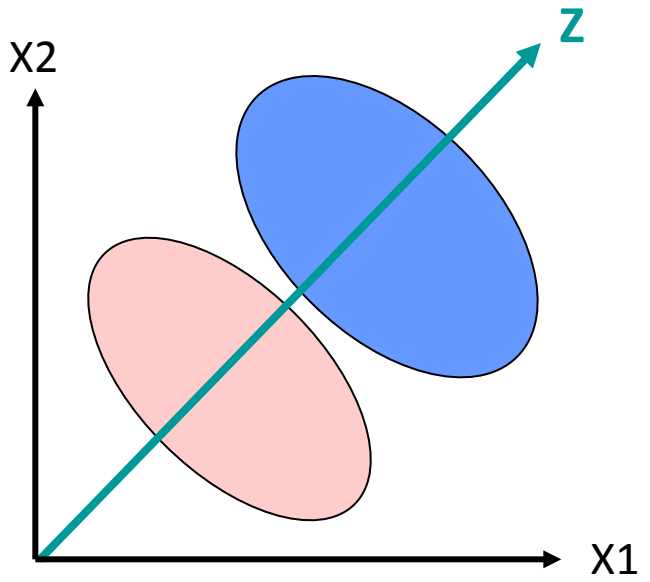
Biologie → différences morphologiques entre groupes (populations, espèces)
Médecine → appartenance à des groupes de patients (« malade » *vs* « sain »)
Nombreux domaines variés (production, informatique)
→ contrôle qualité d'un produit (« bon » - « moyen » - « mauvais »)
→ analyse d'image, reconnaissance de formes,…

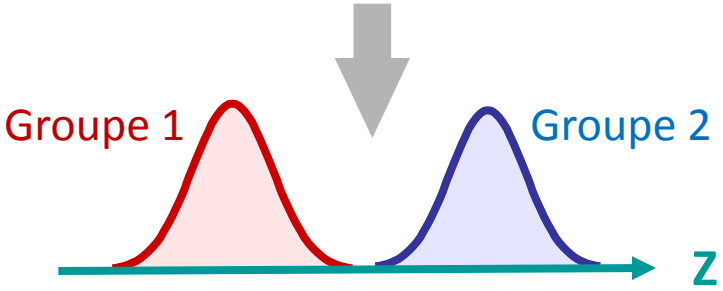
Principe

Discriminer des groupes



Une variable seule ne suffit pas pour séparer les deux groupes

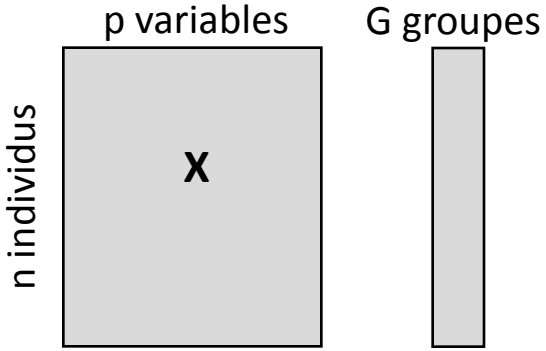




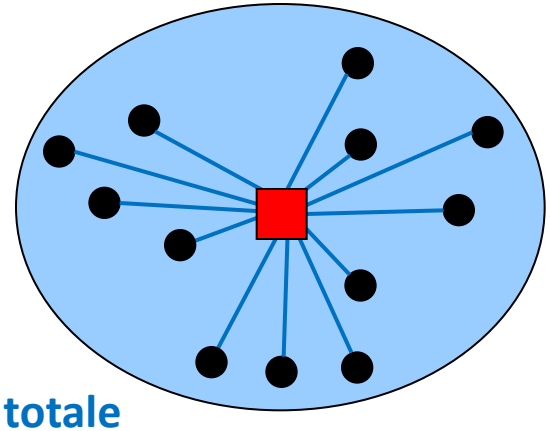
(fonction discriminante)

Principe

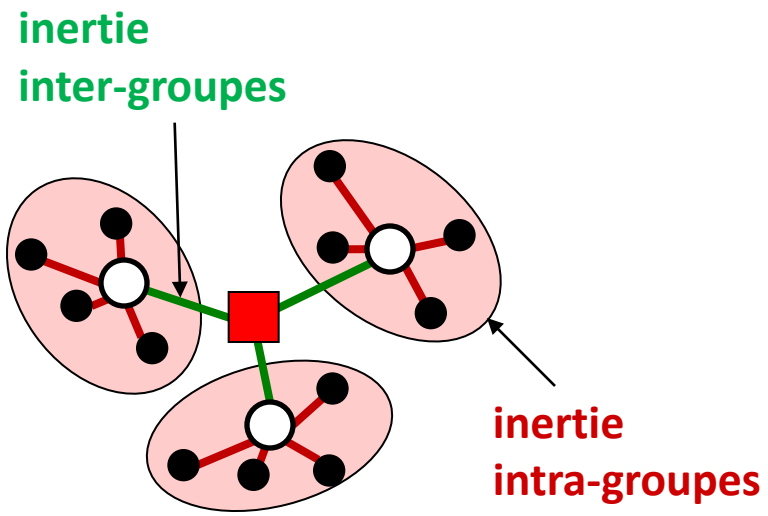
Décomposition de l'inertie



On peut décomposer la variance de la matrice X



inertie totale



Théorème de Huygens

inertie totale = inertie inter + inertie intra

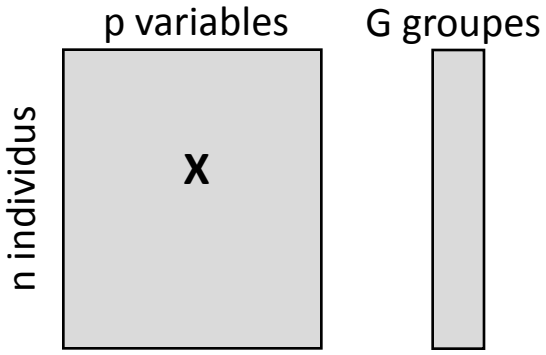
=

distance pondérée entre les centres de gravité des groupes et celui du nuage

+

distance pondérée entre chaque point d'un groupe et son centre de gravité

Principe



Les **fonctions discriminantes** sont des **combinaisons linéaires** des variables X :

$$Z_i = \sum_{j=1}^{p} a_{ij} X_j$$

a_{ij}= coef. de la fonction discriminante

Compromis entre deux objectifs distincts :

- représenter les groupes comme **bien séparés** (maximiser l'inertie inter-groupes)
- représenter les groupes comme **homogènes** (minimiser l'inertie intra-groupes)

La recherche des fonctions discriminantes Z_i revient à trouver une combinaison linéaire qui maximise le rapport variance inter-groupes / variance intra-groupes

$$Z_1 \rightarrow \frac{\mathrm{Var}_{\mathrm{inter}}}{\mathrm{Var}_{\mathrm{intra}}} = \mathrm{Max}$$

- Le rapport des variances sera plus petit pour Z_2 que pour Z_1
- Z_1 et Z_2 sont orthogonales (non corrélées)
- Au final, on obtient min(G-1, p) fonctions discriminantes

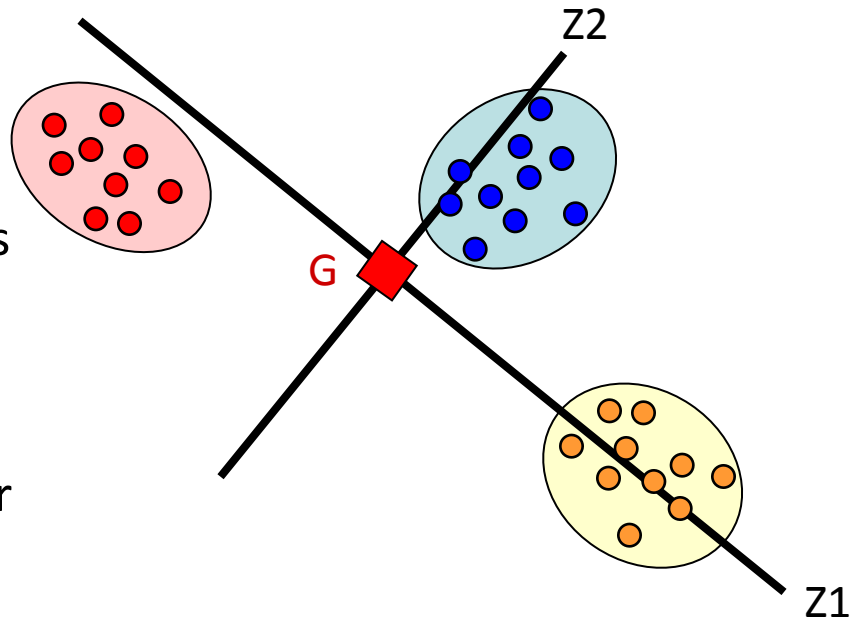
Représentation géométrique

- Variables centrées → G est à l'origine

- Le 1er facteur détermine un axe dans le nuage de points tels que les projections des points sur cet axe aient une variance inter-classe maximale

- Le 2ème facteur est orthogonal au premier
- *etc…*

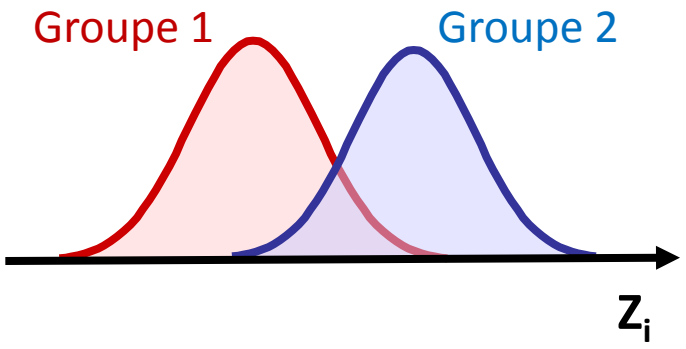
- Si 2 groupes → un seul facteur discriminant



Analyse discriminante = ACP sur le nuage des centres de gravités
(pondérés par les effectifs des groupes)

Interprétation

Valeurs propres



Pour la fonction discriminante Z_i
deux expressions de la valeur propre :

- **Valeur propre μ** : sur l'intervalle [0;+∞[
= rapport de la variance inter-groupes sur la variance intra-groupes

- **Valeur propre λ** : sur l'intervalle [0;1]

$$\lambda = \frac{\mu}{1+\mu}$$

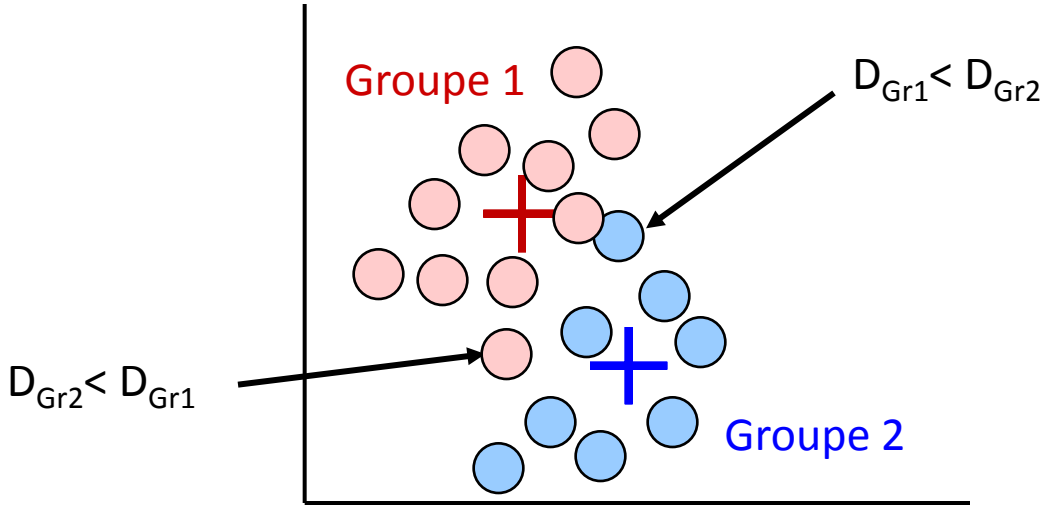
→ λ_i est le **pouvoir discriminant** de Z_i

- $\lambda = 1$ → dispersion intra-groupes nulle et discrimination parfaite si les centres de gravités des groupes se projettent en des points distincts de l'axe
- $\lambda = 0$ → projections des centres de gravités des groupes confondues sur l'axe

Application de la méthode ***Classement de nouveaux individus***

L'AFD est l'analyse d'un nuage de points caractérisée par la distance de Mahalanobis (distance entre les observations et les centres des groupes)

- Permet de classer une nouvelle observation dans le groupe pour lequel cette distance est minimale
- Permet de renseigner sur la qualité de la discrimination (% de mauvais classement)



Ici, deux objets sont mal classés

Conclusion

Analyse Factorielle Discriminante :

- les groupes sont prédéfinis
- le principe est identique à la classification (maximisation de la variance inter-groupe et minimisation de variance intra-groupe)
- méthode factorielle → on cherche à condenser l'information

Les données peuvent être :

- **normées** → les fonctions discriminantes décrivent la contribution relative de chaque variable dans la caractérisation des groupes. Elles permettent le classement de nouveaux objets.

- **brutes** → permet seulement de classer les nouveaux objets.