Analyse de données

L. BELLANGER

Master 1 Ingénierie Statistique
Dpt de Mathématiques - Université de Nantes

Plan

0. Introduction
I. Outils de représentation d'un échantillon
II. Analyse en Composantes Principales (*ACP*)
III. Analyse Factorielle des Correspondances (*AFC*)
IV. Classification et classement
V. Conclusion

IV. Classification et classement

Plan

0. Introduction
I. Classification
   1. Idées générales
   2. Classification par partition
II. Classement par AFD

Introduction :

Qu'est-ce que la classification ?

- **Regrouper des objets** en groupes, ou classes, ou familles, ou segments, ou clusters, de sorte que :
  - 2 objets d'un même groupe se ressemblent le + possible
  - 2 objets de groupes distincts diffèrent le + possible
  - le nombre des groupes est parfois fixé
- **Méthode descriptive :**
  - pas de variable cible privilégiée
  - décrire de façon simple une réalité complexe en la résumant
- **Les objets à classifier sont :**
  - des individus
  - des variables

Introduction :

Classification ≠ Classement

- La **classification** consiste à regrouper les individus d'une population en un nombre limité de classes qui :
  - ne sont pas prédéfinies mais déterminées au cours de l'opération, contrairement aux classes du classement ;
  - regroupent les individus ayant des caractéristiques similaires et séparent les individus ayant des caractéristiques différentes.
- Le **classement** consiste à placer chaque individu de la population dans une classe, parmi plusieurs classes prédéfinies, en fonction des caractéristiques de l'individu indiquées comme variables explicatives.
- Le résultat du classement permet d'affecter chaque individu à la meilleure classe.
- Très souvent, il y a 2 classes prédéfinies (« sain » et « malade », par exemple).

Introduction :

Classification ≠ Classement (1a)

Le vocabulaire s'appuie aussi sur les mots suivants:

- La **Classification** est une méthode d'**analyse non-supervisée**, ce qui sous-entend que le tableau de données *n'est pas structuré* par opposition au
- **Classement** qui est une méthode d'**analyse supervisée**, ce qui sous-entend que le tableau de données est structuré. Le classement est toujours associé à une discrimination préalable, même si ce n'est pas précisé de façon explicite: on ne peut classer des individus que dans des classes préalablement définies.
- De surcroît, les dictionnaires ne sont pas très clairs :
  - **Classer** peut aussi bien vouloir dire **diviser en classes** que **ranger dans une catégorie**.
  - Par contre **classifier** signifie faire ou établir des classifications.

Introduction :

Classification ≠ Classement (1 b)

Donc, pour être précis:

- Dans une **classification**, on **classifie**,
- dans un **classement**, on **classe**.

Enfin de manière générale :

- on classe ou on classifie des individus (= des objets = des observations) ; mais
- on peut tout aussi bien réaliser ces opérations sur des variables.

Introduction :

Classification ≠ Classement (2)

- **Point de terminologie** : 3 techniques de ***data mining***
  …3 terminologies ≠ dans la littérature !

Auteurs anglo-saxons	Certains auteurs francophones	Analyse des données à la française
Clustering	Segmentation	Classification
Classification	Classification	Classement, analyse discriminante
Decision trees	Arbres de décision	segmentation

Introduction :

Structure des **classes** obtenues

- Soit 2 classes sont toujours **disjointes** :
  méthodes de partitionnement :
  - généralement, le nombre de classes est défini a priori ;
  - certaines méthodes permettent de s'affranchir de cette contrainte (analyse relationnelle, méthodes paramétriques par estimation de densité).
- Soit 2 classes sont **disjointes ou l'une contient l'autre** :
  méthodes hiérarchiques :
  - ascendantes (agglomératives : agglomération progressive d'éléments 2 à 2) ;
  - descendantes (divisives).
- Soit 2 classes peuvent avoir plusieurs objets en commun (classes « **empiétantes** » ou « **recouvrantes** ») :
  - analyse « floue », où chaque objet a une certaine probabilité d'appartenir à une classe donnée.

Introduction :

Les différentes méthodes

- **Méthodes de partitionnement**
  - k-means : centres mobiles et nuées dynamiques
  - k -modes, k -prototypes, k -représentants (k -medoids)
  - réseaux de Kohonen
  - méthodes basées sur une notion de densité
  - méthode « de Condorcet » (analyse relationnelle)
- **Méthodes hiérarchiques**
  - ascendantes (agglomératives)
    - basées sur une notion de distance ou de densité
  - descendantes (divisives)
- **Méthodes mixtes**
- **Analyse floue** (fuzzy clustering)

Introduction :

classification des individus

- Il faut choisir une **mesure de ressemblance** entre individus, le plus souvent la distance euclidienne ; mais il en existe de nombreuses ! Cf. après ….
- Nécessité de **standardiser les variables** si elles ne sont pas toutes mesurées dans la même unité et ont des moyennes ou des variances dissemblables
- Préférable d'**isoler les « outliers »** (individus hors-norme)
- Quand on a des variables qualitatives ⇒ se ramener à une classification de variables continues par une *AFCM*

I. Classification

1. IDEES GENERALES

1.1 Mesures de ressemblance (« *similarity* »)

Les mesures de ressemblance entre objets à classer dépendent de la **nature des variables mesurées** qui peuvent être binaires, nominales, ordinales ou numériques.

Définitions générales:

- distance
- similarité
- dissimilarité

1. IDEES GENERALES

1.1 Mesures de ressemblance (« *similarity* »)

- On définit d'abord une *distance* sur un ensemble $\boldsymbol{E}$ de n objets, comme l'application de $\boldsymbol{E} \times \boldsymbol{E}$ dans $\mathbb{R}^+$ vérifiant :
  - $d(i,j) \geq 0$ et $d(i,j) = 0 \Leftrightarrow i = j$
  - $d(i,j) = d(j,i)$
  - $d(i,j) \leq d(i,k) + d(k,j)$ inégalité triangulaire

- Une *distance* est dite **euclidienne** si elle est engendrée par un produit scalaire.

- Une *distance* est dite **ultramétrique** si :

$$d(i,j) \leq sup\big(d(i,k); d(j,k)\big)$$

1. IDEES GENERALES

1.1 Mesures de ressemblance (« *similarity* »)

- Si inégalité triangulaire pas vérifiée : ***dissemblance* ou *dissimilarité*** $\boldsymbol{D}$ sur un ensemble $\boldsymbol{E}$ de n objets, est une application de $\boldsymbol{E} \times \boldsymbol{E}$ dans $\mathbb{R}^+$ vérifiant :
  - $D(i,j) \geq 0$ et $D(i,j) = 0 \Leftrightarrow i = j$
  - $D(i,j) = D(j,i)$

- On parle de ***ressemblance*** ou de **similarité** si on a une application s telle que :
  - $s(i,j) \geq 0$
  - $s(i,j) = s(j,i)$
  - $s(i,i) \geq s(i,j)\ \forall i,j$

1. IDEES GENERALES

1.1 Mesures de ressemblance entre individus x_i

Exemples: x_{ij} : $i = 1, \dots n$ (indiv) et $j = 1, .., p$ (variables)

a/ Données numériques (cf ex 1)

Tableau individus x variables quantitatives

- **Distance de Minkowski** (1896)

$$d(i,i') = d(x_i, x_{i'}) = \left\{ \sum_{j=1}^{p} \alpha_j \left| x_{ij} - x_{i'j} \right|^{\lambda} \right\}^{\frac{1}{\lambda}} \text{ où } \lambda \text{ et } \alpha_j \in R^+$$

- **Cas particuliers:**

Si $\lambda = 1$ *et* $\alpha_j = 1$: **distance de Manhattan** : $d(i,i') = \sum_{j=1}^{p} \left| x_{ij} - x_{i'j} \right|$

Si $\lambda = 2$ *et* $\alpha_j = 1$: **distance euclidienne classique** :

$$d(i,i') = \left\{ \sum_{j=1}^{p} \left| x_{ij} - x_{i'j} \right|^2 \right\}^{\frac{1}{2}}$$

1. IDEES GENERALES

1.1 Mesures de ressemblance entre individus x_i

Exemples: x_{ij} : $i = 1, \dots n$ et $j = 1, .., p$

b/ Données de fréquences (cf ex 4)

Tableau de contingence

- **Distance entre 2 lignes =Distance du Chi-deux**

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2$$

1. IDEES GENERALES

1.1 Mesures de ressemblance entre individus x_i

Exemples: x_{ij} : $i = 1, \dots n$ et $j = 1, .., p$

c/ Données binaires

i 01100001010010...

i' 01010001100010...

Les n indiv. à classer sont caractérisés par p variables binaires codées 0 ou 1. La ressemblance ou similarité entre 2 individus i et i' $s(i,i')$ se calcule à partir des informations du tableau de contingence suivant:

		Sujet i: 1	Sujet i: 0	Tot
Sujet i'	1	a	b	$a+b$
	0	c	d	$c+d$
	Tot	$a+c$	$b+d$	n

- a : nb de concordances communes 11,
- b : nb de concordances 10, c : nb de concordances 01,
- d : nb de concordances 00.

Ces 4 nbs définissent des **indices de similarités** entre individus, par exemple :

$S_1 = \dfrac{a}{a+b+c}$	Indice de communauté de Jaccard
$S_2 = \dfrac{a+d}{n}$	Indice de Sokal & Michener
$S_3 = \dfrac{a}{a+2(b+c)}$	Indice de Sokal & Sneath
$S_4 = \dfrac{a+d}{a+2(b+c)+d}$	Indice de Rogers et Tanimoto
$S_5 = \dfrac{2a}{2a+b+c}$	Indice de Sorensen
$S_6 = \dfrac{a-b-c+d}{n}$	Indice de Gower & Legendre
$S_7 = \dfrac{a}{\sqrt{(a+b)(a+c)}}$	Indice de Ochiai
$S_8 = \dfrac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	Indice de Sockal & Sneath
$S_9 = \dfrac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	Phi de Pearson

Ces indices sont tous ≤ 1 et la dissimilarité **associée** est définie par :

$$d_k = (1 - s_k)^{1/2}$$

$\Rightarrow$ **Faire exemple Td TP Classification, page 2** *Td/TP ch 4* : *indices de Jaccard*

1. IDEES GENERALES

1.1 Mesures de ressemblance entre individus x_i

Exemples: x_{ij} : $i = 1, \dots n$ et $j = 1, \dots, p$

c/ Données binaires

On peut les calculer par la fonction **dist.binary** dans **ade4** qui demandera de choisir :

```
1 = JACCARD index (1901) S3 coefficient of GOWER & LEGENDRE
s1 = a/(a+b+c) --> d = sqrt(1 - s)
2 = SOCKAL & MICHENER index (1958) S4 coefficient of GOWER & LEGENDRE
s2 = (a+d)/(a+b+c+d) --> d = sqrt(1 - s)
3 = SOCKAL & SNEATH(1963) S5 coefficient of GOWER & LEGENDRE
s3 = a/(a+2(b+c)) --> d = sqrt(1 - s)
4 = ROGERS & TANIMOTO (1960) S6 coefficient of GOWER & LEGENDRE
s4 = (a+d)/(a+2(b+c)+d) --> d = sqrt(1 - s)
5 = CZEKANOWSKI (1913) or SORENSEN (1948) S7 coefficient of GOWER & LEGENDRE
s5 = 2*a/(2*a+b+c) --> d = sqrt(1 - s)
6 = S9 index of GOWER & LEGENDRE (1986)
s6 = (a-(b+c)+d)/(a+b+c+d) --> d = sqrt(1 - s)
7 = OCHIAI (1957) S12 coefficient of GOWER & LEGENDRE
s7 = a/sqrt((a+b)(a+c)) --> d = sqrt(1 - s)
8 = SOKAL & SNEATH (1963) S13 coefficient of GOWER & LEGENDRE
s8 = ad/sqrt((a+b)(a+c)(d+b)(d+c)) --> d = sqrt(1 - s)
9 = Phi of PEARSON = S14 coefficient of GOWER & LEGENDRE
s9 = ad-bc)/sqrt((a+b)(a+c)(b+d)(d+c)) --> d = sqrt(1 - s)
10 = S2 coefficient of GOWER & LEGENDRE
s10 = a/(a+b+c+d) --> d = sqrt(1 - s) and unit self-similarity
Select an integer (1-10): 0
```


1. IDEES GENERALES

1.2 Qualité d'une classification

- Détecter les structures présentes dans les données
- Permettre de déterminer le nombre optimal de classes
- Fournir des classes bien différenciées
- Fournir des classes stables vis-à-vis de légères modifications des données
- Traiter efficacement les grands volumes de données

- Traiter tous les types de variables (quantitatives et qualitatives)
  - *Ce point est rarement obtenu sans transformation*
- Conduire à une interprétation et une utilisation facile des résultats

1. IDEES GENERALES

1.3 Concepts courants en classification

Deux idées complémentaires :

- ***cohésion interne*** des classes
- ***isolement*** entre classes.

A ces deux idées s'ajoute celle de ***hiérarchie*** possible entre classes.

Certaines techniques peuvent permettre un certain ***recouvrement*** des classes.

1. IDEES GENERALES

1.4 *Considérations combinatoires*

$B_{n,k}$: nb de partitions en k classes de n objets = **nb de Stirling**

Propriétés :

- $B_{n,1} = B_{n,n} = 1$ et $B_{n,n-1} = C_n^2$
- $B_{n,k} = B_{n-1,k-1} + kB_{n-1,k}$, (récurrence)

Exemple : $B_{12,5} = 1\,379\,400$

- B_n= nb total de partitions de n objets (**nb de Bell**)

$$B_n = \sum_{k=1}^{n} B_{n,k} = \frac{1}{e}\sum_{i=1}^{\infty}\frac{i^n}{i!}$$

Exemple : $B_{12} = 4\,213\,597$

⇒ Nécessité d'algorithmes pour trouver une « bonne » partition.

Comment définir la qualité d'une partition ?

1. IDEES GENERALES

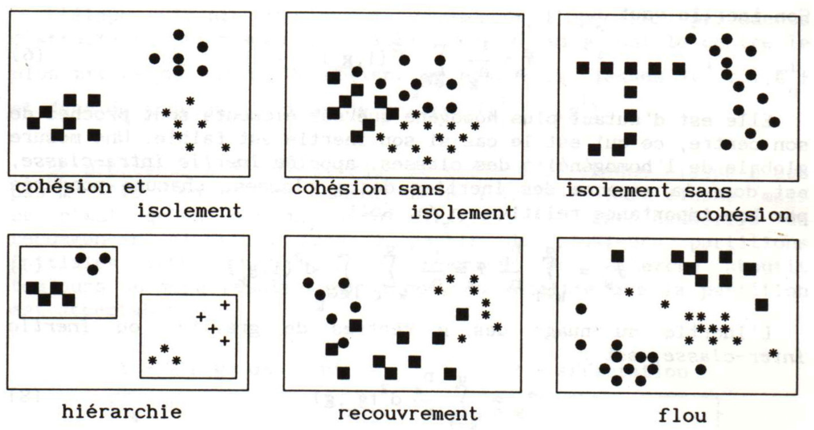


Schéma de quelques concepts courants en classification.

1. IDEES GENERALES

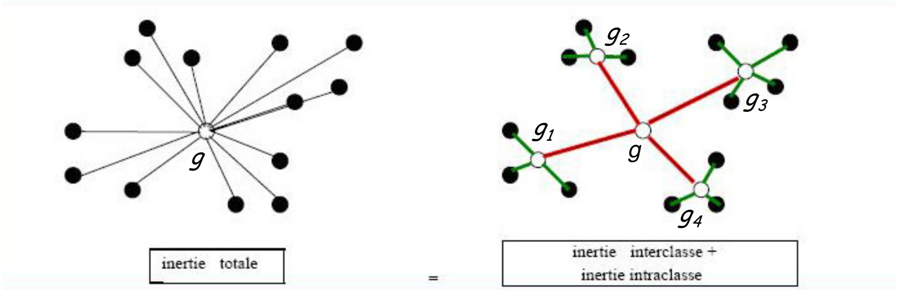
1.5 Inertie inter-classe et inertie intra-classe

- n points dans un espace euclidien ; $d^2(i,i')$ distance euclidienne
- Soit une partition en K classes de poids $p_i = 1/n$
- $g_1, g_2, \dots, g_K$: centres de gravité
- $I_1, I_2, \dots, I_K$: inerties associées

Inertie totale	$I_T = \frac{1}{n}\sum_{i=1}^{n} d^2(\boldsymbol{i},\boldsymbol{g})$ où $\boldsymbol{g} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$
Inertie d'une classe C_k $(k = 1,\dots,K)$	$I_k = \frac{1}{n_k}\sum_{i\in C_k} d^2(\boldsymbol{i},\boldsymbol{g_k})$ où $\boldsymbol{g_k} = \frac{1}{n_k}\sum_{i\in C_k} \boldsymbol{x_i}$
Inertie intra-classe (W= within)	$I_W = \sum_{k=1}^{K} \frac{n_k}{n} I_k = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k} d^2(\boldsymbol{i},\boldsymbol{g}_k)$
Inertie inter-classe (B= between)	$I_B = \sum_{k=1}^{K} \frac{n_k}{n} d^2(\boldsymbol{g}_k,\boldsymbol{g})$
Théorème de Huyghens : $T = B + W$	$I_T = I_B + I_W$

1. IDEES GENERALES

Illustration du Théorème de Huyghens : $T = B + W$



Comparaison de deux partitions en k classes :

La meilleure est celle qui a l'inertie I_W la plus faible (ou l'inertie I_B la plus forte).

Remarque : Ce critère ne permet pas de comparer des partitions à nombres différents de classes.

1. IDEES GENERALES

Critère de classification

- **Comparaison de deux partitions en K classes** :
  - **La meilleure** est celle qui a l'inertie I_W la plus faible (ou l'inertie I_B la plus forte).
  - **Critère global** de qualité de la classification : $C_K^2 = \frac{I_B}{I_T}$
    - Indique la part de la variabilité totale exprimée par la partition (souvent exprimé en %).
    - Exemple : Si C_K^2=0.88 pour une partition en $K = 3$ classes de 6 individus ; 88% de la variabilité des individus est prise en compte par la partition en 3 classes.
    - Tenir compte du nb de classes au regard du nb d'individus ! nb de classes $K \nearrow \Rightarrow C_K^2 \nearrow$

Remarque : critère permettant de comparer des partitions ayant un même nombre de classes !!!

2. CLASSIFICATION PAR PARTITION

2.2 Regroupement d'observations autour de centres mobiles : méthodes k-means

Algorithme de Lloyd

Fixer le nombre K de classes, puis :

- **Etape 1 : Choix des K *centres* $g_k^{(0)}$** (par ex par tirage pseudo-aléatoire) et **1ère partition associée** $C_k^{(0)}$ $(k = 1, \cdots, K)$
  La classe $C_k^{(0)}$ est formée de tous les points plus proches de $g_k^{(0)}$ que de tout autre centre.
- **Etape 2 : Calcul des centres de gravité de chaque classe $g_k^{(1)}$**
  $\Longrightarrow$ définition d'une nouvelle **partition** $C_k^{(1)}$.
- **Etape 3 : Itérations successives de ces étapes**
  $\Longrightarrow$ jusqu'à stabilisation du critère de classification retenu, i.e. quand le contenu des classes n'est plus modifié.

RÉSULTAT FONDAMENTAL

L'inertie intra-classe I_W diminue à chaque étape.

2. CLASSIFICATION PAR PARTITION

2.2 Regroupement d'observations autour de centres mobiles :

RÉSULTAT FONDAMENTAL

L'inertie intra-classe I_W diminue à chaque étape qd K est fixé.

On définit le critère : $I_W^{(m)} = \frac{1}{n} \sum_{k=1}^{k=K} \sum_{i \in C^{(m)}_k} d^2(\mathbf{i}, \mathbf{g_k}^{(m)})$

associé à la partition $C_k^{(m)}$ $(k = 1, ..., K)$ de centre de gravité $\mathbf{g_k}^{(m+1)}$

- Il suffit de montrer que : $I_W^{(m+1)} = \frac{1}{n} \sum_{k=1}^{k=K} \sum_{i \in C^{(m)}_k} d^2(\mathbf{i}, \mathbf{g_k}^{(m+1)}) \leq I_W^{(m)}$
- A l'étape $m+1$, on associe chaque pt au centre le plus proche donc

$$I_W^{(m+1)} \leq I_W^{(m)}$$

- Le nuage de pts étant fini et la suite $\left(I_W^{(m)}\right)$ étant décroissante et > 0, l'algorithme converge vers une valeur minimale I_W^{lim} .

2. CLASSIFICATION PAR PARTITION

Il existe de nombreuses méthodes k-means puisque :

Un *centre mobile* peut être :

- une observation unique,
- quelques observations ou
- leur centre de gravité ou
- tout élément résumant la position d'un certain nombre d'observations.

Le *choix initial* peut aussi être fait:

- par le classificateur lui-même en fonction, par exemple, de ses connaissances *a priori*,
- suite à une autre analyse statistique préalable, comme des points très éloignés sur un plan d'analyse en composantes principales,
- au hasard, faute de mieux!

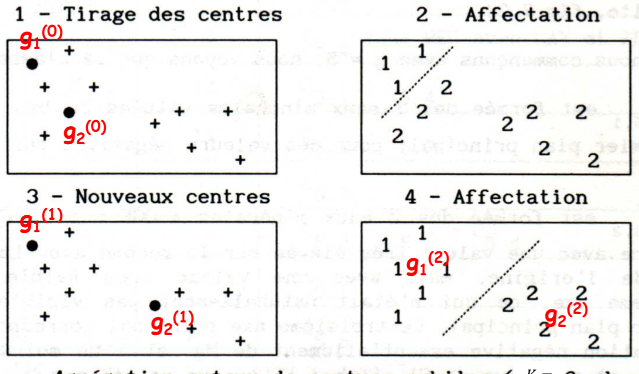
2. CLASSIFICATION PAR PARTITION

Segmentation par centres mobiles

- **Principe**
  Regrouper les individus en fonction de leur distance au « centre » des différentes classes .
- **Variante**: **nuées dynamiques**
  Une classe est caractérisée par un noyau (ensemble formé de q pts appelés étalons) .

2. CLASSIFICATION PAR PARTITION

Schématisation de la **méthode des centres mobiles**



Agrégation autour de centres mobiles (K = 2 classes).
Légende : ● : centre de classe, + : point courant

2. CLASSIFICATION PAR PARTITION

Avantages des méthodes k-means :

- algorithmes simples
- applicables à des corpus de données de grande quantité d'observations.

Inconvénients :

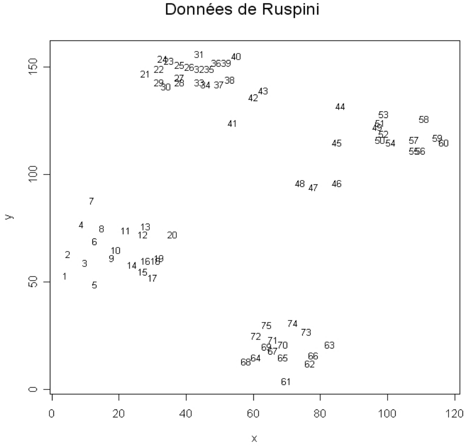
- le résultat dépend fortement du tirage initial des pts représentant les centres des classes.
  - **Remèdes** : rechercher les individus partageant les mêmes groupes lors de partitions répétées (**formes fortes**) ; combiner avec CAH (***classification mixte***).
- méthodes ne permettant pas de détecter la présence d'outliers.

Faire exemple 2 page 3 Td/TP ch 4

2. CLASSIFICATION PAR PARTITION

2.3 Exemple sous R (E. H. Ruspini (1970): Numerical methods for fuzzy clustering. *Inform. Sci.*, **2**, 319–350.)

Données de Ruspini

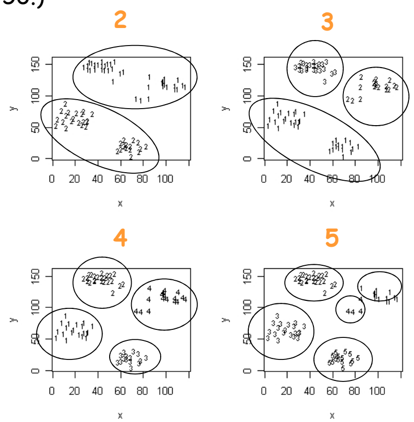


2. CLASSIFICATION PAR PARTITION

2.3 Exemple sous R (E. H. Ruspini (1970): Numerical methods for fuzzy clustering. *Inform. Sci.*, **2**, 319–350.)

Sous R :

```
> library(stats)
> Ruspkmeans.2<-kmeans(ruspini,2)
> Ruspkmeans.3<-kmeans(ruspini,3)
> Ruspkmeans.4<-kmeans(ruspini,4)
> Ruspkmeans.5<-kmeans(ruspini,5)
```



Ce qu'il faut retenir

- La classification automatique ou non supervisée permet d'organiser un ensemble d'objets ou d'individus en classes homogènes.
- Il existe un grand nombre de méthodes en fonction
  - de la nature des observations et
  - du type de classes que l'on cherche.
  - Attention :
    - Par partition : Kmeans : algorithme CV mais instable !
- On retiendra la plus appropriée en fonction de l'objectif recherché :
  - une partition sera bien souvent jugée satisfaisante si elle est composée de *classes interprétables*.

Exercices Td/TP ch IV: classification « à la main », fonction sous R

Données test supplémentaires

Données	Description
`Eaux1.txt`	Corpus 24 eaux minérales décrites par 6 variables
`Eaux2010,txt`	Corpus 113 eaux minérales décrites par 9 variables
`ChaZeb-a.txt`	Corpus de 23 bovins (Charolais & Zébus) décrits par 6 variables pondérales.
`Loup.txt`	Description de 43 crânes Chien/Loup par 6 variables. Identification d'un crane d'origine inconnue.
`CamRiz.xls`	données agronomiques concernant la culture du riz
`Iris.xls`	caractéristiques de 3 variétés d'iris

II. Classement par AFD

L'analyse discriminante : introduction

Objet d'étude de l'analyse discriminante

- Technique statistique visant à décrire, expliquer et/ou prédire l'appartenance à des groupes prédéfinis d'un ens. d'obs. (individus,…) caractérisées par une variable à expliquer **Y** qualitative à partir de variables explicatives $\mathbf{X}^j$.
- **Cas particulier** de l'*ACC* pour lequel **X** décrit un ensemble de variables quantitatives et **Y** représente les variables indicatrices associées aux K modalités d'une variable qualitative.

L'analyse discriminante : introduction

L'analyse discriminante est *utilisée dans de* ***nombreux domaines*** :

- **Médecine** : détection de groupes à hauts risques cardiaques à partir de caractéristiques telles que l'âge, l'alimentation, le fait de fumer ou pas, les antécédents familiaux, etc.
- **Domaine bancaire** : évaluation de la fiabilité d'un demandeur de crédit à partir de ses revenus, du nb de personnes à charge, des encours de crédits qu'il détient, etc.
- **Biologie** : affectation d'un objet à sa famille d'appartenance à partir de ses caractéristiques physiques.
  - Ex très fameux des iris de Fisher, à l'origine de cette méthode. Il s'agit de reconnaître le type d'iris (setosa, virginica, et versicolor) à partir de la longueur/largeur de ses pétales et sépales (4 variables explicatives).

Introduction

2 approches différentes selon les objectifs

Y variable à expliquer qualitative à K catégories

$X^1, X^2, \dots, X^p$ variables explicatives centrées

- **Objectif 1 : Décrire** (**Analyse discriminante descriptive** ou **analyse factorielle discriminante**)
  - Étude de la distribution des X^j/ Y
  - Géométrie : Analyse Factorielle Discriminante (AFD)
    - trouver une représentation graphique dans un espace réduit qui permette de discerner le plus possible les groupes d'individus (ie liaison entre Y et les X^j).

    En ce sens, elle se rapproche de l'analyse factorielle .
  - Tests : Analyse de variance multidimensionnelle MANOVA

Introduction

- **Objectif 2 : Classer** (Analyse discriminante prédictive ou décisionnelle) : construire une fonction de classement (règles d'affectation des individus,...) pour prédire le groupe y d'appartenance d'un individu à partir des valeurs des X^j.
  - repose sur un **cadre probabiliste**.

    Le plus connu : ***distribution multinormale*** (loi Normale) + ***homoscédasticité***, les nuages de points conditionnels ont la même forme, nous aboutissons à **l'analyse discriminante linéaire**.
    - très séduisante dans la pratique car la fonction de classement s'exprime comme une combinaison linéaire des X^j, facile à analyser et à interpréter.
- **Distinction** entre ces 2 approches n'est pas aussi tranchée.
  - possible de dériver des règles géométriques d'affectation à partir de l'analyse factorielle discriminante.

Les différentes formes d'analyse discriminante

	Méthode descriptive (représenter les groupes)	**Méthode prédictive** (prédire l'appartenance à un groupe)
Approche géométrique	Oui analyse factorielle discriminante	Oui analyse discriminante linéaire
Approche probabiliste (bayésienne)	Non	Oui analyse discriminante linéaire a. d. quadratique a. d. non paramétrique régression logistique

multinormalité
homoscédasticité
équiprobabilité

I. L' analyse factorielle discriminante

(canonical discriminant analysis en anglais)

1. Principe et notations
2. Les axes et les variables discriminantes
3. Méthodes géométriques de classement

I. L' analyse factorielle discriminante

1. Principe et notations

- **Y** variable cible qualitative à K modalités correspondant à K groupes G_k ;
- $\mathbf{X}^j$ $j = 1, \ldots, p$.
  p variables explicatives continues **centrées** (cas courant $K < p < n$);
- $\mathbf{X}_i$ individu i défini dans $\mathbb{R}^p$

But AFD est de répondre à:

« les K classes diffèrent-elles sur l'ensemble des caractères quantitatifs ? »

I. L' analyse factorielle discriminante

1. Principe et notations

L'*AFD* vise à produire un nouveau système de représentation, constitué de ***combinaisons linéaires des variables initiales*** $\mathbf{X}^j$, qui permet de séparer au mieux les K catégories.

Pour cela, il faudra :

- Remplacer les $\mathbf{X}^j$ par $d \leq \min(K-1; p)$ ***variables discriminantes*** $F^{(k)}$ $\cdot\, k = 1, \ldots, d$
  - $F^{(k)} = u_1^{(k)} \mathbf{X}^1 + \ldots u_p^{(k)} \mathbf{X}^p$ **combinaisons linéaires des** $\mathbf{X}^j$ (**centrées**) ;
  - prenant des valeurs les plus ≠ possibles pour des individus différents sur la variable cible **Y**.
- Trouver les $d \leq \min(K-1; p)$ vecteurs **u** normalisés (**facteurs ou** fonctions linéaires discriminantes) et orthogonaux. *d* est la ***dimension de la représentation des groupes.***

⟹ Il existe une grande analogie avec l'*ACP*.

I. L' analyse factorielle discriminante

1. Principe et notations

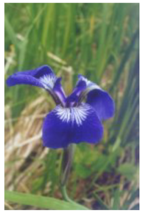
- technique descriptive : obtention d'une représentation graphique permettant de visualiser les proximités entre les obs, appartenant au même groupe ou non.
- technique explicative : possibilité d'interpréter les axes factoriels, combinaisons linéaires des variables initiales, et ainsi de comprendre les caractéristiques qui distinguent les ≠ groupes.

- Contrairement à l'analyse discriminante prédictive, ne repose sur aucune hypothèse probabiliste :
  - méthode essentiellement **géométrique**.

I. L' analyse factorielle discriminante

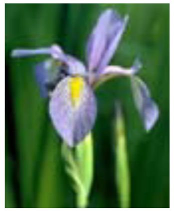
Exemple historique : Les iris de Fisher

(http://cs-people.bu.edu/mdassaro/pp3/)



setosa

versicolor

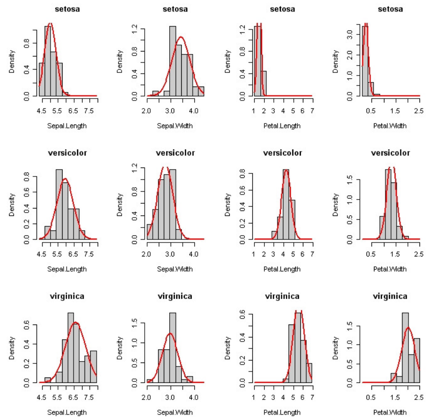


virginica

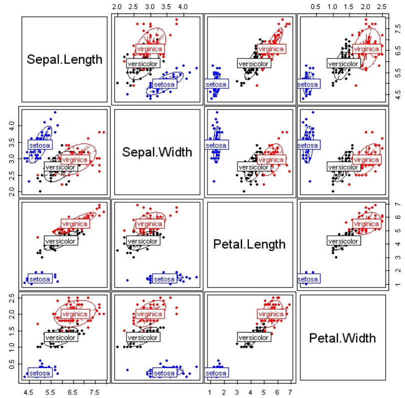
Problème *: reconnaître les 3 types d'iris (setosa, virginica, et versicolor) à partir de la longueur/largeur de ses pétales et sépales (4 variables explicatives). Ici* $d = \min(4; 3-1) = 2$

I. L' analyse factorielle discriminante

Les iris de Fisher `data(iris)`



Histogrammes par espèce et par variable



Représentation des nuages bivariés

I. L' analyse factorielle discriminante

Les iris de Fisher `data(iris)`

La valeur discriminante d'un plan varie fortement dans $\mathbb{R}^4$!

Une mesure varie-t-elle entre espèces, plus généralement entre groupe ?

En dimension 3, on peut encore voir ....

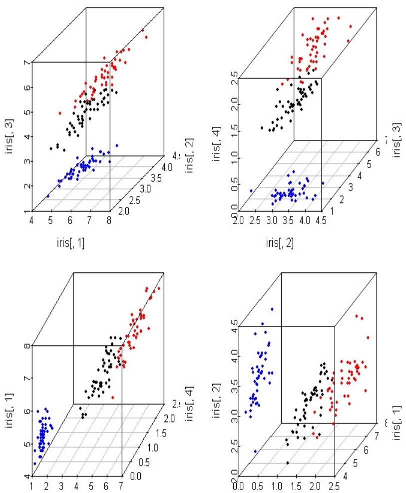
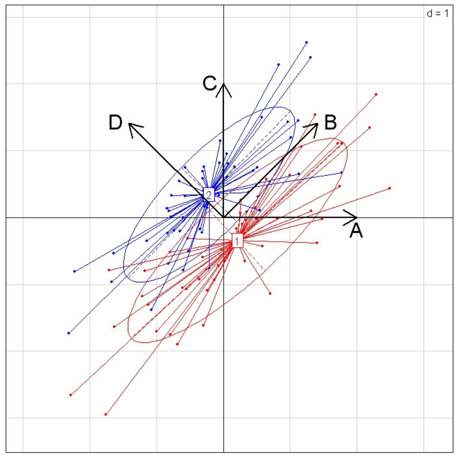


Illustration de la problématique descriptive

Données simulées :
2 populations Normales de 50 obs chacune définies par 2 variables.

⇒ **Choisir la direction** D et projeter les pts sur l'axe ainsi défini permet une meilleure séparation les obs de chacun des 2 groupes (rouge et bleu).
Par contre la direction B ne permet aucune séparation entre elles !

⇒ Suivant la direction de projection, les 2 populations apparaîtront un peu, bcp ou pas du tout différentes.



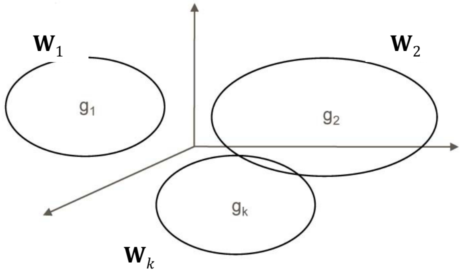
I. L' analyse factorielle discriminante

1. Principe et notations

Les n individus forment un nuage de n points dans $\mathbb{R}^p$, formé des K sous-nuages G_k **à différencier.**

On construit une 1ère variable F^1, **combinaison linéaire des p variables initiales** qui :

- minimise la variance intra $\mathbf{W}_k$ $k = 1, \ldots, K \Longrightarrow$ dispersion intra groupe ; *Within*
- maximise la variance inter $\mathbf{B} \Longrightarrow$ dispersion inter groupe. *Between*



I. L' analyse factorielle discriminante

1. Principe et notations

Calcul de W et B :

les n observations $\boldsymbol{x}_i$

- ont chacune un poids p_i $(i = 1, \ldots, n)$ défini dans la matrice diagonale $\mathrm{D}_{n \times n}$ et
- forment un nuage de pts de $\mathbb{R}^p$, formé des K sous-nuages G_k $(k = 1, \ldots, K)$ qui ont chacun un poids $q_k = \sum_{i \in G_k} p_i$

- **2 niveaux de variabilité :**
  - Variance intraclasse (« within ») $\mathbf{W}$ = moyenne des matrices de covariance $\mathbf{W}_k$ des classes G_k

$$\boxed{\mathbf{W_k} = \frac{1}{q_k} \sum_{i \in G_k} p_i (\mathbf{x_i} - \mathbf{g_k})(\mathbf{x_i} - \mathbf{g_k})^{\mathrm{T}}}$$

  - D'où la **matrice de covariance intraclasse** $\boxed{\mathbf{W} = \sum_{k=1}^{k=K} q_k \mathbf{W_k}}$

I. L' analyse factorielle discriminante

1. Principe et notations

Calcul de W et B :

- Variance interclasse (« between ») $\mathbf{B}$ = variance des barycentres $\boldsymbol{g}_k$ des classes G_k, $k = 1, \ldots, K$.

$$\boxed{\mathbf{B} = \sum_{k=1}^{k=K} q_k (\mathbf{g_k} - \overline{\mathbf{g}})(\mathbf{g_k} - \overline{\mathbf{g}})^{\mathrm{T}}}$$

matrice de covariance « between »

- **Théorème de Huyghens** : Si $\mathbf{T}$ est la matrice de covariance totale

$$\boxed{\mathbf{B} + \mathbf{W} = \mathbf{T}}$$

I. L' analyse factorielle discriminante

1. Principe et notations

Calcul de W et B :

Matriciellement, supposant les var. explicatives **centrées**, ie $\overline{\mathbf{g}} = \mathbf{0}$

Et notant $\mathbf{X_K}\,(n \times K)$ la matrice indicatrice des classes :

$$\begin{array}{cc} k \rightarrow & \begin{array}{cccc} 1 & 2 & \cdots & K \end{array} \\ \mathbf{X_K} = & \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & & \cdots & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & & \cdots & \\ 0 & 0 & \cdots & 1 \end{bmatrix} \end{array}$$

- Les K **barycentres** $\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_K$ sont les lignes de la matrice :

$$\boxed{\left(\mathbf{X_K^T D X_K}\right)^{-1}\left(\mathbf{X_K^T D X}\right)}$$

où $\mathbf{X}_K^{\mathrm{T}}\mathbf{D}\mathbf{X}_k$ est la matrice diagonale $(K \times K)$ des poids q_k des sous-nuages et $\mathbf{D} = \mathrm{diag}(p_i\,;\, i = 1, \ldots, n)$.

I. L' analyse factorielle discriminante

1. Principe et notations

- La matrice de variance interclasse s'écrit (si $\bar{\mathbf{g}} = \mathbf{0}$) :

$$\begin{aligned} \mathbf{B} &= \left(\left(\mathbf{X}_\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{X}_\mathbf{K}\right)^{-1}\left(\mathbf{X}_\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{X}\right)\right)^\mathbf{T}\mathbf{X}_\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{X}_\mathbf{K}\left(\left(\mathbf{X}_\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{X}_\mathbf{K}\right)^{-1}\left(\mathbf{X}_\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{X}\right)\right) \\ &= \mathbf{X}^\mathbf{T}\mathbf{D}\mathbf{X}_\mathbf{K}\left(\mathbf{X}_\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{X}_\mathbf{K}\right)^{-1}\mathbf{X}_\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{X} = \left(\mathbf{X}^\mathbf{T}\mathbf{D}\mathbf{X}_\mathbf{K}\right)\left(\mathbf{X}_\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{X}_\mathbf{K}\right)^{-1}\left(\mathbf{X}_\mathbf{K}^\mathbf{T}\mathbf{D}\mathbf{X}\right) \end{aligned}$$

- Dans le cas où $p_i = 1/n$, en notant les effectifs des K sous-nuages $n_1, n_2, \ldots, n_K$, on montre que l'on a :

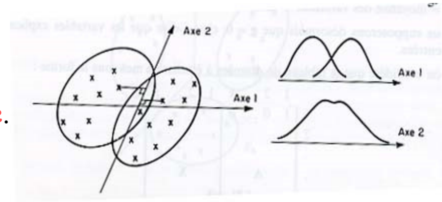
$$\begin{cases} \mathbf{B} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{k=K} n_k \left(\mathbf{g_k} - \bar{\mathbf{g}}\right)\left(\mathbf{g_k} - \bar{\mathbf{g}}\right)^\mathbf{T} \\ \mathbf{W} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{k=K} n_k \mathbf{W_k} \end{cases}$$

I. L' analyse factorielle discriminante

2. Les axes et les variables discriminantes

Soit $\mathbb{R}^p$ (espace des obs.) muni de la métrique $\mathbf{Q}_{p\times p}$ (cf. ACP).
On notera :

- $\boldsymbol{a}_{p\times 1}$ l'axe discriminant,
- $\boldsymbol{u}_{p\times 1}$ le facteur associé $\boldsymbol{u} = \mathbf{Q}\boldsymbol{a}$,
- $\boldsymbol{F} = \mathbf{X}\boldsymbol{u}$ la variable discriminante.



En projection sur l'axe $\boldsymbol{a}$,

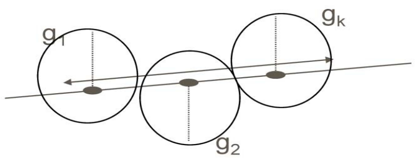
- les K centres de gravité doivent être le + plus **séparés possible**, tandis que
- chaque sous-nuage doit se projeter de manière groupée autour de la projection de sous centre de gravité.

I. L' analyse factorielle discriminante

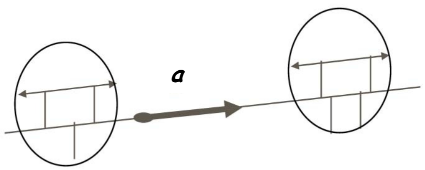
2. Les axes et les variables discriminantes

Axes discriminants $\boldsymbol{a}$ (Q-normé à 1): 2 objectifs simultanés

- *Dispersion inter classe* maximale : $\max \boldsymbol{a}^\mathrm{T}\mathbf{B}\boldsymbol{a}$



- *Dispersion intra classe* minimale : $\min \boldsymbol{a}^\mathrm{T}\mathbf{W}\boldsymbol{a}$



I. L' analyse factorielle discriminante

2. Les axes et les variables discriminantes

Axes discriminants : 2 objectifs simultanés

Géométriquement ceci signifie que :

- la matrice d'inertie des barycentres $\boldsymbol{g}_k$, $\mathbf{QBQ}$ doit être maximale en projection sur $\boldsymbol{a}$,
  cette inertie vaut : $\boxed{\boldsymbol{a}^\mathrm{T}\mathbf{QBQ}\boldsymbol{a} \text{ si } \boldsymbol{a} \text{ est } \mathbf{Q}\text{-normé à } 1}$

- Pour qu'un sous-nuage reste bien groupé il faut, qu'en projection sur $\boldsymbol{a}$, $\boldsymbol{a}^\mathrm{T}\mathrm{Q}\mathbf{W}_k\mathbf{Q}\boldsymbol{a}$ soit la plus faible possible pour toutes les classes, $k = 1, \ldots, K$.

  On cherche donc à minimiser la somme de ces inerties soit :

$$\sum_{k=1}^{k=K} \mathbf{a}^\mathbf{T}\mathbf{Q}\mathbf{W}_\mathbf{k}\mathbf{Q}\mathbf{a} = \mathbf{a}^\mathbf{T}\mathbf{QWQa}$$

I. L' analyse factorielle discriminante

2. Les axes et les variables discriminantes

Axes discriminants : 2 objectifs simultanés

- Simultanéïté impossible
  - $\max_{a} a^{T}Ba \Longrightarrow a$ tq $Ba = \alpha a$, α max
  - $\min_{a} a^{T}Wa \Longrightarrow a$ tq $Wa = \beta a$, β min
- Compromis : On reformule l'objectif

Le théorème de Huyghens entraîne: $\boxed{a^{T}QTQa = a^{T}QBQa + a^{T}QWQa.}$

Avec $u = Qa$ le facteur associé à a, on a donc :

$$\boxed{u^{T}Tu = \underbrace{u^{T}Bu}_{max} + \underbrace{u^{T}Wu}_{min}}$$

=> On peut alors prendre comme **critère à maximiser** soit

le rapport « ***inertie interclasse/inertie intraclasse*** »
ou
le rapport « ***inertie interclasse/inertie totale*** ».

I. L' analyse factorielle discriminante

2. Les axes et les variables discriminantes

Axes discriminants : 2 objectifs simultanés

On prendra comme critère soit :

(a) inter/intra ⇒ maximiser $F_{v} = v^{T}Bv/v^{T}Wv$
sous la contrainte $v^{T}Wv = 1$
⇔
(b) inter/totale ⇒ maximiser $F_{u} = u^{T}Bu/u^{T}Tu$ (Huyghens)
sous la contrainte $u^{T}Tu = 1$.

I. L' analyse factorielle discriminante

2. Les axes et les variables discriminantes

On montre que dans le cas :

- **(a)** v^{1} : 1er vecteur propre de $W^{-1}B$,de valeur propre $\mu_{1} = \lambda_{1}/(1-\lambda_{1})$ (contrainte $v^{T}Wv = 1$).
  ⇔
- **(b)** u^{1} est le 1er vecteur propre de $T^{-1}B$ associé à $\lambda_{1} \in [0;1]$ la plus grande valeur propre de $T^{-1}B$ (contrainte $u^{T}Tu = 1$) tq F_{u} est max .

  u^{1} est le 1er **facteur discriminant**, λ_{1} son **pouvoir discriminant.**

  la 1ère variable discriminante $F^{1} = Xu^{1}$ obtenue, on cherche $F^{2} = Xu^{2}$ non corrélé à F^{1} tq le rapport F_{u} soit maximum et ainsi de suite …

  $\Longrightarrow$ λ a les caractéristiques d'un R^{2} en régression.

I. L' analyse factorielle discriminante

2. Les axes et les variables discriminantes

On montre que :

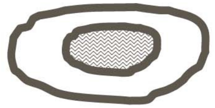
- Les vecteurs propres u et v sont liés par la relation :
  $$\boxed{u = (\sqrt{1-\lambda})v}$$
- Il existe $d \leq \min(K-1, p)$ **axes factoriels discriminants** correspondants aux d valeurs propres de $W^{-1}B$ (ou de $T^{-1}B$) et aux d vecteurs propres associés.

I. L' analyse factorielle discriminante

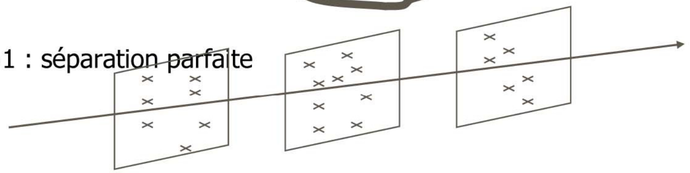
2. Les axes et les variables discriminantes

Les différents cas selon $\lambda_1 \epsilon [0; 1]$: cas (b)diag de $\mathbf{T}^{-1}\mathbf{B}$

1. $\lambda_1 = 0$: aucune séparation linéaire n'est possible, groupes concentriques



2. $\lambda_1=1$: séparation parfaite



3. Mais $0 < \lambda_1 < 1$: séparation possible avec groupes non recouvrants

I. L' analyse factorielle discriminante

2. Les axes et les variables discriminantes

Remarques :

- Les **pratiques anglaise** et **française** diffèrent un peu, et naturellement les logiciels qui en découlent :
  - les anglais : souvent le 1er rapport **(a)**, dans « l'esprit » du modèle linéaire classique (ANOVA, avec 1 seule variable le rapport $\frac{inter/(K-1)}{intra/(n-K)}$ est strict† une statistique F utilisée dans le modèle linéaire :
    - un F élevé traduit une différence importante entre les traitements;
  - les français préfèrent le 2nd **(b)**, lié à la relation entre le tableau des variables indicatrices des classes $\mathbf{X}_k$ et le tableau de données $\mathbf{X}$.
- **Sous R**

`lda(MASS)` utilise $\mathbf{W}^{-1}\mathbf{B}$ et $\boldsymbol{v}^{\mathrm{T}}\mathbf{W}\boldsymbol{v} = 1 \Longrightarrow$ (a)

`discrimin(ade4)` utilise $\mathbf{T}^{-1}\mathbf{B}$ et $\boldsymbol{u}^{\mathrm{T}}\mathbf{T}\boldsymbol{u} = 1 \Longrightarrow$ (b)

I. L' analyse factorielle discriminante

2. Les axes et les variables discriminantes

Rappel : $d \leq \min(K-1; p)$ **axes factoriels discriminants** correspondants aux d valeurs propres μ_k de $\mathbf{W}^{-1}\mathbf{B}$, et aux d vecteurs propres associés. **Choix du nb ????**

Des tests sont possibles sous réserve d'accepter **l'hypothèse de Normalité** (ou de ne pas en être « trop » éloigné).

1/ Test global de la dimension d de représentation ≈ MANOVA 1 : H_0: $\mu_1 = \mu_2 = \cdots = \mu_d$

on calcule le **Lambda de *Wilks*** :

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} = \frac{|\mathbf{W}|}{|\mathbf{W}+\mathbf{B}|} = \frac{1}{|\mathbf{W}^{-1}\mathbf{B}+\mathbf{I}|} = \prod_{k=1}^{k=d}(1-\lambda_k) = \prod_{k=1}^{k=d}\left(\frac{1}{1+\mu_k}\right)$$

I. L' analyse factorielle discriminante

2. Les axes et les variables discriminantes

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} = \frac{|\mathbf{W}|}{|\mathbf{W}+\mathbf{B}|} = \frac{1}{|\mathbf{W}^{-1}\mathbf{B}+\mathbf{I}|} = \prod_{k=1}^{k=d}(1-\lambda_k) = \prod_{k=1}^{k=d}\left(\frac{1}{1+\mu_k}\right)$$

Sous H_0 , Λ suit la loi du même nom, à 3 paramètres $(p, n-K, K-1)$

On utilise généralement l'approximation :

$$-\left[n-1-\frac{1}{2}(p+K)\right]\ln(\Lambda) \approx \chi^2_{p(K-1)}$$

Il existe **3 autres tests** que l'on peut utiliser, en option dans *R*, dans la directive `summary.manova` :

- ***Lawley-Hotelling*** : $U^{(d)} = trace(\mathbf{W}^{-1}\mathbf{B}) = \sum_{k=1}^{d} \mu_k$
- ***Pillai*** : $V^{(d)} = trace(\mathbf{T}^{-1}\mathbf{B}) = \sum_{k=1}^{d} \frac{\mu_k}{1-\mu_k}$
- La plus grande valeur propre de ***Roy*** : $\theta = \mu_1$

I. L' analyse factorielle discriminante

2. Les axes et les variables discriminantes

2/ Détermination du nombre d'axes $d-q$ suffisants pour discriminer les nuages de points :

=> repose sur le Lambda de Wilks suivant :

$$\Lambda_q = \prod_{k=d-q}^{d} (1-\lambda_k) = \prod_{k=d-q}^{d} \left(\frac{1}{1+\mu_k}\right)$$

H_0: non significativité simultanée des q derniers axes discriminants

Introduction pas à pas de variables dans la règle d'après leur capacité à faire baisser le Lambda de Wilks :

Test de variation du Lambda de Wilks

$$\frac{n-K-q}{K-1}\left(\frac{\Lambda_q}{\Lambda_{q+1}}-1\right) \cong F_{(K-1;n-K-q)} \; ; q=1,\dots,K-1 \; ss \; H_0 \text{ " NS de l'axe } q+1\text{"}$$

dès que la statistique précédente n'est plus significative, on décide que la dimension de représentation est $d-q$.

I. L' analyse factorielle discriminante

3. Une ACP particulière

$$\begin{array}{c|cccc|cccc}
 & 1 & 2 & \dots & k & 1 & 2 & j & p \\
1 & 1 & 0 & \dots & 0 & X_1^1 & X_1^2 & X_1^j & X_1^p \\
2 & 1 & 0 & \dots & 0 & & & & \\
 & & & \dots & & & & & \\
i & 0 & 0 & \dots & 1 & X_i^1 & X_i^2 & X_i^j & X_i^p \\
n & 0 & 0 & \dots & 1 & X_n^1 & X_n^2 & X_n^j & X_n^p
\end{array}$$

indicatrices des groupes — variables explicatives

définissons la matrice indicatrice X_K $(n \times K)$ des classes.

*Matrice **X** tableau initial $n \times p$, centré*

I. L' analyse factorielle discriminante

3. Une ACP particulière

- **AFD⇔ ACP (X_{Gk}, Q, D) du nuage X_{Gk} des K centroïdes g_k** où
  - La métrique $D_{K\times K}$ sur $\mathbb{R}^K$ **(espace des variables)** :
    - la matrice diagonale des poids $q_k = n_k/n$ des classes
  - la métrique $Q_{p\times p}$ sur $\mathbb{R}^p$ **(espaces des individus)** :
    - $\mathbf{T}^{-1}$ ou $\mathbf{W}^{-1}$ dite métrique de *Mahalanobis*.

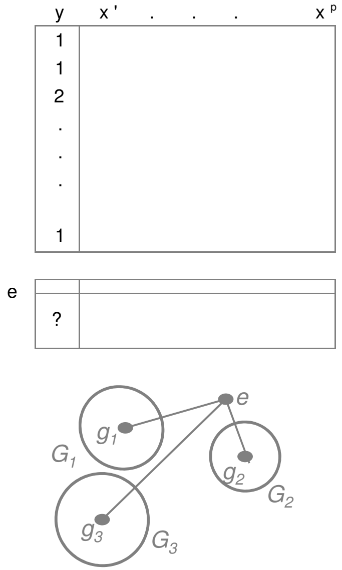
  Remarques:
  - L'utilisation de $\mathbf{T}^{-1}$ et $\mathbf{W}^{-1}$ comme métrique est donc indifférente, on dit qu'elles sont *équivalentes.*
  - La métrique $\mathbf{W}^{-1}$ (métrique **de Mahalanobis**) est plus utilisée par les Anglo-saxons et les éditeurs de logiciels.
  - Distance d de 2 indiv $\boldsymbol{x}$ et $\boldsymbol{y}$: $d^2(\boldsymbol{x},\boldsymbol{y}) = (\mathbf{x}-\mathbf{y})^T\mathbf{W}^{-1}(\mathbf{x}-\mathbf{y})$
- Ces métriques correspondent à une **projection oblique.** Sans cette oblicité, il s'agirait d'une simple *ACP* ; mais les groupes seraient mal séparés.
- Nombre d'axes discriminants est au plus égal à $K-1$ dans le cas courant où $n > p > K$.

I. L' analyse factorielle discriminante

- **Conséquences construction *AFD***
  - Lien avec d'autres méthodes (cf. Lebart & al. (1995), p. 259) :
    - ACP
      - les variables discriminantes sont non corrélées 2 à 2
      - On pourra interpréter les variables discriminantes au moyen du cercle de corrélation
    - ACC
  - Pas de test, mais … sous réserve de ne pas rejeter l'hypothèse de Normalité ….
  - Pas d'erreurs standard sur les coefficients
  - MAIS possibilité d'utiliser les méthodes de type « pas à pas » comme en régression. Sous R : **`stepclass(klaR)`**.

I. L' analyse factorielle discriminante

4. Méthodes géométriques de classement



- Échantillon d'apprentissage
- e observation de groupe inconnu
- **Règle géométrique** de classement :
  - e classé dans le groupe k tel que $d(e\,;\,g_k)$ soit minimale

L' analyse factorielle discriminante

4. Méthodes géométriques de classement

- **Règles géométriques** : e classé ds le groupe G_k pour lequel la distance (définie par $\mathbf{W}^{-1}$) à g_k est minimale :

Cte ne dépendant pas de k

$$d^2(e,g_k) = (e-g_k)^{\mathrm{T}}\mathbf{W}^{-1}(e-g_k) = e^{\mathrm{T}}\mathbf{W}^{-1}e - 2g_k^{\mathrm{T}}\mathbf{W}^{-1}e + g_k^{\mathrm{T}}\mathbf{W}^{-1}g_k$$

- D'où

$$\boxed{\textbf{Minimiser } d^2(e,g_k) \Leftrightarrow \textbf{ maximiser } \left(2g_k^{\mathrm{T}}\mathbf{W}^{-1}e - g_k^{\mathrm{T}}\mathbf{W}^{-1}g_k\right)}$$

=> **règle linéaire** par rapport aux coordonnées de e

- Pour chacun des K groupes G_k, on a une **fonction discriminante de Fisher** (fonction de classement !) obtenue après inversion de la matrice $\mathbf{W}$:

$$\alpha_k + \beta_{k,1}\mathbf{X}^1 + \beta_{k,2}\mathbf{X}^2 + \cdots + +\beta_{k,p}\mathbf{X}^p$$

=> e classé dans le groupe k où la fonction est maximale.
Sous R, utiliser la fonction `predict`

L'analyse discriminante linéaire

Les iris de Fisher `data(iris)`

```
> library(MASS)
> ir.lda<-lda(Species ~ ., iris); ir.lda #va et vp de W-1B
[…]
Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026

Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8293776  0.02410215
Sepal.Width   1.5344731  2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width  -2.8104603  2.83918785

Proportion of trace:
   LD1    LD2
0.9912 0.0088
```

Coefficients des 2=3-1 fonctions discriminantes.

Coordonnées des indiv sur les 2 variables discriminantes (non corrélées) = combinaisons linéaires des variables initiales centrées.

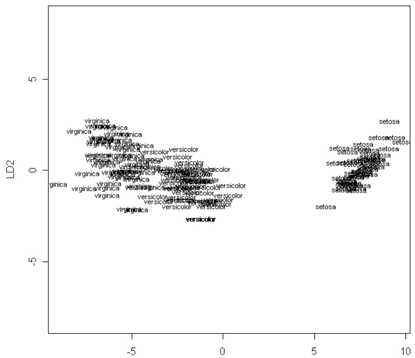
99.1% de la variabilité interclasse est expliquée par le 1er axe discriminant !

L'analyse discriminante linéaire

Les iris de Fisher `data(iris)`

```
> plot(ir.lda)
```

```
# règles géométriques
> pred<-predict(ir.lda)$class
> table(iris$Species,pred)
              pred
             setosa versicolor virginica
setosa           50          0         0
versicolor        0         48         2
virginica         0          1        49
```



Représentation des individus dans le plan 1-2 de l'AFD

L'analyse discriminante linéaire

Les iris de Fisher `data(iris)`

- Utilisation de la fonction `discrimin(ade4)`

```
#va et vp de T-1B
```

ACP du nuage des centroïdes g_k

```
> dis1 <- discrimin(dudi.pca(iris[, 1:4], scan = F),
  iris$Species, scan = F)
> dis1
Discriminant analysis
call: discrimin(dudi = dudi.pca(iris[, 1:4], scan = F), fac = iris$Species, scannf = F)
class: discrim
$nf (axis saved) : 2
eigen values: 0.9699 0.222 (cf résultats précédents)
          data.frame nrow ncol content
1 $fa          4    2    loadings / canonical weights u
2 $li        150    2    canonical scores      F=Xu
3 $va          4    2    cos(variables, canonical scores)
4 $cp          4    2    cos(components, canonical scores)
5 $gc          3    2    class scores
```


II Analyse discriminante probabiliste

- **Approche géométrique de classement** ne prend pas en compte les proba *a priori* des différentes classes, qui peuvent être très inégales !
- **Modèle bayésien d'affectation** :
  - Pour tout $i \leq k$, soient :
    - $P(G_i/x)$ = proba *a posteriori* d'appartenance à G_i sachant x (connaissant les caractéristiques de x, son « dossier »)
    - $p_i = P(G_i)$ = proba *a priori* d'appartenance à G_i (proportion de G_i dans la population)
    - $f_i(x) = P(x/G_i)$ = densité conditionnelle de la loi de x connaissant son groupe G_i
  - D'après le **théorème de Bayes** :

$$\boxed{P(G_i/\mathbf{x}) = \frac{p_i f_i(\mathbf{x})}{\sum_{i=1}^{k} p_i f_i(\mathbf{x})}}$$

  - **Règle de classement bayésienne** :
    - on classe x dans le groupe G_i où $P(G_i/x)$ est **maximum**

=> <u>**Pb**</u> = estimer $P(G_i/x)$!

II Analyse discriminante probabiliste :

3 possibilités pour estimer $P(G_i/x)$

En commençant par calculer $P(x/G_i)$

1. Selon une **méthode paramétrique** (on suppose la **multinormalité** de $P(x/G_i)$ avec éventuellement égalité des $\mathbf{\Sigma}_i$, donc le nb de paramètres du problème est fini : AD Linéaire ou AD Quadratique)
2. Selon une **méthode non paramétrique** (pas d'hypothèse sur la densité $P(x/G_i)$: méthode du noyau ou des plus proches voisins)
3. Directement par une **approche semi-paramétrique** (régression logistique) où on écrit $P(G_i/x)$ sous la forme :

$$P(G_i/x) = \frac{e^{\alpha' x+\beta}}{1+e^{\alpha' x+\beta}}$$

II Analyse discriminante probabiliste :

La règle bayésienne naïve dans le cadre Normal

- La densité d'une loi multinormale $N(\mu_i,\Sigma_i)$ est :

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma_i)}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_i)'\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right]$$

- D'après Bayes, maximiser $P(G_i/\mathbf{x})$ ⇔ maximiser $p_i f_i(\mathbf{x})$ ie attribuer $\mathbf{x}$ au groupe le plus probable a posteriori

$$\max_i \left[Log(p_i) - \frac{1}{2}(\mathbf{x}-\mu_i)'\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{1}{2}Log(\det(\Sigma_i))\right]$$

=> On obtient une **règle quadratique** en x !

II Analyse discriminante probabiliste : Hypothèses Normalité + homoscédasticité

Hypothèse simplificatrice : $\Sigma_1 = \Sigma_2 \quad \ldots \quad = \Sigma$

On attribue x au groupe j tel que :

$$\max \left[\mathrm{Ln}\ \mathrm{p_j} - \underbrace{\frac{1}{2} x'\Sigma^{-1} x}_{\substack{\textit{indépendant} \\ \textit{du groupe}}} - \frac{1}{2}\ \mu_j' \Sigma^{1} \mu_j + x'\Sigma^{-1} \mu_j \right]$$

$$\textit{donc} : \max \left[\underbrace{\mathrm{Ln}\ \mathrm{p_j} - \frac{1}{2} \mu_j' \Sigma^{-1} \mu_j}_{a_j} + x'\Sigma^{-1} \mu_j \right]$$

Règle linéaire équivalente à la règle géométrique si équiprobabilité, après estimation de μ_j par g_j et de Σ par W.

Hypothèses Normalité + homoscédasticité + équiprobabilité
=> équivalence des règles géométrique (maximiser la fct de Fisher) et bayésienne.

L'analyse discriminante linéaire
Les iris de Fisher data(iris)

```
# erreur d'apprentissage : analyse disc probabiliste
> table(iris[,"Species"],predict(ir.lda,iris)$class)
# matrice de confusion
```

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	$\boxed{2}$
virginica	0	$\boxed{1}$	49

Classes obtenues par resubstitution

Classes observées n1=n2=n3 =50

L'analyse discriminante linéaire
Les iris de Fisher data(iris)

```
> pred<-predict(ir.lda)$class
# classement à partir des 2 fonctions de score LD1 et LD2
> pred.ld1<-predict(ir.lda,dimen=1)$class
# classement à partir de la fonction de score LD1 seult
> table(Species,pred.ld1)
             pred.ld1
Species       setosa versicolor virginica
  setosa          50          0         0
  versicolor       0         48         2
  virginica        0          0        50
```

qda (règle quadratique) existe sous R

L'analyse discriminante linéaire
Les iris de Fisher data(iris)

Autre évaluation des fonctions discriminantes = test ANOVA pour voir si les groupes considérés diffèrent pour les valeurs moyennes de LD1 et LD2.

```
> ld1 <- predict(ir.lda)$x[,1] # valeurs de LD1
> ld2 <- predict(ir.lda)$x[,2] # valeurs de LD2 pour les 150
> anova(lm(ld1 ~ Species))
Analysis of Variance Table
Response: ld1
           Df Sum Sq Mean Sq F value    Pr(>F)
Species     2 4732.2  2366.1  2366.1 < 2.2e-16 ***
Residuals 147  147.0     1.0
---
> anova(lm(ld2 ~ Species))
Analysis of Variance Table
Response: ld2
           Df  Sum Sq Mean Sq F value   Pr(>F)
Species     2  41.952  20.976  20.976 9.68e-09 ***
Residuals 147 147.000   1.000
> ?qlda #règle quadratique
```

les 2 fonctions de scores discriminent le facteurs Species

Comment estimer un tx d'erreur non biaisé ?

- Les performances « par défaut » de la règle sont optimistes !
- La règle est évaluée à partir des données même qui ont conduit à son élaboration

(méthode dite de resubstitution)
- Il faudrait pouvoir **l'évaluer sur de nouveaux individus** !

Les solutions :
- Méthode d'échantillon / test
- Validation croisée (Leave One Out ou LOO)
- Technique du bootstrap

Avantages de l'analyse discriminante

- Problème à solution analytique directe (calcul des vecteurs propres de $\mathbf{W}^{-1}\mathbf{B}$)
- Optimale quand les hypothèses de non colinéarité des variables, homoscédasticité et multinormalité sont vérifiées
- Les coef. des combinaisons linéaires constituent un résultat simple qui peut s'interpréter par la corrélation avec les variables de départ, pratiquement comme dans une régression
- Modélise très bien les phénomènes linéaires
- Ne nécessite pas un gros ensemble d'apprentissage
- Rapidité de calcul du modèle
- Possibilité de sélection pas à pas
- Facilité d'intégrer des coûts d'erreur de classement
- Technique implémentée dans de nombreux logiciels
- Rééchantillonnage simple en particulier le **jackknife**.

Inconvénients de l'analyse discriminante

- **Ne détecte que les phénomènes linéaires**, mais il existe une analyse discriminante quadratique qui, tout en s'appuyant sur les mêmes principes introduit davantage de paramètres.
- **Ne s'applique pas à tout type de données** (données numériques sans valeurs manquantes)
- **Hypothèses contraignantes**, et pour s'en rapprocher :
  - normaliser les variables
  - sélectionner soigneusement les variables les + discriminantes
  - éliminer les variables colinéaires
  - éliminer les individus hors norme
  - s'il reste de l'hétéroscédasticité, mieux vaut avoir des classes de tailles comparables
  - travailler sur des populations homogènes

Références bibliographiques

- L. Bellanger, R. Tomassone, *Exploration de données et méthodes statistiques : Data analysis & Data mining avec R. Collection Références Sciences, Editions* Ellipses, Paris, 2014.
- A. Bouchier, Documents et supports de cours disponibles sur le site : http://rstat.ouvaton.org/
- B.S. Everitt, S. Landau, L. Morven. *Cluster Analysis*, 4th ed., Oxford University Press Inc., Oxford, 2001..
- A.D., Gordon, A. D., *Classification*. 2nd Edition. London: Chapman and Hall / CRC, 1999.
- F. Husson, S. Lê & J. Pagès, *Analyse de données avec R*. PUR, Rennes, 2009.
- L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York, 1990.
- L. Lebart, A. Morineau, M. Piron, *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 2006.
- J.-P. Nakache, J. Confais, *Approche pragmatique de la Classification*. Editions Technip, Paris, 2005.
- G. Saporta, *Probabilités, Analyse des données*. Editions Technip, Paris, 2006.
- Statistics with R : http://zoonek2.free.fr/UNIX/48_R/all.html
- S. Tufféry, *Data mining et statistique décisionnelle : L'intelligence dans les bases de données*. Editions Technip, Paris, 2005.