

Analyse factorielle discriminante

Michaël Genin

Université de Lille 2

EA 2694 - Santé Publique : Epidémiologie et Qualité des soins

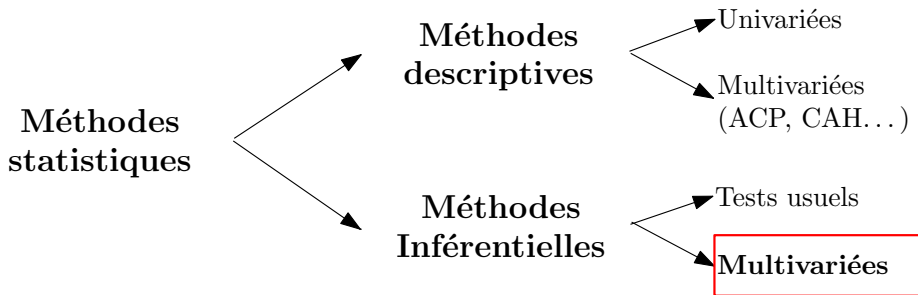
michael.genin@univ-lille2.fr

Master 1 Biologie Santé - Parcours C

Plan

- 1 Introduction
- 2 Principe général
- 3 Principe d'interprétation

Analyse factorielle discriminante (AFD)



2 familles de méthodes de classification

Classification non-supervisée (clustering)

- Partitionner les observations en groupes différents (classes, catégories) mais les plus homogènes possible au regard de variables décrivant les observations.
- **Le nombre de classes n'est pas connu à l'avance**
- Méthodes : Classification hiérarchique, K-plus-proches voisins, Classification bayésienne naïve. . .

Classification supervisée (discrimination)

- Obtenir un critère de séparation afin de prédire l'appartenance à une classe ($Y = f(X) + \epsilon$).
- **Le nombre de classes est connu à l'avance (Variable à expliquer)**
- Méthodes : Régression logistique, **Analyse discriminante**, Arbres de décision, Réseaux de neurones, Réseaux bayésiens, Support Vector Machine...

Méthodes de discrimination

2 objectifs principaux :

- Etude du lien entre Y (Variable à expliquer **qualitative**) et les X_j (Variables explicatives **quantitatives ou binaires**) \Rightarrow Facteurs prédictifs
- Prédiction (système d'aide à la décision (scores cliniques, crédit scoring, ...))

2 catégories de méthodes de discrimination :

- 1 Méthodes explicatives : règles de prédiction claires (**AFD**, Reg Log, Arbres de décision)
- 2 Méthodes non explicatives : règles de prédiction floues (RN, RB, SVM...)

En pratique en médecine

- 2 classes \Rightarrow Régression logistique
- > 2 classes : **Analyse discriminante**, Arbres de décision

En résumé

L'Analyse Factorielle Discriminante est une méthode de discrimination, explicative qui a pour but :

- Etude du lien entre Y (Variable à expliquer **qualitative**) et les X_j (Variables explicatives **quantitatives ou binaires**) \Rightarrow Facteurs prédictifs
- Prédiction de l'appartenance à une classe

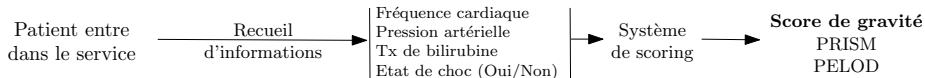
Modèle linéaire

On considère des combinaisons linéaires entre les X_j

$$\text{Score} = \lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_p X_p = \sum_{j=1}^p \lambda_j X_j$$

Ce (ou ces) score va permettre de prédire l'appartenance des individus à une classe (Y).

Exemple 1 : score en réanimation



PRISM : Pediatric RISK of Mortality

PELOD : PEdiatric Logistic Organ Dysfunction

Exemple 2 : Score de Framingham

Prédiction d'un évènement cardio-vasculaire dans les 10 ans.

Construit à partir de la cohorte de Framingham (5 209 individus)

Age (classes quinquennales)

[55-59 ans] → + 4

Tx de cholesterol LDL

si $\in [100 - 160]$: 0

si < 100 : -3 (Protecteur)

si ≥ 160 : +2 (Risque)

PA diastolique (PAD) et PA systolique (PAS) en mm de mercure

SI PAD < 80 ET PAS < 120 : 0

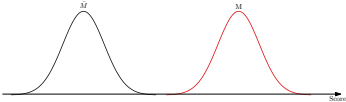

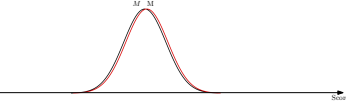
SI PAD ≤ 89 ET PAS $\in [130 - 139]$: +1

Si $S \geq 14 \rightarrow 56\%$ de risque d'évènement CV dans les 10 ans.

Lien avec la notion de score linéaire :

$$\text{Score Framingham} = \lambda_1 X_1 + \cdots + \underbrace{\text{Age}[55-59]}_{0/1} \underbrace{\lambda}_{=4} + \cdots + \lambda_p X_p$$

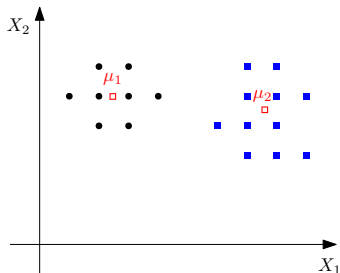
$k=2$ et score déjà connu

Cas	Représentation graphique	Qualité de séparation
Cas 1		Bonne
Cas 2		Moyenne
Cas 3		Mauvaise

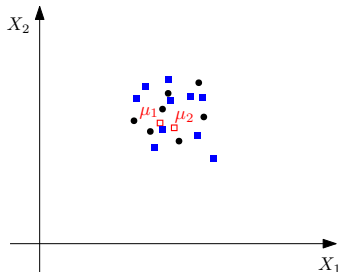
Cas 3 : impossibilité de trouver un score discriminant les 2 groupes.

Condition nécessaire

Les groupes doivent être séparables (non-superposés)

Exemple : X_1, X_2 et $K = 2$ 

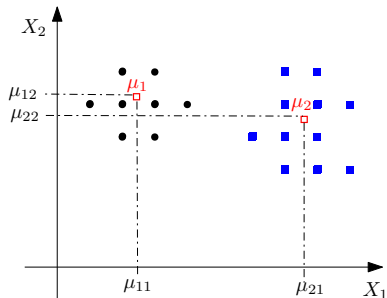
Les centres de gravité μ_1 et μ_2 sont séparés (*i.e.* les groupes sont séparés)



Les centres de gravité μ_1 et μ_2 ne sont pas séparés (*i.e.* les groupes ne sont pas séparés)

k=2 et score déjà connu

Point d'entrée de l'analyse : tester la séparabilité des groupes en utilisant les coordonnées des centres de gravités :



X_1 et X_2

$$\mu_1 = \begin{pmatrix} \mu_{11} \\ \mu_{12} \end{pmatrix} \quad \mu_2 = \begin{pmatrix} \mu_{21} \\ \mu_{22} \end{pmatrix}$$

X_1, \dots, X_p

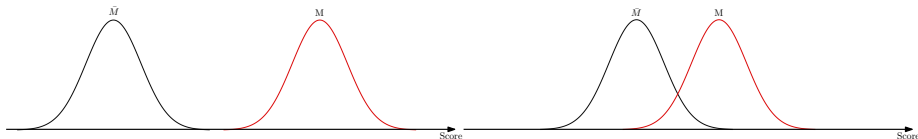
$$\mu_1 = \begin{pmatrix} \mu_{11} \\ \vdots \\ \mu_{1p} \end{pmatrix} \quad \mu_2 = \begin{pmatrix} \mu_{21} \\ \vdots \\ \mu_{2p} \end{pmatrix}$$

MANOVA : Multivariate ANALysis Of VArance

$$\begin{cases} \mathcal{H}_0 : & \mu_1 = \mu_2 & \text{Groupes confondus} \\ \mathcal{H}_1 : & \mu_1 \neq \mu_2 & \text{Groupes séparés} \end{cases}$$

k=2 et score déjà connu

Si les groupes sont séparés (MANOVA) \Rightarrow Retour aux scores discriminants



Cas 1 : le score discrimine bien les deux groupes

Cas 2 : le score n'est pas assez discriminant pour réaliser des prédictions

Nécessité

Pour les scores \Rightarrow utilisation d'un critère de qualité de discrimination

k=2 et score déjà connu

Idée : ANOVA sur le score

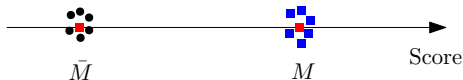
En utilisant le théorème de Huygens

$$\underbrace{S_T^2}_{\text{Variance totale}} = \underbrace{S_B^2}_{\text{Variance inter-classes}} + \underbrace{S_W^2}_{\text{Variance intra-classe}}$$

Indicateur de qualité de séparation entre les groupes

$$R^2 = \frac{S_B^2}{S_T^2} \in [0, 1]$$

Remarque : si $R^2 \approx 1 \rightarrow$ variance intra quasi-inexistante :



Cas de 2 groupes ($k = 2$)

Objectif de l'AFD

Déterminer parmi toutes les combinaisons linéaires des X_j ($\sum_{j=1}^p \lambda_j X_j$), les pondérations λ_j qui **maximisent le R^2** .

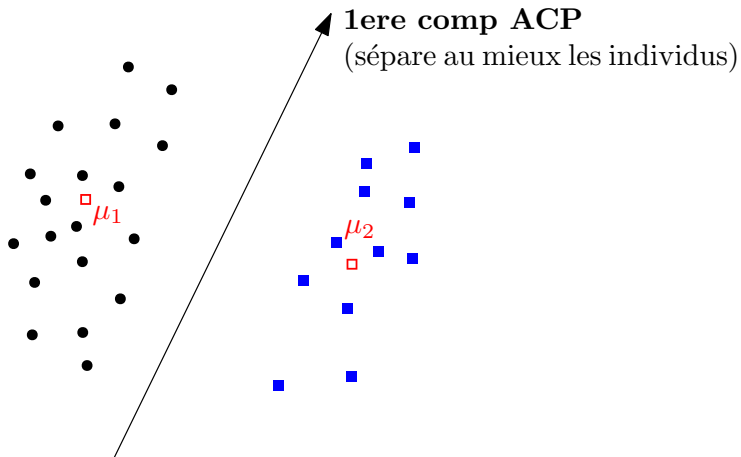
Problème : il existe une infinité de combinaisons de λ_j . Comment déterminer les λ_j optimaux ?

Théorème

Si les groupes sont séparés (MANOVA) alors il existe une combinaison linéaire (score discriminant, composante discriminante) **unique** qui maximise le R^2 .

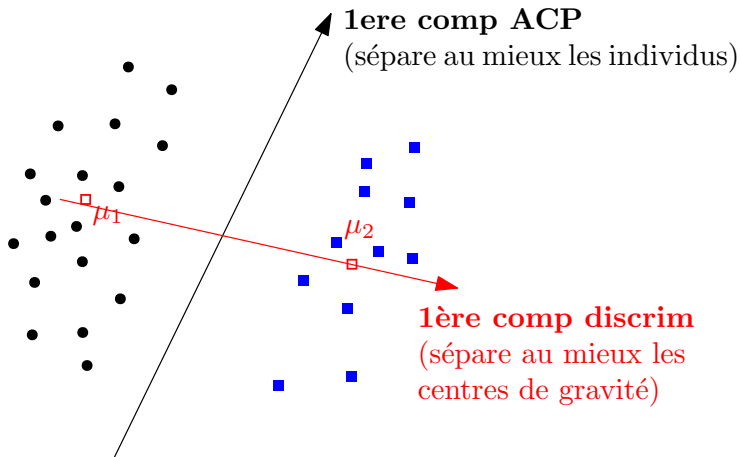
Cas de 2 groupes ($k = 2$)

Lien avec l'ACP



Cas de 2 groupes ($k = 2$)

Lien avec l'ACP



Cas de 2 groupes ($k = 2$)

Détermination des λ_j

AFD : ACP particulière sur les centres de gravité :

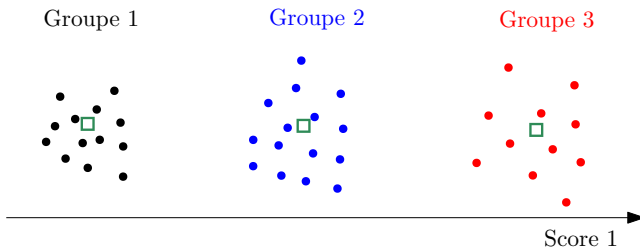
	X_1	X_2	\dots	X_j	\dots	X_p
G_1	μ_{11}	μ_{12}	\dots	μ_{1j}	\dots	μ_{1p}
G_2	μ_{21}	μ_{22}	\dots	μ_{2j}	\dots	μ_{2p}

Distance particulière : distance de Mahalanobis

- Maximise l'inertie inter-classe projetée sur l'axe
- Minimise l'inertie intra-classe projetée sur l'axe

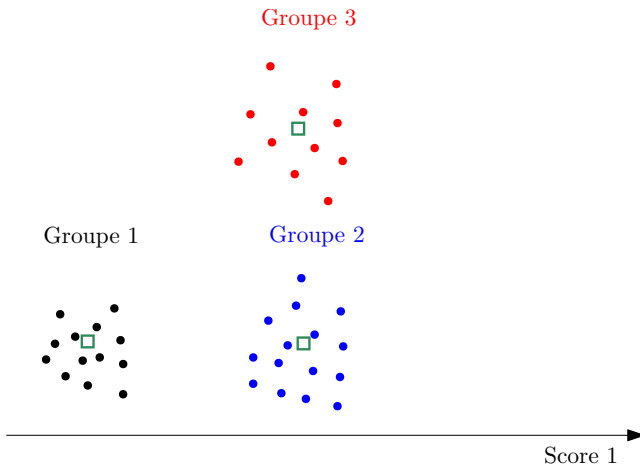
Cas de 3 groupes ($k = 3$)

Situation rare :



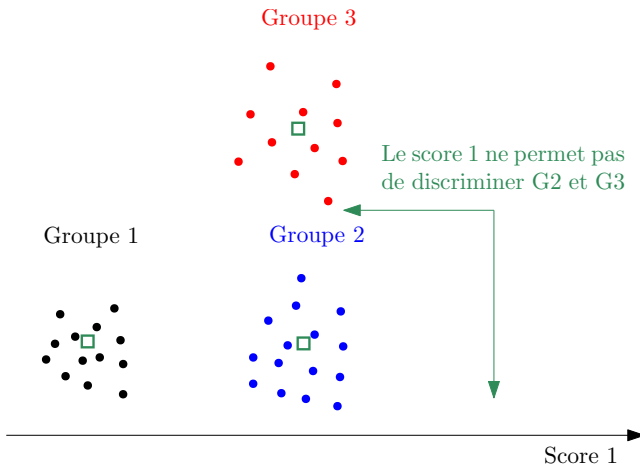
Cas de 3 groupes ($k = 3$)

Situation plus fréquente :



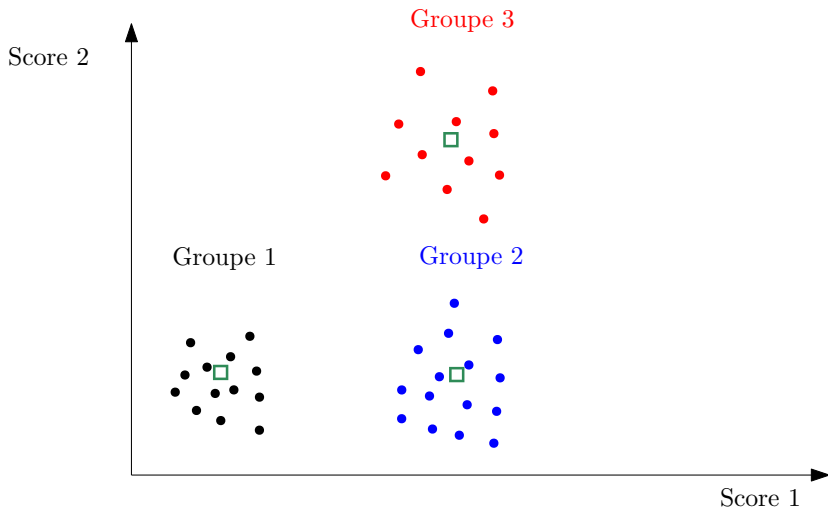
Cas de 3 groupes ($k = 3$)

Situation plus fréquente :



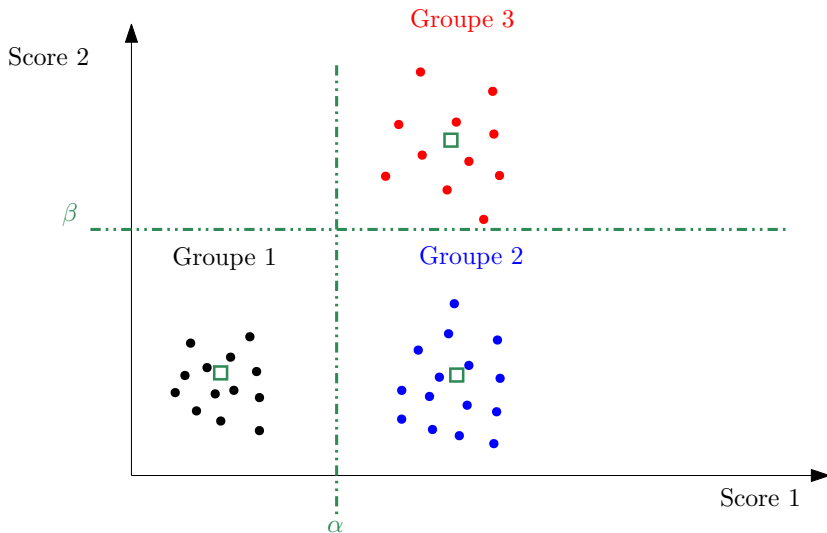
Cas de 3 groupes ($k = 3$)

Situation plus fréquente :



Cas de 3 groupes ($k = 3$)

Situation plus fréquente :



Cas de k groupes ($k > 2$)

Théorème

Soit Y qui définit k groupes. Si les groupes sont séparés, alors

Il existe $k - 1$ | composantes discriminantes | tels que
scores discriminants

1^{er} score S_1 rend maximal le R^2

2^{ème} score S_2 est orthogonal à S_1 et maximise le R^2

⋮

$(k - 1)$ ^{ème} score S_{k-1} est orthogonal à S_{k-2} et maximise le R^2

Résumé

AFD : méthode explicative de discrimination

- Une variable à expliquer **qualitative** Y à k groupes (classes)
- p variables explicatives X_j quantitatives ou binaires
- Etudier les variables discriminantes des groupes
- Prédire l'appartenance à un groupe
- Méthode linéaire : scores linéaires qui vont prédire l'appartenance aux classes
- Les classes doivent être séparées (MANOVA)
- Les scores : issus d'une ACP particulière sur les centres de gravités (composantes)
- Toujours $k - 1$ scores discriminants

Principe d'interprétation

3 étapes clés :

- ❶ Est-ce que, mathématiquement, la discrimination est bonne ?
 - Est-ce que les groupes sont bien séparés par les scores ?
- ❷ Est-ce que les scores ont une interprétation clinique ?
 - Cohérence par rapport à l'expertise clinique. . .
- ❸ Construction de règles de classement
 - Règle d'affectation d'un nouvel individu à une classe

Principe d'interprétation - Données exemple

Données "insectes" de Lubischew ($n = 72$)¹.

- **Variable à expliquer** : espèce d'insecte (species)
 - Concinna (con) (codée 1)
 - Heikertingeri (hei) (codée 2)
 - Heptapotamica (hep) (codée 3)
 - $Y = \{\text{con,hei,hep}\}$
- **Variables explicatives**
 - Largeur de l'appareil reproducteur (aedeagus) (μm) (width)
 - Angle de l'appareil reproducteur (aedeagus) (degré) (angle)

Objectifs

- Déterminer quelle sont les variables discriminant les groupes d'insectes
- Etablir des règles de classement

1. Lubischew, A.A. (1962) On the use of discriminant functions in taxonomy. Biometrics, 18, 455-477

Principe d'interprétation - Interprétation mathématique

A - Vérification de la condition nécessaire

MANOVA : Multivariate ANalysis Of VAriance

$$\begin{cases} \mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 & \text{Groupes confondus} \\ \mathcal{H}_1 : \exists \text{ au moins } (i,j) / \mu_i \neq \mu_j & \text{Groupes séparés} \end{cases}$$

Sous SPSS

Wilks' Lambda				
Test of Function (s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,047	215,101	4	,000
2	,250	97,622	1	,000

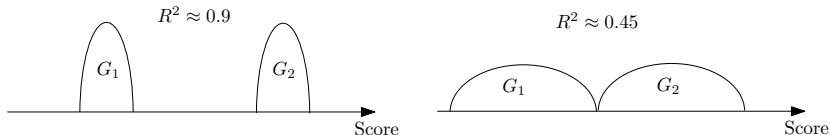
Principe d'interprétation - Interprétation mathématique

B - Utilisation de plusieurs critères

- $R^2 \rightarrow$ autant que de scores discriminants

Proche de 1 ?

Exemple



Pourtant le score discrimine bien dans les 2 cas

\rightarrow Pas forcément de seuil sur le R^2

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	4,293 ^a	58,9	58,9	,901
2	2,994 ^a	41,1	100,0	,866

$$0.901^2 = 0.811 = R^2$$

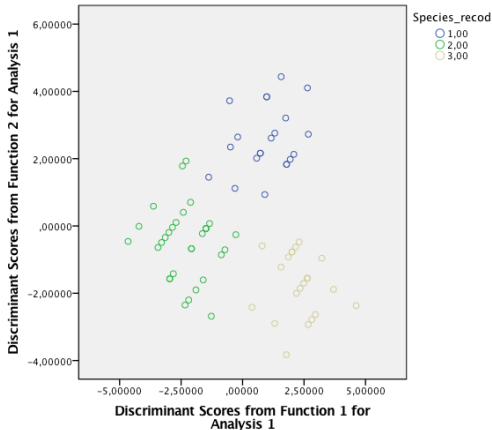
$$0.866^2 = 0.749 = R^2$$

Principe d'interprétation - Interprétation mathématique

B - Utilisation de plusieurs critères

- Représentations graphiques

Rep. des individus sur l'espace des scores discriminants

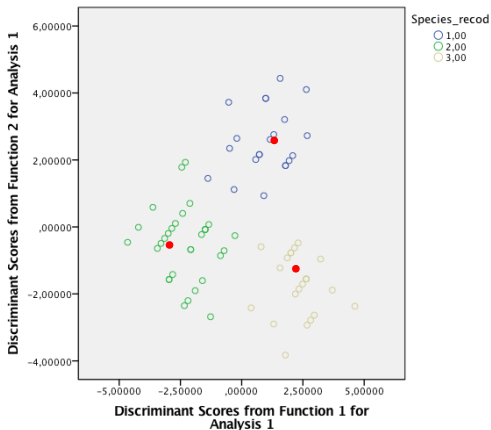


Principe d'interprétation - Interprétation mathématique

B - Utilisation de plusieurs critères

- Classements automatiques

Méthode des médiatrices

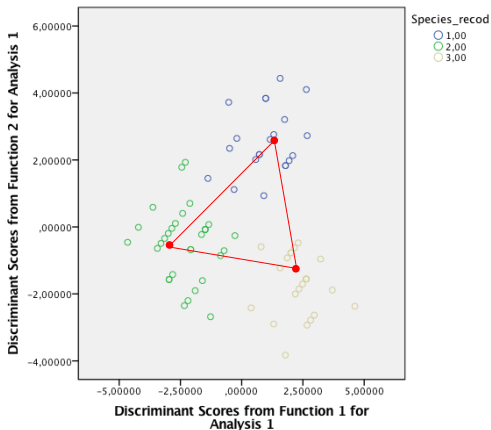


Principe d'interprétation - Interprétation mathématique

B - Utilisation de plusieurs critères

- Classements automatiques

Méthode des médiatrices

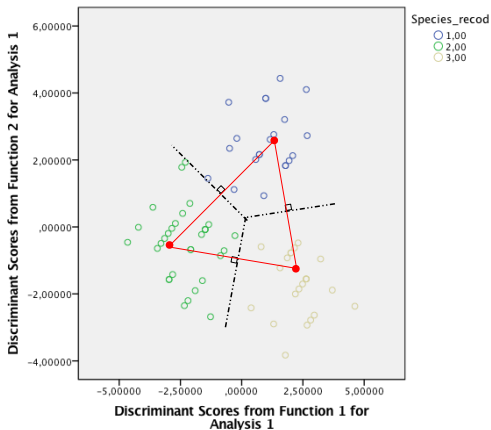


Principe d'interprétation - Interprétation mathématique

B - Utilisation de plusieurs critères

- Classements automatiques

Méthode des médiatrices

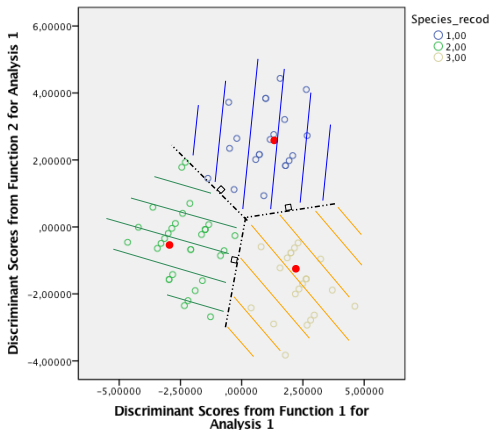


Principe d'interprétation - Interprétation mathématique

B - Utilisation de plusieurs critères

- Classements automatiques

Méthode des médiatrices



Principe d'interprétation - Interprétation mathématique

B - Utilisation de plusieurs critères

- Classements automatiques

Matrice de confusion

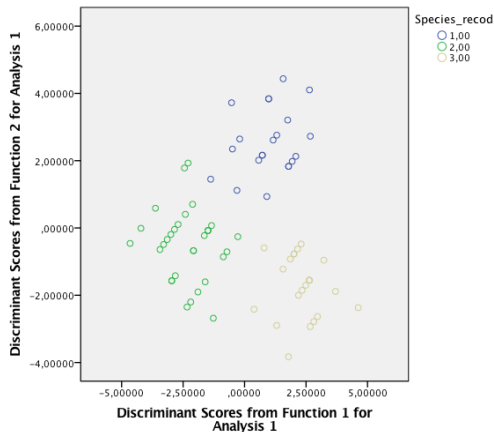
Classification Results^a

			Predicted Group Membership			Total
			1,00	2,00	3,00	
Original	Count	1,00	20	1	0	21
		2,00	0	31	0	31
		3,00	0	0	22	22
	%	1,00	95,2	4,8	,0	100,0
		2,00	,0	100,0	,0	100,0
		3,00	,0	,0	100,0	100,0

a. 98,6% of original grouped cases correctly classified.

En pratique : $\geq 80\%$ d'observations bien classées

Principe d'interprétation - Interprétation clinique



Valeurs élevées de $S_2 \rightarrow$ Groupe 1

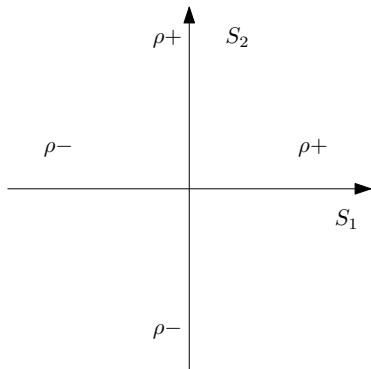
Principe d'interprétation - Interprétation clinique

Idée : corrélation entre les X_j et chacun des scores

Règle

$$|\rho(X_j, S_k)| > 0.5$$

Rq : si X_j est binaire (0/1) : ANOVA $\equiv \rho(X_j, S_k)$

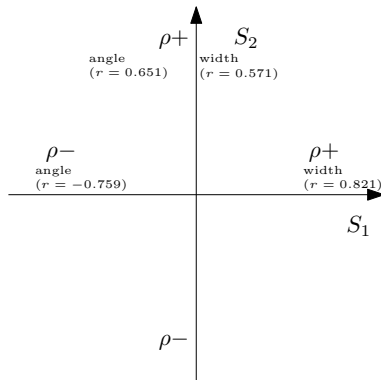


Principe d'interprétation - Interprétation clinique

Idée : corrélation entre les X_j et chacun des scores

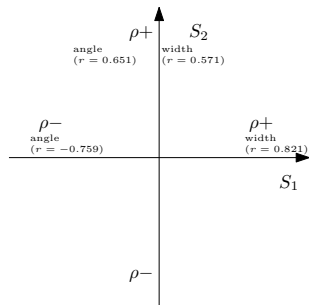
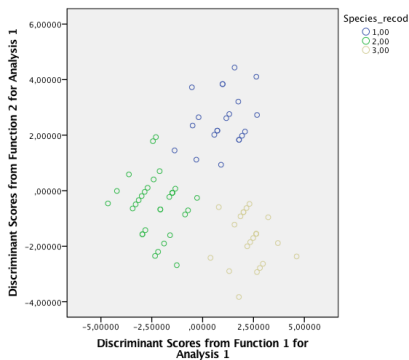
Correlations

	Discriminant Scores from Function 1 for Analysis 1	Discriminant Scores from Function 2 for Analysis 1
Width	,821 ,000 74	,571 ,000 74
Angle	-,759 ,000 74	,651 ,000 74



Principe d'interprétation - Interprétation clinique

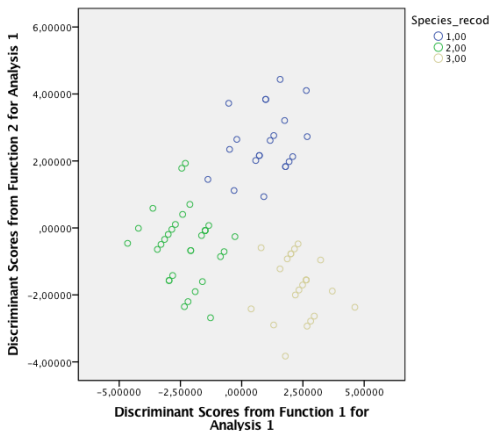
Idée : corrélation entre les X_j et chacun des scores



Construction de règles de classement

3 solutions :

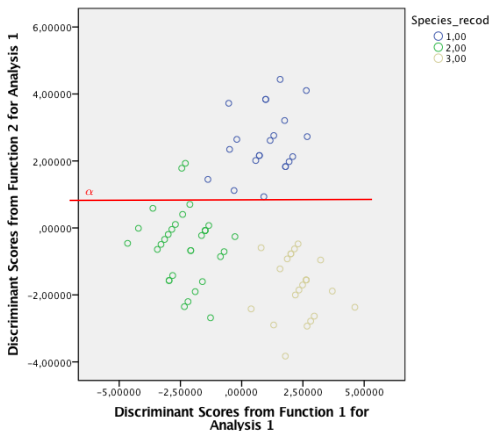
- ① Utiliser les classes prédites par le logiciel (Méthode des médiatrices)
 - **Problème** : "boîte noire"
 - Pas de règle explicite
- ② Méthode graphique



Construction de règles de classement

3 solutions :

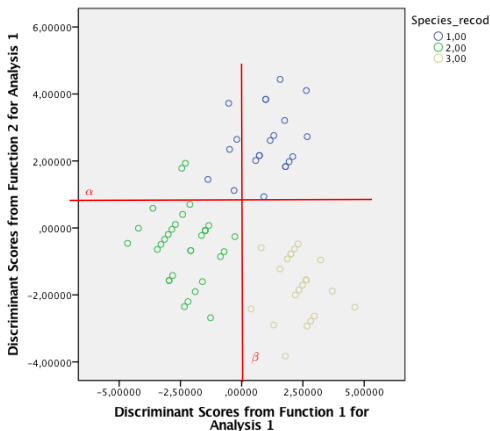
- ① Utiliser les classes prédites par le logiciel (Méthode des médiatrices)
 - **Problème** : "boîte noire"
 - Pas de règle explicite
- ② Méthode graphique



Construction de règles de classement

3 solutions :

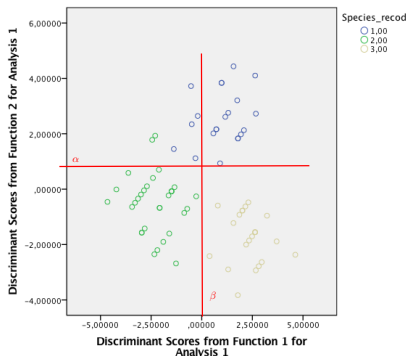
- 1 Utiliser les classes prédites par le logiciel (Méthode des médiatrices)
 - **Problème** : "boîte noire"
 - Pas de règle explicite
- 2 Méthode graphique



Construction de règles de classement

3 solutions :

- ① Utiliser les classes prédites par le logiciel (Méthode des médiatrices)
 - **Problème** : "boîte noire"
 - Pas de règle explicite
- ② Méthode graphique



Règle :

SI $S_2 > \alpha$ ALORS Groupe 1
 SINON

SI $S_1 > \beta$ ALORS Groupe 3
 SINON Groupe 2

FSI

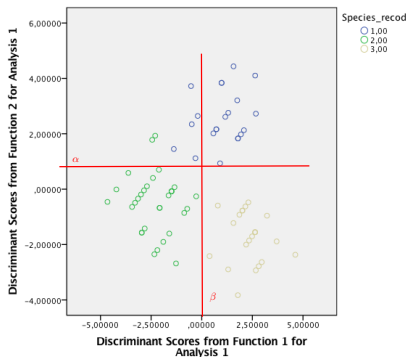
FSI

Seuils optimaux ?

Construction de règles de classement

3 solutions :

- ③ Courbe Roc pour déterminer α et β



Pour S_2 :

- ① Créer une variable binaire (G_1 vs G_2, G_3)
- ② Courbe ROC sur S_2 avec nouvelle variable
→ α optimal pour S_2

Pour S_1 :

- ① Sous-échantillon : uniquement G_2 et G_3
- ② Courbe ROC sur S_1 avec species
→ β optimal pour S_1

Construction de règles de classement

Classement d'un nouvel individu : angle=14 ; width=144

Calcul de S_1 et S_2 pour l'individu :

Canonical Discriminant
Function Coefficients

	Function	
	1	2
Width	,147	,149
Angle	-,625	,780
(Constant)	-11,752	-30,258

Unstandardized coefficients

$$S_1 = 0.147 \times \underbrace{\text{width}}_{=144} - 0.625 \times \underbrace{\text{angle}}_{=14} - 11.752 = 0.666$$

$$S_2 = 0.149 \times \underbrace{\text{width}}_{=144} + 0.780 \times \underbrace{\text{angle}}_{=14} - 30.258 = 2.118$$

Construction de règles de classement

Classement d'un nouvel individu : angle=14 ; width=144, $S_1 = 0.666$, $S_2 = 2.118$

Posons $\alpha = 1$ et $\beta = 0$

Règle :

SI $S_2 > \alpha$ ALORS Groupe 1

SINON

SI $S_1 > \beta$ ALORS Groupe 3

SINON Groupe 2

FSI

FSI

Ici $S_2 > \alpha$ donc le nouvel individu est affecté au groupe 1