RD7 : Analyse factorielle discriminante

1 Introduction

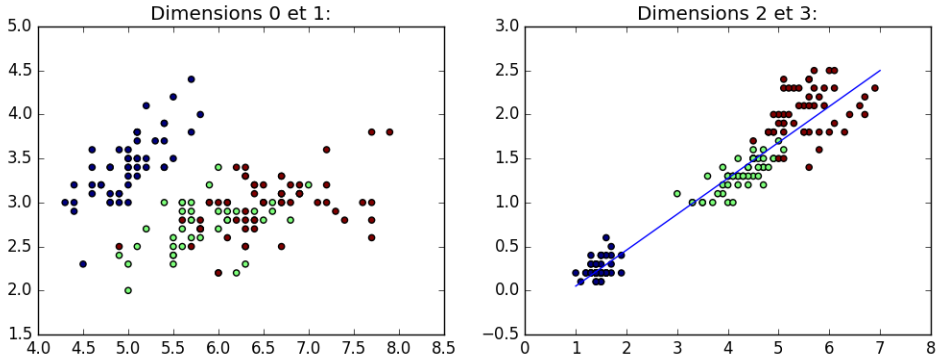
On considère ici q classes d'individus sur lesquels p variables quantitatives sont mesurées. On note T la matrice des indicatrices et $X = (X^1, \ldots, X^p)$ le tableau des variables quantitatives.

L'objectif de l'analyse factorielle discriminante (AFD) est double :

- **descriptif**, en cherchant les combinaisons linéaires des p variables permettant de séparer au mieux les individus et en donner une représentation graphique satisfaisant cet objectif,
- **décisionnel**, en permettant de classer de nouveaux individus dans les classes préexistantes connaissant les p variables.

Pour illustrer le cours, nous utiliserons l'exemple iris caractérisant $q = 3$ variétés d'iris par $p = 4$ variables quantitatives, longueur et klargeur des sépales et des pétales.

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn import datasets
iris = datasets.load_iris()
Xiris = iris.data
yiris = iris.target
plt.figure(figsize=(12,4))
plt.subplot(1, 2, 1)
plt.title("Longueur et largeur des sépales")
plt.scatter(Xiris[:, 0], Xiris[:, 1], c=yiris)
plt.subplot(1, 2, 2)
plt.title("Longueur et largeur des pétales")
plt.scatter(Xiris[:, 2], Xiris[:, 3], c=yiris)
plt.plot([1,7],[0.05,2.5])
plt.show()
```



L'AFD est une méthode très utilisée et diversifiée. On la rencontre en contrôle qualité, en diagnostic, en prévision des risques. L'approche décisionnelle conduit à la construction de scores permettant la construction de règles de classement. Par exemple, on peut ainsi prévoir le risque d'avalanche à partir de mesures météo, d'exposition, de pentes, faire un diagnostic médical à partir de mesures accessibles.

http ://cedric.cnam.fr/vertigo/Cours/ml/tpAfd.html

https ://archive.ics.uci.edu/ml/index.php

2 Détermination des fonctions linéaires discriminantes

Principe : Le principe général est de construire une première variable dite discriminante comme combinaison linéaire des variables initiales. Cette variable doit minimiser la variance intra-classe et maximiser la variance inter-classes (critère d'ajustement). La seconde variable discriminante est construite comme non corrélée à la première et vérifiant le même critère, et ainsi de suite.

2.1 Notation

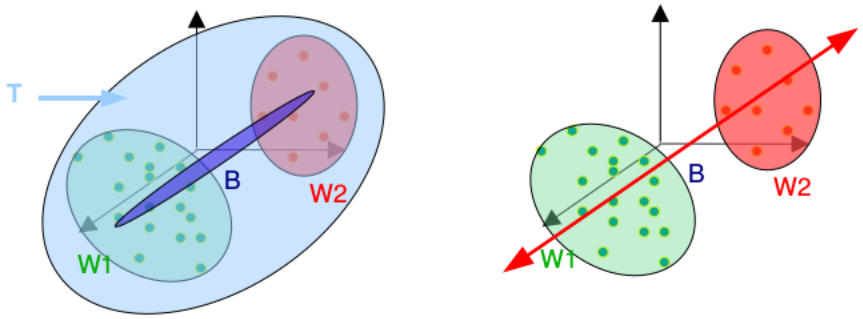
On considère un n échantillon constitué de n individus appartenant à des classes données, 1 à q, sur lesquels sont mesurés p variables quantitatives.

Les n individus sont représentés par des points $X_i^T = (x_i^1, \ldots, x_i^p)$ dans l'espace $\mathbb{R}^p$.

Pour simplifier les calculs tout en préservant la généralité de la démonstration, nous considérerons que tous les individus ont le même poids $\omega_i = \frac{1}{n}$ et que le tableau a été préalablement **centré**.

On note :

- ω_i le poids d'un individu i, D la matrice diagonale des poids,
- $I(k)$ représente l'ensemble des indices i des n_k individus de la classe k,
- n_k le nombre d'individus dans la classe k et $\omega_k = \sum_{i \in I(k)} \omega_i = \frac{n_k}{n}$ le poids de la classe k,
- $g = \sum_{i=1}^{n} \omega_i X_i$ le centre de gravité du nuage (l'origine ici car centré),
- $g_k = \sum_{i \in I(k)} \frac{\omega_i}{\omega_k} X_i$ le centre de gravité du sous-nuage $\mathcal{N}_k$ constitué des individus de la classe k,
- $W_k = \sum_{i \in I(k)} \frac{\omega_i}{\omega_k} (X_i - g_k)(X_i - g_k)^T)$ la matrice de variance-covariance du nuage $\mathcal{N}_k$,
- $B = \sum_{k=1}^{q} \omega_k (g_k - g)(g_k - g)^T$ la matrice de variance-covariance inter-classes (between),
- $W = \sum_{k=1}^{q} \omega_k W_k$ la matrice de variance-covariance intra-classes (within),
- $T = \sum_{i=1}^{n} \omega_i (X_i - g)(X_i - g)^T$ la matrice de variance-covariance totale.



2.2 Propriétés du nuage de points

Proposition 1 *Théorème de Huygens*

La matrice de variance covariance totale, T, est égale à la somme de la matrice de variance covariance intra groupe, W, et de la matrice de variance covariance inter groupe, B On obtient ainsi la relation : $T = B + W$.

Preuve Théorème de Huygens (RD1)

Proposition 2 *Cas d'une population multinormale*

Dans le cas d'une population multinormale, chaque individu suit une loi multinormale $\mathcal{N}(\mu_k, \Sigma_k)$.

Sous cette hypothèse, les estimateurs non biaisés utilisés sont :

- $\hat{\mu}_k = g_k$,
- $\hat{\Sigma}_k = \dfrac{n_k}{n_k - 1} W_k$,
- $\hat{\Sigma} = \dfrac{n}{n - q} W_k$,
- $\hat{T} = \dfrac{n}{n - 1} T$
- $\hat{B} = \dfrac{n}{q} T$

Remarque : On n'a plus l'égalité $\hat{T} = \hat{B} + \hat{B}$.

Les hypothèses de loi multinormale et d'homoscédasticité sont à la base des principaux tests statistiques utilisés.

Définition 1 *La métrique $M = \Sigma^{-1}$ engendre la distance de Mahalanobis. On note $\Delta_k^2 = (X_i - g_k)^T \Sigma^{-1} (X_i - g_k)$ la distance de Mahalanobis entre un point et le centre de gravité. Cette distance tient compte de la forme du nuage de points dans le calcul de distance.*

Sous l'hypothèse gaussienne, Δ_k^2 suit un χ^2 à p ddl. On en déduit ainsi la forme de l'ellipsoïde de confiance.

$$(X_i - g_k)^T \Sigma^{-1} (X_i - g_k) = \chi^2_{1-\alpha}(p).$$

Cette distance est estimé par $D_p^2 = (X_i - g_k)^T \hat{\Sigma}^{-1} (X_i - g_k)$.

Exemple :

On étudie une population constituée de 2 classes (T=1 ou 2) de 5 individus chacune caractérisés par deux variables quantitatives X^1 et X^2.

X^1	0	2	4	6	8	5	7	9	11	13
X^2	3	1	5	9	7	2	0	4	8	6
T	1	1	1	1	1	2	2	2	2	2

Déterminer $g, g_1, g_2, T, B, W_1, W_2, W, \hat{W}$

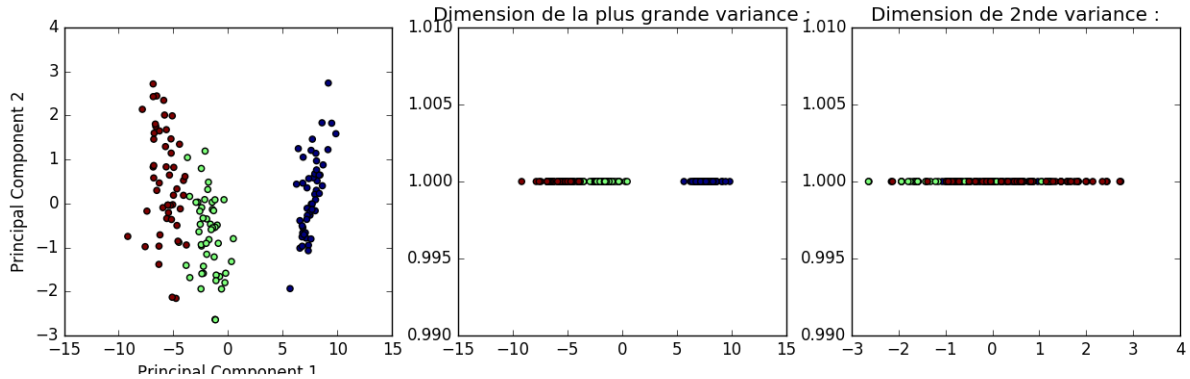
```
X=np.array([[0,3],[2,1],[4,5],[6,9],[8,7],[5,2],[7,0],[9,4],[11,8],[13,6]])
C=np.array([[1,0],[1,0],[1,0],[1,0],[1,0],[0,1],[0,1],[0,1],[0,1],[0,1]])
n=len(C);n
In=np.ones([n,1]);In $\#$ vecteur 1n
g=np.transpose(X).dot(In)/n;g # gravité
X0=X-In.dot(np.transpose(g));X0 # centrée
Dq=np.transpose(C).dot(C)/n;Dq # poids des groupes
nq=np.transpose(C).dot(C);nq # effectif groupes
Xq=np.linalg.inv(nq).dot(np.transpose(C).dot(X0));Xq
# centre des groupes
B=np.transpose(Xq).dot(Dq.dot(Xq));B
T=np.transpose(X0).dot(X0)/10;T
W=T-B;W
Xc=C.dot(Xq);Xc $\#$matrice des espérances
XX=X0-Xc;XX $\#$ Ecart à gk pour calculer Wk
X1c=XX[0:nq[1,1],:];X1c
W1=np.transpose(X1c).dot(X1c)/nq[0,0];W1
X2c=XX[nq[1,1]:(n+1),:];X2c
W2=np.transpose(X2c).dot(X2c)/nq[1,1];W2
```

2.3 Critère d'ajustement

Définition 2 *Une variable discriminante, F_D, est une combinaison linéaire des variables initiales, soit Xa, cette variable avec a un vecteur colonne. a est la fonction linéaire discriminante correspondante (forme linéaire).*

$$F_D = Xa, \quad F_D(i) = \sum_{j=1}^{p} a_j X_i^j.$$

```
lda = LinearDiscriminantAnalysis(n_components=2)
XirisLDA = lda.fit(Xiris, yiris).transform(Xiris)
def graph.acp2(XPC2, y):
    plt.figure(figsize=(15,4))
    plt.subplot(1, 3, 1)
    plt.xlabel('Principal Component 1')
    plt.ylabel('Principal Component 2')
    plt.scatter(XPC2[:, 0], XPC2[:, 1], c=2*y)
    plt.subplot(1, 3, 2)
    plt.title("Dimension de la plus grande variance :")
    plt.scatter(XPC2[:, 0], np.ones(XPC2.shape[0]), c=y)
    plt.subplot(1, 3, 3)
    plt.title("Dimension de 2nde variance :")
    plt.scatter(XPC2[:, 1], np.ones(XPC2.shape[0]), c=y)
    plt.show()
graph.acp2(XirisLDA, Yiris)
```



Proposition 3 *Les variables discriminantes sont centrés et de somme des carrés :*

$$F_D^T D F_D = \sum_{i=1} n\omega_i F_D(i)^2 = a^T T a$$

La somme des carrés totale $a^T T a$ se décompose alors en une somme inter $a^T B a$ et une somme intra $a^T W a$:

$$a^T T a = a^T B a + a^T W a.$$

L'objectif de l'analyse discriminante est de définir de nouvelles variables à partir de combinaisons linéaires des variables initiales et :

- rendre maximale la variance inter-classe $a^T Ba$ et minimales l'inertie intra-classe $a^T Wa$,
- ou de façon équivalente, rendre maximale la variance interclasse $a^T Ba$ par rapport à l'inertie totale $a^T Ta$.

Ces deux critères définissent des variables discriminantes équivalentes.

Proposition 4 *Critère d'ajustement*
Les fonctions discriminantes a sont les fonctions qui permettent de maximiser le quotient $\frac{a^T Ba}{a^T Wa}$ *équivalent à maximiser* $\frac{a^T Ba}{a^T Ta}$.

2.4 Fonctions linéaire discriminantes

Proposition 5 *Le quotient* $\frac{a^T Ba}{a^T Wa}$ *est invariant si a est changé en* λa, $\lambda > 0$. *Maximiser* $\frac{a^T Ba}{a^T Wa}$ *revient à maximiser* $a^T Ba$ *sous la contrainte* $a^T Wa = 1$ *par exemple. Dans la pratique, on utilisera la normalisation de lda (package MASS, R) :*

$$a^T \hat{W} a = 1 \quad \text{équivalent à } a^T W a = \frac{n-q}{n}.$$

Remarque : Dans la littérature, les fonctions discriminantes peuvent être définies différemment suivant :

- le critère utilisé $\frac{a^T Ba}{a^T Wa}$ ou $\frac{a^T Ba}{a^T Ta}$,
- la normalisation de a.

Proposition 6 *Ajustement*
L'ajustement revient à rechercher les valeurs propres non nulles, λ_s *de* $W^{-1}B$ *et ses vecteurs propres* $\hat{W}$ *normés* a_s.

Preuve

Retour à l'exemple : Trouver a et F_D.

$W^{-1}B = \begin{bmatrix} 2.52 & -0.50 \\ -2.17 & 0.43 \end{bmatrix}$

```
WiB=np.linalg.inv(W).dot(B);WiB # matrice W**-1B
vap=np.linalg.eig(WiB)[0];vap # une seule vap non nulle
vep=np.linalg.eig(WiB)[1];vep # a non normé
np.diag(np.transpose(vep).dot(Wc.dot(vep)))**(-0.5)
a=vep.dot(np.diag(np.diag(np.transpose(vep).dot(Wc.dot(vep)))**(-0.5)))
np.transpose(a).dot(Wic.dot(a))
a=a[:,0] # on ne retient qu'une variable discriminante
y=np.array([1,1,1,1,1,2,2,2,2,2])
# avec scikrit learn
lda = LinearDiscriminantAnalysis(n_components=2)
XLDA=lda.fit(X0,y).transform(X0)
lda.scalings_
a
```

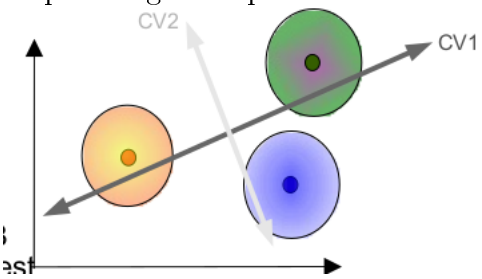
Déterminer les fonctions et variables discriminantes dans l'exemple.

2.5 Nombre de fonctions discriminantes

Proposition 7 *$W^{-1}B$ est une matrice de rang $r = min(p, q-1)$.*

Preuve

Dans le graphique suivant, on dispose de 3 classes soit 2 axes discriminants possibles dans l'espace engendré par les 3 centres de gravité.



Interprétation des valeurs propres : Soit $\lambda_1 \geq \cdots \geq \lambda_r > 0$ les r valeurs propres non nulles :

- si λ est très grand, l'inertie inter est très grandes par rapport à l'inertie intra donc les q nuages sont dans des plans M-orthogonaux
- si $\lambda \approx 0$, les q nuages se projettent en un même point (concentrique) et l'axe n'a aucun intérêt pour la discrimination (axe orthogonal au plan formé par les q centres g_k).

2.6 Equivalence des critères utilisés

Proposition 8 *Equivalence entre les critères d'ajustement*

En notant λ_s et a_s, les valeurs propres et fonctions discriminantes obtenues avec $\frac{a^T Ba}{a^T Wa}$ et μ_s et u_s, les valeurs propres et fonctions discriminantes obtenues avec $\frac{a^T Ba}{a^T Ta}$. On a alors les équivalences :

- $\lambda_s = \frac{\mu_s}{1-\mu_s}$ *et* $\mu_s = \frac{\lambda_s}{1+\lambda_s}$,
- a_s *et* u_s *sont égaux à un facteur près.*

Remarque : Dans les calculs, nous choisirons le critère et la normalisation $t\hat{W}a = 1$ pour normaliser a et ainsi être en adéquation avec la fonction lda de R. La fonction de ade4, discrimin, utilise au contraire le critère $\frac{a^T Ba}{a^T Ta}$ et la normalisation $u^T \hat{T} u = 1$. Les interprétations changent mais le résultat final est le même.

3 AFD et ACP

Soit X le tableau initial décrivant n individus en fonction de p variables quantitatives.

On note C la matrice des indicatrices de l'appartenance aux q groupes. $C = (c_{ik})$, $c_{ik} = 1$ si l'individu i appartient au groupe k, 0 sinon. On note D la matrice diagonale des poids des individus.

Proposition 9 *On a :*

- $g^T = 1_n^T DX$, *on centre alors* $X = X - 1_n G^T$, X **est centrée à partir de là**,
- *le poids de chaque classe est* $D_q = C^T DC$,
- *la matrice* X_q *des barycentres des* q *classes est* $X_q = C^T DX$,
- *on note* $\hat{X}$ *la matrice qui associe à chaque individu le centre de gravité de la classe correspondante, c'est l'espérance conditionnelle de* X_i *sachant sa classe :*
  - $\hat{X} = CX_q = HX$ *avec* $H = C(C^T DC)^{-1} C^T D$ *l'opérateur de projection des variables sur les indicatrices de classes,*
  - *on en déduit que* $X = HX + (I_n - H)X$ *et :*

$$T = X^T DX = X^T (H^T + (I_n - H)^T) D(H + I_n - H)X = \hat{X}^T D\hat{X} + (X - \hat{X})^T D(X - \hat{X}) = B + W = X$$

Proposition 10 *L'AFD de* $(Y|C, D)$ *est l'ACP du triplet* (X_q, W^{-1}, D_q) *équivalent au triplet* (HX, W^{-1}, D).

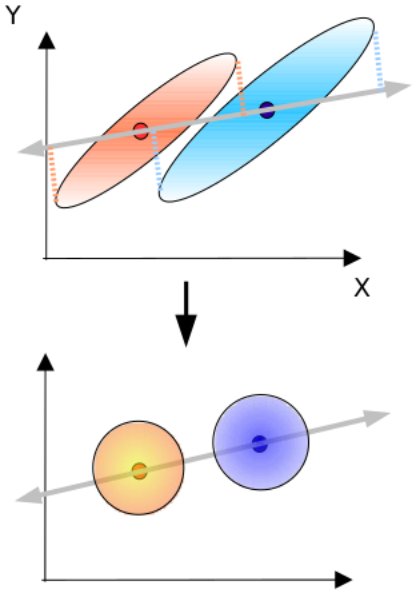
Preuve TD

Remarques :

Réaliser une AFD revient à réaliser l'ACP du nuage des centre de gravité des classes (dimension $q - 1$ maximum) avec la métrique W^{-1}.

On réalise ensuite la projection W^{-1} orthogonale des points sur les axes factorielles ainsi définis. On examine alors la discrimination des individus et des centres de gravité dans les différents espaces ainsi définis.

Géométriquement, l'utilisation de la métrique de Mahalanobis revient à normaliser les matrice de variance covariance des groupes et ainsi de passer de nuages ellipsoides à des nuages sphériques permettant de discriminer au mieux les nuages :



4 Cas particuliers de deux classes : fonction de Fisher

Dans ce cas il n'existe qu'une valeur propre non nulle ($q-1=1$) et donc une seule fonction discriminante. On obtient alors pour B :

Proposition 11 *On considère ici deux groupes. On a alors :*

- $n_1 g_1 + n_2 g_2 = 0$,
- $B = \frac{n_1 \times n_2}{n^2}(g_2 - g_1)(g_2 - g_1)^T$,
- $W^{-1}B$ *admet une unique valeur propre non nulle* $\lambda = \frac{n_1 \times n_2}{n^2} D_p^2$ *avec*

$$D_p^2 = (g_2 - g1)^T \hat{W}^{-1}(g_2 - g_1)$$

l'estimation de la distance de Mahalanobis entre les deux centres de gravité,

- *le facteur discriminant, a, appelée fonction de Fisher, est :*

$$a = \hat{W}^{-1}(g_2 - g_1).$$

Preuve TD

5 Inférence dans le cas de populations suivant une loi multinormale

http ://eric.univ-lyon2.fr/∼ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf
https ://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-m-modmixt5-manova.pdf

5.1 Estimation des matrices de variances

Supposons maintenant que dans chaque groupe, les individus suivent une loi multinormale. Sous l'hypothèse où les lois multinormales ont toute la même matrice de covariance, Σ, un estimateur non biaisé de Σ est donné par $\hat{W}$.

5.2 Pseudo F

Pour une fonction discriminante a donnée, permettant le calcul des coordonnées $F_D = Xa$, il est courant de calculer le pseudo F, $F*$, correspondant au test F d'analyse de variance qui teste l'égalité des moyennes :

$$F* = \frac{\frac{a^T B a}{q-1}}{\frac{a^T W a}{n-q}}.$$

Il permet d'apprécier la qualité de discrimination de l'axe en comparaison des qualités des variables initiales. F* ne suit pas exactement F.

5.3 Egalité des matrices de variances intra groupes

Il est possible de tester l'hypothèse d'égalité des Σ_k, avec le test de Kullback, de Cox, de Bartlett. Cette égalité est nécessaire pour le test de Bartlett et le classement des individus.

5.4 Test de Bartlett (différences entre groupes)

L'AFD repose sur l'existence d'une différence des moyennes μ_k entre classes. L'hypothèse H_0 est alors ""$\mu_1 = \cdots = \mu_q$".

Proposition 12 *Le test de Bartlett permet de tester H_0 : Il repose sur la statistique*

$$[n-1-\frac{p+q}{2}]\sum_{s=1}^{r} ln(1+\lambda_s) = -[n-1-\frac{p+q}{2}]\ln\Lambda$$

qui suit asymptotiquement un χ^2 avec $p(q-1)$ ddl.

On appelle lambda de Wilks la statistique $\Lambda = \frac{|W|}{|B|}$ avec $\Lambda^{-1} = \prod_{s=1}^{r}(1+\lambda_s)$.

Un test analogue permet de déterminer parmi les fonctions discriminantes celles significatives.

Pour tester l'apport des fonctions r' à r, on utilise la statistique

$$[n-1-(p+q)/2]\sum_{s=r'}^{r} ln(1+\lambda_s)$$

qui suit asymptotiquement un χ^2 avec $(p-r'+1)(q-r')$ ddl.

Certaines procédures permettent également de choisir les variables initiales permettant la meilleure discrimination et limitant ainsi le nombre de variables initiales nécessaires.

5.5 Cas de deux groupes : distance de Mahalanobis

Dans le cas de deux groupes, on a le carré de la distance de Mahalanobis qui est :

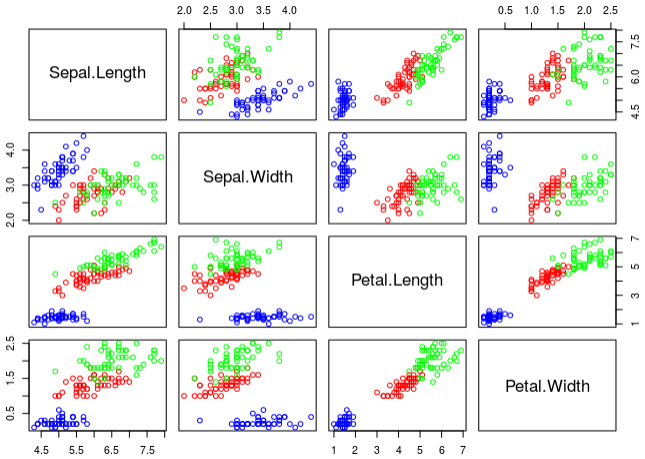
$$\delta_p^2 = (\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1) \text{ estimé par} D_p^2\frac{n_1+n_2-2}{n^2}(g_2-g_1)^T\hat{W}^{-1}(g_2-g_1)$$

Proposition 13 *Sous H_0 : "$\mu_1 = \mu_2$,* $\frac{n_1 n_2}{n^2}\frac{n-p-1}{p(n-2)}D_p^2$ *suit la loi $F(p; n-p-1)$.*

6 Interprétation

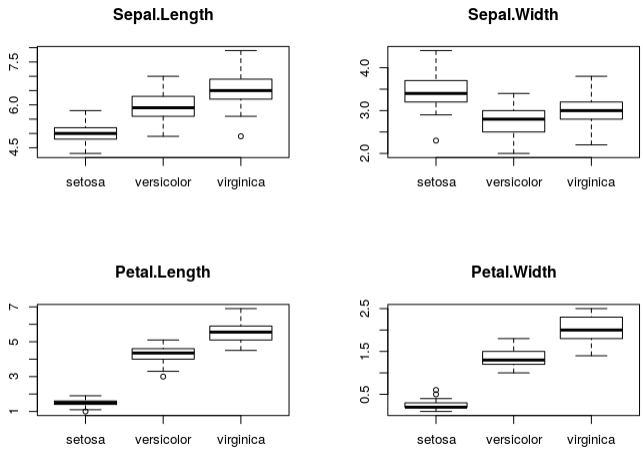
Reprenons l'exemple iris mais avec R, MASS et ade4 :

```
library(ade4)
data(iris)
plot(iris[,1:4],col=c("blue","red","green")[iris$Species])
```

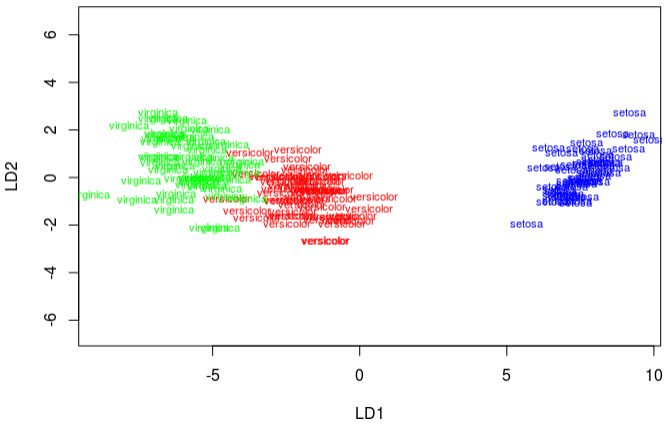


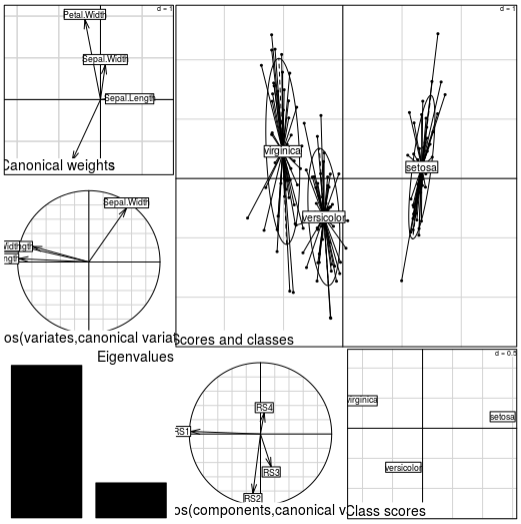
Recherche de la variable la plus discriminante

```
for (i in 1:4) print(anova(lm(iris[,i]~iris$Species)))
Response: iris[, 1]
             Df Sum Sq Mean Sq F value    Pr(>F)
iris$Species   2 63.212  31.606  119.26 < 2.2e-16 ***
Residuals    147 38.956   0.265
Response: iris[, 2]
             Df Sum Sq Mean Sq F value    Pr(>F)
iris$Species   2 11.345  5.6725   49.16 < 2.2e-16 ***
Residuals    147 16.962  0.1154
Response: iris[, 3]
             Df Sum Sq Mean Sq F value    Pr(>F)
iris$Species   2 437.10 218.551  1180.2 < 2.2e-16 ***
Residuals    147  27.22   0.185
Response: iris[, 4]
             Df Sum Sq Mean Sq F value    Pr(>F)
iris$Species   2 80.413  40.207  960.01 < 2.2e-16 ***
Residuals    147  6.157   0.042  \\
par(mfrow=c(2,2))
for (i in 1:4) boxplot(iris[,i]~iris$Species,main=names(iris)[i])
```




```
par(mfrow=c(1,1))
# Existe-t-il une différence entre les centres de gravités?
summary(manova(as.matrix(iris[,1:4])~iris$Species), test="Wilks")
             Df    Wilks approx F num Df den Df    Pr(>F)
iris$Species  2 0.023439   199.15      8    288 < 2.2e-16 ***
Residuals   147
# Analyse discriminante
library(MASS)
iris.lda=lda(iris$Species~.,data=iris[,1:4])
iris.lda$scaling
                    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
names(iris.lda)
[1] "prior"   "counts"  "means"   "scaling" "lev"     "svd"     "N"
"call"    "terms"   [10] "xlevels"
library(ade4)
iris.dis=discrimin(dudi.pca(iris[,1:4],scan=FALSE),iris$Species,scan=FALSE)
names(iris.dis)
[1] "eig"  "nf"   "fa"   "li"   "va"   "cp"   "gc"   "call"\\
#Qualité de l'AFD F*
anova(lm(iris.dis$\$$li[,1]~iris$\$$Species))\\
Response: iris.dis$li[, 1]
             Df  Sum Sq Mean Sq F value    Pr(>F)
iris$Species  2 145.481  72.740  2366.1 < 2.2e-16 ***
Residuals   147   4.519   0.031
plot(iris.lda,col=c("blue","red","green")[iris$\$$Species])
plot(iris.dis)
```


LD2
LD1
6
4
2
0
-2
-4
-6
-5
0
5
10
setosa
versicolor
virginica


Petal.Width
Sepal.Width
Sepal.Length
Canonical weights
virginica
setosa
versicolor
Eigenvalues
Class scores