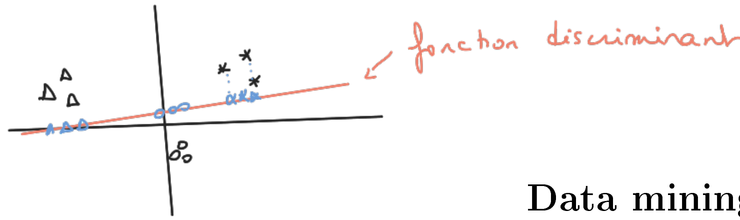


Université d'Angers **Data mining**
master DS1

TD RD4 : Analyse factorielle disciminante

Exercice 1 Cas particuliers de deux classes : fonction de Fisher

On considère un tableau $X_{n\times p}$ constitué de p variables quantitatives centrées (X centrée), n observations réparties en 2 classes. On note n_k et g_k les effectifs et centres de gravités des classes k de 1 à 2.

1. Montrer que :

a. $n_1 g_1 + n_2 g_2 = 0$,

$n\vec{gg_1} + n_2\vec{gg_2} = 0$

$(n_1+n_2)\vec{og} = n_1\vec{og_1} + n_2\vec{og_2}$

Si centrée $g = 0 \Rightarrow n_1 g_1 + n_2 g_2 = 0$

$\Rightarrow g_2 = -\frac{n_1}{n_2} g_1$

b. $B = \frac{n_1 \times n_2}{n^2}(g_2 - g_1)(g_2 - g_1)^T$, on pourra développer et retrouver $B = \frac{1}{n}\sum_i g_{k(i)} g_{k(i)}^T$ (nuage centré). On rappelle que les vecteurs sont en colonne.

$n_1 + n_2 = n$

$$B = \frac{n_1}{n} g_1^T g_1 + \frac{n_2}{n} g_2^T g_2 = \frac{1}{n}\left(n_1 g_1^T g_1 + \frac{n_1^2}{n_2} g_1 g_1^T\right) = \frac{n_1}{\cancel{n}}\frac{\cancel{n}}{n_2} g_1 g_1^T$$

$$\frac{n_1 \times n_2}{n^2}(g_2 - g_1)(g_2 - g_1)^T = \frac{n_1 \times n_2}{n^2}\left(-\frac{n}{n_2} g_1\right)\left(-\frac{n}{n_2} g_1\right)^T = \frac{n_1}{n_2} g_1 g_1^T$$

$$F_D = X . a \qquad \begin{array}{c|c} & 1 \cdots p \\ \hline \begin{array}{c} 1 \\ \vdots \\ n \end{array} & \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \end{array} \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}$$

$a^T B a$ max (inter)

et $a^T W a$ min (intra)

1 seule valeur propre.

2. Justifier que $W^{-1}B$ est de rang 1, qu'en déduit-on ?

inversible + de rang 1 donc $W^{-1}B$ de rang 1

$B = [\ \ ]_{g_r} [\ \ ]_{g_1^T} = [\ \ \ ]$ → chaque colonne colinéaire à g donc B de rang 1.

3. Monter que l'unique valeur propre non nulle est $\lambda = \frac{n_1 \times n_2}{n^2}(g_2 - g1)^T W^{-1}(g_2 - g_1)$. Montrer que $v = W^{-1}(g_2 - g_1)$ est un vecteur propre.

$W^{-1}B\, v = \lambda v$

$W^{-1} \frac{n_1 n_2}{n^2}(g_2 - g_1)(g_2 - g_1)^T\, v = \lambda v.$

$\frac{n_1 n_2}{n^2}(g_2 - g_1)^T\, W^{-1}(g_2 - g_1) \cdot (g_2 - g_1)^T v = \lambda (g_2 - g_1)^T v$

donc $\lambda =$

pour $v = W^{-1}(g_2 - g_1)$

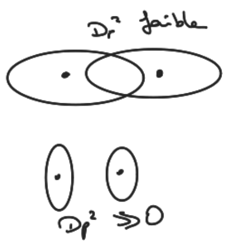
$\frac{n_1 n_2}{n^2} W^{-1}(g_1 - g_2)(g_2 - g_1)^T \boxed{W^{-1}(g_2 - g_1)}$

$= \lambda \boxed{W^{-1}(g_2 - g_1)}$

distance entre les groupes

estimation non biaisé matrice variance-covariance intra-groupe

4. On pose $D_p^2 = (g_2 - g_1)^T \hat{W}^{-1}(g_2 - g_1)$ avec $\hat{W} = \frac{n}{n-2}W$. A quoi correspond $\hat{W}$ et D_p^2?



5. On appelle fonction discriminante de Fisher la fonction $\hat{W}^{-1}(g_2 - g_1)$. Montrer que la fonction discriminante de lda (scaling), notée a, $\hat{W}$ normée, est $a = \frac{1}{D_p}\hat{W}^{-1}(g_2 - g_1)$.

En déduire que le pseudo F est $F* = \frac{n-2}{2-1}\lambda = \frac{n_1 \times n_2}{n} D_p^2$

$\hat{W} = \frac{n}{n-2} W \Rightarrow \hat{W}^{-1} = \frac{n-2}{n} W^{-1}$

$\lambda = \frac{n_1 n_2}{n} D_p^2$

$F^* = \frac{SCM/(2-1)}{SCR/(n-2)} = \frac{n a^T B a/1}{n a^T W a/(n-2)}$
$= (n-2)\lambda = \frac{n_1 n_2}{n} D_p^2.$

Dans lda on a $\hat{v}$ normé $a^T \hat{W} a = 1$ et $W^{-1}Ba = \lambda a$

a et v colinéaires

$v^T W v = (g_2 - g_1)^T W^{-1} W W^{-1} g_2 - g_1 = (g_2 - g_1)^T W^{-1}(g_2 - g_1)$

$$v^T \hat{W}^{-1} v = (g_2 - g_1)^T \hat{W}^{-1} (g_2 - g_1)^T$$ → carré

$$\hat{W} = \frac{n}{n-2} W \qquad \frac{n-2}{n} W^{-1} = \hat{W}^{-1}$$

Exercice 2 Etude de l'exemple charolais/zebu

Les exemples sont tirés des ouvrages de Tomassone (Comment interpréter ?, ITCF, 1988) et Tomassone et al. (Discrimination et classement, Masson, 1988).

Le fichier chazeb est constitué de deux populations (charolais :cha et zebu :zeb) sur lesquelles sont mesurées 6 poids en kg : vif, carcasse, première qualité, viande totale, gras, os. variables quantitatives.

```
library(ade4)
data(chazeb);chazeb
tab=chazeb$tab
cla=chazeb$cla
```

regroupe+ des viandes par groupes (zebra et charolais)

1. Interpréter les résultats suivants.

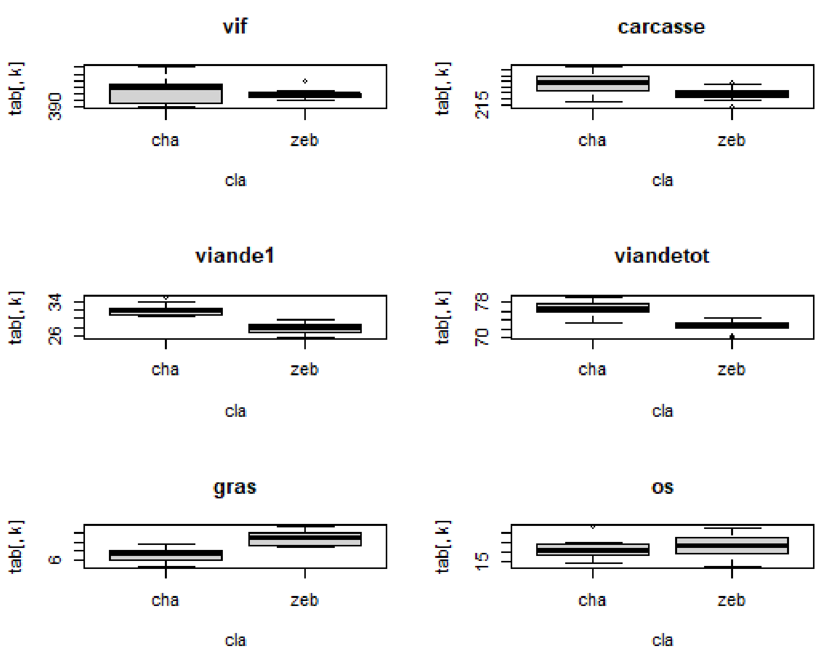
```
> for (k in 1:6) print(c(names(tab)[k],anova(lm(tab[,k]~cla))$F[1]))
[1] "vif"                "0.739376448881091"
[1] "carcasse"           "7.57276918762758"
[1] "viande1"            "59.5127780781559"
[1] "viandetot"          "47.1496369693344"
[1] "gras"               "28.8594870550728"
[1] "os"                 "0.243615358592007"
> summary(manova(as.matrix(tab)~cla),test="Wilks")
           Df   Wilks approx F num Df den Df   Pr(>F)
cla         1 0.15492   14.547      6     16 1.11e-05 ***
Residuals  21
```

→ le plus grand F donc c'est la variable le plus discriminant.

m pour multiple

différence significative entre cha et zeb.

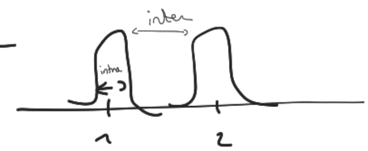
pas de valeurs communs



2. Comment sont déterminées les variables discriminantes ? Combien de variables discriminantes sont attendues ? 1 seul car on a 2 groupes

$F_0 = Xa = \sum a_j X^j$

On veut (l'inertie inter la plus grande possible
L'inertie intra la plus petit possible



3. Interpréter les coefficients ci-dessous donnés par lda du package MASS. La normalisation est faite avec l'inertie intra.

linear discriminant analysize

```
> afd1=lda(cla~., tab)
> round(afd1$scaling,2)
            LD1
vif        -0.07
carcasse   -0.04
viande1    -0.72
viandetot  -0.79
gras       -0.62
os         -0.49
```

→ on le fait avec tout les variables

→ F_0 dans le cours

4. Indiquer en quoi les résultats suivants nous informent sur l'intérêt de cette variable discriminante.

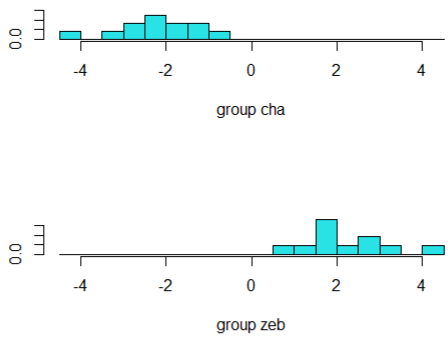
```
> afd1$svd # racine carrée de la valeur propre
[1] 10.7031
> round(afd1$svd^2,2)
[1] 114.56
> anova(lm(as.matrix(tab)%*%afd1$scaling~cla))
Analysis of Variance Table

Response: as.matrix(tab) %*% afd1$scaling
          Df Sum Sq Mean Sq F value    Pr(>F)
cla        1 114.56  114.56  114.56 5.819e-10 ***
Residuals 21  21.00    1.00
```

F^*

la valeur max la valeur initiale était 59 et là c'est 114 donc oui on a amélioré la situation

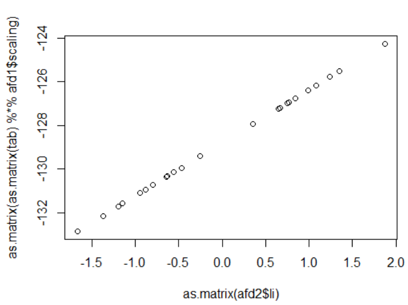
5. Interpréter le graphique ci-dessous. Comment utiliser ce résultat pour une classification supervisée ?

```
>plot(afd1)
```



6. Interpréter les coefficients ci-dessous donnés par discrimin du package ade4. La normalisation est faite avec l'inertie totale.

```
> afd2=discrimin(dudi.pca(tab,scan=FALSE),cla,scan=FALSE)
> names(afd2)
[1] "eig"  "nf"   "fa"   "li"   "va"   "cp"   "gc"   "call"
> afd2$fa
                 DS1
vif        -0.2012722
carcasse   -0.1319587
viande1    -0.7467706
viandetot  -0.7878439
gras       -0.6038732
os         -0.2212261
>plot(as.matrix(afd2$li),as.matrix(as.matrix(tab)%*%afd1$scaling)
)
```



7. Interpréter les résultats suivants.

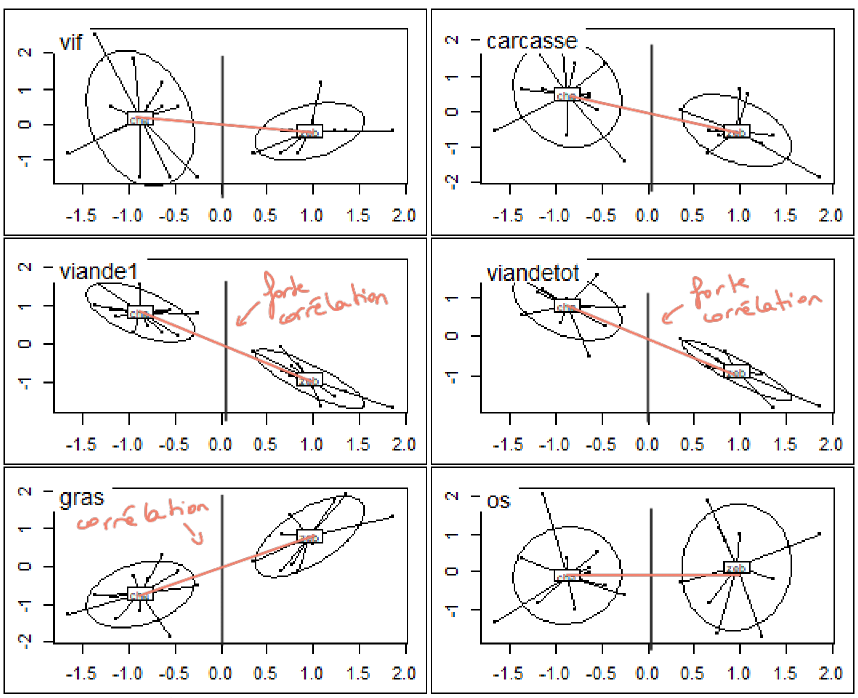
```
> anova(lm(as.matrix(afd2$li)~cla))
Analysis of Variance Table

Response: as.matrix(afd2$li)
          Df  Sum Sq  Mean Sq  F value     Pr(>F)
cla        1 19.4369  19.4369  114.56  5.819e-10 ***
Residuals 21  3.5631   0.1697
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
> afd2$eig;1-afd2$eig;afd2$eig/(1-afd2$eig)
[1] 0.845083
[1] 0.154917
[1] 5.455068
```

la m̃ valeur F* que le précedent.

8. Interpréter le graphique suivant.

```
plot(afd2)
```

la variable discriminante discrime parfaitement



→ 2 régression négative

↑ zebu plus de gras que cha
charolais plus de viande que zeb

9. Interpréter les corrélations entre la variable discriminantes et les variables initiales (va) et les composantes principales (cp). Comment retrouver ces coefficients ?

```
> afd2$va
                   DS1
vif        -0.2006132
carcasse   -0.5600174
viande1    -0.9352399 ) corrélation ⊖
viandetot  -0.9048107 )
gras        0.8276005 → corréla° ⊕
os          0.1164899
> afd2$cp
            CS1
RS1  0.94954572
RS2  0.20780580
RS3  0.07505468
RS4 -0.17661015
RS5  0.07905467
RS6 -0.11002585
```

10. li représente la variable discriminante et gc sa moyenne par classe. Comparer avec le résultat de lda.

```
> afd2$li[1:5,]
[1] -1.6702012 -0.6426701 -0.6313576 -0.4690064 -0.2533122
> afd2$gc
           DS1
cha -0.8801474
zeb  0.9601608
> afd1$means
         vif carcasse  viande1 viandetot      gras       os
cha 402.5000 233.0000 31.99167  76.60000  7.258333 16.30833
zeb 399.7273 224.2727 27.66364  72.56364 10.845455 16.54545
```

parenthèse

X = iris [, 1:4]

y = iris [, 5]

plot (X[, 1:2], col = y)

3:4 c'est mieux → plus discriminant.

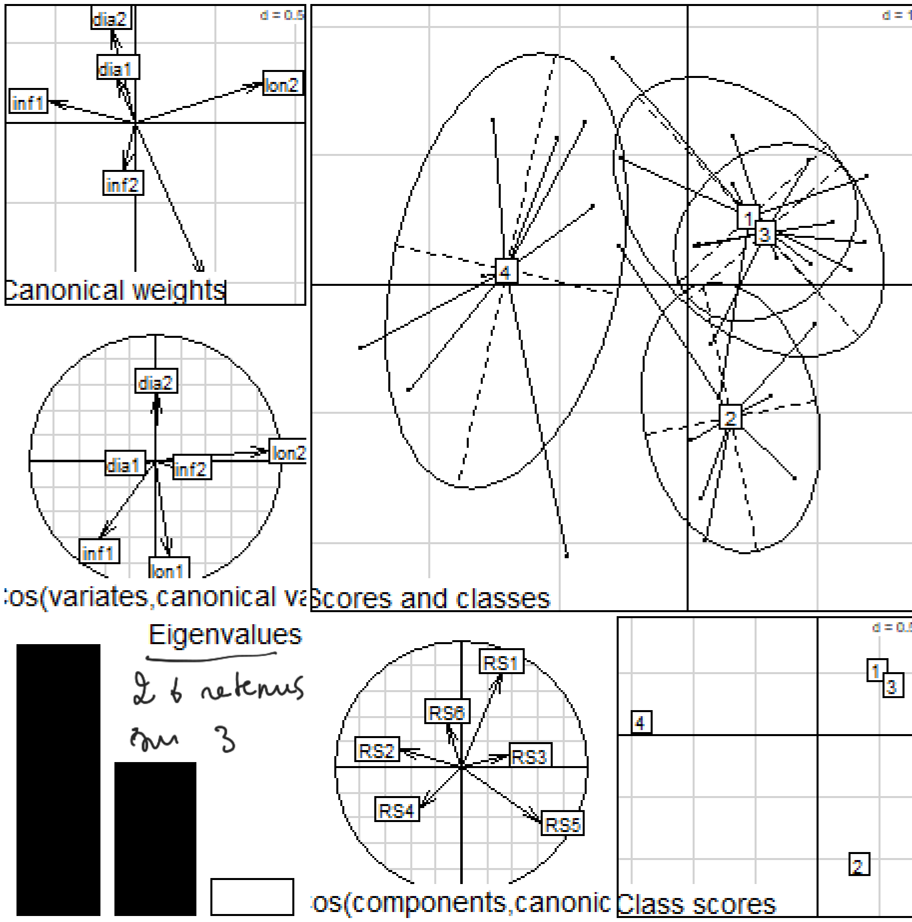
plot (X, col = y) → donne toutes les combinaisons possibles

afd1 = lda (y ~ ., data = X)

plot (afd1) → 4 axes retenus (max = 4

afd2 = discrimin (dudi.pca(X), y) → 2 fonctions retenu (max = 2 car 3 grps) plot (afd2)

11. En présence de plus de deux classes, ade4 propose le graphique bilan suivant. Interpréter le :



point moyen de chaque population

Iris