Fiche TD avec le logiciel : tdr63

Analyse discriminante linéaire

A.B. Dufour, D. Chessel & J.R. Lobry

La fiche introduit à l'usage de la MANOVA et de l'analyse discriminante. Les iris de Fisher. La variable discriminante. Comparaison de `lda` dans `MASS` et `discrimin` dans `ade4`. Tests classiques de signification. Discrimination descriptive et discrimination prédictive.

Table des matières

1 Les Iris de Fisher

C'est un des jeux de données les plus célèbres de la statistique (Anderson [1935], Fisher [1936]).

– On lit le data frame `iris`.

```
data(iris)
names(iris)
```

```
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

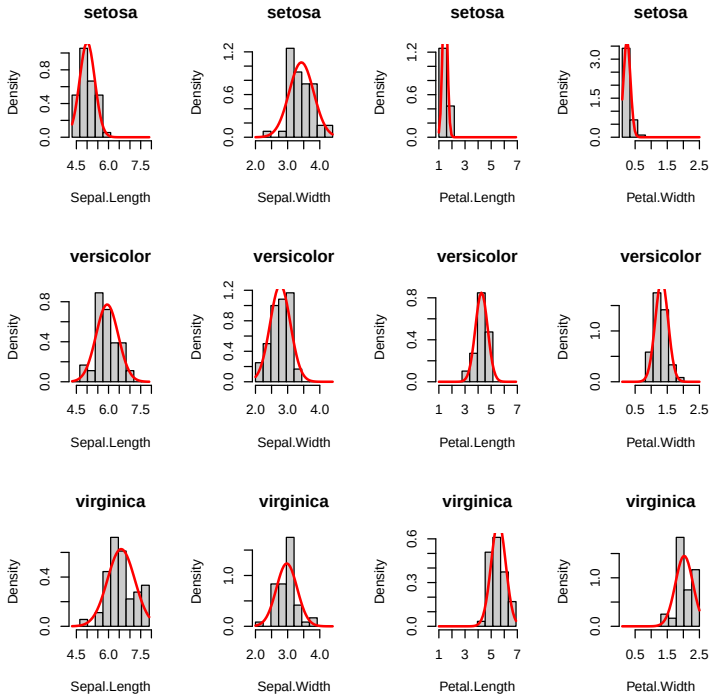
```
dim(iris)
```

```
[1] 150   5
```

setosa versicolor virginica

[1]

– Approche univariée : on construit les histogrammes par espèce et par variable :




```
par(mfcol = c(3, 4))
for (k in 1:4) {
    j0 <- names(iris)[k]
    br0 <- seq(min(iris[, k]), max(iris[, k]), le = 11)
    x0 <- seq(min(iris[, k]), max(iris[, k]), le = 50)
    for (i in 1:3) {
        i0 <- levels(iris$Species)[i]
```

[1] http://cs-people.bu.edu/mdassaro/pp3/

```
        x <- iris[iris$Species == i0, j0]
        hist(x, br = br0, proba = T, col = grey(0.8), main = i0,
            xlab = j0)
        lines(x0, dnorm(x0, mean(x), sd(x)), col = "red", lwd = 2)
    }
}
```

Noter qu'on peut retrouver ces histogrammes sur le site :

http://www.uib.no/med/avd/miapr/arvid/MOD3_2002/Moenstergjenkjenning/iris_histograms.gif

- Approche bivariée : on réalise les analyses de la variance à un facteur, variable par variable.
  On prend par exemple la recherche d'une relation entre la longueur des sépales et les espèces.
  - les moyennes par groupe

```
tapply(iris$Sepal.Length, iris$Species, mean)
```

```
    setosa versicolor  virginica
     5.006      5.936      6.588
```

  - les écarts-type par groupe

```
tapply(iris$Sepal.Length, iris$Species, sd)
```

```
    setosa versicolor  virginica
0.3524897  0.5161711  0.6358796
```

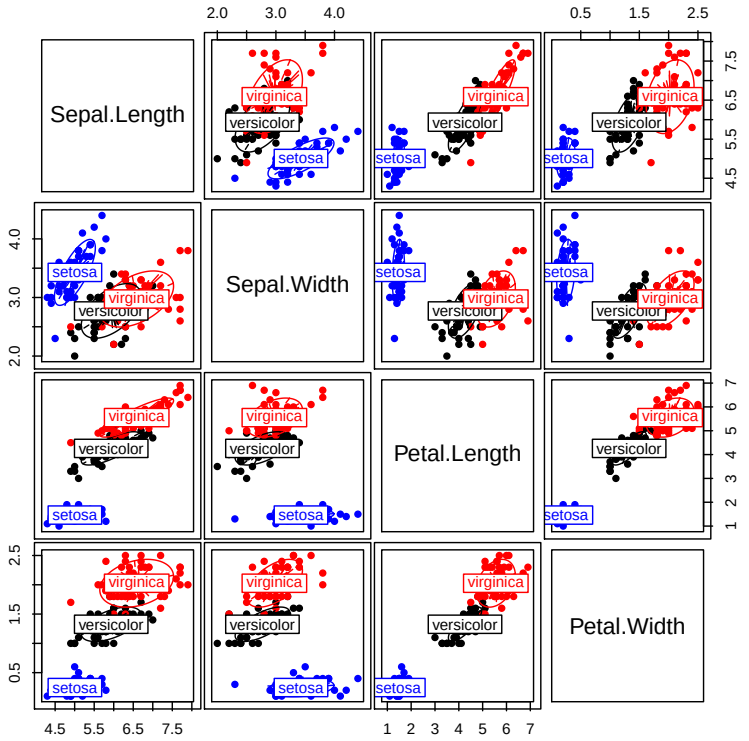
  - l'analyse de la variance à un facteur

```
options(show.signif.stars = FALSE)
anova(lm(iris$Sepal.Length ~ iris$Species))
```

```
Analysis of Variance Table
Response: iris$Sepal.Length
              Df Sum Sq Mean Sq F value    Pr(>F)
iris$Species   2 63.212  31.606  119.26 < 2.2e-16
Residuals    147 38.956   0.265
```

- Approche bivariée : on représente tous les nuages bivariés.

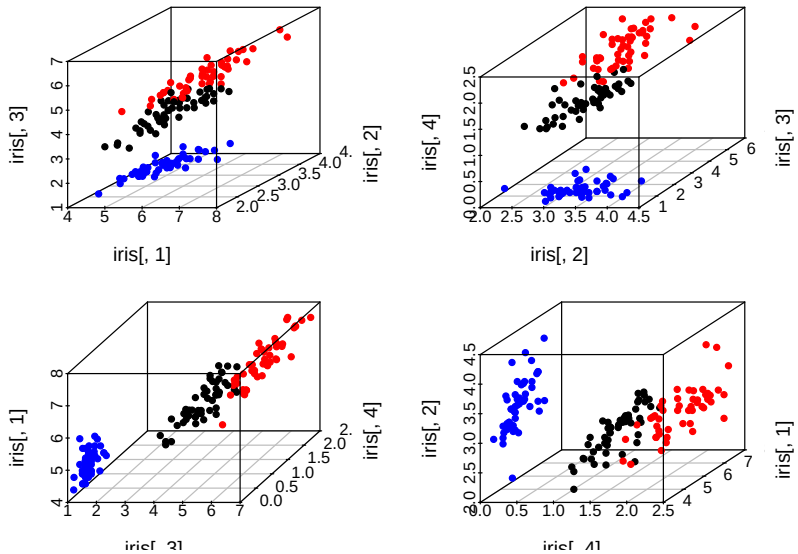
```
library(ade4)
par(mar = c(0, 0, 0, 0))
pan1 <- function(x, y, ...) {
    xy <- cbind.data.frame(x, y)
    s.class(xy, iris$Species, include.ori = F, add.p = T, clab = 1.5,
        col = c("blue", "black", "red"), cpoi = 2, csta = 0.5)
}
pairs(iris[, 1:4], panel = pan1)
```



La valeur discriminante d'un plan varie fortement en dimension 4. Le problème est donc clairement posé. **Une** mesure varie-t'elle entre espèces (*i.e.* entre groupes) ? C'est une question d'histogramme et de modèle linéaire (analyse de la variance à un facteur). Pour **deux** mesures, on a la même question et plus de configurations possibles. Pour ***p*** mesures, la question est vaste.

– Approche en dimension 3 : on peut encore visualiser en trois dimensions.

```
library(scatterplot3d)
par(mfrow = c(2, 2))
mar0 = c(2, 3, 2, 3)
scatterplot3d(iris[, 1], iris[, 2], iris[, 3], mar = mar0, color = c("blue",
    "black", "red")[iris$Species], pch = 19)
scatterplot3d(iris[, 2], iris[, 3], iris[, 4], mar = mar0, color = c("blue",
    "black", "red")[iris$Species], pch = 19)
scatterplot3d(iris[, 3], iris[, 4], iris[, 1], mar = mar0, color = c("blue",
    "black", "red")[iris$Species], pch = 19)
scatterplot3d(iris[, 4], iris[, 1], iris[, 2], mar = mar0, color = c("blue",
    "black", "red")[iris$Species], pch = 19)
```

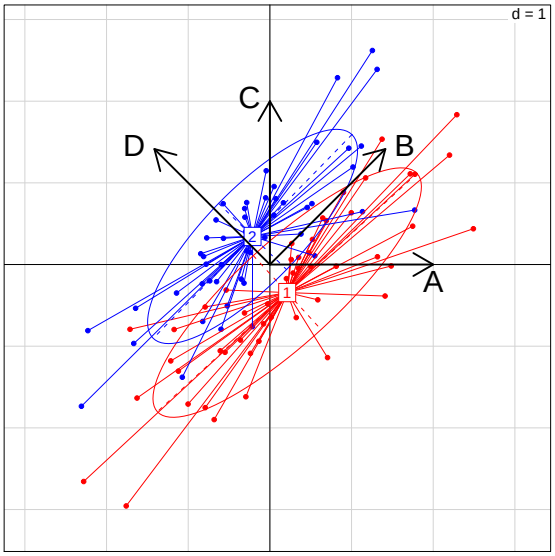


Chercher à mesurer ce qui sépare des groupes connus est ce qu'on appelle discriminer. En statistique, la discrimination n'est pas positive ou négative, elle est ou elle n'est pas ! Pourquoi discriminer ? C'est essentiellement pour affecter un nouvel individu dont on ne connaît pas le groupe mais uniquement les mesures qui ont généré la discrimination.
On dit alors qu'on a un problème de **discrimination descriptive** quand la question est : qu'est-ce qui sépare les groupes ?
et un problème de **discrimination prédictive** quand la question est : à quel groupe est-ce que je peux affecter un nouvel individu et avec quel type d'erreur ?
Une part importante de la statistique a été consacrée à ce thème et la littérature est considérable. Il y avait déjà 400 pages de références en 1973 dans Cacoullos and Styan [1973].

2 La variable discriminante

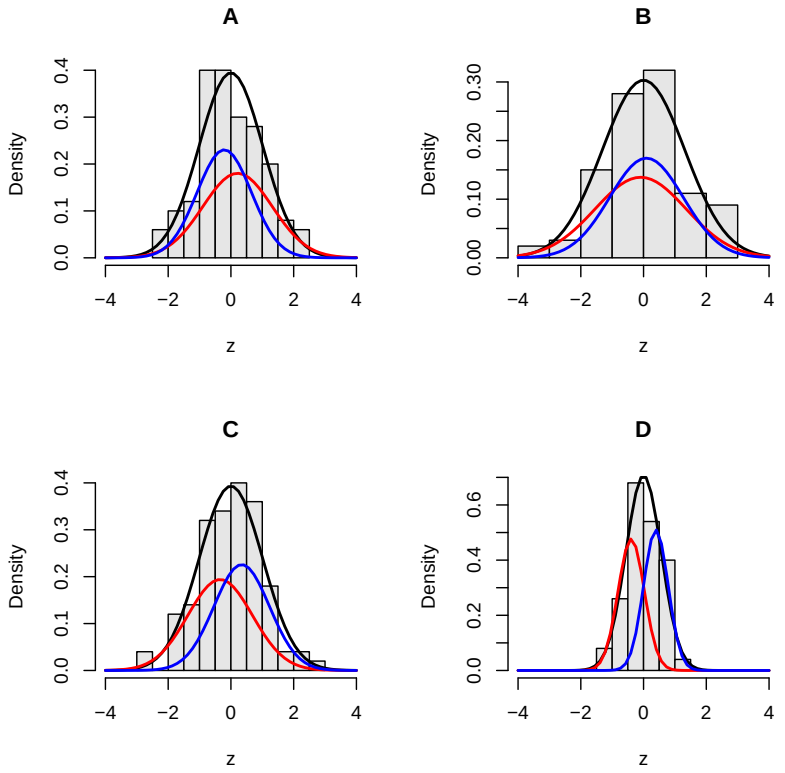
On comprend rapidement le problème avec la figure :

```
set.seed(24122006)
library(MASS)
library(ade4)
s <- matrix(c(1, 0.8, 0.8, 1), 2)
x1 <- mvrnorm(50, c(0.3, -0.3), s)
x2 <- mvrnorm(50, c(-0.3, 0.3), s)
x <- rbind.data.frame(x1, x2)
x <- scalewt(x, scale = F)
fac <- factor(rep(1:2, rep(50, 2)))
s.class(x, fac, col = c("red", "blue"))
arrows(0, 0, 2, 0, lwd = 2)
text(2, 0, "A", pos = 1, cex = 2)
arrows(0, 0, sqrt(2), sqrt(2), lwd = 2)
text(sqrt(2), sqrt(2), "B", pos = 4, cex = 2)
arrows(0, 0, 0, 2, lwd = 2)
text(0, 2, "C", pos = 2, cex = 2)
arrows(0, 0, -sqrt(2), sqrt(2), lwd = 2)
text(-sqrt(2), sqrt(2), "D", pos = 2, cex = 2)
```



Deux populations sont définies par deux variables. On représente le plan des variables centrées. Dans ce plan, choisir une direction et projeter les points sur l'axe ainsi défini, c'est faire une combinaison linéaire des variables centrées. Suivant la direction, les deux populations apparaîtront un peu, beaucoup ou pas du tout différentes.

```
par(mfrow = c(2, 2))
f1 <- function(a, b, cha) {
    z <- a * x[, 1] + b * x[, 2]
    z0 <- seq(-4, 4, le = 50)
    z1 <- z[fac == 1]
    z2 <- z[fac == 2]
    hist(z, proba = TRUE, col = grey(0.9), xlim = c(-4, 4), main = cha)
    lines(z0, dnorm(z0, mean(z), sd(z)), lwd = 2)
    lines(z0, 0.5 * dnorm(z0, mean(z1), sd(z1)), col = "red", lwd = 2)
    lines(z0, 0.5 * dnorm(z0, mean(z2), sd(z2)), col = "blue", lwd = 2)
}
f1(1, 0, "A")
f1(1/sqrt(2), 1/sqrt(2), "B")
f1(0, 1, "C")
f1(-1/sqrt(2), 1/sqrt(2), "D")
```



Cette figure indique un autre élément essentiel. Entre les positions, la capacité discriminante varie mais varie également la ... variabilité. On peut faire beaucoup de variance (B), beaucoup de variance inter-classe ou beaucoup de variance inter-classe en proportion (D). On peut discriminer avec peu de variabilité : ce qui est alors d'abord une variabilité inter-classe). On peut faire beaucoup de variance - c'est le propre de l'ACP - sans rien discriminer.

```
f2 <- function(a, b) {
    z <- a * x[, 1] + b * x[, 2]
    a1 <- var(z) * 99/100
    a2 <- var(predict(lm(z ~ fac))) * 99/100
    a3 <- a2/a1
    round(c(a1, a2, a3), 3)
}
f2(1, 0)
```

```
[1] 1.012 0.045 0.044
```

```
f2(1/sqrt(2), 1/sqrt(2))
```

```
[1] 1.716 0.009 0.005
```

```
f2(0, 1)
```

```
[1] 1.019 0.118 0.116
```

```
f2(-1/sqrt(2), 1/sqrt(2))
```

```
[1] 0.315 0.154 0.490
```

La première quantité est la variance descriptive totale. La deuxième quantité est la variance intergroupe. La dernière est le rapport de la variance intergroupe sur la variance totale.

La première quantité ne peut pas excéder la première valeur propre de l'ACP centrée et ne peut pas être moindre que la seconde valeur propre.

```
w <- dudi.pca(x, scal = F, scannf = F)
w$eig
```

```
[1] 1.7161192 0.3150369
```

```
w$c1
```

```
         CS1        CS2
V1 0.7054885 -0.7087214
V2 0.7087214  0.7054885
```

La position B est presque celle qui maximise la variance (elle est en fait l'axe principal théorique) : toute la variabilité est intra-population.
La position D est presque celle qui minimise la variance (elle est en fait le second axe principal théorique). On y trouve quatre fois moins de variance mais la variabilité y est pour moitié inter-groupe.

L'**analyse inter-classe** cherche les axes sur la base de la variabilité inter-classe :

```
wbet <- between(w, fac, scannf = F)
wbet$eig
```

```
[1] 0.1630112
```

```
wbet$c1
```

```
          CS1
V1 -0.5250083
V2  0.8510971
```

La position D est presque celle qui maximise la variance inter-classe (elle est en fait l'axe inter-classe théorique) : on ne peut pas trouver plus de variance inter-groupe.

L'**analyse discriminante** est celle qui maximisera le rapport :

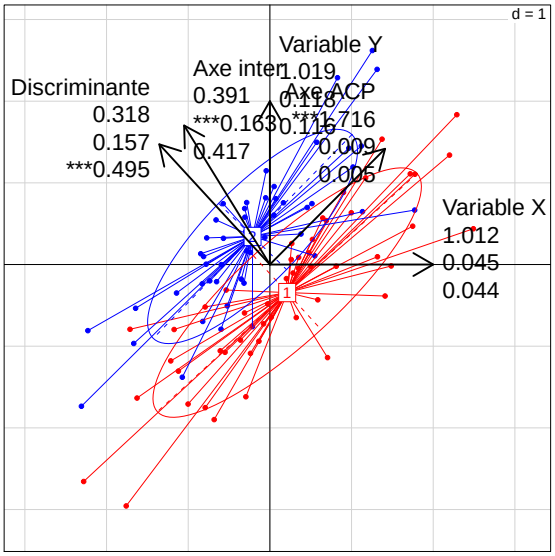
```
wdis <- discrimin(w, fac, scannf = F)
wdis$fa/sqrt(sum(wdis$fa^2))
```

```
          DS1
V1 -0.6770657
V2  0.7359226
```

```
wdis$eig
```

```
[1] 0.4945397
```

On résumera alors la situation par la figure :



On y voit les variations de trois quantités en fonction de la direction du plan.

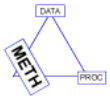
1. La première est la variance de la coordonnée de la projection sur l'axe. C'est la variance des variables quand on est sur les axes. Elle est maximisée (***) avec le premier axe principal et ne peut dépasser la première valeur propre de l'ACP.
2. La seconde est la variance inter-classe ou expliquée par le facteur groupe. Elle peut varier fortement d'une variable à l'autre. Elle est optimisée (***) avec le premier axe de l'ACP inter-classe (ici, celui qui relie les centres de gravité). Elle ne peut dépasser la première valeur propre de l'ACP inter-classe.
3. La troisième est le rapport de la seconde sur la première. Elle est optimisée (***) par l'analyse discriminante. Elle ne peut dépasser la première valeur propre de l'analyse discriminante.

Quand le nombre de dimensions augmente, la diversité des situations possibles augmente et les difficultés d'interprétation apparaissent. Il est possible de très bien discriminer avec très peu de variabilité et la prudence impose l'usage des tests de signification.

On commence toujours par caractériser la valeur discriminante de chacune des variables :

```
apply(iris[, 1:4], 2, function(x) summary(lm(x ~ iris[, 5])))
```

Le rôle des variables est ici trop explicite pour que la partie multivariée soit un apport décisif. Mais l'exemple a valeur d'illustration. On suppose, dans cet exposé, que la méthode ACP est bien connue. L'ACP inter-classe est l'ACP du nuage des centres de gravité (ici il y a deux points et c'est vite fait). Le calcul de la variable discriminante est nouveau.



Soit $\mathbf{X}$ le tableau de données (n individus et p variables) et $\mathbf{q}$ la variable qualitative qui définit les groupes. $\mathbf{X}^k$ est la $k^{\text{ième}}$ colonne de $\mathbf{X}$.

$\mathbf{D}$ est la diagonale des poids des points. On peut prendre $\mathbf{D} = (1/n)\mathbf{I}_n$ mais cela ne simplifie rien. $m_k = \langle \mathbf{X}^k | \mathbf{1}_n \rangle_{\mathbf{D}}$ est la moyenne de $\mathbf{X}^k$.

$\mathbf{X}_0^k = \mathbf{X}^k - m_k \mathbf{1}_n$ est la variable centrée ; $v_k = \|\mathbf{X}_0^k\|_{\mathbf{D}}^2$ est la variance de $\mathbf{X}^k$; $c_{jk} = \langle \mathbf{X}_0^j | \mathbf{X}_0^k \rangle_{\mathbf{D}}$ est la covariance entre les variables j et k.

$\mathbf{X}_0$ est le tableau des variables centrées et $\mathbf{C} = [c_{jk}] = \mathbf{X}_0^T \mathbf{D} \mathbf{X}_0$ est la matrice des covariances de $\mathbf{X}$.

La variable $\mathbf{q}$ a g modalités et définit donc g groupes. $\widehat{\mathbf{X}_0^k}$ est la projection de $\mathbf{X}_0^k$ sur le sous-espace engendré par les indicatrices des classes de $\mathbf{q}$: c'est la variable centrée où chaque composante est remplacée par la moyenne des composantes de la même classe.

$b_k = \|\widehat{\mathbf{X}_0^k}\|_{\mathbf{D}}^2$ est la variance inter-classe de $\mathbf{X}^k$: c'est la variance des moyennes par classe pondérée par les poids des classes.
$w_k = \|\mathbf{X}_0^k - \widehat{\mathbf{X}_0^k}\|_{\mathbf{D}}^2$ est la variance intra-classe de $\mathbf{X}^k$: c'est la moyenne des variances dans chaque classe pondérée par les poids des classes.

$$v_k = \left\|\mathbf{X}_0^k\right\|_{\mathbf{D}}^2 = \|\widehat{\mathbf{X}_0^k}\|_{\mathbf{D}}^2 + \|\mathbf{X}_0^k - \widehat{\mathbf{X}_0^k}\|_{\mathbf{D}}^2 = b_k + w_k$$

est la décomposition de la variance totale en variance inter-classe et variance intra-classe par le théorème de Pythagore.

On montre que ce résultat s'étend à deux variables par :

$$c_{jk} = \langle \mathbf{X}_0^j | \mathbf{X}_0^k \rangle_{\mathbf{D}} = \langle \hat{\mathbf{X}}_0^j | \hat{\mathbf{X}}_0^k \rangle_{\mathbf{D}} + \langle \mathbf{X}_0^j - \hat{\mathbf{X}}_0^j | \mathbf{X}_0^k - \hat{\mathbf{X}}_0^k \rangle_{\mathbf{D}} = b_{jk} + w_{jk}$$

b_{jk} est la covariance inter-classe des variables $\mathbf{X}^j$ et $\mathbf{X}^k$: c'est la covariance des moyennes par classe pondérée par les poids des classes.
w_{jk} est la covariance intra-classe des variables $\mathbf{X}^j$ et $\mathbf{X}^k$: c'est la moyenne des covariances dans chaque classe pondérée par les poids des classes.

On en déduit que :

$$\mathbf{C} = [c_{jk}] = \mathbf{X}_0^T \mathbf{D} \mathbf{X}_0 = [b_{jk}] + [w_{jk}] = \mathbf{B} + \mathbf{W}$$

L'équation d'analyse de la variance s'étend à une matrice de covariances. On obtient directement ce résultat en utilisant le projecteur sur les indicatrices de classes. On note $\mathbf{H}$, le tableau disjonctif complet associé à $\mathbf{q}$:

$$\mathbf{P} = \mathbf{H} \left(\mathbf{H}^T \mathbf{D} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{D}$$

Avec $\mathbf{Q} = \mathbf{I}_n - \mathbf{P}$:

$$\mathbf{C} = [c_{jk}] = \mathbf{X}_0^T \mathbf{D} \mathbf{X}_0 = \mathbf{X}_0^T \left(\mathbf{P}^T + \mathbf{Q}^T\right) \mathbf{D} \left(\mathbf{P} + \mathbf{Q}\right) \mathbf{X}_0 = \mathbf{B} + \mathbf{W}$$

La variable $\mathbf{X}^k$ a une valeur discriminante qui se mesure par :

$$\eta^2_{(\mathbf{X}^k,\mathbf{q})} = \frac{b_k}{v_k} = \frac{b_k}{b_k + w_k} = \frac{1}{1 + \frac{w_k}{b_k}} = \frac{v_k - w_k}{v_k} = 1 - \frac{w_k}{v_k}$$

C'est un jeu d'écriture simple qui a une grande importance pour comparer les programmes car on peut parler de grande valeur discriminante par *une grande valeur de inter/totale* ou *une petite valeur de intra/inter* ou *une grande valeur de inter/intra* ou *une petite valeur de intra/totale.*

Une combinaison linéaire des variables de départ s'écrit $\mathbf{y} = \mathbf{X}_0\mathbf{a}$. Elle est centrée par combinaison linéaire de variables centrées et sa propre valeur discriminante dérive de :

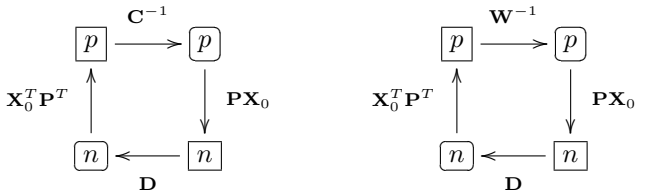
$$v_\mathbf{y} = \mathbf{a}^T\mathbf{X}_0^T\mathbf{D}\mathbf{X}_0\mathbf{a} = \mathbf{a}^T\mathbf{C}\mathbf{a} = \mathbf{a}^T(\mathbf{B} + \mathbf{W})\mathbf{a} = b_\mathbf{y} + w_\mathbf{y}$$

La **variable discriminante** est, si elle existe, celle qui maximisera :
- $b_\mathbf{y}/v_\mathbf{y}$ ou,
- $b_\mathbf{y}/w_\mathbf{y}$ ou,
- $b_\mathbf{y}$ sous la contrainte $v_\mathbf{y} = 1$ ou,
- $b_\mathbf{y}$ sous la contrainte $w_\mathbf{y} = 1$.

La solution, si elle existe, est celle d'un des deux problèmes :

$$\mathbf{a}^T\mathbf{C}\mathbf{a} = 1 \quad \mathbf{a}^T\mathbf{B}\mathbf{a} = \mathbf{a}^T\mathbf{X}_0^T\mathbf{P}^T\mathbf{D}\mathbf{P}\mathbf{X}_0\mathbf{a} \text{ Maximum}$$
$$\mathbf{a}^T\mathbf{W}\mathbf{a} = 1 \quad \mathbf{a}^T\mathbf{B}\mathbf{a} = \mathbf{a}^T\mathbf{X}_0^T\mathbf{P}^T\mathbf{D}\mathbf{X}_0\mathbf{P}\mathbf{a} \text{ Maximum}$$

Cette solution existe comme premier facteur d'un des deux schémas :



Le premier schéma diagonalise $\mathbf{B}\mathbf{C}^{-1}$ et le second diagonalise $\mathbf{B}\mathbf{W}^{-1}$.

3 Comparaison de deux fonctions

1. Utiliser **lda** et consulter la documentation (dans `MASS`) :

```
lda1 <- lda(as.matrix(iris[, 1:4]), iris$Species)
lda1
```

```
Call:
lda(as.matrix(iris[, 1:4]), grouping = iris$Species)
Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

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
```

2. Utiliser **discrimin** et consulter la documentation (dans `ade4`) :

```
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: discrimin
$nf (axis saved) : 2

eigen values: 0.9699 0.222

  data.frame nrow ncol content
1 $fa        4    2    loadings / canonical weights
2 $li        150  2    canonical scores
3 $va        4    2    cos(variables, canonical scores)
4 $cp        4    2    cos(components, canonical scores)
5 $gc        3    2    class scores
```

Sans en avoir l'air, les deux fonctions sont cohérentes. La première donne une combinaison linéaire de variables de départ avec les coefficients qui sont dans la colonne `LD1`. Calculer cette combinaison :

```
w1 <- as.vector(as.matrix(iris[, 1:4]) %*% lda1$scaling[, 1])
w1[1:10]
```

```
 [1] 5.956693 5.023581 5.384722 4.708094 6.027203 5.596840 5.107511 5.500187 4.455445
[10] 5.237953
```

```
w1[141:150]
```

```
[1] -8.758193 -7.210666 -7.612586 -8.901126 -8.952466 -7.750110 -7.284671 -7.072847
[9] -7.991252 -6.788261
```

La seconde donne une combinaison linéaire de variables normalisées (en $1/n$) avec les coefficients qui sont dans la composante `fa` (`fa` pour facteur, dans le vocabulaire du schéma de dualité). Calculer cette combinaison et comparer :

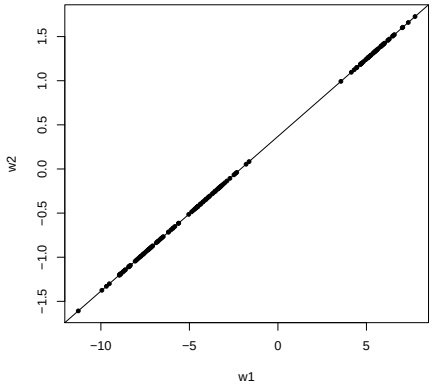
```
w2 <- as.vector(scalewt(iris[, 1:4]) %*% dis1$fa[, 1])
w2[1:10]
```

```
 [1] 1.413522 1.249914 1.313235 1.194598 1.425885 1.350427 1.264630 1.333481 1.150300
[10] 1.287502
```

```
w2[141:150]
```

```
[1] -1.1665246 -0.8951876 -0.9656587 -1.1915858 -1.2005876 -0.9897715 -0.9081633
[8] -0.8710231 -1.0320523 -0.8211248
```

```
plot(w1, w2, pch = 20)
abline(lm(w2 ~ w1))
```



La seconde donne une combinaison linéaire de variance totale 1 (en $1/n$) :

```
var(w2) * 149/150
```
```
[1] 1
```

qui maximise la variance inter-classe (la première valeur propre) :

```
dis1$eig
```
```
[1] 0.9698722 0.2220266
```
```
summary(lm(w2 ~ iris[, 5]))$r.squared
```
```
[1] 0.9698722
```

ou encore :

```
var(predict(lm(w2 ~ iris[, 5]))) * 149/150
```
```
[1] 0.9698722
```

La première donne une combinaison linéaire de variance intra-classe unité :

```
tapply(w1, iris[, 5], var)
```
```
    setosa versicolor  virginica
0.7181898  1.0736485  1.2081617
```
```
mean(tapply(w1, iris[, 5], var))
```
```
[1] 1
```

qui maximise la variance inter-classe.

4 Tester les valeurs propres

4.1 Liens entre les deux procédures

Le lien entre les valeurs propres de `discrimin` et les contributions à la trace de `lda` est plus caché. Notons que :

$$\mathbf{BC}^{-1}\mathbf{u} = \lambda\mathbf{u} \Rightarrow \mathbf{CB}^{-1}\mathbf{BC}^{-1}\mathbf{u} = \lambda\mathbf{CB}^{-1}\mathbf{u} \Rightarrow \mathbf{u} = \lambda\left(\mathbf{B}+\mathbf{W}\right)\mathbf{B}^{-1}\mathbf{u}$$
$$\Rightarrow \mathbf{u} = \lambda\mathbf{u} + \lambda\mathbf{WB}^{-1}\mathbf{u} \Rightarrow \tfrac{\lambda}{1-\lambda}\mathbf{WB}^{-1}\mathbf{u} = \mathbf{u} \Rightarrow \mathbf{BW}^{-1}\mathbf{u} = \tfrac{\lambda}{1-\lambda}\mathbf{u}$$

Les deux matrices ont mêmes vecteurs propres. Si les valeurs propres de $\mathbf{BC}^{-1}$ sont λ_k et les valeurs propres de $\mathbf{BW}^{-1}$ sont μ_k, alors :

$$\mu_k = \frac{\lambda_k}{1-\lambda_k} \Leftrightarrow \lambda_k = \frac{\mu_k}{1+\mu_k}$$

Les valeurs propres de `discrimin` sont :

```
w1 <- dis1$eig
w1
[1] 0.9698722 0.2220266
```

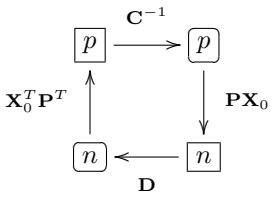
Les valeurs propres de l'autre diagonalisation sont donc :

```
w2 <- w1/(1 - w1)
w2
[1] 32.1919292  0.2853910
w2/sum(w2)
[1] 0.991212605 0.008787395
```

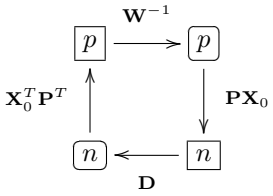
On retrouve les contributions à la trace dans `lda` à l'affichage et dans :

```
lda1$svd^2/sum(lda1$svd^2)
[1] 0.991212605 0.008787395
```

La procédure `discrimin` utilise le schéma :



La procédure `lda` utilise le schéma :



Les deux procédures donnent, à une constante près, la même fonction discriminante et les deux procédures ont des valeurs propres liées par une fonction simple. Sur cette base commune, les deux procédures ont des applications différentes.

4.2 Objectifs des tests

Les deux procédures permettent de tester l'existence d'une réelle différence entre groupes. Mais elles sont complémentaires. L'hypothèse nulle peut être que chaque groupe est un échantillon aléatoire simple de la même loi normale multivariée. L'écart entre les moyennes par groupe n'est alors que la conséquence du hasard. C'est l'anova étendue au cas multivarié.

$$\underset{\text{ANOVA}}{\mu_1 = \mu_2 = ... = \mu_g} \quad \Longrightarrow \quad \underset{\text{MANOVA}}{\begin{pmatrix} \mu_1^1 \\ \mu_1^2 \\ \vdots \\ \mu_1^p \end{pmatrix} = \begin{pmatrix} \mu_2^1 \\ \mu_2^2 \\ \vdots \\ \mu_2^p \end{pmatrix} = ... = \begin{pmatrix} \mu_g^1 \\ \mu_g^2 \\ \vdots \\ \mu_g^p \end{pmatrix}}$$

Si l'hypothèse nulle est fausse, la valeur discriminante de la variable discriminante, la première valeur propre de l'analyse discriminante est anormalement grande. Les schémas ci-dessus indiquent l'existence d'une seconde variable discriminante, puis d'une troisième, ... qui maximisent successivement le même critère sous contrainte d'orthogonalité. La contrainte a une forte signification.

Dans le premier schéma, le premier facteur $\mathbf{a}_1$ est $\mathbf{C}^{-1^{-1}} = \mathbf{C}$ normé ce qui signifie qu'il donne une combinaison de variance 1 :

$$\|\mathbf{a}_1\|_{\mathbf{C}}^2 = \mathbf{a}_1^T \mathbf{C} \mathbf{a}_1 = 1 = \mathbf{a}_1^T \mathbf{X}_0^T \mathbf{D} \mathbf{X}_0 \mathbf{a}_1 = \text{var}\left(\mathbf{X}_0 \mathbf{a}_1\right) = 1$$

Le second facteur $\mathbf{a}_2$, également $\mathbf{C}$-normé, donne une autre combinaison de variance 1 et les deux variables discriminantes sont alors non corrélées car :

$$\langle \mathbf{a}_1 \mid \mathbf{a}_2 \rangle_{\mathbf{C}} = 0 \Rightarrow \mathbf{a}_1^T \mathbf{C} \mathbf{a}_2 = 1 = \mathbf{a}_1^T \mathbf{X}_0^T \mathbf{D} \mathbf{X}_0 \mathbf{a}_2 = \text{cov}\left(\mathbf{X}_0 \mathbf{a}_1, \mathbf{X}_0 \mathbf{a}_2\right) = 0$$

La deuxième combinaison est non corrélée avec la première et maximise à nouveau la variance inter-classe, le maximum atteint étant la deuxième valeur propre, etc, jusqu'au nombre maximum possible - le plus petit du nombre de variables et du nombre de classes.

Les valeurs propres de l'analyse permettent alors de tester la valeur discriminante de une ou toutes les combinaisons ainsi construites. C'est la MANOVA (Multivariate Analysis Of VAriance) qui généralise l'ANOVA. Evidemment, ici la MANOVA est extrêmement significative puisque les anova simples sont très significatives. Il y a plusieurs variantes.

4.3 Test de Pillai

```
size <- as.matrix(iris[, 1:4])
spec <- iris[, 5]
m1 <- manova(size ~ spec)
summary(m1, test = "Pillai")
```

```
           Df Pillai approx F num Df den Df    Pr(>F)
spec        2 1.1919   53.466      8    290 < 2.2e-16
Residuals 147
```

Le critère de Pillai est la somme des valeurs propres de l'analyse discriminante du schéma [1] :

$$P = \sum_{k=1}^{r} \lambda_k$$

```
w1 <- dis1$eig
sum(w1)
```

```
[1] 1.191899
```

4.4 Test de Wilks

```
summary(m1, test = "Wilks")
```

```
          Df    Wilks approx F num Df den Df    Pr(>F)
spec       2 0.023439   199.15      8    288 < 2.2e-16
Residuals 147
```

Le critère de Wilks est le produit des pourcentages de variance intra-classes :

$$W = \prod_{k=1}^{r} (1 - \lambda_k)$$

qu'on teste avec $-log(W)$:

```
prod(1 - w1)
```

```
[1] 0.02343863
```

4.5 Test de Hotelling-Lawley

```
summary(m1, test = "Hotelling-Lawley")
```

```
          Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
spec       2           32.477   580.53      8    286 < 2.2e-16
Residuals 147
```

Le critère de Hotelling-Lawley est la somme des valeurs propres de l'analyse discriminante du second schéma :

$$W = \sum_{k=1}^{r} \mu_k$$

```
w2 <- w1/(1 - w1)
w2
```

```
[1] 32.1919292  0.2853910
```

```
sum(w2)
```

```
[1] 32.47732
```

4.6 Le test de Roy

```
summary(m1, test = "Roy")
```

```
          Df    Roy approx F num Df den Df    Pr(>F)
spec       2 32.192   1167.0      4    145 < 2.2e-16
Residuals 147
```

Le critère de Roy est la plus grande valeur propre du schéma 2 :

$$W = max(\mu_k)$$

```
w2
```

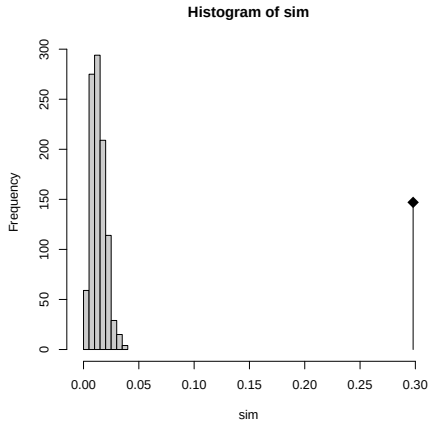
```
[1] 32.1919292  0.2853910
```

Ces tests s'étendent aux modèles linéaires quelconques comme l'anova. Pour en savoir plus, consulter Tomassone et al. [1988] ou Krzanowski and Marriot [1994]. Un exemple est préparé dans la fiche :

http://pbil.univ-lyon1.fr/R/pps/pps083.pdf

Quand l'hypothèse de normalité est intenable, on a une version non paramétrique du test de Pillai avec `randtest.discrimin`.

```
plot(randtest.discrimin(dis1))
```



La statistique observée est le critère de Pillai divisé par le rang de l'analyse de départ :

```
sum(dis1$eig)/4
```

```
[1] 0.2979747
```

4.7 Conclusion

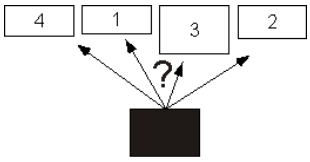
Le vieux débat de la statistique exploratoire contre la statistique confirmatoire, de la description contre la décision, de l'image contre le modèle a beaucoup concerné l'analyse discriminante.

Ce débat s'est nourri de l'hétérogénéité des situations concrètes. Faire un test pour trouver une probabilité critique de 10^-16 est ridicule. Interpréter une figure qui a des chances d'être un aléa l'est tout autant. On trouvera toujours une situation pour ridiculiser son adversaire.

Suivant les situations, l'analyse discriminante est un outil descriptif ; dans d'autres, elle est un outil décisionnel. Cela dépend des données et des objectifs.

5 Discrimination prédictive ou descriptive

La fonction `lda` est centrée sur la question de l'affectation d'un individu à une classe. Peut-on prédire à quelle classe appartient un individu dont on connaît les mesures ?



5.1 Exemple des iris

Le but de l'exercice est de diviser au hasard le tableau de données en deux parties, la première pour chercher une fonction discriminante, la seconde pour déterminer l'espèce à l'aide de cette fonction. On comparera ensuite le résultat obtenu et les vraies valeurs.

```
echa <- sample(1:150, 50)
tabref <- iris[echa, 1:4]
espref <- iris[echa, 5]
tabsup <- iris[-echa, 1:4]
espsup <- iris[-echa, 5]
lda2 <- lda(tabref, espref)
lda2
```

```
Call:
lda(tabref, espref)
Prior probabilities of groups:
    setosa versicolor  virginica
      0.38       0.38       0.24

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         4.915789    3.289474     1.457895   0.2473684
versicolor     5.821053    2.657895     4.105263   1.2789474
virginica      6.375000    2.991667     5.308333   2.0583333

Coefficients of linear discriminants:
                    LD1        LD2
Sepal.Length  1.335947 -0.1300696
Sepal.Width   1.119701  1.7648650
Petal.Length -3.537331 -1.8359582
Petal.Width  -1.818110  4.7215422

Proportion of trace:
   LD1    LD2
0.9845 0.0155
```

```
espestim <- predict(lda2, tabsup)$class
table(espestim, espsup)
```

```
            espsup
espestim     setosa versicolor virginica
  setosa         31          0         0
  versicolor      0         29         0
  virginica       0          2        38
```

A l'aide de 50 plantes, on récupère pratiquement sans erreur le nom des 100 inconnues. Le cas est très favorable.

5.2 Exemple des crânes égyptiens

On étudie ici un cas qui est beaucoup moins favorable avec les données `skulls` proposées par B. Manly [1994] [pp. 6, 13, 51, 64, 72, 107, 112 et 117]. Les mesures concernent 5 groupes de 30 crânes égyptiens. Les classes sont :

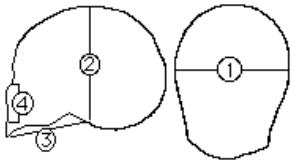
1 période prédynastique ancienne (4000 avant JC) ;

2 période prédynastique récente (3300 avant JC)

3 12 et 13 ème dynastie (1850 avant JC)

4 période de Ptolémée (200 avant JC)

5 période romaine (150 après JC)
Les variables sont définies comme suit :



Les classes sont dans l'attribut `date` du *data frame* :

```
data(skulls)
table(attributes(skulls)$date)
```

```
A{-4000} B{-3300} C{-1850}  D{-200}   E{150}
      30       30       30       30       30
```

```
date <- attr(skulls, "date")
summary(date)
```

```
A{-4000} B{-3300} C{-1850}  D{-200}   E{150}
      30       30       30       30       30
```

1. Représenter les variables par groupe.
2. Faire les analyses de variance univariées. Les crânes sont-ils différents ?
3. Faire l'analyse en composantes principales normée du tableau et l'analyse de variance sur les coordonnées.
4. Faire la MANOVA du tableau des 4 variables sur les groupes :

```
summary(manova(as.matrix(skulls) ~ date), "Pillai")
```

```
          Df  Pillai approx F num Df den Df    Pr(>F)
date       4 0.35331    3.512     16    580 4.675e-06
Residuals 145
```

```
summary(manova(as.matrix(skulls) ~ date), "Roy")
```

```
          Df    Roy approx F num Df den Df    Pr(>F)
date       4 0.4251   15.410      4    145 1.588e-10
Residuals 145
```

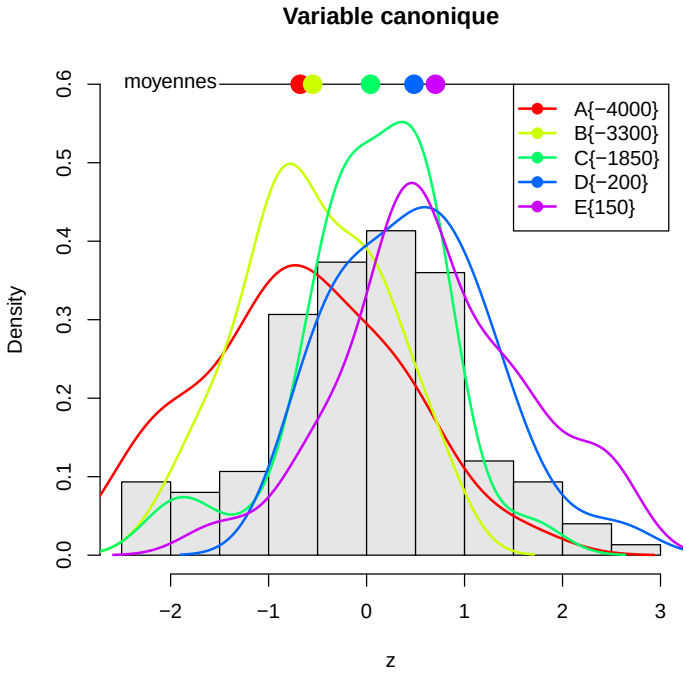
5. Faire l'analyse discriminante et affecter les individus qui ont défini l'analyse à une classe :

```
lda3 <- lda(skulls, date)
table(date, predict(lda3, skulls)$class)
```

```
date       A{-4000} B{-3300} C{-1850} D{-200} E{150}
  A{-4000}       12        8        4       4      2
  B{-3300}       10        8        5       4      3
  C{-1850}        4        4       15       2      5
  D{-200}         3        3        7       5     12
  E{150}          2        4        4       9     11
```

6. Faire une figure expliquant pourquoi le tableau des reallocations n'est pas bon. Ceci est un exemple :

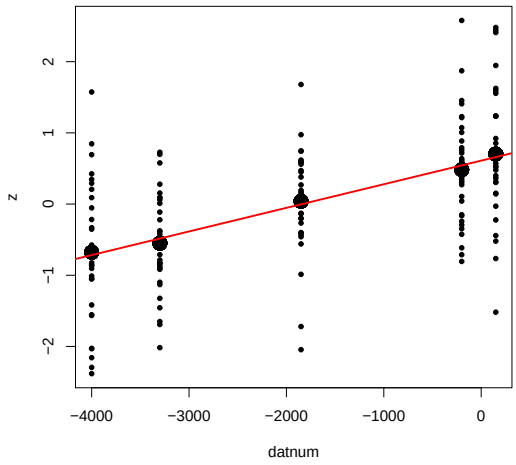
```
dis2 <- discrimin(dudi.pca(skulls, scannf = F), date, scannf = F)
z <- dis2$li[, 1]
hist(z, proba = T, col = grey(0.9), nclass = 15, ylim = c(0, 0.6),
    main = "Variable canonique")
palette(rainbow(5))
text(-2, 0.6, "moyennes")
segments(-1.5, 0.6, 1.5, 0.6)
for (k in 1:5) {
    d0 <- density(z[date == levels(date)[k]])
    lines(d0, col = k, lwd = 2)
    i <- which(d0$y == max(d0$y))
    points(mean(z[date == levels(date)[k]]), 0.6, pch = 19, cex = 2,
        col = k)
}
legend(1.5, 0.6, as.character(levels(date)), lwd = 2, pch = 19,
    col = 1:5)
palette("default")
```



On ne saurait dater un crâne avec quatre mesures. La prédiction est vaine mais la description est inachevée, ou du moins, la question de la prédiction est mal posé.

```
datnum <- c(-4000, -3300, -1850, -200, 150)[as.numeric(date)]
plot(z ~ datnum, pch = 20)
points(predict(lm(z ~ date)) ~ datnum, pch = 19, cex = 2)
abline(lm(z ~ datnum), lwd = 2, col = "red")
summary(manova(as.matrix(skulls) ~ datnum + date))
```

```
           Df   Pillai approx F num Df den Df    Pr(>F)
datnum      1 0.296451  14.9585      4    142 3.171e-10
date        3 0.058693   0.7184     12    432    0.7338
Residuals 145
```

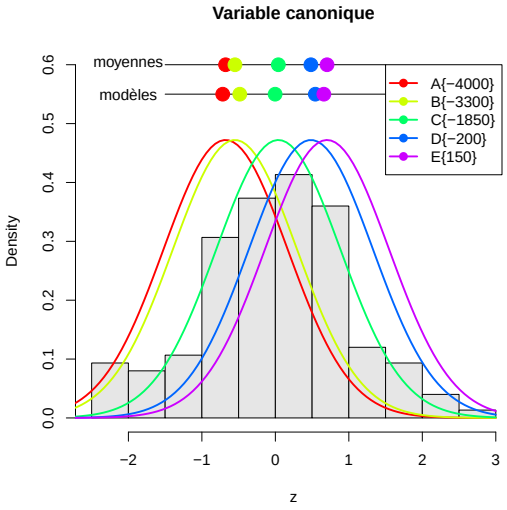


Il ne fallait pas prédire la date avec la mesure mais la mesure avec la date : on est devant un phénomène continu. Quelle signification a le vecteur `z` ? Commenter ce test :

```
bartlett.test(split(z, date))

        Bartlett test of homogeneity of variances
data:  split(z, date)
Bartlett's K-squared = 4.8185, df = 4, p-value = 0.3064
```

Représenter les cinq groupes par une distribution normale. Doit-on prendre une variance commune ?

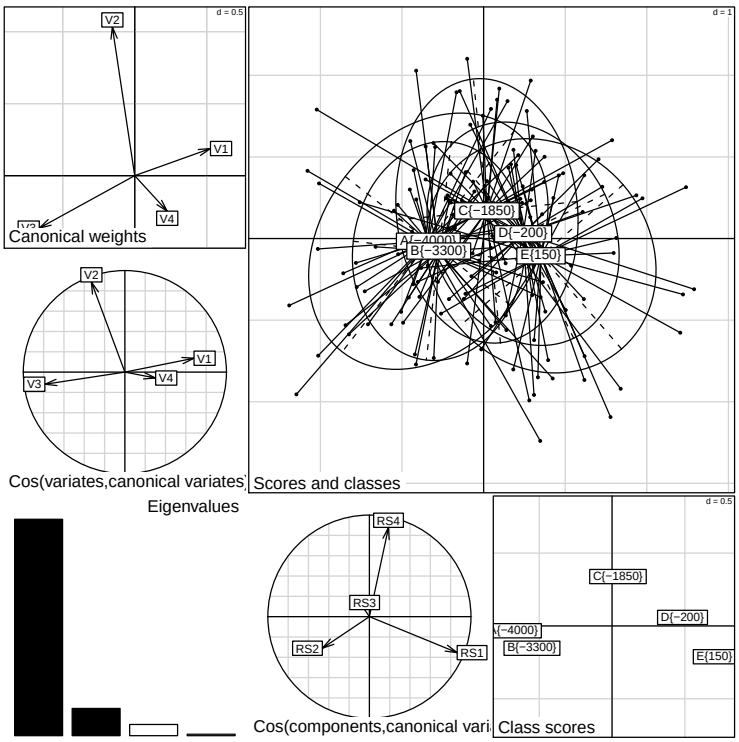


Une analyse discriminante peut être très significative sans qu'elle serve à une allocation d'individus inconnus. Si cet objectif est le bon, la fonction `lda` contient bien d'autres possibilités.

5.3 Représentation graphique liée à `discrimin`

La réalisation d'une analyse discriminante avec la fonction **`discrimin`** permet de visualiser les résultats de façon optimale.

```
plot(dis2)
```



Le graphe associé à **`discrimin`** contient :

1.1 les poids canoniques (coefficients des combinaisons linéaires de variance unité et de variance inter maximales). Les variables utilisées sont les colonnes normalisées de l'analyse en composantes principales préalable.

1.2 les variables canoniques (combinaisons linéaires de variance unité et de variance inter maximales) et les groupes - les ellipses - qui donnent le mode de discrimination opérée.

2.1 les corrélations entre variables canoniques et les variables de départ. Si les graphes 1 et 3 ne sont pas cohérents, c'est l'indice d'une instabilité numérique qui remet en cause l'analyse.

3.1 le graphe des valeurs propres.

3.2 les corrélations entre les variables canoniques et le composantes principales de l'analyse de départ. On peut ainsi savoir si la discrimination se fait dans la partie interprétable de l'analyse préliminaire (sinon il faut être méfiant, des variables discriminantes pouvant être non interprétables).

3.3 les moyennes des variables canoniques par classe.

5.4 Autre exemple

On peut également analyser les données de la fiche :

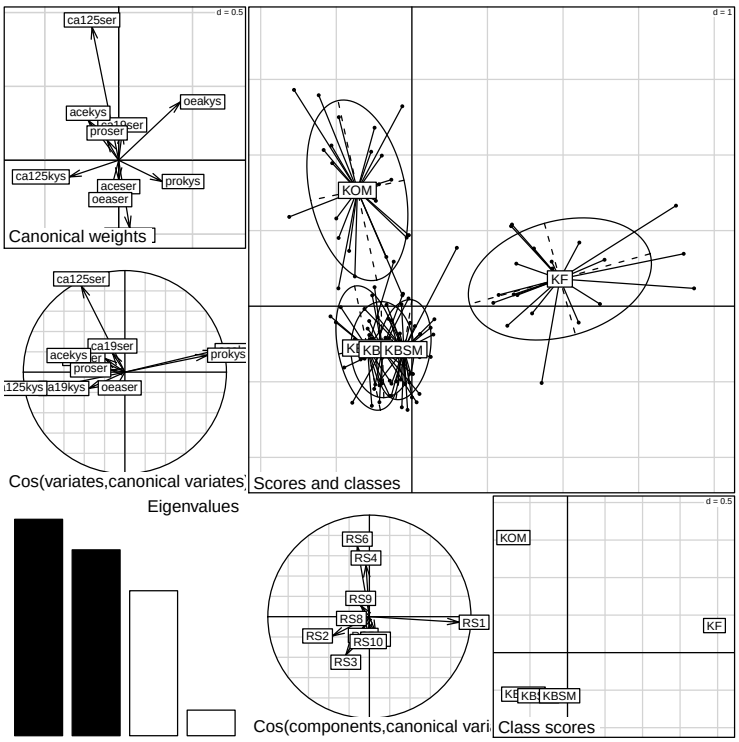
http://pbil.univ-lyon1.fr/R/pdf/pps001.pdf

```
kys <- read.table(url("http://pbil.univ-lyon1.fr/R/donnees/pps001.txt"),
    h = T)
names(kys)
```

```
[1] "cla"      "aceser"   "ca19ser"  "ca125ser" "oeaser"   "proser"   "acekys"
[8] "ca19kys"  "ca125kys" "oeakys"   "prokys"
```

La partition est dans :

```
cla <- kys$cla
tab <- kys[, -1]
kys.pca <- dudi.pca(tab, scannf = F)
kys.dis <- discrimin(kys.pca, cla, scannf = F)
plot(kys.dis)
```



Le plan 1-2 discrimine très bien les kystes organiques, bénins et malins. Le troisième facteur complète la séparation des groupes.

```
lda4 <- lda(tab, cla)
table(cla, predict(lda4, tab)$class)
```

```
cla      KBEnd KBSer KBSM KF KOM
  KBEnd     10     0    1  0   1
  KBSer      1    17   10  0   0
  KBSM       2     4   31  0   0
  KF         0     0    0 20   0
  KOM        0     2    2  0  20
```

Dans cet exemple, le tableau de reallocation est d'une grande importance. La présence d'un kyste cancéreux impose l'ablation. L'erreur est grave pour les faux positifs et l'est encore plus pour les faux négatifs. Commenter.
Peut-on de cette manière discriminer sur des variables qualitatives, sur des données distributionnelles, sur des mélanges ? La réponse est oui. Voir la suite dans :

`file:///d:/pedapdf/fichestd/tdr54.pdf`

Références

E. Anderson. The irises of the gaspe peninsula. *Bulletin of the American Iris Society*, 59 :2–5, 1935.

T. Cacoullos and G.P.H. Styan. A bibliography of discriminant analysis. In T. Cacoullos, editor, *Discriminant analysis and applications*, pages 375–434. Academic Press, New York, 1-434, 1973.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 :179–188, 1936.

W.J. Krzanowski and F.H.C. Marriot. *Multivariate analysis. Part 1 Distributions, ordination and inference.* Statistics and Computing. Edward Arnold, London, 1994.

B.F. Manly. *Multivariate Statistical Methods. A primer. Second edition.* Chapman & Hall, London, 1994.

R. Tomassone, M. Danzard, J.J. Daudin, and J.P. Masson. *Discrimination et classement.* Masson, Paris, 1988.