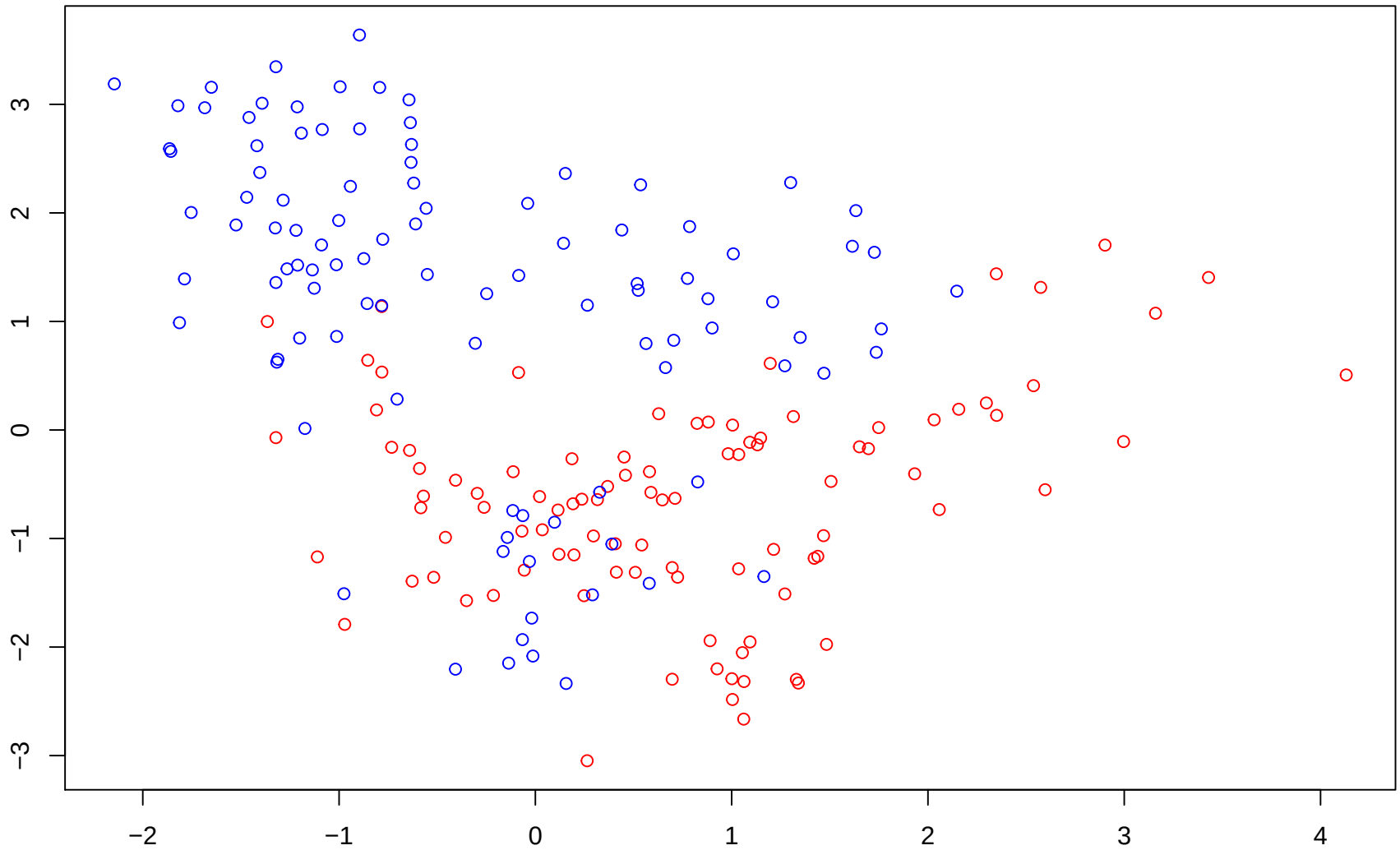
Chapitre 2 : Mélange Gaussien/algorithme LDA - Illustrations sous R

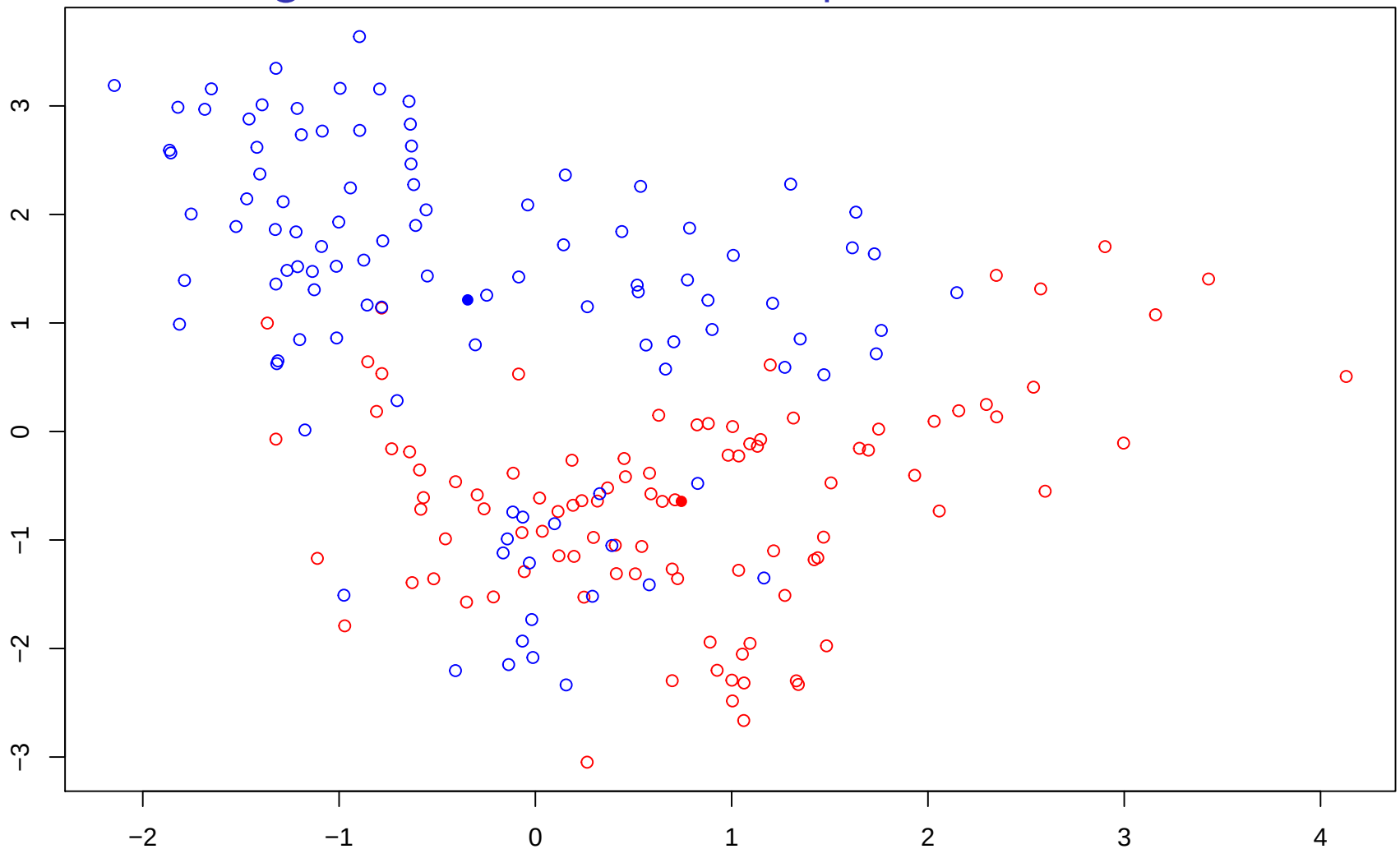
Cours de classification

2022-2023

Retour sur l'exemple du Chapitre 1



Modèle de mélange Gaussien homoscédastique



On représente sur le dessin les moyennes estimées des deux mélanges. On trouve $\hat{\mu}_1 =$ (0.7438, -0.6441) et $\hat{\mu}_2 =$ (-0.3447, 1.2125). La variance commune estimée vaut $\begin{pmatrix} 1.1310 & -0.1588 \\ -0.1588 & 1.5539 \end{pmatrix}$

Modèle de mélange Gaussien homoscédastique - utilisation de la fonction lda

On utilise la fonction `lda` du package `MASS`.

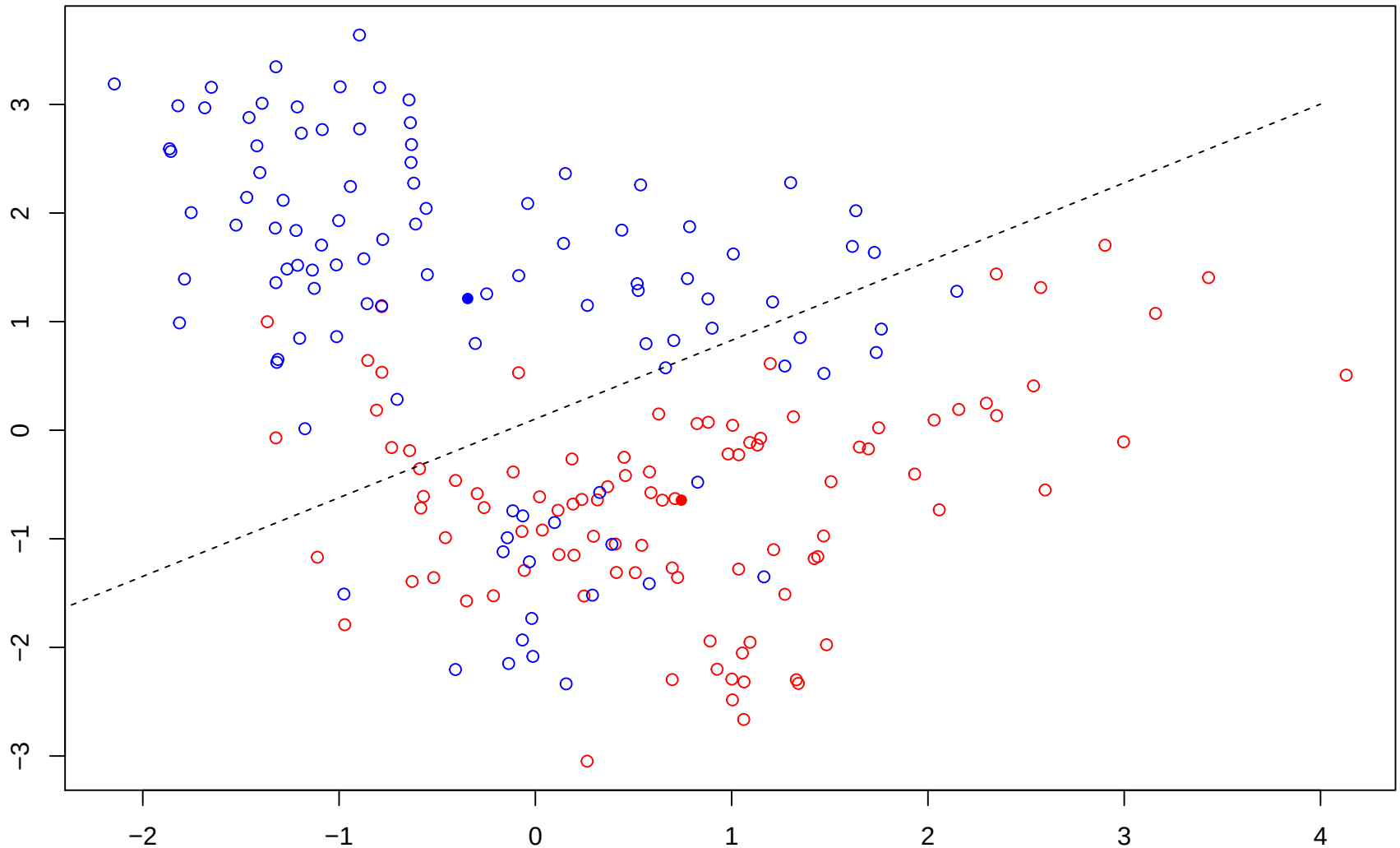
```
require(MASS)
```

```
## Loading required package: MASS
```

```
fit=lda(grouptrain~Xtrain[,1]+Xtrain[,2])
fit
```

```
## Call:
## lda(grouptrain ~ Xtrain[, 1] + Xtrain[, 2])
##
## Prior probabilities of groups:
##    0    1
## 0.51 0.49
##
## Group means:
##   Xtrain[, 1] Xtrain[, 2]
## 0  -0.3447311   1.2125456
## 1   0.7437763  -0.6440869
##
## Coefficients of linear discriminants:
##                     LD1
## Xtrain[, 1]  0.4676205
## Xtrain[, 2] -0.6452162
```

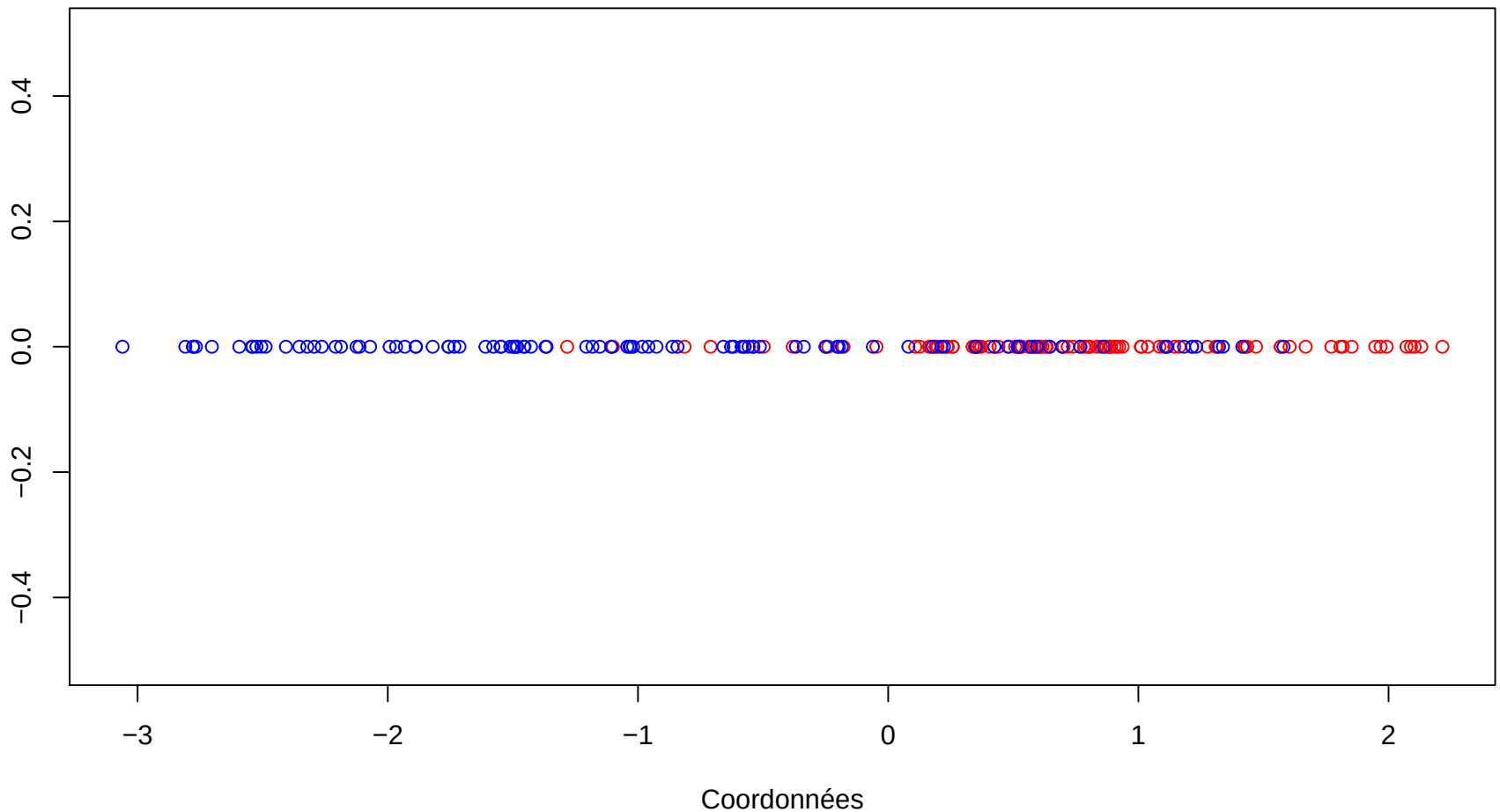
Modèle de mélange Gaussien homoscédastique - frontière de classification



Modèle de mélange Gaussien homoscédastique - projection sur l'axe orthogonal

On applique la transformation $\begin{pmatrix} 0.4676 & -0.6452 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0.4676 x_1 - 0.6452 x_2$

à tous les points $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$.

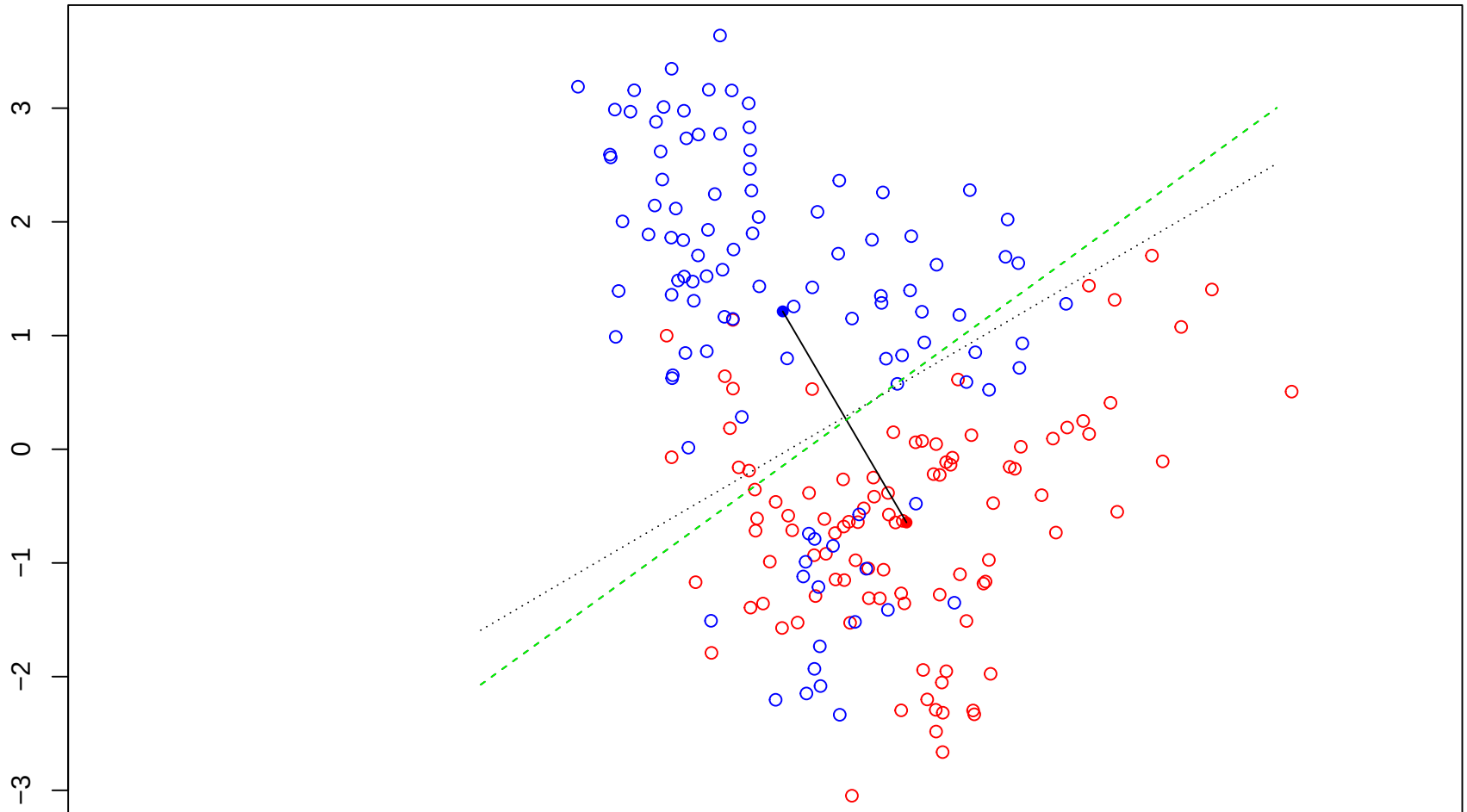


Modèle de mélange Gaussien homoscédastique - remarques

- Quand la matrice de variance Σ est diagonale, le classifieur est simplement la droite orthogonale au segment reliant les deux centroides.
- Ce n'est pas le cas si la covariance entre les coordonnées de X est non-nulle.
- On peut en fait montrer que la méthode **lda** cherche à maximiser la variance *inter-classes* et à minimiser la variance *intra-classes*.
- C'est un algorithme différent de l'ACP, ce dernier cherche uniquement à maximiser la variances des données projetées sur l'axe des composantes (ou autrement dit, à minimiser la variance des données sur l'axe orthogonal à l'axe des composantes).

Modèle de mélange Gaussien homoscédastique - remarques

- On trace en vert la droite de classification basée sur les moindres carrés.
- On obtient une frontière de classification quasiement identique à la méthode **lda** (les deux droites sont confondues sur le dessin. Cela vient du fait que les tailles d'échantillon des groupes bleu et rouge sont quasiment identiques (voir exercice supplémentaire)
- On trace en pointillé noir la droite orhtogonale au segment reliant les deux centroides.



Calcul du taux de missclassification

```
## Loading required package: class
```

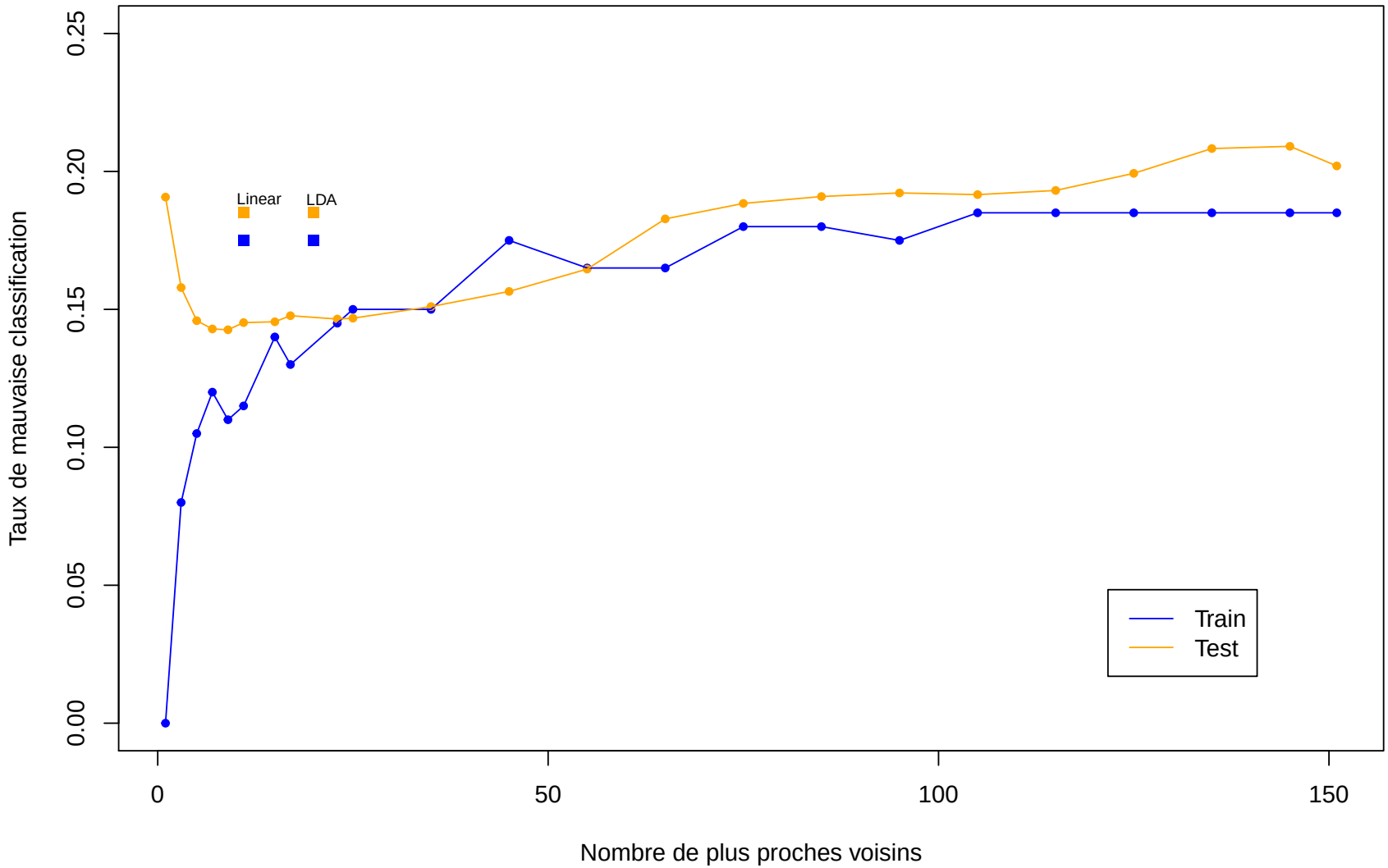


Illustration sur des données d'eau de vie

But : déterminer le type d'eau de vie (Kirsch, Mirab, Poire) en fonction de leur composition (buthanol, méthanol, etc.)

```
require(xlsx)
alcohol.data <- read.xlsx(file="alcohol.xls",sheetIndex=1,header=T)
head(alcohol.data)
```

```
## Loading required package: xlsx
```

```
##     TYPE MEOH ACET BU1 MEPR ACAL LNPRO1
## 1 KIRSCH    3   15 0.2    9  9.0   5.86
## 2 KIRSCH   23   13 0.8    9  2.0   6.67
## 3 KIRSCH   65   96 0.4    9  4.0   5.31
## 4 KIRSCH  279   66 0.9   36  4.8   5.45
## 5 KIRSCH  292  210 1.1   34  8.0   4.08
## 6 KIRSCH  371  414 1.2   39  9.0   6.22
```

Illustration sur des données d'eau de vie

```
alcohol.lda <- lda(TYPE ~ ., data = alcohol.data)
alcohol.lda
```

```
## Call:
## lda(TYPE ~ ., data = alcohol.data)
##
## Prior probabilities of groups:
##    KIRSCH     MIRAB     POIRE
## 0.2337662 0.3766234 0.3896104
##
## Group means:
##              MEOH     ACET       BU1     MEPR     ACAL   LNPRO1
## KIRSCH  378.6944 218.0167  1.511111 32.06667 11.16667 6.231111
## MIRAB   939.1379 247.3448 17.906897 30.55172 12.54138 4.883103
## POIRE  1035.4000 173.3667 19.620000 43.00000 13.27333 5.145667
##
## Coefficients of linear discriminants:
##                  LD1           LD2
## MEOH    3.382089e-03  0.0005710473
## ACET   -4.649248e-05 -0.0066573606
## BU1     1.322048e-01 -0.0162598664
## MEPR   -2.562255e-02  0.0533609640
## ACAL   -4.048757e-02  0.0297883525
## LNPRO1 -2.791911e-01  0.3894400487
##
## Proportion of trace:
##    LD1    LD2
## 0.9168 0.0832
```

Matrice de confusion

```
pred.lda <- predict(alcohol.lda,newdata=alcohol.data)
table(alcohol.data$TYPE,pred.lda$class)
```

```
##
##           KIRSCH MIRAB POIRE
##    KIRSCH     18     0     0
##    MIRAB       0    23     6
##    POIRE       0     9    21
```

Kirsch est parfaitement classifié !! Poire et mirab sont les plus difficiles à classifier avec respectivement 70% et 79% de bonnes classifications.

Projection sur la frontière (un plan)

```
eqscplot(pred.lda$x[,1],pred.lda$x[,2],col=c(rep("red",18)
legend("topleft",col=c("red","green","blue"),lty=c(1,1,1),
       legend=c("Kirsch","Mirab","Poire"),inset=0.1)
```