Discriminant Analysis

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

jan.graffelman@upc.edu

April 21, 2020

Discriminant analysis: Aims

- Group separation
- Dimension reduction: from p variables to k discriminators with $k < p$.
- Classification of new cases

Discriminant Analysis: the data matrix

Ind.	X_1	X_2	$\cdots$	X_p	Group
1	X_{11}	X_{12}	$\cdots$	X_{1p}	1
2	X_{21}	X_{22}	$\cdots$	X_{2p}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_1	$X_{n_1 1}$	$X_{n_1 2}$	$\cdots$	$X_{n_1 p}$	1
1	X_{11}	X_{12}	$\cdots$	X_{1p}	2
2	X_{21}	X_{22}	$\cdots$	X_{2p}	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_2	$X_{n_2 1}$	$X_{n_2 2}$	$\cdots$	$X_{n_2 p}$	2
1	X_{11}	X_{12}	$\cdots$	X_{1p}	m
2	X_{21}	X_{22}	$\cdots$	X_{2p}	m
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_m	$X_{n_m 1}$	$X_{n_m 2}$	$\cdots$	$X_{n_m p}$	m

Some examples

- Which morphological measurements can discriminate between men and women?
- Given various biochemical measurements, is this person healthy or diseased?
- Given the variables of this wheat kernel, to which of the known varieties does it belong?
- ...
- One can distinguish between two-group and multiple group problems.

Two-group linear discriminant analysis

Criteria for designing a classification rule:

- small probability of misclassification
- take prevalence into account (prior probabilities)
- take the cost of misclassification into account

Two-group linear discriminant analysis

Some basic definitions:

- π_1 and π_2 represent population 1 and 2.
- $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ represent the multivariate probability densities for each population.
- $\Omega = R_1 \cup R_2$ is the partitioned sample space for outcome $\mathbf{x}$.
- If $\mathbf{x}$ falls in R_1, the case is classified as π_1, else in π_2.
- p_1 is the prior probability of pertaining to π_1, p_2 the prior probability of pertaining to π_2 (prevalence)
- Misclassification probabilities:
  1. $P(2|1) = P(\mathbf{X} \in R_2|\pi_1) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x}$
  2. $P(1|2) = P(\mathbf{X} \in R_1|\pi_2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x}$

Cost matrix

		Predicted class	
		π_1	π_2
True	π_1	0	$c(2\|1)$
Class	π_2	$c(1\|2)$	0

$c(1|2)$ and $c(2|1)$ are not necessarily equal

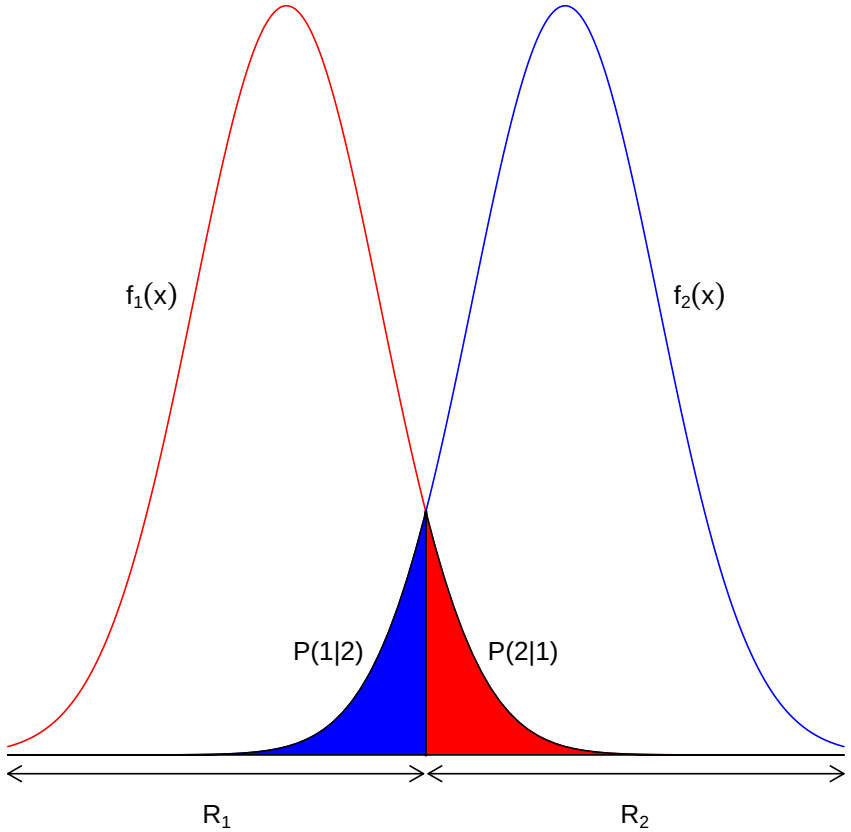
ECM = Expected Cost of Misclassification

$$P\left(\text{from } \pi_1 \cap \text{classified } \pi_2\right) = P\left(2|1\right) \cdot p_1$$

$$P\left(\text{from } \pi_2 \cap \text{classified } \pi_1\right) = P\left(1|2\right) \cdot p_2$$

$$\text{ECM} = c(1|2)P(1|2)p_2 + c(2|1)P(2|1)p_1$$

Classification rule: minimizing ECM



ECM Rule

The regions that minimize the ECM are:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

If there is differential prevalence:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1}$$

If there is differential cost:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)}$$

And if we have both differential prevalence and differential cost:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1} \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}$$

Two normal populations with equal covariance matrices

For continuous $\mathbf{X}$, we assume multivariate normality:

$$f_1(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}$$

$$f_2(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}$$

Two-group linear discriminant analysis

Sample based ECM Rule: assign observation $\mathbf{x}$ to population 1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} \mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln \left(\left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right) \right)$$

where $\mathbf{S}_p$ is the pooled covariance matrix:

$$\mathbf{S}_p = \frac{n_1 - 1}{n_1 + n_2 - 2} \mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} \mathbf{S}_2$$

Two-group linear discriminant analysis

Define:

$$\mathbf{a} = \mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \qquad y = \mathbf{a}'\mathbf{x}$$

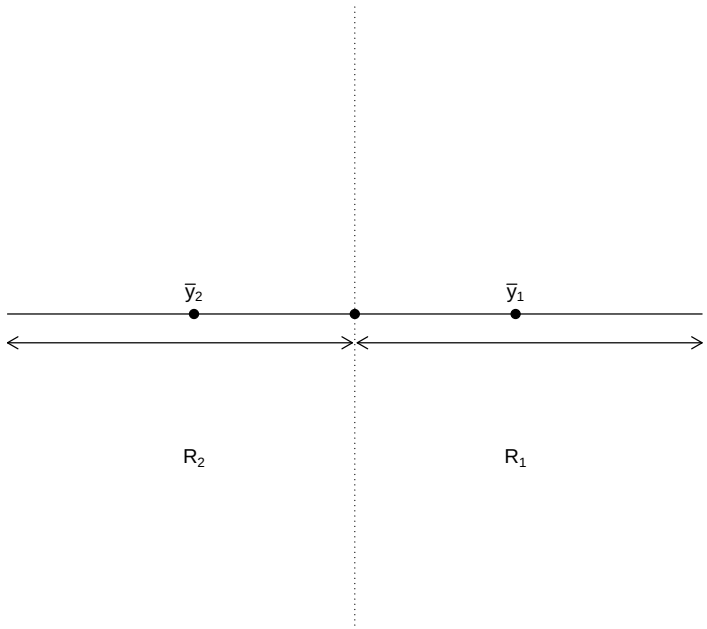
Note that:

$$y_i = \mathbf{a}'\mathbf{x}_i \qquad \overline{y}_1 = \mathbf{a}'\bar{\mathbf{x}}_1 \qquad \overline{y}_2 = \mathbf{a}'\bar{\mathbf{x}}_2$$

With equals costs and priors, the ECM rule for R_1 boils down to the univariate rule:

$$y_i > \frac{1}{2}(\overline{y}_1 + \overline{y}_2)$$

y is the classifier or linear discriminant function.



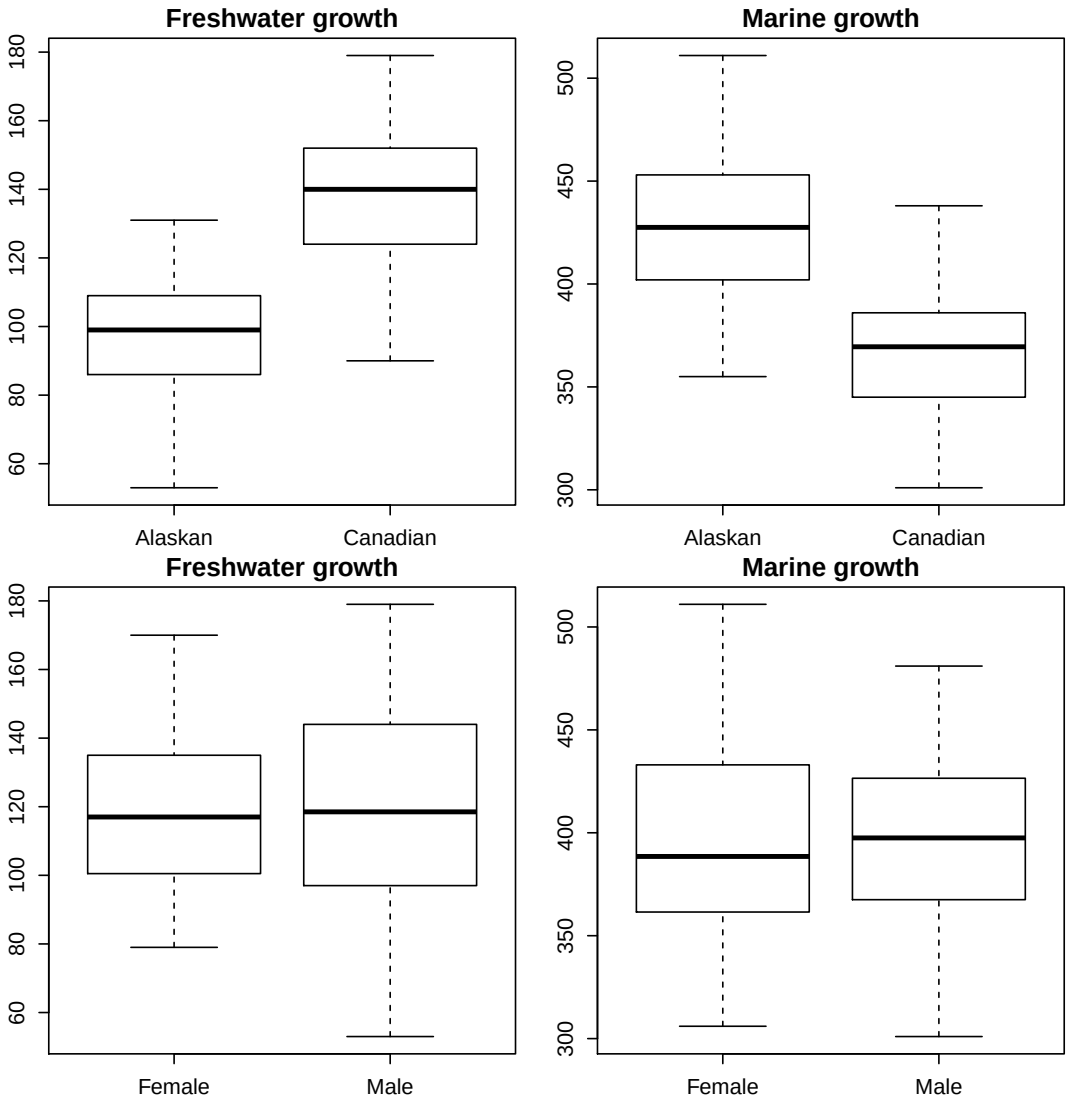
Example: the Salmon data

	Origin	Gender	Fresh	Marine
1	Alaskan	Female	108	368
2	Alaskan	Male	131	355
3	Alaskan	Male	105	469
4	Alaskan	Female	86	506
5	Alaskan	Male	99	402
6	Alaskan	Female	87	423
⋮	⋮	⋮	⋮	⋮
95	Canadian	Female	140	388
96	Canadian	Female	150	339
97	Canadian	Female	124	341
98	Canadian	Male	125	346
99	Canadian	Male	153	352
100	Canadian	Male	108	339

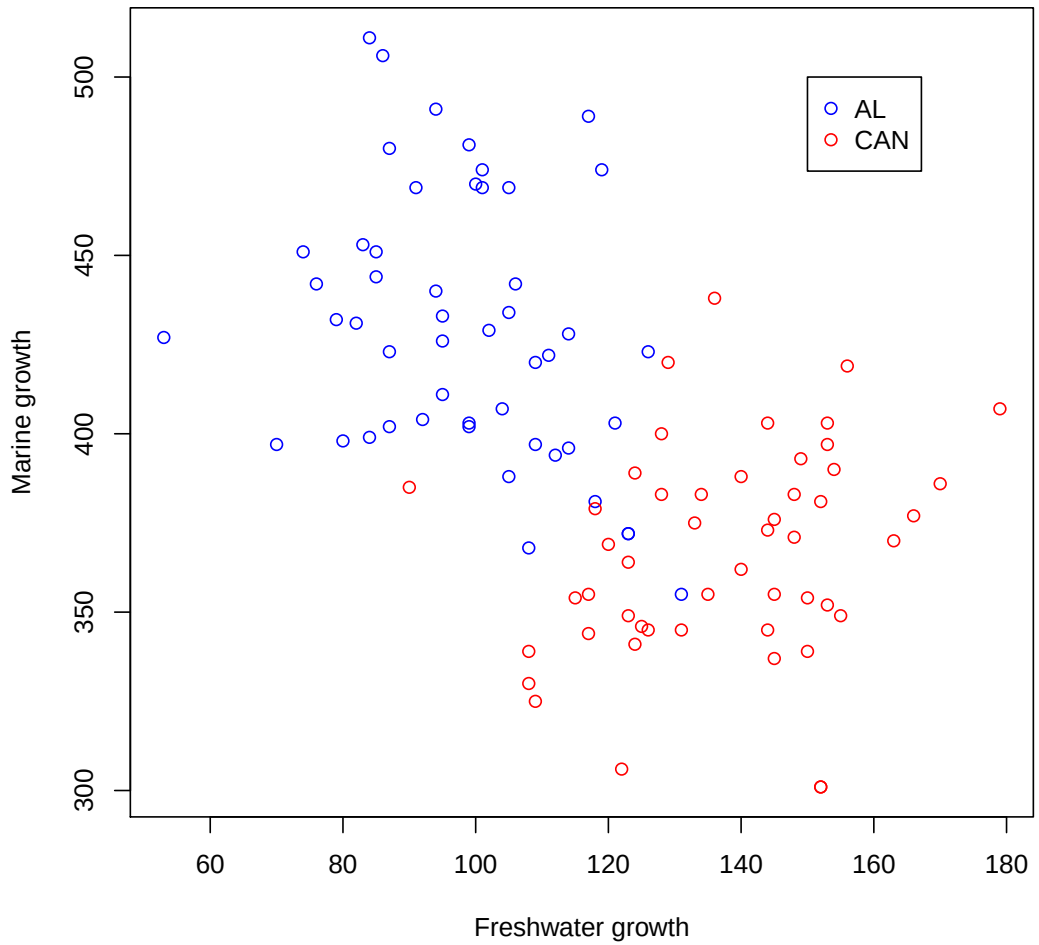
Download Salmon.dat

```
Two-sample Hotelling T2 (unequal covariance matrices)
T2 =  207.2967 TCrit =  5.991465 p-value = 0
```

Exploring


Freshwater growth
60
80
100
120
140
160
180
Alaskan
Canadian
Marine growth
300
350
400
450
500
Alaskan
Canadian
Freshwater growth
60
80
100
120
140
160
180
Female
Male
Marine growth
300
350
400
450
500
Female
Male

Exploring

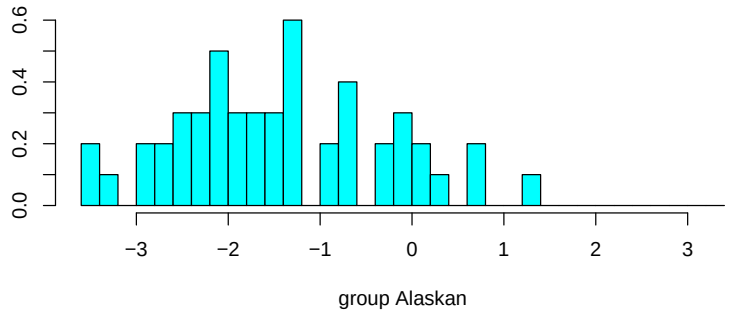

AL
CAN
Marine growth
Freshwater growth
300
350
400
450
500
60
80
100
120
140
160
180

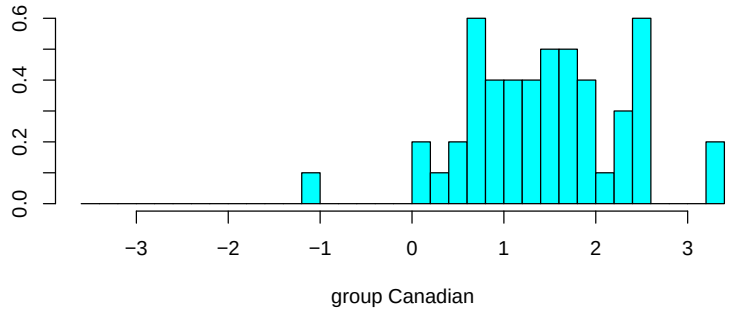
Calculations

```
mean vectors

          Fresh Marine
Alaskan   98.38 429.66
Canadian 137.46 366.62

> S1 # Alaskan
           Fresh    Marine
Fresh   260.6078 -188.0927
Marine -188.0927 1399.0861
> S2 # Canadian
          Fresh   Marine
Fresh  326.0902 133.5049
Marine 133.5049 893.2608

> Spool <- ((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
> Spool
           Fresh     Marine
Fresh  293.34898  -27.29388
Marine -27.29388 1146.17347

> a <- solve(Spool)%*%(m1-m2)
> a
              [,1]
Fresh  -0.12838726
Marine  0.05194311
```

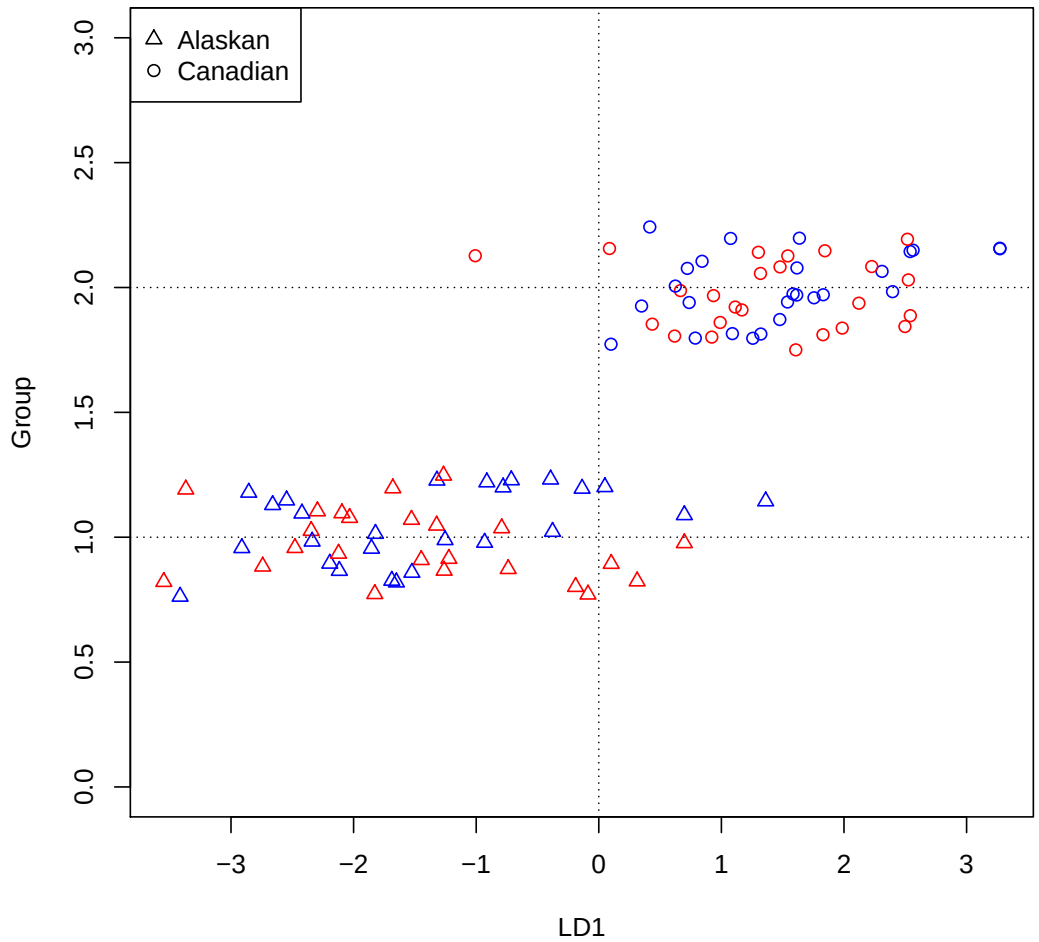
LDA in R

```
> out <- lda(Origin~Fresh+Marine,data=X)
> out
Call:
lda(Origin ~ Fresh + Marine, data = X)

Prior probabilities of groups:
 Alaskan Canadian
     0.5      0.5

Group means:
          Fresh Marine
Alaskan   98.38 429.66
Canadian 137.46 366.62

Coefficients of linear discriminants:
               LD1
Fresh   0.04458572
Marine -0.01803856
>
> plot(out)
```





Graphical representation



Two-group QDA

Under the assumption of multivariate normality with $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, using the same ECM principle, a quadratic classification rule is obtained.

Sample based ECM Rule: assign observation $\mathbf{x}$ to population 1 if

$$-\frac{1}{2}\mathbf{x}'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x} + (\overline{\mathbf{x}}_1{}'\mathbf{S}_1^{-1} - \overline{\mathbf{x}}_2{}'\mathbf{S}_2^{-1})\mathbf{x} - k \geq \ln\left(\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right)$$

with

$$k = \frac{1}{2}\ln\left(\frac{|\mathbf{S}_1|}{|\mathbf{S}_2|}\right) + \frac{1}{2}(\overline{\mathbf{x}}_1'\mathbf{S}_1^{-1}\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2'\mathbf{S}_2^{-1}\overline{\mathbf{x}}_2)$$

Error rates and Confusion matrix

- It is of interest to evaluate the performance of a classification rule.
- There are several criteria to do so.
- Actual error rate (AER, density dependent)

$$\text{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x})d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x})d\mathbf{x}$$

- Apparent error rate (APER, not density dependent) based on the confusion matrix

		Predicted class	
		π_1	π_2
True	π_1	n_{11}	n_{12}
Class	π_2	n_{21}	n_{22}

- APER obtained as

$$\text{APER} = \frac{n_{12} + n_{21}}{n_1 + n_2}$$

- APER underestimates the AER.

Jackknife or hold-one-out

Procedure:

- Take the data from group π_1. Omit the ith observation, build the classifier with $n_1 - 1 + n_2$ observations.
- Classify the ith observation using the classifier.
- Repeat for all observations in π_1.
- Calculate n_{1M}^H, the number of observations that were held out and misclassified.
- Do the same for group π_2 and calculate n_{2M}^H.
- Obtain an estimate of the expected actual error rate

$$E\,(\text{AER}) = \frac{n_{1M}^H + n_{2M}^H}{n_1 + n_2}$$

Salmon data revisited

```
> out <- lda(Origin~Fresh+Marine,data=X)
> pre <- predict(out)
> confusion <- table(X$Origin,pre$class)
> confusion

          Alaskan Canadian
  Alaskan       44        6
  Canadian       1       49
> aper <- (confusion[1,2]+confusion[2,1])/sum(confusion)
> aper
[1] 0.07
>

> n <- nrow(X)
> nmisclas <- 0
> for(i in 1:n) {
+   ho <- X[i,]
+   out <- lda(Origin~Fresh+Marine,data=X[-i,])
+   aaa <- predict(out,newdata = ho)
+   if(aaa$class!=ho$Origin) nmisclas<-nmisclas+1
+ }
> nmisclas/n
[1] 0.07
>
```

```
> out <- qda(Origin~Fresh+Marine,data=X)
> pre <- predict(out)
> confusion <- table(X$Origin,pre$class)
> confusion

          Alaskan Canadian
  Alaskan       45        5
  Canadian       2       48
> aper <- (confusion[1,2]+confusion[2,1])/sum(confusion)
> aper
[1] 0.07
>
>
> n <- nrow(X)
> nmisclas <- 0
> for(i in 1:n) {
+   ho <- X[i,]
+   out <- qda(Origin~Fresh+Marine,data=X[-i,])
+   aaa <- predict(out,newdata = ho)
+   if(aaa$class!=ho$Origin) nmisclas<-nmisclas+1
+ }
> nmisclas/n
[1] 0.08
>
```

LDA with multiple groups

- The ECM rule can be extended to k groups
- Fisher's discriminant analysis

ECM rule

ECM rule with k groups (equal costs)

Assign $\mathbf{x}$ to π_k if

$$p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}) \quad \forall \quad i \neq k$$

Fisher's linear discriminant analysis

- Searches for an optimal linear combination:

$$Z_1 = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p$$

- Maximizes the ratio of variability between groups to variability within groups
- Objective function

$$\frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}$$

- $\mathbf{W}$ is the matrix with within-group sums-of-squares
- For a single group i

$$\mathbf{W}_i = (\mathbf{X}_i - \mathbf{1}\mathbf{m}_i')'(\mathbf{X}_i - \mathbf{1}\mathbf{m}_i')$$

- $\mathbf{W} = \sum_{i=1}^{k} \mathbf{W}_i$
- $\mathbf{B}$ is the matrix with between-group sums-of-squares
- $\mathbf{T}$ is the matrix with total sums-of-squares

$$\mathbf{T} = (\mathbf{X} - \mathbf{1}\mathbf{m}')'(\mathbf{X} - \mathbf{1}\mathbf{m}') \qquad \mathbf{T} = \mathbf{W} + \mathbf{B}$$

Solution

- The optimal weights are found by solving an eigenvector-eigenvalue problem

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}$$

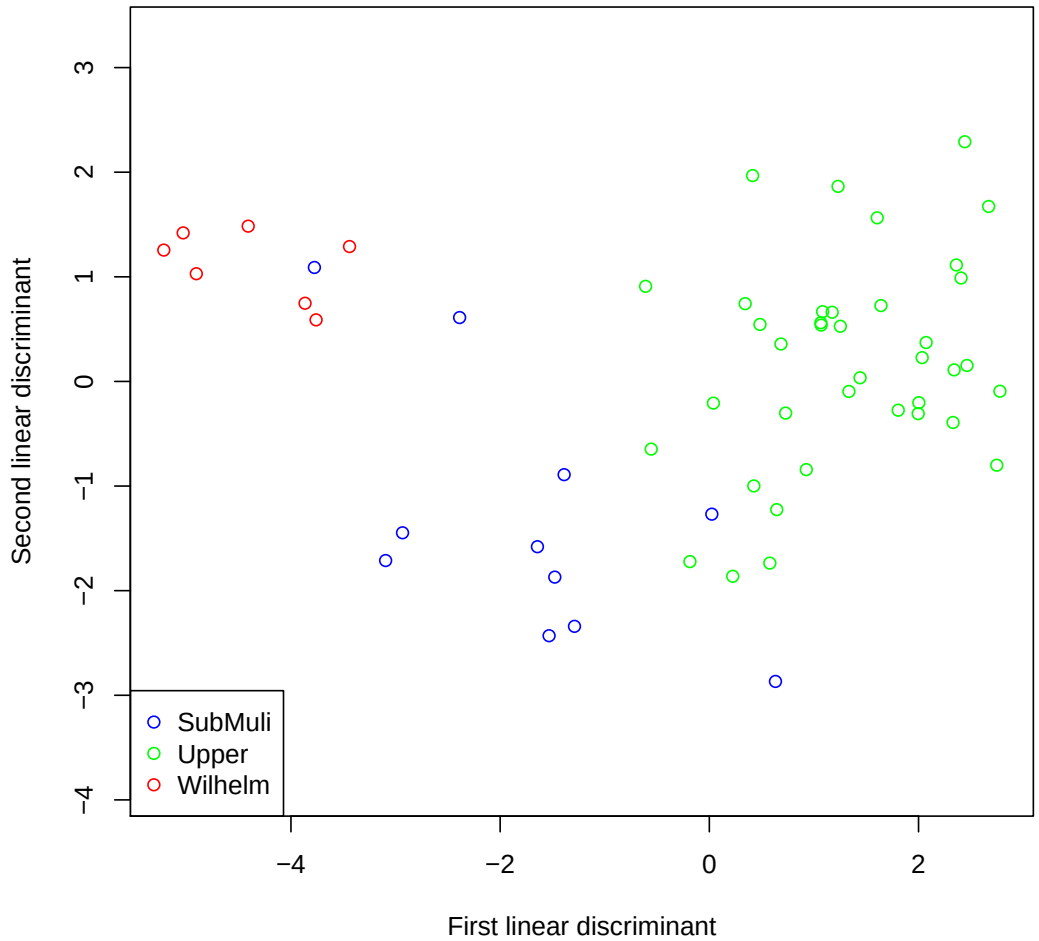
- The number of dimensions d in the solution is given by $\min(k-1, p)$

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{A} = \mathbf{A}\mathbf{D}_{\lambda}$$

- Eigenvectors scaled to satisfy $\mathbf{A}'\mathbf{S}_p\mathbf{A} = \mathbf{I}$
- Selecting the first two eigenvalues and eigenvectors allows for dimension reduction

Crude-oil data

	Vanadium	Iron	Beryllium	InvHydrocarbon	Aromatic	Oiltype
1	3.90	7.14	0.45	0.14	12.19	Wilhelm
2	2.70	7.00	0.26	0.14	12.23	Wilhelm
3	2.80	6.00	0.55	0.14	11.30	Wilhelm
4	3.10	6.71	0.28	0.14	13.01	Wilhelm
5	3.50	6.78	0.32	0.13	12.63	Wilhelm
6	3.90	6.56	0.26	0.16	10.42	Wilhelm
7	2.70	5.92	0.00	0.20	9.00	Wilhelm
8	5.00	6.86	0.26	0.14	6.10	SubMuli
9	3.40	5.66	0.45	0.17	4.69	SubMuli
10	1.20	3.46	0.00	0.18	3.15	SubMuli
11	8.40	4.12	0.26	0.16	4.55	SubMuli
12	4.20	6.00	0.71	0.11	4.95	SubMuli
13	4.20	5.92	0.71	0.18	2.22	SubMuli
14	3.90	6.40	0.32	0.18	2.94	SubMuli
⋮	⋮	⋮	⋮	⋮	⋮	
47	9.00	4.47	0.71	0.17	11.17	Upper
48	6.20	4.00	0.22	0.24	4.18	Upper
49	7.30	4.47	0.71	0.23	3.50	Upper
50	3.60	3.87	0.84	0.14	4.82	Upper
51	6.20	5.83	0.26	0.21	2.37	Upper
52	7.30	4.69	0.00	0.24	2.70	Upper
53	4.10	5.39	0.84	0.17	7.76	Upper
54	5.40	5.39	0.45	0.22	2.65	Upper
55	5.00	5.83	0.84	0.24	6.50	Upper
56	6.20	5.20	0.55	0.25	2.97	Upper


Second linear discriminant
First linear discriminant
SubMuli
Upper
Wilhelm
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2

Some R code

```
> colnames(X) <- c("Vanadium","Iron","Beryllium",
+                  "InvHydrocarbon","Aromatic","Oiltype")
> X$Iron <- sqrt(X$Iron)
> X$Beryllium <- sqrt(X$Beryllium)
> X$InvHydrocarbon <- 1/X$InvHydrocarbon
> out.lda <- lda(Oiltype~Vanadium+Iron+Beryllium+InvHydrocarbon+Aromatic,data=X)
> lda.pred <- predict(out.lda)
> LD <- lda.pred$x

> out.lda
Call:
lda(Oiltype ~ Vanadium + Iron + Beryllium + InvHydrocarbon +
    Aromatic, data = X)

Prior probabilities of groups:
  SubMuli     Upper   Wilhelm
0.1964286 0.6785714 0.1250000

Group means:
        Vanadium     Iron Beryllium InvHydrocarbon  Aromatic
SubMuli 4.445455 5.666848 0.3439707      0.1571001  5.483636
Upper   7.226316 4.633666 0.5981250      0.2231776  5.767895
Wilhelm 3.228571 6.586497 0.3033081      0.1495973 11.540000
```

Some R code

```
Coefficients of linear discriminants:
                     LD1        LD2
Vanadium        0.3121837 -0.1694498
Iron           -0.7099884  0.2454856
Beryllium       2.7638171  2.0456035
InvHydrocarbon 11.8090852 24.4533141
Aromatic       -0.2354662  0.3778283

Proportion of trace:
   LD1    LD2
0.8862 0.1138
> confusion <- table(X$Oiltype,lda.pred$class)
> confusion

          SubMuli Upper Wilhelm
  SubMuli       8     2       1
  Upper         1    37       0
  Wilhelm       0     0       7
>
> colvec <- rep(NA,nrow(X))
> colvec[X$Oiltype=="SubMuli"] <- "blue"
> colvec[X$Oiltype=="Upper"] <- "green"
> colvec[X$Oiltype=="Wilhelm"] <- "red"
> plot(LD[,1],LD[,2],asp=1,xlab="First linear discriminant",
+      ylab="Second linear discriminant",col=colvec,pch=1)
> legend("bottomleft",c("SubMuli","Upper","Wilhelm"),pch=1,col=c("blue","green","red"))
```

Alternative statistical techniques

- An alternative technique for two-group DA is logistic regression
- An alternative technique for multi-group DA is the multinomial logit model

References

- Hand, D.J. (1981) Discrimination and Classification. Wiley, New York.
- Johnson & Wichern, (2002) Applied Multivariate Statistical Analysis, 5th edition, Prentice Hall, Chapter 11.
- Lachenbruch, P.A. (1975) Discriminant Analysis. Hafner Press, New York.
- Peña, D. (2002) Análisis de datos multivariantes. McGraw-Hill, Madrid.