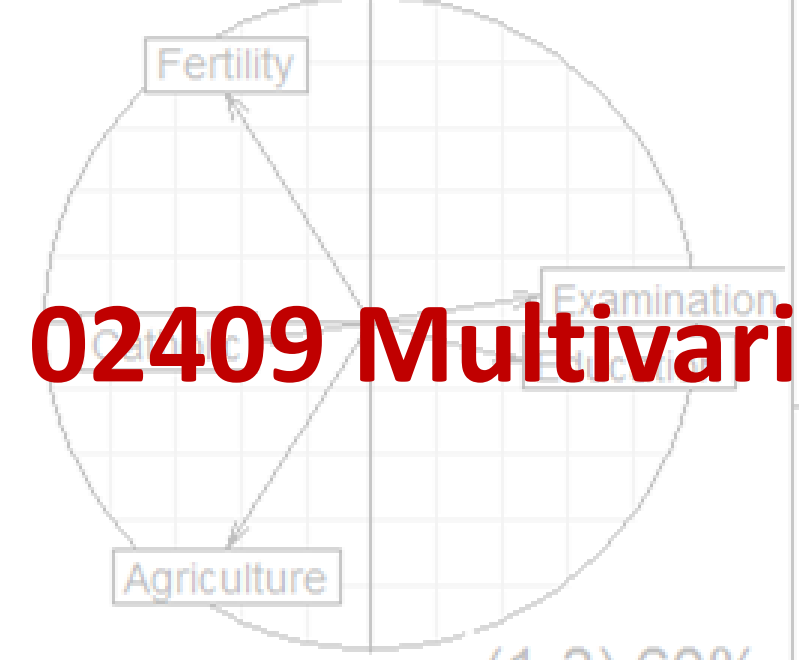
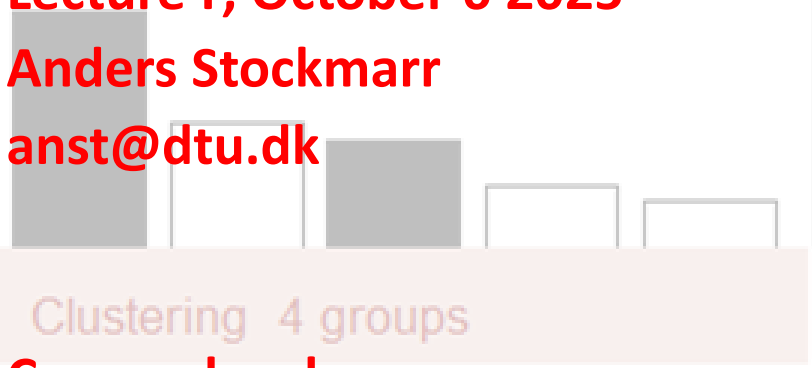


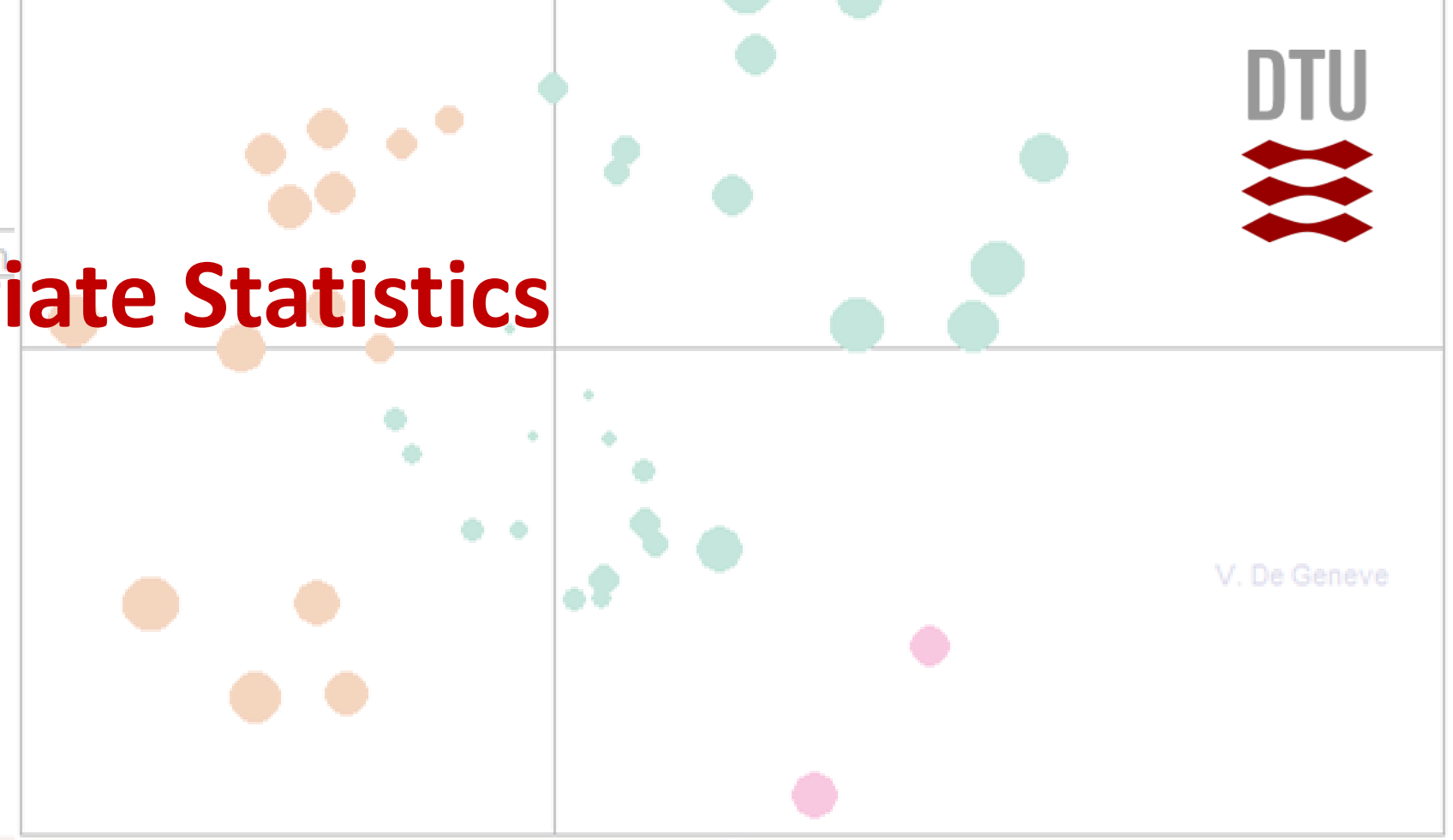
02409 Multivariate Statistics

Lecture F, October 6 2025

Anders Stockmarr

anst@dtu.dk





Clustering 4 groups

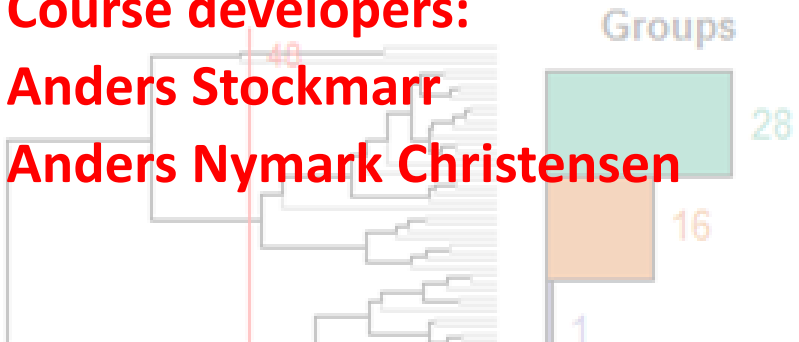
Factor 1 [41%]

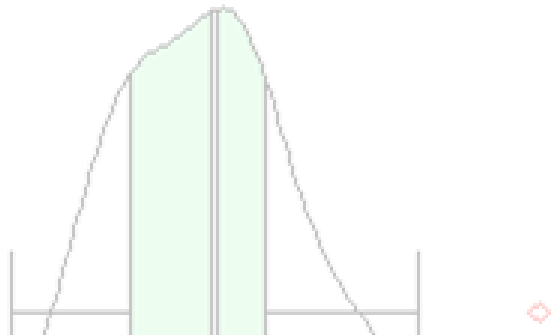
Factor 3 [19%]

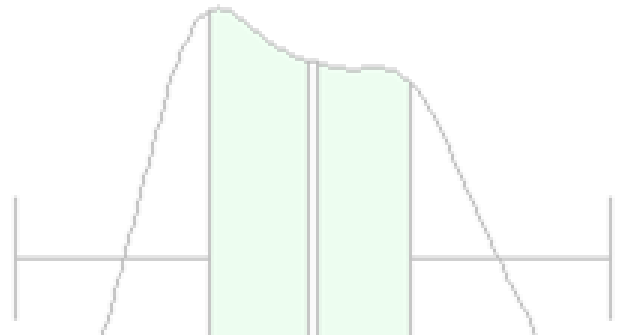
Course developers:

Anders Stockmarr

Anders Nymark Christensen







Agenda

- CCA – recap

- Discrimination
  - The elements of discrimination and classification
  - Bayes and MINIMAX solutions
  - The Normal Case
    - Linear Discriminant Analysis
  - Mahalanobis' distance

Canonical variables and correlations

We consider a random variable

$$Z \sim N_{p+q}(\mu, \Sigma),$$

where $p \leq q$ and Z and the parameters have been partitioned as follows:

$$Z = \begin{bmatrix} Y \\ X \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}.$$

Definition 6.11

Consider Z as above. Then the first ***pair of canonical variables*** is the pair of linear combinations

$$V_1 = a_1^T Y \text{ and } W_1 = b_1^T X$$

each having variance 1 that maximize the correlation $\rho\left(a^T Y, b^T X\right)$ for all (a, b). The maximum correlation ϱ_1 is the ***first canonical correlation***. For $r \leq p$ we define the ***r'th pair of canonical variables*** as the pair of linear combinations

$$V_r = a_r^T Y \text{ and } W_r = b_r^T X ,$$

which each has the variance 1, which are uncorrelated with the previous $r-1$ pairs of canonical variables, and which maximizes the correlation $\rho\left(a^T Y, b^T X\right)$ under those constraints. The maximum correlation ϱ_r is the ***r'th canonical correlation***.

CCA – salesdata

- Sales Performance:
  - Sales Growth
  - Sales Profitability
  - New Account Sales
- Test Scores as a Measure of Intelligence
  - Creativity
  - Mechanical Reasoning
  - Abstract Reasoning
  - Mathematics

Obs	growth	profit	new	create	mech	abs	math
1	93.0	96.0	97.8	9	12	9	20
2	88.8	91.8	96.8	7	10	10	15
3	95.0	100.3	99.0	8	12	9	26
4	101.3	103.8	106.8	13	14	12	29
5	102.0	107.8	103.0	10	15	12	32
6	95.8	97.5	99.3	10	14	11	21
7	95.5	99.5	99.0	9	12	9	25
8	110.8	122.0	115.3	18	20	15	51
9	102.8	108.3	103.8	10	17	13	31
10	106.8	120.5	102.0	14	18	11	39

CCA – salesdata

To what extent is sales performance correlated with measures of intelligence?

We will use canonical correlation analysis to investigate that.

CCA – salesdata

We take

$$\boldsymbol{Y = (growth, profit, new)}$$
$$\boldsymbol{X = (create, mech, abs, mat)}$$

```
salesdata<-read.csv2("Data/Salesdata.csv")
names(salesdata)
[1] "growth" "profit" "new"    "create" "mech"   "abs"    "math"

Y<-salesdata[,1:3]
X<-salesdata[,4:7]
```

CCA – salesdata

Constructing hte matrices E_1, E_2:

```
Sigmayy<-var(Y)
Sigmaxx<-var(X)
Sigmayx<-cov(Y,X)
Sigmaxy<-t(Sigmayx)
```

$$E_1 = \Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$
$$E_2 = \Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$$

E_1: 3×3 matrix

```
E1<-solve(Sigmayy)%*%Sigmayx%*%solve(Sigmaxx)%*%Sigmaxy
E2<-solve(Sigmaxx)%*%Sigmaxy%*%solve(Sigmayy)%*%Sigmayx
E1
          growth    profit         new
growth 0.5241851 0.1481294  0.33088173
profit 0.1273726 0.8098589 -0.05362605
new    0.4729584 0.1452355  0.57317616
```

E_2: 4×4 matrix

```
E2
           create        mech         abs      math
create  0.3936913  0.04077489  0.21007893 0.3414371
mech   -0.1095057  0.20397352 -0.09896763 0.6466981
abs     0.4429526 -0.08171082  0.58337701 0.1264595
math    0.1160912  0.19365385  0.02750066 0.7261784
```

CCA – salesdata

Extracting the canonincal correlations:

We find the canonical correlation coefficients as the eigenvalues for E_1:

```
(my.cor<-sqrt(eigen(E1)$values))
[1] 0.9944827 0.8781065 0.3836057
```

Thus,

$$\hat{\varrho} = (0.9944827, 0.8781065, 0.3836057)$$

Note that

```
sqrt(round(eigen(E2)$values,digits=6))
[1] 0.9944828 0.8781065 0.3836053 0.0000000
```

E_2 is singular, but the first three eigenvalues are equal to the eigenvalues of E_1.

CCA – salesdata

Finding the canonical vectors:

We find the canonical vectors as the eigenvectors for E_1 and E_2 (the first 3):

```
A<-eigen(E1)$vectors
B<-eigen(E2)$vectors[,1:3]
```

Checking that the signs are right, so that we have positive correlations:

```
round(cov(as.matrix(Y)%*%A,as.matrix(X)%*%B),digits=4)
        [,1]    [,2]   [,3]
[1,] 72.9489  0.0000 0.0000
[2,]  0.0000 -4.0158 0.0000
[3,]  0.0000  0.0000 1.7533
```

E_1 and E_2 are singular

Correcting (arbitrarily choosing `A`):

```
A[,2]<--A[,2]
```

CCA – salesdata

Standardizing the canonical vectors:

We must have $A^T \Sigma_{yy} A = I$, $B^T \Sigma_{xx} B = I$. Current values:

```
round(t(A)%*%Sigmayy%*%A,digits=4)
        [,1]   [,2]   [,3]
[1,] 95.6643 0.0000 0.0000
[2,]  0.0000 6.8621 0.0000
[3,]  0.0000 0.0000 3.3337

round(t(B)%*%Sigmaxx%*%B,digits=4)
        [,1]   [,2]   [,3]
[1,] 56.2462 0.0000 0.0000
[2,]  0.0000 3.0479 0.0000
[3,]  0.0000 0.0000 6.2665
```

Diagonal matrix but not the identity. Reason: eigenvectors are given as **unit vectors**. We standardize:

```
A<-A%*%diag(1/sqrt(diag(t(A)%*%Sigmayy%*%A)))
B<-B%*% diag(1/sqrt(diag(t(B)%*%Sigmaxx%*%B)))
```

- Because we multiply with a diagonal matrix, we don't change the direction of the vectors, but only the length. Compare with the approach in the book page 383.

CCA – salesdata

The canonical vectors:

In standardized form

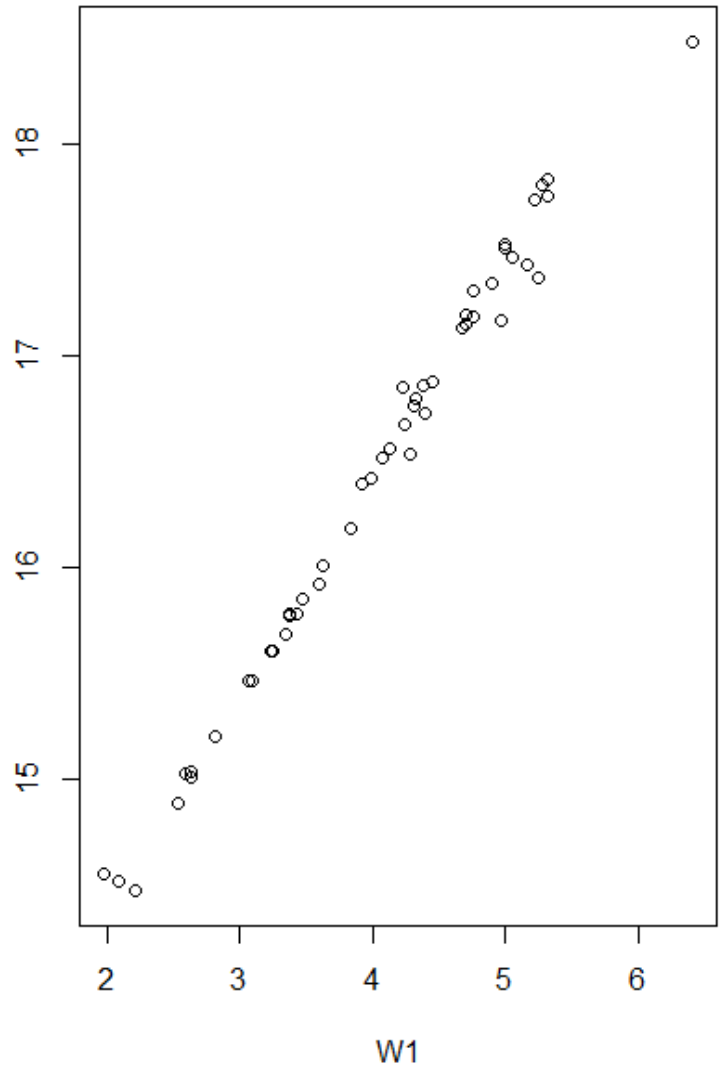
```
A
           [,1]       [,2]       [,3]
[1,] 0.06237788 -0.1740703 -0.3771529
[2,] 0.02092564  0.2421641  0.1035150
[3,] 0.07825817 -0.2382940  0.3834151
B
           [,1]        [,2]        [,3]
[1,] 0.06974814 -0.19239132  0.24655659
[2,] 0.03073830  0.20157438 -0.14189528
[3,] 0.08956418 -0.49576326 -0.28022405
[4,] 0.06282997  0.06831607  0.01133259
```

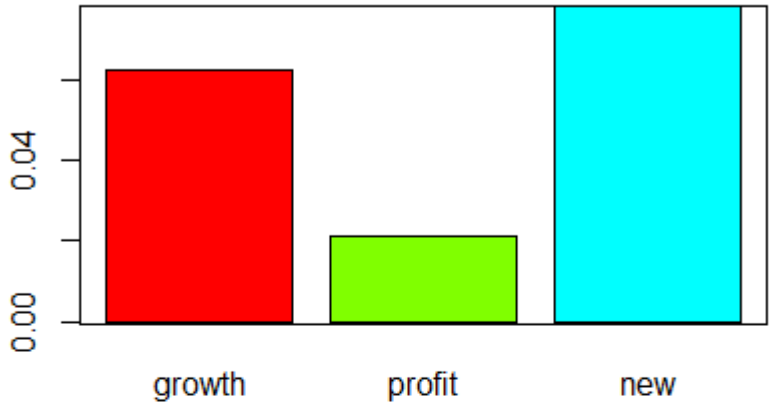
The linar combinations of sales performance and intelligence measures that correlates the most:

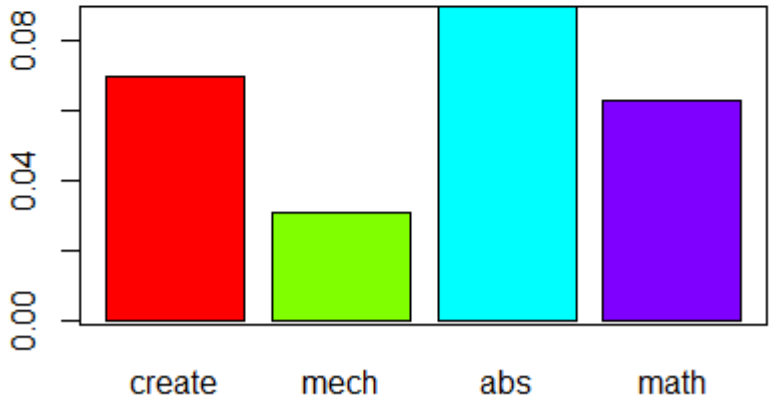
$$V_1 = 0.06 * growth + 0.02 * profit + 0.08 * new$$

$$W_1 = 0.07 * create + 0.03 * mech + 0.09 * abs + 0.06 * math$$

CCA – salesdata







- New account sales and sales growth correlate with measures of intelligence, lesser so for sales profitability.
- In particular Abstract reasoning, Creativity and to some extent Mathematics correlate with sales performance, less so for mechanical thinking.

CCA – salesdata

Testing: Assume that (Y, X) are normally distributed.

$$V\begin{pmatrix} Y \\ X \end{pmatrix} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}, \qquad Cor\begin{pmatrix} V \\ W \end{pmatrix} = \begin{pmatrix} I_3 & \begin{pmatrix} \varrho_1 & & \\ & \varrho_2 & \\ & & \varrho_3 \end{pmatrix} \\ * & I_3 \end{pmatrix}$$

$rank(\Sigma_{yx}) \leq p$: Testing for no canonical correlation corresponds to testing that $V\begin{pmatrix} Y \\ X \end{pmatrix}$ is a diagonal matrix.

- Test for no canonical correlation:

$$\mathrm{H}_0\colon \Sigma = V\begin{pmatrix} Y \\ X \end{pmatrix} = \begin{pmatrix} \Sigma_{vv} & \\ & \Sigma_{ww} \end{pmatrix} \quad \text{vs.} \quad \mathrm{H}\colon \Sigma = V\begin{pmatrix} Y \\ X \end{pmatrix} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}$$

Test statistic the p canonical correlations are 0:

$$Q_p = LR^{2/n} = \frac{\det(\hat{\Sigma})}{det(\hat{\Sigma}_{yy})det(\hat{\Sigma}_{xx})}$$

CCA – salesdata

$$\boldsymbol{V(AX) = AV(X)A^T}$$

$$Q = \frac{\det(\hat{\Sigma})}{det(\hat{\Sigma}_{yy})det(\hat{\Sigma}_{xx})}$$

Note that

$$V\begin{pmatrix} Y - \Sigma_{yx}\Sigma_{xx}^{-1}X \\ X \end{pmatrix} = \begin{pmatrix} I & -\Sigma_{yx}\Sigma_{xx}^{-1} \\ 0 & I \end{pmatrix}\begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}\begin{pmatrix} I & 0 \\ -\Sigma_{xx}^{-1}\Sigma_{xy} & I \end{pmatrix} = \begin{pmatrix} \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} & 0 \\ 0 & \Sigma_{xx} \end{pmatrix} = \begin{pmatrix} \Sigma_{y|x} & 0 \\ 0 & \Sigma_{xx} \end{pmatrix}$$

and that

\begin{pmatrix} Y - Sigma_yx Sigma_xx^-1 X \\ X \end{pmatrix = A \begin{pmatrix} Y \\ X \end{pmatrix}

$$det\begin{pmatrix} I & -\Sigma_{yx}\Sigma_{xx}^{-1} \\ 0 & I \end{pmatrix} = 1$$

A = \begin{pmatrix} I & -\Sigma_yx Sigma_xx^-1 \\ 0 & I \end{pmatrix}

V(AZ) = AV(Z)A^T

Thus,

$$det(\Sigma) = det(\Sigma_{y|x})det(\Sigma_{xx})$$

and

$$Q_p = \frac{\det(\hat{\Sigma})}{det(\hat{\Sigma}_{yy})det(\hat{\Sigma}_{xx})} = \frac{det(\hat{\Sigma}_{y|x})det(\hat{\Sigma}_{xx})}{det(\hat{\Sigma}_{yy})det(\hat{\Sigma}_{xx})} = \frac{det(\hat{\Sigma}_{y|x})}{det(\hat{\Sigma}_{yy})} = det(\hat{\Sigma}_{yy}^{-1}\hat{\Sigma}_{y|x}) = det(I - \hat{\Sigma}_{yy}^{-1}\hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}) = det(I - \hat{E}_1)$$

$$= \det\left(I - \begin{pmatrix} \hat{\varrho}_1^2 & & \\ & \ddots & \\ & & \hat{\varrho}_p^2 \end{pmatrix}\right) = \prod_{i=1}^{p}(1 - \hat{\varrho}_i^2)$$

follows.

CCA – salesdata

$$Q_0 = \frac{\det(\hat{\Sigma})}{det(\hat{\Sigma}_{vv})det(\hat{\Sigma}_{ww})} = \prod_{i=1}^{p}(1-\hat{\varrho}_i^2)$$

Note that $rank(\Sigma_{yx}) = rank(E_1) = \#\ nonzero\ canonical\ correlations$. From the estimation procedure it is immediate that the test statistic for

$$\mathrm{H}_0\colon\ \Sigma = V\begin{pmatrix}Y\\X\end{pmatrix} = \begin{pmatrix}\Sigma_{vv} & \\ & \Sigma_{ww}\end{pmatrix} \quad \text{vs.} \quad \mathrm{H}_\mathrm{r}\colon \Sigma = V\begin{pmatrix}Y\\X\end{pmatrix} = \begin{pmatrix}\Sigma_{yy} & \Sigma_{yx}\\ \Sigma_{xy} & \Sigma_{xx}\end{pmatrix}\ with\ rank(\Sigma_{yx}) = r$$

is maximizing $det(I - \hat{E}_1)$ under the assumption that $rank(E_1) = r$, ie.

$$\prod_{i=p-r+1}^{p}(1-\hat{\varrho}_i^2)$$

Thus, from simple division, the test statistic for that r canonical correlations are equal to 0:

$$\mathrm{H}_\mathrm{r}\colon \Sigma = V\begin{pmatrix}Y\\X\end{pmatrix} = \begin{pmatrix}\Sigma_{yy} & \Sigma_{yx}\\ \Sigma_{xy} & \Sigma_{xx}\end{pmatrix}\ with\ rank(\Sigma_{yx}) = p - r \qquad vs \qquad \mathrm{H}:\ \Sigma = V\begin{pmatrix}Y\\X\end{pmatrix} = \begin{pmatrix}\Sigma_{yy} & \Sigma_{yx}\\ \Sigma_{xy} & \Sigma_{xx}\end{pmatrix}$$

is

$$Q_r = \prod_{i=1}^{r}(1-\hat{\varrho}_i^2)$$

CCA – salesdata

Testing

$\boldsymbol{dim(\Sigma_{xx}) = q}$

$\boldsymbol{dim(\Sigma_{yy}) = p}$

$$H_0\colon \Sigma = \begin{pmatrix} \Sigma_{yy} & \\ & \Sigma_{xx} \end{pmatrix} \quad vs. \quad H\colon \Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}$$

is

$$Q_0 = \frac{\det(\hat{\Sigma})}{det(\hat{\Sigma}_{yy})det(\hat{\Sigma}_{xx})}$$

Under H_0, Q_0 is Wilks distributed, $\Lambda\,(p, n-1-q, q)$.

df from unbiased estimation of variances

CCA – salesdata

Heuristics for the df for Q_r:

Let the columns of $W^{\perp}$ span the orthogonal complement of W. $W^{\perp}$ will have $q-p$ columns.

Taking $Y' = V_{p-r+1;\ldots;p}$ (r variables), $X' = \left(W_{p-r+1;\ldots;p}; W^{\perp}\right)$ (q-p+r variables),

then

$$cor(Y', X') = \begin{pmatrix} \varrho_{p-r+1} & & 0 & \\ & \ddots & & \ddots \\ & & \varrho_p & & 0 \end{pmatrix}$$

Then H_r vs. H corresponds to testing that $V\begin{pmatrix} Y' \\ X' \end{pmatrix}$ is a diagonal matrix, with blocks of dimensions $r \times r, (q-p+r) \times (q-p+r)$.

CCA – salesdata

Hypothesis	$Q_{1:3}$	Value	Distribution
H_3: $\varrho_1 = \varrho_2 = \varrho_3 = 0$	$(1-\hat{\varrho}_1^2)(1-\hat{\varrho}_2^2)(1-\hat{\varrho}_3^2)$	0.8528	$\wedge(3, n-5, 4)$
H_2: $\varrho_2 = \varrho_3 = 0$	$(1-\hat{\varrho}_2^2)(1-\hat{\varrho}_3^2)$	0.1952	$\wedge(2, n-5, 3)$
H_1: $\varrho_3 = 0$	$(1-\hat{\varrho}_3^2)$	0.0021	$\wedge(1, n-5, 2)$

$$\wedge(r, n-1-q, q-p+r), r = 1, \ldots, 3:$$

Wilks' distribution. Super complicated, and not tabled in base R. Wilks' distributions have **small values critical** for the test.

- Therefore, we transform ourselves to lesser complicated distributions.
- Or use the approximative Barlett's test (lecture E).

CCA – salesdata

Λ	f_1	f_2	$F(f_2, f_1)$	F Exact or approximative?
$\Lambda\,(3, n-5, 4)$	12	$\nu(n-5)-5$	$\frac{1-\Lambda^{1/\nu}}{\Lambda^{1/\nu}}\frac{f_1}{f_2}$	Approximative
$\Lambda\,(2, n-5, 3)$	6	$2(n-5)-2$	$\frac{1-\Lambda^{1/2}}{\Lambda^{1/2}}\frac{f_1}{f_2}$	Exact
$\Lambda\,(1, n-5, 2)$	2	$n-5$	$\frac{1-\Lambda}{\Lambda}\frac{f_1}{f_2}$	Exact

Conversion from $\Lambda\,(r, n-1-q, q-p+r)$:

v is also the transformation ……

$$f_1 = r(q-p+r),$$

$$If\ r=1\ take\ \nu=1; if\ r>1\ take\ \nu=\sqrt{\frac{f_1^2-4}{r^2+(q-p+r)^2-5}}$$

$$f_2 = \nu(n-1-q)-\frac{f_1}{2}+1$$

$$r=2:\ \nu=2. \qquad r=3:\nu=\sqrt{7}.$$

CCA – salesdata

- **Barlett's test:**
- In general,

$$-\left(n-\frac{p+q+3}{2}\right)log(Q_i)\sim\chi^2_{r(q-p+r)}$$

The same df as f_1 on the previous slide. Note that if $p=q$, the degrees if freedom is equal to r^2;.

CCA – salesdata

Hypothesis	$Q_{3:1}$	Value	f_1	f_2	$F(f_1, f_2)$	p_F	$\chi^2_{2,6,12}$	p_{chisq}
H_3: $\varrho_1 = \varrho_2 = \varrho_3 = 0$	$(1-\hat{\varrho}_1^2)(1-\hat{\varrho}_2^2)(1-\hat{\varrho}_3^2)$	0.0021	12	108.0588	87.39	< 0.001	276.43	< 0.001
H_2: $\varrho_2 = \varrho_3 = 0$	$(1-\hat{\varrho}_2^2)(1-\hat{\varrho}_3^2)$	0.1952	6	88	17.53	< 0.001	73.51	< 0.001
H_1: $\varrho_3 = 0$	$(1-\hat{\varrho}_3^2)$	0.8528	2	45	3.88	0.03	7.16	0.03

- Data do not support deletion of any canonical correlations (p=0.03).

CCA – salesdata

- Same values using `p.asym`:

```
p.asym(my.cor,n,p,q, tstat="Wilks")
Wilks' Lambda, using F-approximation (Rao's F):
                stat    approx df1      df2      p.value
1 to 3:  0.002148472 87.391525  12 114.0588 0.000000e+00
2 to 3:  0.195241267 18.526265   6  88.0000 8.248957e-14
3 to 3:  0.852846693  3.882233   2  45.0000 2.783536e-02
```

Exactly the same values, but without the Chisquare approximation.

CCA – salesdata

Correlation between Y, X, V, W:

	growth	profit	new	create	mech	abs	math	V1	V2	V3	W1	W2	W3
growth	1.00	0.93	0.88	0.57	0.71	0.67	0.93	0.98	0.00	-0.20	0.97	0.00	-0.08
profit	0.93	1.00	0.84	0.54	0.75	0.47	0.94	0.95	0.32	0.01	0.94	0.28	0.00
new	0.88	0.84	1.00	0.70	0.64	0.64	0.85	0.95	-0.19	0.24	0.95	-0.16	0.09
create	0.57	0.54	0.70	1.00	0.59	0.15	0.41	0.63	-0.19	0.25	0.64	-0.22	0.65
mech	0.71	0.75	0.64	0.59	1.00	0.39	0.57	0.72	0.21	-0.03	0.72	0.24	-0.07
abs	0.67	0.47	0.64	0.15	0.39	1.00	0.57	0.64	-0.44	-0.22	0.65	-0.50	-0.57
math	0.93	0.94	0.85	0.41	0.57	0.57	1.00	0.94	0.17	-0.04	0.94	0.20	-0.09
V1	0.98	0.95	0.95	0.63	0.72	0.64	0.94	1.00	0.00	0.00	0.99	0.00	0.00
V2	0.00	0.32	-0.19	-0.19	0.21	-0.44	0.17	0.00	1.00	0.00	0.00	0.88	0.00
V3	-0.20	0.01	0.24	0.25	-0.03	-0.22	-0.04	0.00	0.00	1.00	0.00	0.00	0.38
W1	0.97	0.94	0.95	0.64	0.72	0.65	0.94	0.99	0.00	0.00	1.00	0.00	0.00
W2	0.00	0.28	-0.16	-0.22	0.24	-0.50	0.20	0.00	0.88	0.00	0.00	1.00	0.00
W3	-0.08	0.00	0.09	0.65	-0.07	-0.57	-0.09	0.00	0.00	0.38	0.00	0.00	1.00

Canonocal Correlation analysis

What are the strengths and weaknesses of CCA?

- **Best:**

  Questions concerning the number and nature of mutually independent relations between two sets of variables. For the sales data: 3 dimensions.

- **Mediocre:**

  Questions concerning the degree of overlap or redundancy between two sets of variables.

- **Not Very Well:**

  Questions concerning the similarity between two within-set correlation or covariance matrices.

Discriminant Analysis

- So far we have studied relations between multivariate data through the variance matrix.
- We have seperated parts of the data, and studied correlaitons between subsets.

- But what if the correlation come from a (n unknown) substructure in the data, dividing the study population into discrete groups, each having the same distribution of the data?

- Dicriminant analysis is about modeling and uncovering groupings within the data.

Discrimination and Classification

1. The elements of discrimination and classification
2. Bayes and MINIMAX solutions
3. The Normal Case:
    1. Linear Discriminant Analysis
    2. Quadratic Discriminant Analysis
4. Mahalanobis' distance
5. Estimation of classification errors in LANDSAT data
6. Canonical Discriminant Analysis

First next week

Discriminant Analysis

- Multivariate data in a population with two subgroups π_1, π_2:

$$X_j = \begin{pmatrix} X_{j1} \\ \vdots \\ X_{jp} \end{pmatrix}, j = 1, \dots, n.$$

Density for X_j:

$$f_1(x), if\ X_j \in \pi_1;$$ First population

$$f_2(x), if\ X_j \in \pi_2.$$ Secund population

Decision functions

- We seek a function to determine, based on observational data, if a subject belongs to π_1 or π_2:

$$d: \mathbb{R}^p \rightarrow \{\pi_1, \pi_2\}$$

- Since d can only attain two values, it will necessarily subdivide $\mathbb{R}^p$ into two disjoint subsets R_1, R_2:

$$d(x) = \begin{cases} \pi_1 & if\ x \in R_1 \\ \pi_2 & if\ x \in R_2 \end{cases}$$

- If we choose R_1, we choose d!

Loss

- We shall assume that it is not without consequences if we make a mistake. If X really belongs to π_1, we suffer a **loss** $L(\pi_1, \pi_2)$ if we classify X_j to π_2 via the decision function.

From		Classify as	
		π_1	π_2
Nature	π_1	0	$L(\pi_1, \pi_2) = L_{12}$
	π_2	$L(\pi_2, \pi_1) = L_{21}$	0

- A good decision function **minimizes the loss**.

A Bayesean framework

- Let us suppose that we have an idea about the distribution of the subpopulations π_1, π_2. Maybe we have seen other observations from the same subpopulations.
- Let us therefore assume that, prior to any observations if the values of X_j, there is a probability distribution Π on the subclasses in the population, such that for all individuals it holds that

$$P(\Pi = \pi_1) = g(\pi_1) = 1 - P(\Pi = \pi_2) = 1 - g(\pi_2)$$

- In the following, we will write p_i for $g(\pi_i), i = 1,2$.

Posterior Probability

We define the *posterior probability*

$$k(\pi_i|x) := \frac{p_i f_i(x)}{p_1 f_1(x) + p_2 f_2(x)}, i = 1,2$$

density function = P (X = x I \pi = j) = ~ P (X \in x+ dx I \pi = j) dx

P(\pi =3,142 i I X = x) = P(\pi = i , X = x) / P(X = x)
= P(\pi = i , X = x) / P (X = x, \pi = 1) + P (X = x, \pi = 2)
= P(\pi = i , X = x) / [P (X = x, \pi = 1) P (\pi = 1) + P (X = x, \pi = 2) + P (X = x, \pi = 2) P (\pi = 2)]
= f_1(x i P_1) / f_1(x) P_i + f_2 (x) P_2)

- In contrast to the a priori distribtion, the posterior distribution k holds the probability of the classes AFTER observation of the data X_j.

Expected loss

- The class $\Pi(j)|X_j = x$ now has a distribution.
- Then the loss, $L\left(\Pi(i), d\left(X_j\right)\right)|X_j = x$ has a distribution.
- Let us calculate the average loss given $X_j = x$:

\ pi is binary, it can only contain 1 or 2

$$
\begin{aligned}
& E\left(L\left(\Pi(j), d\left(X_j\right)\right)\middle|X_j = x\right) \\
&= E\left(L\left(\Pi(j), d(x)\right)\middle|X_j = x\right) \\
&= \begin{cases} L(\pi_2, \pi_1)k(\pi_2|x) & if\ x \in R_1 \\ L(\pi_1, \pi_2)k(\pi_1|x) & if\ x \in R_2 \end{cases}
\end{aligned}
$$

\pm [2] = \Sigma z P(Z=z)

The loss

L(\pi_2, \pi_1) k (\pi_2 i x) + L(\pi_1, \pi_1) k (\pi_1 I x) if x \in R

L(\pi_1, \pi_1) = 0

The Bayes solution

First population $L(\pi_2, \pi_1)k(\pi_2|x) \quad if\ x \in R_1$

Second population $L(\pi_1, \pi_2)k(\pi_1|x) \quad if\ x \in R_2$

- **Loss minimiza**tion: Choose R_1 to be the the area where the top line loss is less than the bottom line loss:

$$
\begin{aligned}
R_1 &= \{x \in \mathbb{R}^p \mid L(\pi_1, \pi_2)k(\pi_2|x) \leq L(\pi_2, \pi_1)k(\pi_1|x)\} \\
&= \left\{x \in \mathbb{R}^p \,\middle|\, \frac{L_{12}k(\pi_1|x)}{L_{21}k(\pi_2|x)} \geq 1\right\} \\
&= \left\{x \in \mathbb{R}^p \,\middle|\, \frac{L_{12}p_1f_1(x)}{L_{21}p_2f_2(x)} \geq 1\right\} \\
&= \left\{x \in \mathbb{R}^p \,\middle|\, \frac{f_1(x)}{f_2(x)} \geq \frac{L_{21}p_2}{L_{12}p_1}\right\}
\end{aligned}
$$

The Bayes solution

Theorem 5.1

The *Bayes solution* to the classification problem is given by the region

$$R_1 = \left\{ x \;\middle|\; \frac{f_1(x)}{f_2(x)} \geq \frac{L_{21} p_2}{L_{12} p_1} \right\} .$$

Risk

- The **loss** is what we face if we make a mis-classification – the CONSEQUENCE of a misclassification.
- But if the probability of making a misclassification is small, then the consequences will likely not occur.

- Let us focus on **risk**: $probability \times consequence$.
- Instead of mimimizing the loss, we could minimize the risk that we run;
- We can do so by selecting the region R_1 so that the risks in R_1 and R_2 are the same:

Risk

- If $\Pi(j) = \pi_1$ (ie. Individual j belongs to π_1), the expected risk is

$$R = consequence \times probability = L_{12}P(X \in R_2)$$

If $\Pi(j) = \pi_2$ (ie. Individual j belongs to π_2), the expected risk is

$$R = consequence \times probability = L_{21}P(X \in R_1)$$

Minimax solution

Theorem 5.2

The *minimax solution* for the classification problem is given by the region

$$R_1 = \left\{ x \,\middle|\, \frac{f_1(x)}{f_2(x)} \geq c \right\} .$$

where c is determined by

$$L_{12} P\left\{ \frac{f_1(x)}{f_2(x)} < c \,\middle|\, \pi_1 \right\} = L_{21} P\left\{ \frac{f_1(x)}{f_2(x)} \geq c \,\middle|\, \pi_2 \right\} .$$

- It is seen that the last line exactly expresses that $\text{risk} = consequence \times P(misclassification)$ is independent of the true state Π.

The Normal Distribution

Assume that individuals in π_1, π_2 have normally distributed data, $N_p(\mu_1, \Sigma), N_p(\mu_2, \Sigma)$, respectively, with Σ invertible.

Introduce the inner product $<,>_{\Sigma^{-1}}$ and the corresponding norm $||\cdot||_{\Sigma^{-1}}$ on $\mathbb{R}^p$ by

$$< x, y >_{\Sigma^{-1}} = x^T \Sigma^{-1} y$$

$$f_1(x) = \frac{1}{(2\pi)^{p/2}\sqrt{det(\Sigma)}} exp(-)$$

The Normal Distribution

Densities:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}\sqrt{det(\Sigma)}} exp\left(-\frac{1}{2}\|x-\mu_i\|^2_{\Sigma^{-1}}\right), i = 1,2.$$

Thus:

$$\frac{f_{1(x)}}{f_{2(x)}} = \exp\left(-\frac{1}{2}\left(\|x-\mu_1\|^2_{\Sigma^{-1}} - \|x-\mu_2\|^2_{\Sigma^{-1}}\right)\right)$$

The Normal Distribution

Second polynomia by taking the log on both side

$$\frac{f_{1(x)}}{f_{2(x)}} \geq c \Leftrightarrow -\|x-\mu_1\|^2_{\Sigma^{-1}} + \|x-\mu_2\|^2_{\Sigma^{-1}} \geq 2\log(c)$$

$$\Leftrightarrow\ 2 < x, \mu_1 - \mu_2 >_{\Sigma^{-1}} -\|\mu_1\|^2_{\Sigma^{-1}} + \|\mu_2\|^2_{\Sigma^{-1}} \geq 2\log(c)$$

Theorem 5.4

Let $\pi_1 \sim N(\mu_1, \Sigma)$ and $\pi_2 \sim N(\mu_2, \Sigma)$.Then we have

$$\frac{f_1(x)}{f_2(x)} \geq c \Leftrightarrow x^T\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 \geq \log c$$

$$\Leftrightarrow \left[x^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right] - \left[\ x^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2\right] \geq \log c\,.$$

The Normal Distribution

Note that the decision function is a *separating hyperplane*:

The hyperplane satisfies

$$2 < x, \mu_1 - \mu_2 >_{\Sigma^{-1}} - \|\mu_1\|^2_{\Sigma^{-1}} + \|\mu_2\|^2_{\Sigma^{-1}} = 2\log(c)\,,$$

ie. it has the form $x = \tilde{c}\nu + \beta$, with $\nu = \Sigma^{-1}(\mu_1 - \mu_2)$, and $\beta \perp \nu$ with β varying. The hyperplane has dimension $p - 1$.

Linear Discriminant Functions

- Bayes: $\mathrm{R}_1 = \left\{ x \in \mathbb{R}^p \middle| \frac{f_1(x)p_1}{f_2(x)p_2} \geq \frac{L_{21}}{L_{12}} \right\}$
- Minimax: $\mathrm{R}_1 = \left\{ x \in \mathbb{R}^p \middle| \frac{f_1(x)}{f_2(x)} \geq c \right\}$

Linear discriminant functions:

Bayes: $$< x, \mu_i >_{\Sigma^{-1}} - \frac{1}{2} \|\mu_i\|^2_{\Sigma^{-1}} + \log(p_i), \ i = 1,2;$$

Minimax: $$< x, \mu_i >_{\Sigma^{-1}} - \frac{1}{2} \|\mu_i\|^2_{\Sigma^{-1}}, \ i = 1,2.$$

Linear Discriminator

Linear discriminator between π_1 and π_2:

$$< x, \mu_1 - \mu_2 >_{\Sigma^{-1}} - \frac{1}{2} \|\mu_1\|^2_{\Sigma^{-1}} + \frac{1}{2} \|\mu_2\|^2_{\Sigma^{-1}} + \log(c)$$

Separating hyperplane

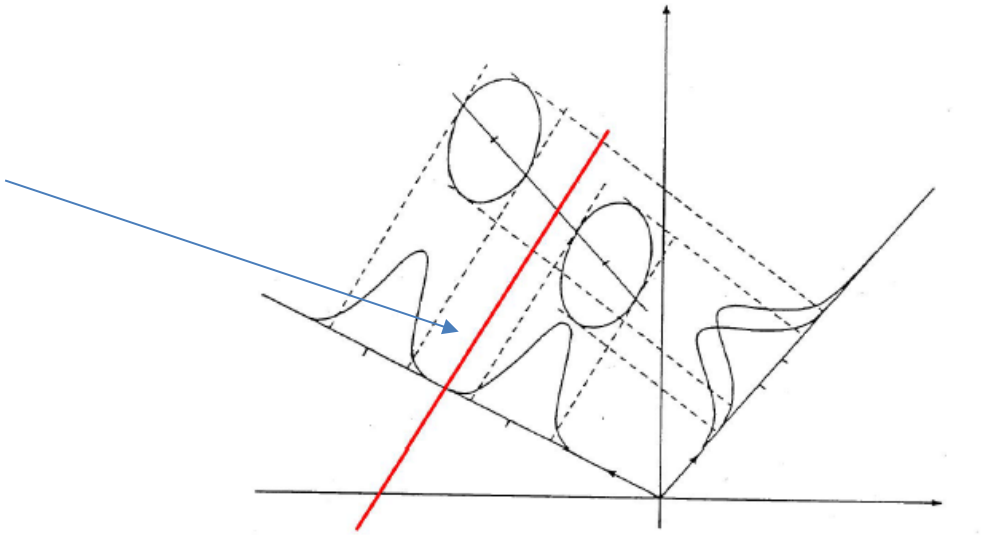


Figure 5.4 – Classification example

Linear Discriminator

Linear discriminator between π_1 and π_2:

$$< x, \mu_1 - \mu_2 >_{\Sigma^{-1}} - \frac{1}{2}\|\mu_1\|^2_{\Sigma^{-1}} + \frac{1}{2}\|\mu_2\|^2_{\Sigma^{-1}} - \log(c)$$

Separating hyperplane

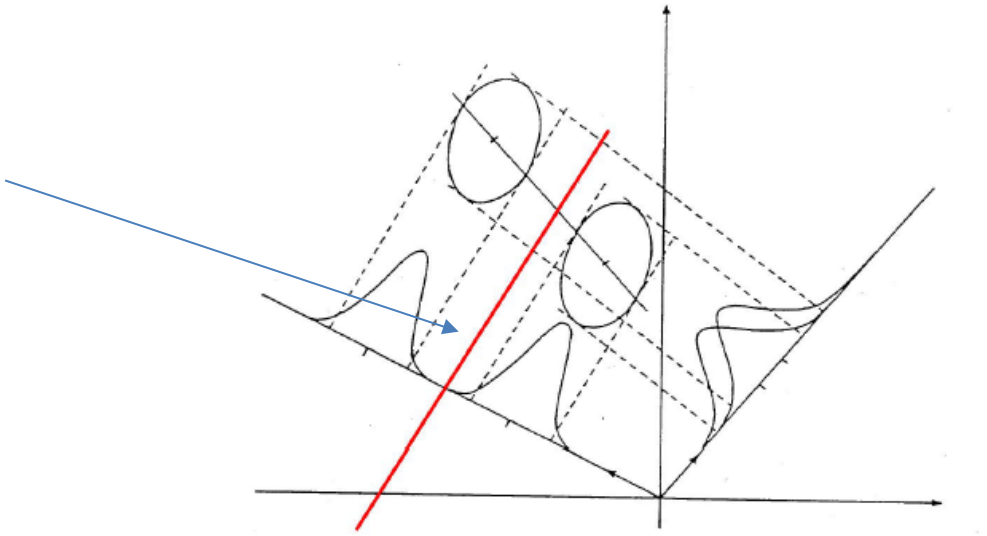


Figure 5.4 – Classification example

Selecting the Optimal Discriminator

- Take

$$\delta = \Sigma^{-1}(\mu_1 - \mu_2)$$

\MU = Sigma^1/2 d

\eta = Sigma^1/2 (\mu_1 - \mu_2

g(\MU) = [< \eta , \MU > / II \MU II]^2

\MU = < \MU, \eta > / II \eta^2 II * \eta + \beta
= \aplha II \eta II^-1 \eta + \beta

\alpha^2 0 II \betaII^2 = 1

< \eta , 1/ II \MU II \MU > = \aplha II \eta II^-1 \eta

\MU = c \Sigma ^-1/2 (\mu_1-\mu_2)

\delta = \Sigma^-1/2 \MU = c \Sigma^-1 (\mu_1 - \mu2)

Theorem 5.6

The vector δ has the property that it maximizes the function

$$g(d) = \frac{[\mathrm{E}_1(X^T d) - \mathrm{E}_2(X^T d)]^2}{\mathrm{V}(X^T d)} = \frac{\left[(\mu_1 - \mu_2)^T d\right]^2}{d^T \Sigma d}$$

Distribution of the Linear Discriminator

Theorem 5.8

We consider the random variable defined by the linear discriminator (omitting the term $-\log c$), i.e.

$$Z = X^T \Sigma^{-1} (\mu_1 - \mu_2) - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 .$$

Then

$$Z \sim \begin{cases} \mathrm{N}\left(+\frac{1}{2} \|\mu_1 - \mu_2\|^2_{\Sigma^{-1}},\ \|\mu_1 - \mu_2\|^2_{\Sigma^{-1}} \right) & \text{if } \pi_1 \text{ is true} \\ \mathrm{N}\left(-\frac{1}{2} \|\mu_1 - \mu_2\|^2_{\Sigma^{-1}},\ \|\mu_1 - \mu_2\|^2_{\Sigma^{-1}} \right) & \text{if } \pi_2 \text{ is true} \end{cases} .$$

Z = AX + B

E[Z] = A E[X]+B

V(Z) = AV(X)A^T
= (\mu_1 - \mu_2)^T \Sigma^-1 \Sigma \Sigma^-1 (\mu_1 - \mu_2)
= (\mu_1 - \mu_2)^T \Sigma^-1 (\mu_1 - \mu_2)
= || \mu_1 - \mu_1 ||^2_\Sigma

E[X] = \cases \mu_1 \pi_1 true, \mu_2 \pi_2 true

Example

Suppose that

$$\pi_1 \leftrightarrow N(\mu_1, \Sigma), \qquad \pi_2 \leftrightarrow N(\mu_2, \Sigma)$$

with

$$\mu_1 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, \mu_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \Sigma^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$$

Losses:

From		Classify as	
		π_1	π_2
Nature	π_1	0	$L_{12} = 2$
	π_2	$L_{21} = 1$	0

We seek the minimax solution.

Example

$$\mu_1 = \begin{pmatrix}4\\2\end{pmatrix}, \mu_2 = \begin{pmatrix}1\\1\end{pmatrix}, \Sigma = \begin{pmatrix}1 & 1\\1 & 2\end{pmatrix}, \Sigma^{-1} = \begin{pmatrix}2 & -1\\-1 & 1\end{pmatrix}, L_{12} = 2, L_{21} = 1$$

Note that

$$\|\mu_1 - \mu_2\|^2_{\Sigma^{-1}} = \begin{bmatrix}3 & 1\end{bmatrix}\begin{bmatrix}2 & -1\\-1 & 1\end{bmatrix}\begin{bmatrix}3\\1\end{bmatrix} = 13$$

Must find c so that

$$2P\left(\frac{f_1(x)}{f_2(x)} < c \middle| \pi_1\right) = P\left(\frac{f_1(x)}{f_2(x)} \geq c \middle| \pi_2\right)$$

Example

$$\|\mu_1 - \mu_2\|^2_{\Sigma^{-1}} = 13$$

Theorem 5.8:

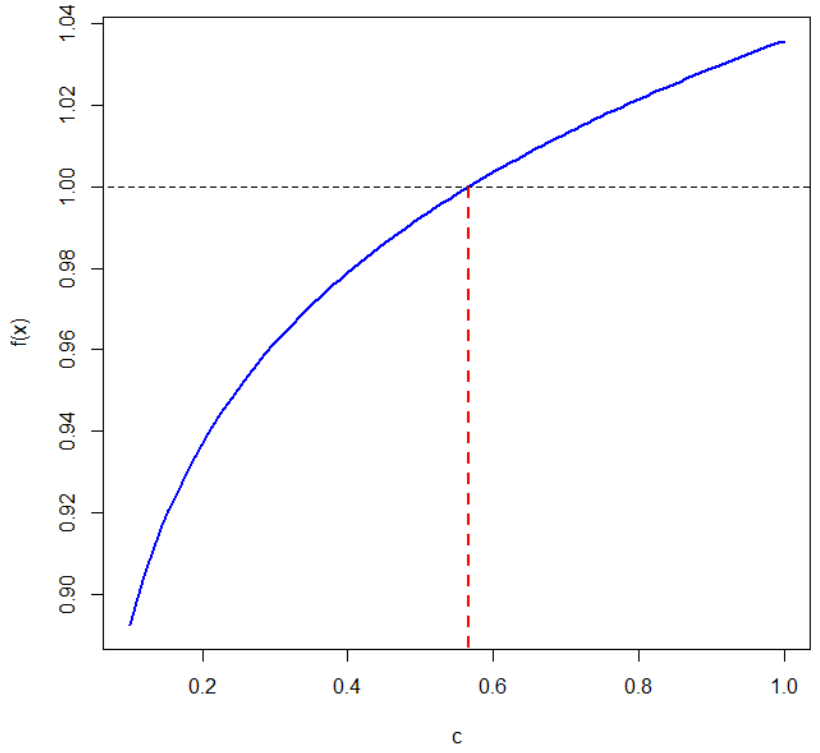
$$2P\left(\frac{f_1(x)}{f_2(x)} < c \middle| \pi_1\right) = P\left(\frac{f_1(x)}{f_2(x)} \geq c \middle| \pi_2\right) \Leftrightarrow$$
$$2P(Z < \log(c)|\pi_1) = P(Z \geq \log(c)|\pi_2) \Leftrightarrow$$
$$2P\left(N\left(\frac{1}{2}\cdot 13,13\right) < \log(c)\right) = P\left(N\left(-\frac{1}{2}\cdot 13,13\right) \geq \log(c)\right) \Leftrightarrow$$
$$2P\left(N(0,1) < \frac{\log(c) - 6.5}{\sqrt{13}}\right) = P\left(N(0,1) \geq \frac{\log(c) + 6.5}{\sqrt{13}}\right) \Leftrightarrow$$
$$2\Phi\left(\frac{\log(c) - 6.5}{\sqrt{13}}\right) = 1 - \Phi\left(\frac{\log(c) + 6.5}{\sqrt{13}}\right) \Leftrightarrow$$
$$2\Phi\left(\frac{\log(c) - 6.5}{\sqrt{13}}\right) + \Phi\left(\frac{\log(c) + 6.5}{\sqrt{13}}\right) = 1$$

with Φ the standard normal distribution function.

Example

$$2\Phi\left(\frac{\log(c)-6.5}{\sqrt{13}}\right)+\Phi\left(\frac{\log(c)+6.5}{\sqrt{13}}\right)=1$$

$$f(c)=2\Phi\left(\frac{\log(c)-6.5}{\sqrt{13}}\right)+\Phi\left(\frac{\log(c)+6.5}{\sqrt{13}}\right)$$



We read off that
$c = 0.5666$

Example

- Miclassification probabilities:

If π_1:$\Phi\left(\frac{(\log(0.5666)-6.5)}{\sqrt{13}}\right) = 0.025$

If π_2:$1-\Phi\left(\frac{(\log(0.5666)+6.5)}{\sqrt{13}}\right) = 0.05$

Linear discriminator:

$$< x, \mu_1 - \mu_2 >_{\Sigma^{-1}} - \frac{1}{2}\|\mu_1\|^2_{\Sigma^{-1}} + \frac{1}{2}\|\mu_2\|^2_{\Sigma^{-1}} - \log(c)$$
$$= 5x_1 - 2x_2 - 8.93,$$

Separates at 0.

Example

Linear discriminator:

$$5x_1 - 2x_2 - 8.93,$$

Separates at 0.

Suppose that we observe $X_j = \begin{pmatrix} 9 \\ 0 \end{pmatrix}$. The linear discriminator is

$$45 - 0 - 8.93 = 36.93 > 0.$$

Participant j is therefore classified to π_1.

Discrimination With Unknown Parameters

Assume that individuals in π_1, π_2 have normally distributed data, $N_p(\mu_1, \Sigma), N_p(\mu_2, \Sigma)$, respectively, with Σ invertible. This time, with unknown parameters.

- We use the estimated values

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} = \bar{X}_1, \hat{\Sigma}_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)(X_{1i} - \bar{X}_1)^T$$

$$\hat{\mu}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i} = \bar{X}_2, \hat{\Sigma}_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)(X_{2i} - \bar{X}_2)^T$$

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2\right)$$

Discrimination With Unknown Parameters

- Estimated decision rule (Theorem 5.4):

$$x^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) - \frac{1}{2}\|\hat{\mu}_1\|_{\hat{\Sigma}^{-1}} + \frac{1}{2}\|\hat{\mu}_2\|_{\hat{\Sigma}^{-1}}$$

- For large sample sizes we can use the distribution of Z from Theorem 5.8 to estimate the separating hyperplane.

Discrimination With Unknown Parameters

- Theorem 5.8 utilizes the so-called *Mahanalobis distance*

$$\Delta^2_{\Sigma^{-1}}(\mu_1, \mu_2) = \|\mu_1 - \mu_2\|^2_{\Sigma^{-1}}$$

We define the *empirical Mahanalobis distance*

$$\mathrm{D}^2_{\hat{\Sigma}^{-1}}(\hat{\mu}_1, \hat{\mu}_2) = \|\hat{\mu}_1 - \hat{\mu}_2\|^2_{\hat{\Sigma}^{-1}}$$

D^2 is linked in distribution to Hotellings T^2, in that

$$D = \frac{n_1 + n_2}{n_1 n_2} T^2$$

Testing the Optimal Discriminator

- The optimal discriminator is

$$\hat{\delta} = \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$$

found from optimizing

$$\hat{g}(d) = \frac{\left((\hat{\mu}_1 - \hat{\mu}_2)^T d\right)^2}{\|d\|^2_{\hat{\Sigma}^{-1}}}$$

Optimized value:

$$\hat{g}(\hat{\delta}) = \frac{\left((\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)\right)^2}{(\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)} = (\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) = D^2$$

Testing the Optimal Discriminator

Define for any d_0

$$D_0^2 = \hat{g}(d_0) = \frac{\left((\hat{\mu}_1 - \hat{\mu}_2)^T d_0\right)^2}{\|d_0\|^2_{\hat{\Sigma}^{-1}}}$$

Theorem 5.13

The statistic

$$Z = \frac{n_1 + n_2 - p - 1}{p - 1} \cdot \frac{n_1 n_2 (D^2 - D_0^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_0^2}$$

may be used in testing the hypothesis that the linear projection determined by d_0 is the best discriminator against all alternatives. Z is $F(p-1, n_1+n_2-p-1)$ –distributed under the hypothesis and large values of Z are critical, i.e., the critical region is

$$C = \{x_{11}, \cdots, x_{1n_1}, x_{21}, \cdots, x_{2n_2} \mid z > F(p-1, n_1+n_2-p-1)_{1-\alpha}\}$$

if we use the significance level α. Here z is the observed value of Z.

Exercises

- 6.1: General introduction to Linear Discriminant Analysis
- 6.2: Test of means, test for further information, classification
- 6.3: Test of means, classification