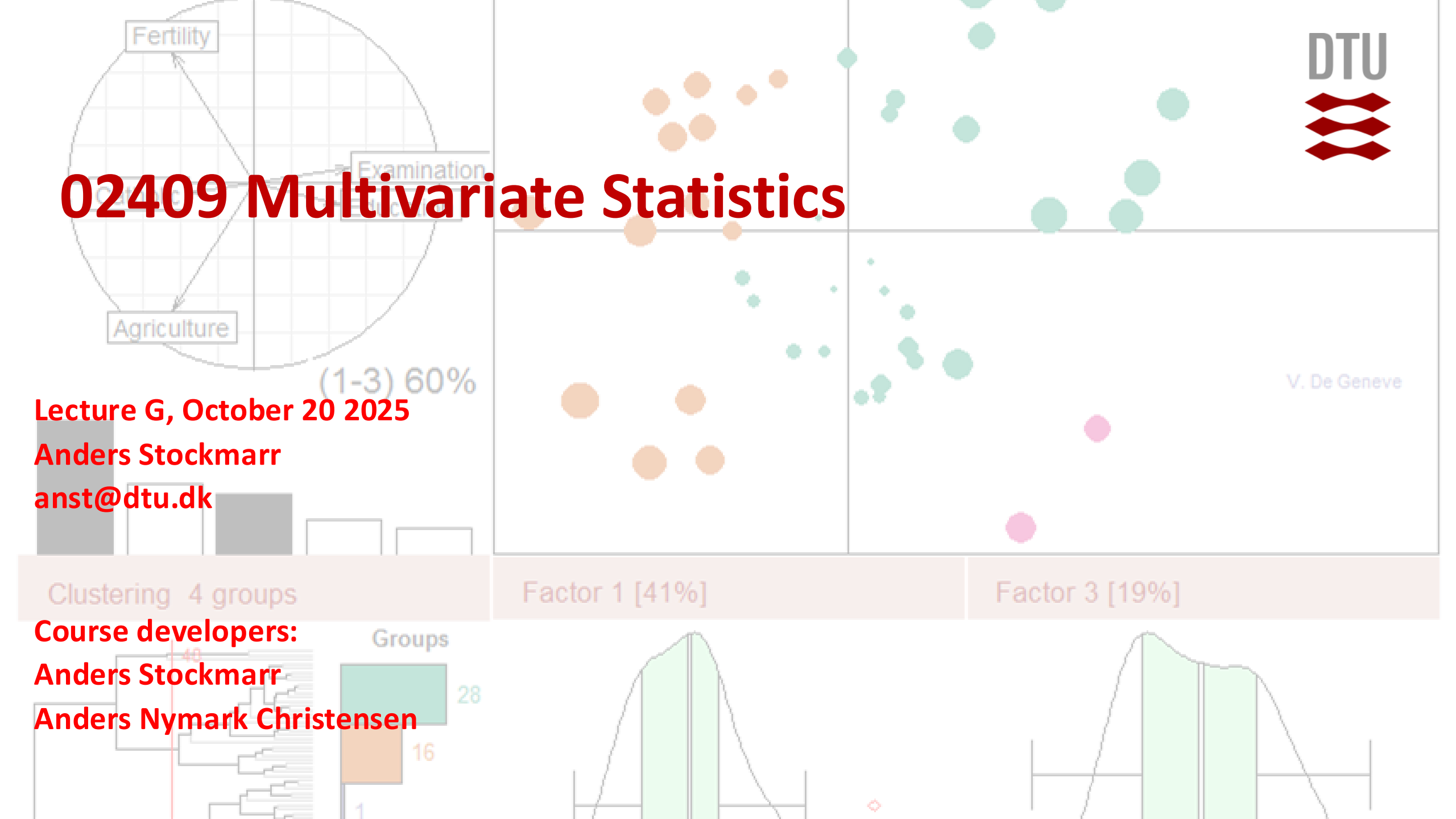


02409 Multivariate Statistics

Lecture G, October 20 2025

Anders Stockmarr

anst@dtu.dk

Course developers:

Anders Stockmarr

Anders Nymark Christensen

Agenda

- Concepts in discriminant analysis:
  - MINIMAX
  - LDF
  - QDF
- Canonical Discriminant analysis
  - Theory
  - Examples of use
- Introduction to GLM
  - The General Linear Model

Discriminant Analysis

- 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!

The Bayes and Minimax Solutions

Theorem 5.1

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

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

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

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

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

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

Prerequisite: π_1 and π_2 individuals have the **SAME** variance Σ

$$c = \frac{L_{21}p_2}{L_{12}p_1} \ / \ c = \frac{L_{21}}{L_{12}}$$

Example

Suppose that

$$\pi_1 \leftrightarrow N(\mu_1, \Sigma_1), \qquad \pi_2 \leftrightarrow N(\mu_2, \Sigma_2)$$

with

$$\mu_1 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, \mu_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \Sigma_1 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \Sigma_1^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_2^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Note that $det(\Sigma_1) = det(\Sigma_2) = 1$ so that

$$f_i(x) = \frac{1}{2\pi} exp\left(-\frac{1}{2}\|x - \mu_i\|^2_{\Sigma_i^{-1}}\right), \qquad i = 1,2.$$

Example

$$\mu_1 = \begin{pmatrix}4\\2\end{pmatrix}, \mu_2 = \begin{pmatrix}1\\1\end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix}1 & 1\\1 & 2\end{pmatrix}, \Sigma_1^{-1} = \begin{pmatrix}2 & -1\\-1 & 1\end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}, \Sigma_2^{-1} = \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$$

$$\frac{f_1(x)}{f_2(x)} \geq c \Leftrightarrow -(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) \geq 2\log(c)$$

$$\Leftrightarrow -\left(\begin{bmatrix}x_1\\x_2\end{bmatrix} - \begin{bmatrix}4\\2\end{bmatrix}\right)^T \begin{pmatrix}2 & -1\\-1 & 1\end{pmatrix}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix} - \begin{bmatrix}4\\2\end{bmatrix}\right) + \left(\begin{bmatrix}x_1\\x_2\end{bmatrix} - \begin{bmatrix}1\\1\end{bmatrix}\right)^T \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix} - \begin{bmatrix}1\\1\end{bmatrix}\right) \geq 2\log(c)$$

$$\Leftrightarrow -x_1^2 + 2x_1x_2 + 10x_1 - 6x_2 - 18 \geq 2\log(c)$$

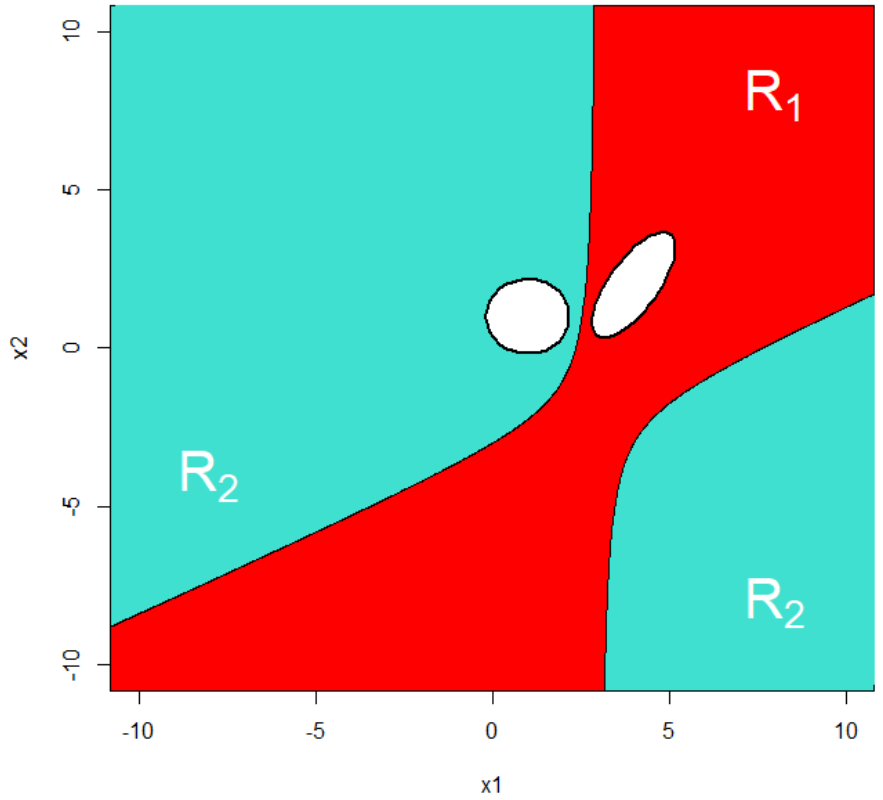
The discriminator is no longer linear in x!

Example

- Choose $c = 1$, equal loss

$$-x_1^2 + 2x_1x_2 + 10x_1 - 6x_2 - 18 \geq 0$$

- The separation of $\mathbb{R}^2$ into R_1 and R_2 is no longer a hyperplane, but the hyperbola given above:



Working with parameters estimated from data

- Asumme again two populations, with $\Sigma_1 = \Sigma_2 = \Sigma$ regular.

The **Linear Discriminator**

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

is estimated by

$$< x, \hat{\mu}_1 - \hat{\mu}_2 >_{\hat{\Sigma}^{-1}} - \frac{1}{2} \|\hat{\mu}_1\|^2_{\hat{\Sigma}^{-1}} + \frac{1}{2} \|\hat{\mu}_2\|^2_{\hat{\Sigma}^{-1}}$$

More than two populations: Distance Between Populations

- We introduce the *Mahanalobis' distance* between populations π_1, π_2 by

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

The regularized deviations of the means in terms of standard deviations.

The *empirical Mahanalobis' distance* is similarly

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

The Mahanalobis' Distance

The empirical Mahanalobis' distance

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

Can be used to test if $\mu_1 = \mu_2$

2-sample Test In One Dimension

$$X_{1i} \sim N(\mu_1, \sigma^2), \qquad X_{2i} \sim N(\mu_2, \sigma^2), i = 1, \dots, n$$

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N\left(0, \frac{2}{n}\sigma^2\right), \qquad 2(n-1)\hat{\sigma}^2 \sim \sigma \cdot \chi^2_{2(n-1)}$$

$$T = \sqrt{\frac{n}{2}} \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\sqrt{\hat{\sigma}^2}} \sim t\big(2(n-1)\big), \qquad T^2 = \frac{n}{2} \frac{(\hat{\mu}_1 - \hat{\mu}_2)^2}{\hat{\sigma}^2} \sim F(1, 2(n-1))$$

2-sample Test In p Dimensions

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N_p\left(0, \frac{2}{n}\Sigma\right), \qquad 2(n-1)\hat{\Sigma} \sim W_p(2(n-1), \Sigma)$$

$$T^2 = \frac{n}{2}\|\hat{\mu}_1 - \hat{\mu}_2\|^2_{\hat{\Sigma}^{-1}} = \frac{n}{2}(\hat{\mu}_1 - \hat{\mu}_2)^T\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2),$$

$$\frac{2(n-1)-p+1}{2(n-1)p}T^2 \sim F(p, 2(n-1)-p+1)$$

- Follows from Lemma 4.1

2-sample Test In p Dimensions

- T^2 when numbers of observations in the groups are uneven:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \|\hat{\mu}_1 - \hat{\mu}_2\|^2_{\hat{\Sigma}^{-1}},$$

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \sim F(p, n_1 + n_2 - p - 1)$$

- Follows from Lemma 4.1

2-sample Test In p Dimensions

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

Theorem 5.12

Using the significance level α, the critical area for a test of the hypothesis $\mu_1 = \mu_2$ against all alternatives becomes

$$C = \left\{ x_{11}, \cdots, x_{1n_1}, x_{21}, \cdots, x_{2n_2} \;\middle|\; \frac{n_1+n_2-p-1}{p(n_1+n_2-2)} \cdot \frac{n_1 n_2}{n_1+n_2} d^2 > F(p, n_1 + n_2 - p - 1)_{1-\alpha} \right\}$$

Here d^2 is the observed value of D^2.

Discrimination Between Several Populations

- **Bayes Solution:**
- Populations: $\boldsymbol{\pi_1, \ldots, \pi_k}$.
- Prior distribution: $\boldsymbol{g(\pi_i) = p_i}, i = 1, \ldots, k$.
- Loss function:

		Classify as				
		π_1	$\cdots$	π_i	$\cdots$	π_k
Nature	π_1	0	$\cdots$	L_{1i}	$\cdots$	L_{1k}
	$\vdots$	$\vdots$				$\vdots$
	π_ν	$L_{\nu 1}$	$\cdots$	$L_{\nu i}$	$\cdots$	$L_{\nu k}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	π_k	L_{k1}	$\cdots$	L_{ki}	$\cdots$	0

Discrimination Between Several Populations

- Posterior distribution of the population Π:

$$k(\pi_v|X=x) = P(\Pi=\pi_v|X=x)$$

$$= \frac{P(\Pi=\pi_v, X=x)}{P(X=x)}$$

$$= \frac{P(X=x|\Pi=\pi_v)P(\Pi=\pi_v)}{\sum_{i=1}^{k} P(X=x|\Pi=\pi_i)P(\Pi=\pi_i)}$$

$$= \frac{f_v(x)p_v}{\sum_{i=1}^{k} f_i(x)p_i}$$

Discrimination Between Several Populations

- Suppose that we observe $X = x$. The **expected loss** by classifying X to π_i is

$$E[L(\Pi, \pi_i)|X = x] = \sum_{j=1}^{k} L_{ji} P(\Pi = \pi_j | X = x)$$

$$= \sum_{j=1}^{k} L_{ji} k(\pi_j | X = x) = \frac{1}{\sum_{j=1}^{k} f_j(x) p_j} \sum_{j=1}^{k} L_{ji} f_j(x) p_j$$

$$= \frac{1}{h(x)} \sum_{j=1}^{k} L_{ji} f_j(x) p_j$$

Discrimination Between Several Populations

- Minimizing the loss means **maximizing** the *discriminant score*

$$S_i^*(x) = -\sum_{j=1}^{k} L_{ji} f_j(x) p_j$$

- Decision rule:

Select population $\boldsymbol{\pi_\nu}$ if $\quad \forall \boldsymbol{i}: \quad \boldsymbol{S_\nu^* > S_i^*}$

- If all losses are equal: Simplifies to

Select population $\boldsymbol{\pi_\nu}$ if $\quad \forall \boldsymbol{i}: \quad \boldsymbol{f_\nu(x) p_\nu > f_i(x) p_i}$

If all prior probabilities are equal (uniform prior): Simplifies to

Select population $\boldsymbol{\pi_\nu}$ if $\quad \forall \boldsymbol{i}: \quad \boldsymbol{f_\nu(x) > f_i(x)}$

The Case of Normal Distributions

Assume that individuals in $\pi_1, \ldots, \pi_k$ have normally distributed data, $N_p(\mu_i, \Sigma_i)$, with Σ invertible, $i = 1, \ldots, k$.

Quadratic discriminant function S_i^Q:

$$S_i^Q(x) = \log(c \cdot f_i(x) p_i)\,,$$

$$S_i^Q(x) = -\frac{1}{2}\log\big(det(\Sigma_i)\big) - \frac{1}{2}\|x - \mu_i\|^2_{\Sigma_i^{-1}} + \log(p_i)$$

Quadratic in x.

The Case of Normal Distributions

Suppose that $\Sigma_1 = \cdots = \Sigma_k$. The common factors can be incorporated in the c value:

Linear discriminant function S_i^L:

$$S_i^L(x) = \log(c \cdot f_i(x) p_i)\,,$$

Linear in x. Similar to the linear discriminant from Lecture F:

$$S_i^L(x) = < \mathrm{x}, \mu_i >_{\Sigma^{-1}} - \frac{1}{2} + \log(p_i)$$

Example

- Suppose that $k = 3$ with

$$\mu_1 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \mu_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mu_3 = \begin{bmatrix} 2 \\ 6 \end{bmatrix} \text{ and } \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

- Note that $\Sigma^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$.

- Assume equal losses and uniform prior.

Example

- We have

$$S_1^L(x) = [x_1 \quad x_2]\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 4 \\ 2 \end{bmatrix} - \frac{1}{2}[4 \quad 2]\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 4 \\ 2 \end{bmatrix} = 6x_1 - 2x_2 - 10$$

$$S_2^L(x) = [x_1 \quad x_2]\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{1}{2}[1 \quad 1]\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = x_1 - \frac{1}{2}$$

$$S_3^L(x) = [x_1 \quad x_2]\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ 6 \end{bmatrix} - \frac{1}{2}[2 \quad 6]\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ 6 \end{bmatrix} = -2x_1 + 4x_2 - 10$$

Example

- Hyperplane that separates π_1 and π_2 :

$$u_{12}(x) = 0, \qquad u_{12}(x) = S_1^L(x) - S_2^L(x) = 5x_1 - 2x_2 - \frac{19}{2}, \qquad ie.\ x_2 = \frac{5}{2}x_1 - \frac{19}{4}$$

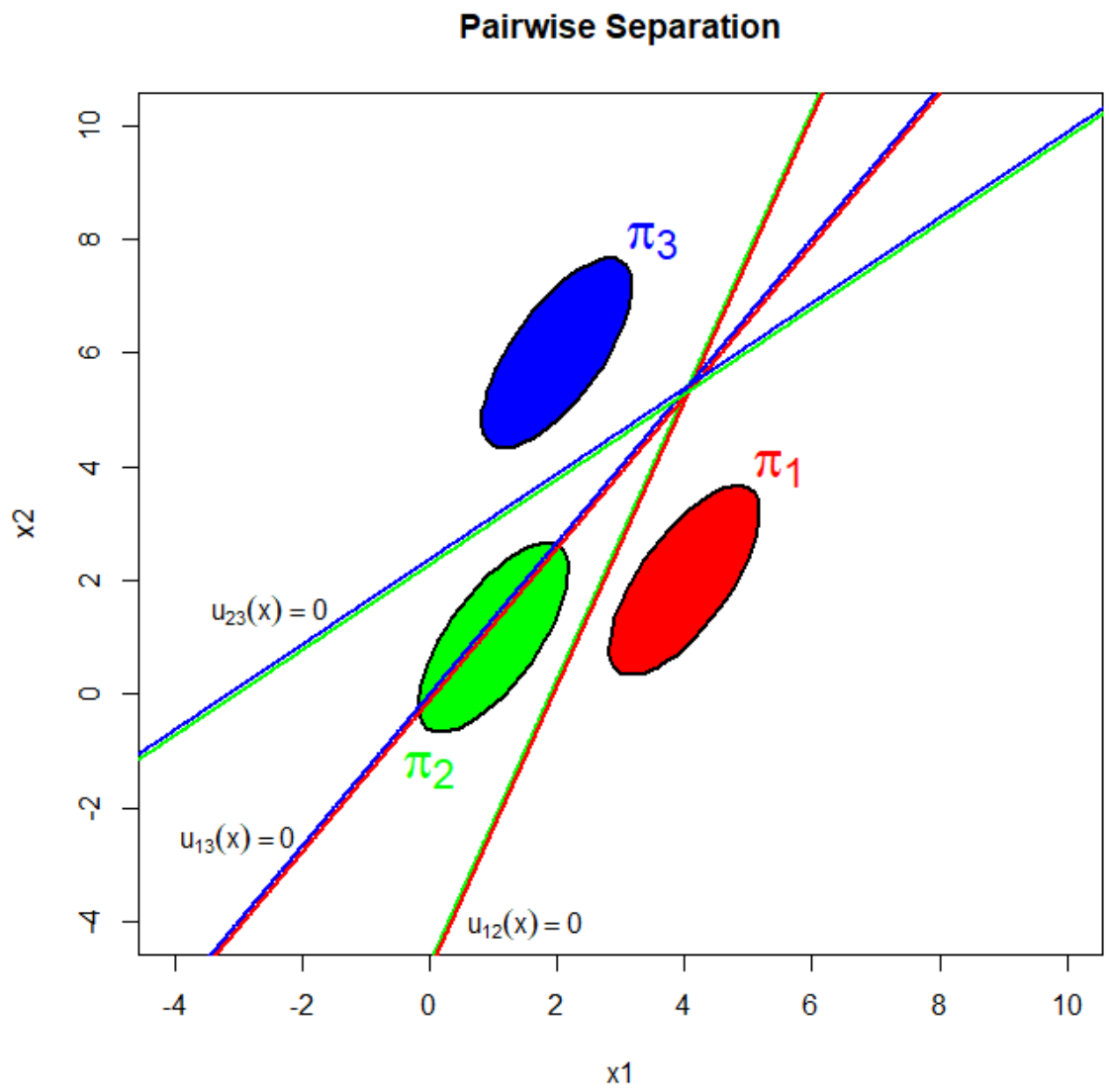
- Hyperplane that separates π_1 and π_3 :

$$u_{13}(x) = 0, \qquad u_{13}(x) = S_1^L(x) - S_3^L(x) = 8x_1 - 6x_2, \qquad ie.\ x_2 = \frac{4}{3}x_1$$

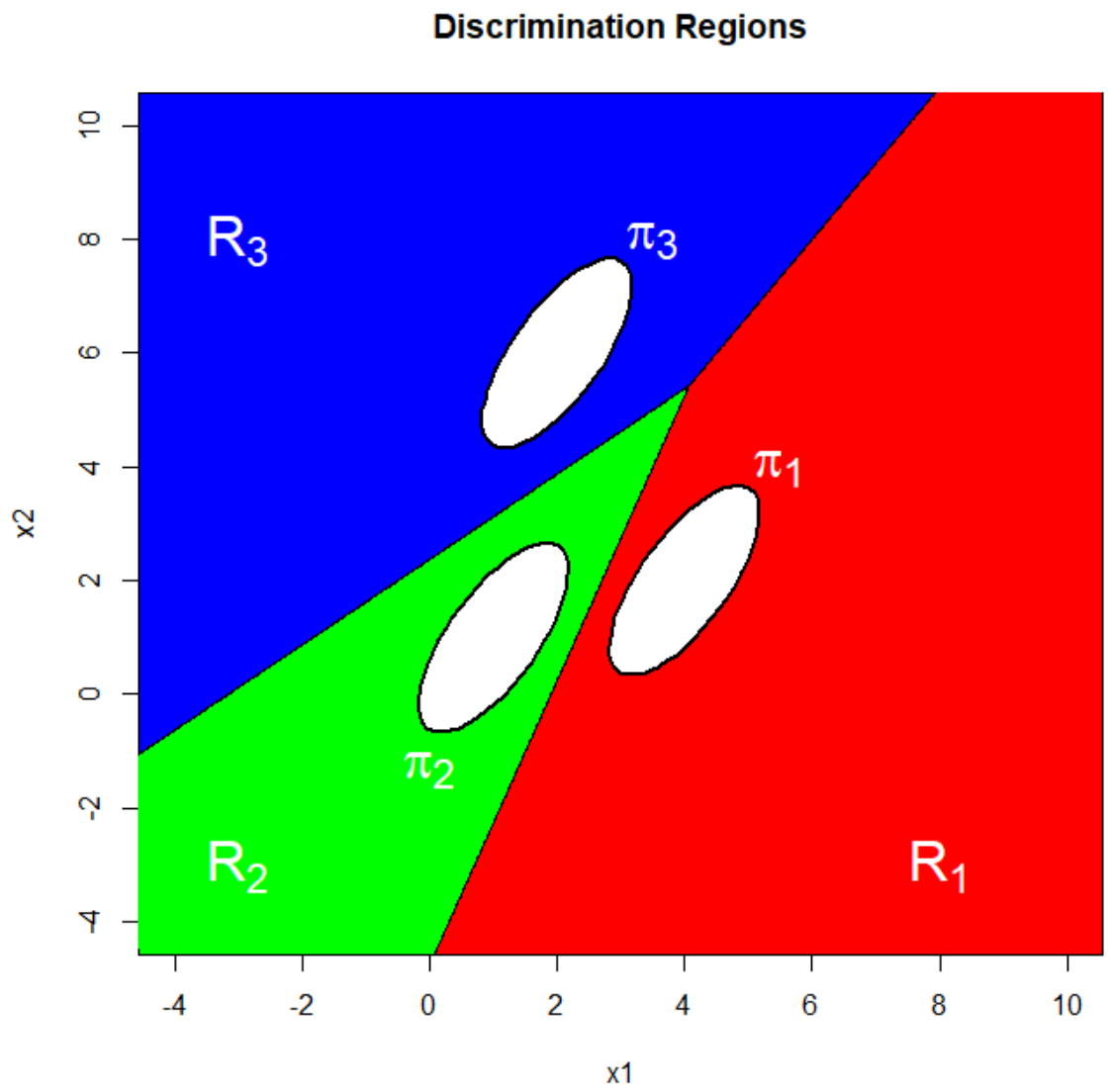
- Hyperplane that separates π_2 and π_3 :

$$u_{23}(x) = 0, \qquad u_{23}(x) = S_2^L(x) - S_3^L(x) = 3x_1 - 4x_2 + \frac{19}{2}, \qquad ie.\ x_2 = \frac{3}{4}x_1 + \frac{19}{8}$$

Example


Pairwise Separation
π3
π1
π2
u23(x) = 0
u13(x) = 0
u12(x) = 0
x1
x2
-4
-2
0
2
4
6
8
10

Example


Discrimination Regions
R3
π3
π1
π2
R2
R1
x2
x1
10
8
6
4
2
0
-2
-4
-4
-2
0
2
4
6
8
10

The Case of Normal Distributions, Unknown Parameters

- Assume that individuals in $\pi_1, \dots, \pi_k$ have normally distributed data, $N_p(\mu_i, \Sigma_i)$, with Σ invertible, $i = 1, \dots, k$. μ_i, Σ_i are **not known**.
- Prior distribution Π:

$$P(\Pi = \pi_i) = p_i, i = 1, \dots, k.$$

- Observations

$$X_{i1}, \dots, X_{in_i}, i = 1, \dots, k$$

The Case of Normal Distributions, Unknown Parameters

- Define the **W**ithin Group, **B**etween Groups and **T**otal sums of square matrices by

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T$$

$$B = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T$$

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})^T$$

Note:

$$T = W + B$$

The Case of Normal Distributions, Unknown Parameters

- Define the **W**ithin Group, **B**etween Groups and **T**otal sums of square matrices by

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T$$

$$B = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T$$

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})^T$$

Note:

$$T = W + B$$

Parameter Estimation

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}\,, \qquad \hat{\Sigma}_i = \frac{1}{n_i - 1} W_i.$$

Under $H: \Sigma_1 = \cdots = \Sigma_k$, we estimate ($N = \sum_{i=1}^{k} n_i$):

$$\hat{\Sigma} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\hat{\Sigma}_i = \frac{1}{N-k} W$$

Distance Between Populations

Definition 5.15

Assuming that the hypothesis $H_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k$ is true, we define *the squared generalized distance from* $\widehat{\boldsymbol{\mu}}_j$ *to population* $\boldsymbol{\pi}_i$ as

$$D_i^2(\widehat{\boldsymbol{\mu}}_j) = \begin{cases} \left(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j\right)^T \widehat{\boldsymbol{\Sigma}}^{-1} \left(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j\right) & \text{if the priors are equal} \\ \left(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j\right)^T \widehat{\boldsymbol{\Sigma}}^{-1} \left(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j\right) - 2\log p_i & \text{if the priors are not all equal} \end{cases}$$

If the hypothesis is *not* true, we define *the squared generalized distance from* $\widehat{\boldsymbol{\mu}}_j$ *to population* $\boldsymbol{\pi}_i$ as

$$D_i^2(\widehat{\boldsymbol{\mu}}_j) = \begin{cases} \left(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j\right)^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \left(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j\right) & \text{if the priors are equal} \\ \left(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j\right)^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \left(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j\right) + \log\det \widehat{\boldsymbol{\Sigma}}_i - 2\log p_i & \text{if the priors are not all equal} \end{cases}$$

Distance Between Populations

- Note that under H_0 and equal priors, $D_i^2(\hat{\mu}_j) = D_j^2(\hat{\mu}_i)$ is the (empirical) Mahanalobis distance between $\hat{\mu}_i$ and $\hat{\mu}_j$.
- Since

$$\frac{n_i n_j}{n_i + n_j} \frac{N - k - p + 1}{(N - k)p} D_i^2(\hat{\mu}_j) \sim F(p, N - k - p + 1)$$

if $\mu_i = \mu_j$. We can thus use $D_i^2(\hat{\mu}_j)$ as a *test statistic* for the hypothesis

$$H_{ij}: \mu_i = \mu_j$$

- (This test can also be obtained in the General Linear Model as a MANOVA hypothesis)

- Which problems occur if H_0 is not true?

Test for Additional Information

- Assume H_0, and subdivide X_{ij}into the first q and the last $p-q$ variables, ie.

$$X_{ij} = \begin{bmatrix} X_{i1} \\ \vdots \\ X_{iq} \\ X_{iq+1} \\ \vdots \\ X_{ip} \end{bmatrix} = \begin{bmatrix} X_i^{(1)} \\ X_i^{(2)} \end{bmatrix} \sim N\left(\begin{pmatrix} \mu_i^{(1)} \\ \mu_i^{(2)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

- We want to investigate the hypothesis

$$H: The\ deviations\ between\ \pi_i\ and\ \pi_j\ may\ be\ determined\ from\ the\ first\ q\ variables$$

- Translation: If we know the first q variables, there is no difference between $\pi_i\ and\ \pi_j$ for the last $p-q$ variables.
- Consequence if true: The last $p-q$ variables do not contribute to the separation of $\pi_i\ and\ \pi_j$, and may be omitted from the analysis!

Test for Additional Information

$$E\left(X_{ij}^{(2)}\middle|X_{ij}^{(1)}=x^{(1)}\right)=\mu_i^{(2|1)}=\mu_i^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\left(x^{(1)}-\mu_i^{(1)}\right)$$
$$=\mu_i^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}\mu_i^{(1)}-\Sigma_{21}\Sigma_{11}^{-1}x^{(1)}$$

Conditional contrasts:

$$\mu_i^{(2|1)}-\mu_j^{(2|1)}=\mu_i^{(2)}-\mu_j^{(2)}-\Sigma_{21}\,\Sigma_{11}^{-1}\left(\mu_i^{(1)}-\mu_j^{(1)}\right)-\Sigma_{21}\Sigma_{11}^{-1}\left(x^{(1)}-x^{(1)}\right)$$
$$=\mu_i^{(2)}-\mu_j^{(2)}-\Sigma_{21}\,\Sigma_{11}^{-1}\left(\mu_i^{(1)}-\mu_j^{(1)}\right)$$

is parametric, does not depend on the actual observation!

- H translates into

$$H':\mu_1^{(2|1)}=\cdots=\mu_k^{(2|1)}$$

A multivariate one-sided analysis of variance hypothesis!

Test for Additional Information

- Subdivide

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}, \qquad T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$$

And construct

$$W_{2|1} = W_{22} - W_{21}W_{11}^{-1}W_{12}, \qquad T_{2|1} = T_{22} - T_{21}T_{11}^{-1}T_{12}$$

Test for Additional Information

Theorem 5.19

The hypothesis that the last $p-q$ variables provide no additional information to the discrimination between the populations $\pi_1, \cdots, \pi_k$ given that we are using the first q variables may be tested using the test statistic

$$\Lambda_{2|1} = \frac{\left|\boldsymbol{W}_{2|1}\right|}{\left|\boldsymbol{T}_{2|1}\right|} = \frac{|\boldsymbol{T}_{11}|}{|\boldsymbol{T}|} \times \frac{|\boldsymbol{W}|}{|\boldsymbol{W}_{11}|} = \frac{\Lambda_p}{\Lambda_q}$$

where Λ_p and Λ_q are values the test statistic Wilks' Lambda for the case with all p variables and for the case using the first q variables. Under the hypothesis the statistic follows a $\mathrm{U}(p, k-1, N-k-q)$ distribution.

$U(p, k-1, N-k-q)$ is the Wilks distribution $\Lambda(p, k-1, N-k-q)$.

- Note that the test is essentially the ratio between the variation **W**ithin groups and the **T**otal variation, given $X^{(1)}$, with small values critical.

- Essentially a consequence of Theorem 4.25, page 304, applied to the conditional distribution given $X^{(1)}$

Test for Additional Information

Theorem 5.20

If $q = p - 1$, i.e. we investigate one variable at a time, we have

$$\frac{N-k-p+1}{k-1} \times \frac{1-\Lambda_{2|1}}{\Lambda_{2|1}} \sim F(k-1, N-k-p+1)$$

F distributions, exact for $p - q = 2$ and approximative for $p - q > 2$, may be obtained as in Lecture F, slide 19.

Test for Additional Information

The simple case: $k = 2$.

Take

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

$$D^2 = (\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2)$$

$$D_1^2 = \left(\hat{\mu}_1^{(1)} - \hat{\mu}_2^{(1)}\right)^T \hat{\Sigma}^{-1}_{11} \left(\hat{\mu}_1^{(1)} - \hat{\mu}_2^{(1)}\right)$$

and note that

$$D^2_{(2|1)} = D^2 + D_1^2$$

Test for Additional Information

The simple case: $k = 2$.

Take

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

$$D^2 = (\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2)$$

$$D_1^2 = \left(\hat{\mu}_1^{(1)} - \hat{\mu}_2^{(1)}\right)^T \hat{\Sigma}^{-1}_{11} \left(\hat{\mu}_1^{(1)} - \hat{\mu}_2^{(1)}\right)$$

and note that

$$D^2_{(2|1)} = D^2 + D_1^2$$

Test for Additional Information

The simple case: $k = 2$:

Theorem 5.21

The critical region for testing the hypothesis that the last $p - q$ variables do not contribute to the discrimination between the populations π_1 and π_2, i.e. the hypothesis that $\Delta^2_{(2|1)} = 0$ against all alternatives is

$$\left\{ x_{11}, \cdots, x_{2n_2} \;\middle|\; \frac{n_1+n_2-p-1}{p-q} \frac{d^2-d_1^2}{(n_1+n_2)(n_1+n_2-2)/(n_1n_2)+d_1^2} > F(p-q, n_1+n_2-p-1)_{1-\alpha} \right\}$$

Here d^2 and d_1^2 are the observed values of D^2 and D_1^2.

Canonical Discriminant Analysis

- We maintain H_0.

- Selecting an optimal discriminator:
- Should maximize the variation between groups, relative to the variation within groups.

- $k = 2$: $\delta = \Sigma^{-1}(\mu_1 - \mu_2)$

Theorem 5.6

The vector δ has the property that it maximizes the function

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

Canonical Discriminant Analysis

- $k \geq 2$:
- Variation Between groups estimated by B;
- Variation Within groups estimated by W;

We seek a (linear) function $y = d^T x$ so that d maximizes

$$\phi(d) = \frac{d^T B d}{d^T W d}$$

Canonical Discriminant Analysis

$$\phi(d) = \frac{d^T B d}{d^T W d}$$

- **Solution:** The eigenvector $\boldsymbol{d_1}$ for the largest eigenvalue for the matrix $W^{-1}B$, subject to that $\|\boldsymbol{d_1}\|_W^2 = \boldsymbol{d_1^T W d_1}$=1.

- Next step: Search for solutions W-perpendicular to $\boldsymbol{d_1}$, ie.

$$< \mathbf{d_2}, \mathbf{d_1} >_{\mathrm{W}} = \mathbf{d_2^T W d_1} = 0$$

- **Solution:** The eigenvector $\boldsymbol{d_2}$ for the second largest eigenvalue for the matrix $W^{-1}B$, subject to that $\|\boldsymbol{d_2}\|_W^2 = \boldsymbol{d_2^T W d_2}$=1.

Canonical Discriminant Analysis

$$\phi(d) = \frac{d^T B d}{d^T W d}$$

- **Solution:** The eigenvector $\boldsymbol{d_1}$ for the largest eigenvalue for the matrix $W^{-1}B$, subject to that $\|\boldsymbol{d_1}\|_W^2 = \boldsymbol{d_1^T W d_1}$=1.

- Next step: Search for solutions W-perpendicular to $\boldsymbol{d_1}$, ie.

$$< \mathbf{d_2}, \mathbf{d_1} >_{\mathrm{W}} = \mathbf{d_2^T W d_1} = 0$$

- **Solution:** The eigenvector $\boldsymbol{d_2}$ for the second largest eigenvalue for the matrix $W^{-1}B$, subject to that $\|\boldsymbol{d_2}\|_W^2 = \boldsymbol{d_2^T W d_2}$=1.

- Etc, etc… We find at most $k-1$ non-zero eigenvalues.

Canonical Discriminant Analysis

The functions $\boldsymbol{d_1^T}x, \boldsymbol{d_2^T}x, \ldots$ Are called the ***canonical discriminant functions***, and deriving them forms the basis of Canonical Discriminant Analysis.

$\boldsymbol{d_1^T}x$ is the discriminator function that is optimal for separating the data into **two subsets** through a separating hyperplane; $\boldsymbol{d_2^T}x$ is the optimal discriminator function for further subdivisions, etc.

- **'Score' plots**: Visualize

$$\boldsymbol{d_1^T}(x_{ij} - \bar{\boldsymbol{x}})\ vs.\ \boldsymbol{d_2^T}(x_{ij} - \bar{\boldsymbol{x}})$$

  Color according to i, and consider higher order plots as well.

- **'Loadings' plots**: Vizualize

$$d_1\ vs.\ d_2$$

  Corresponding to the vectors

$$\begin{bmatrix} d_{11} \\ d_{21} \end{bmatrix}, \ldots, \begin{bmatrix} d_{1p} \\ d_{2p} \end{bmatrix}$$

  Label points $1, \ldots, p$ and consider higher order plots.

- **Apply** the canonical discriminant functions to discriminate the data.

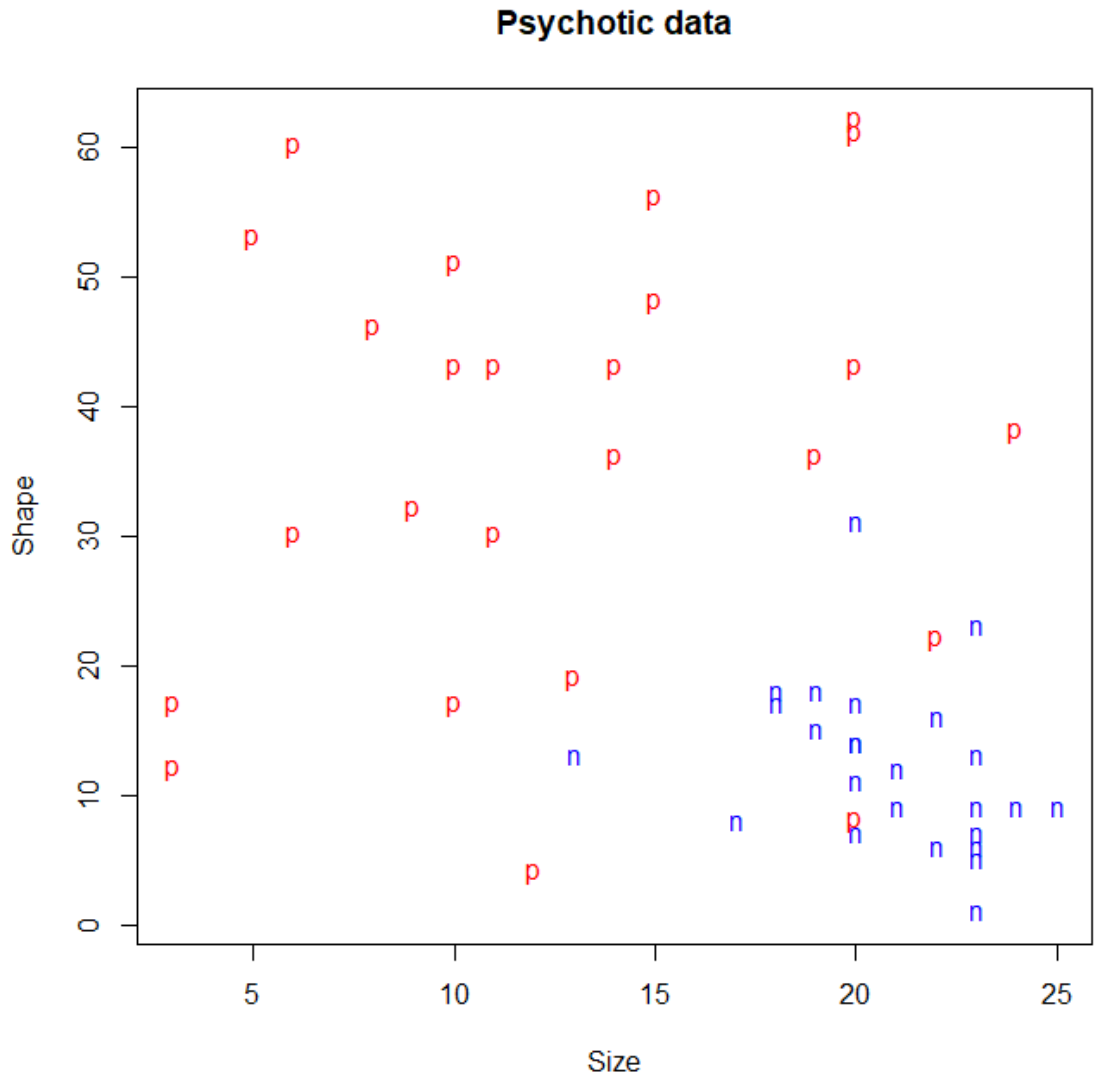
Example – Psychology Data

- A group of 25 NORMAL and 25 PSYCHOTIC individuals were given certain tests.

- Two variables SIZE and SHAPE were calculated as proxies.

```
> psychotic<-read.csv2("Data/psychotic.csv")
> psychotic$group<-as.factor(psychotic$group)
> summary(psychotic)
      group          size            shape
 normal   :25   Min.   : 3.0   Min.   : 1.0
 psychotic:25   1st Qu.:12.2   1st Qu.: 9.5
                Median :19.5   Median :17.0
                Mean   :16.8   Mean   :24.4
                3rd Qu.:21.8   3rd Qu.:37.5
                Max.   :25.0   Max.   :62.0
```

Example – Psychology Data



Example – Psychology Data

- We want to construct a decision rule that **separates psychotic individuals from normal individuals**. Method: canonical discriminant analysis.

- Need to construct the fundamental matrices W, B, T.

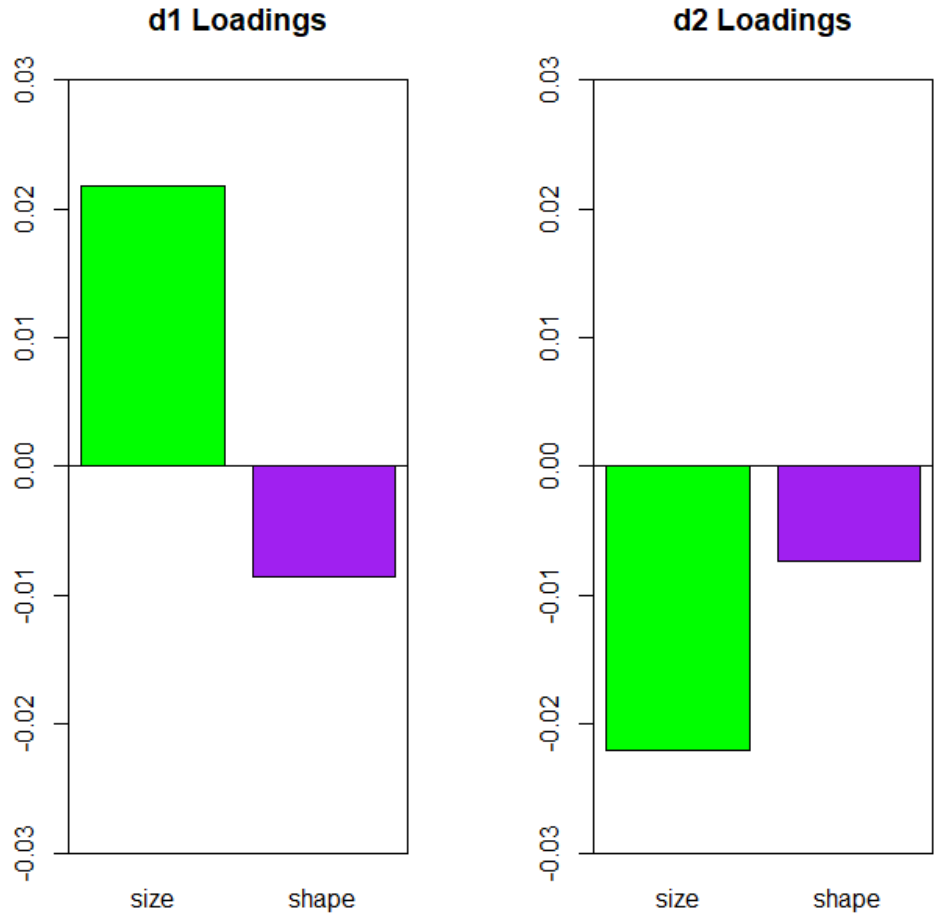
- Data management:

```
>data.all<-as.matrix(psychotic[,2:3])
>data.normal<-as.matrix(psychotic[psychotic$group=="normal",2:3])
>data.psychotic<-as.matrix(psychotic[psychotic$group=="psychotic",2:3])

>xbar.all<-matrix(rep(1,100),ncol=2)%*%diag(colMeans(data.all))
>xbar.normal<-matrix(rep(1,50),ncol=2)%*%diag(colMeans(data.normal))
>xbar.psychotic<-matrix(rep(1,50),ncol=2)%*%diag(colMeans(data.psychotic))
```

Example – Psychology Data

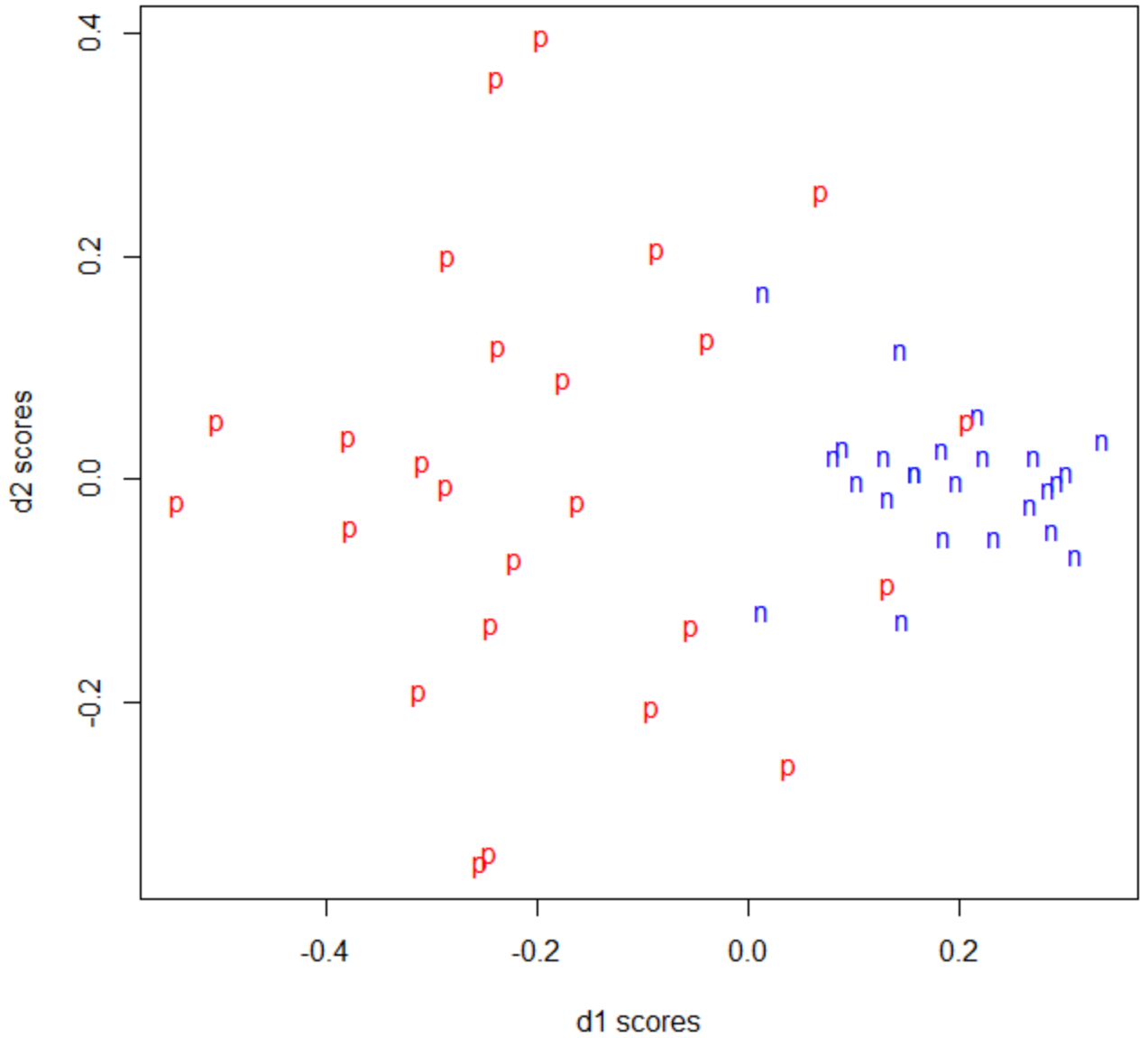
Constructing the canonical discriminators:

```
> d1<-eigen(solve(W)%*%B)$vectors[,1]
> d1<-(1/as.numeric(sqrt(t(d1)%*%W%*%d1)))*d1
>
> d2<-eigen(solve(W)%*%B)$vectors[,2]
> d2<-(1/as.numeric(sqrt(t(d2)%*%W%*%d2)))*d2
>
>
> d1
[1]  0.021749 -0.008586
> d2
[1] -0.022049 -0.007325
```



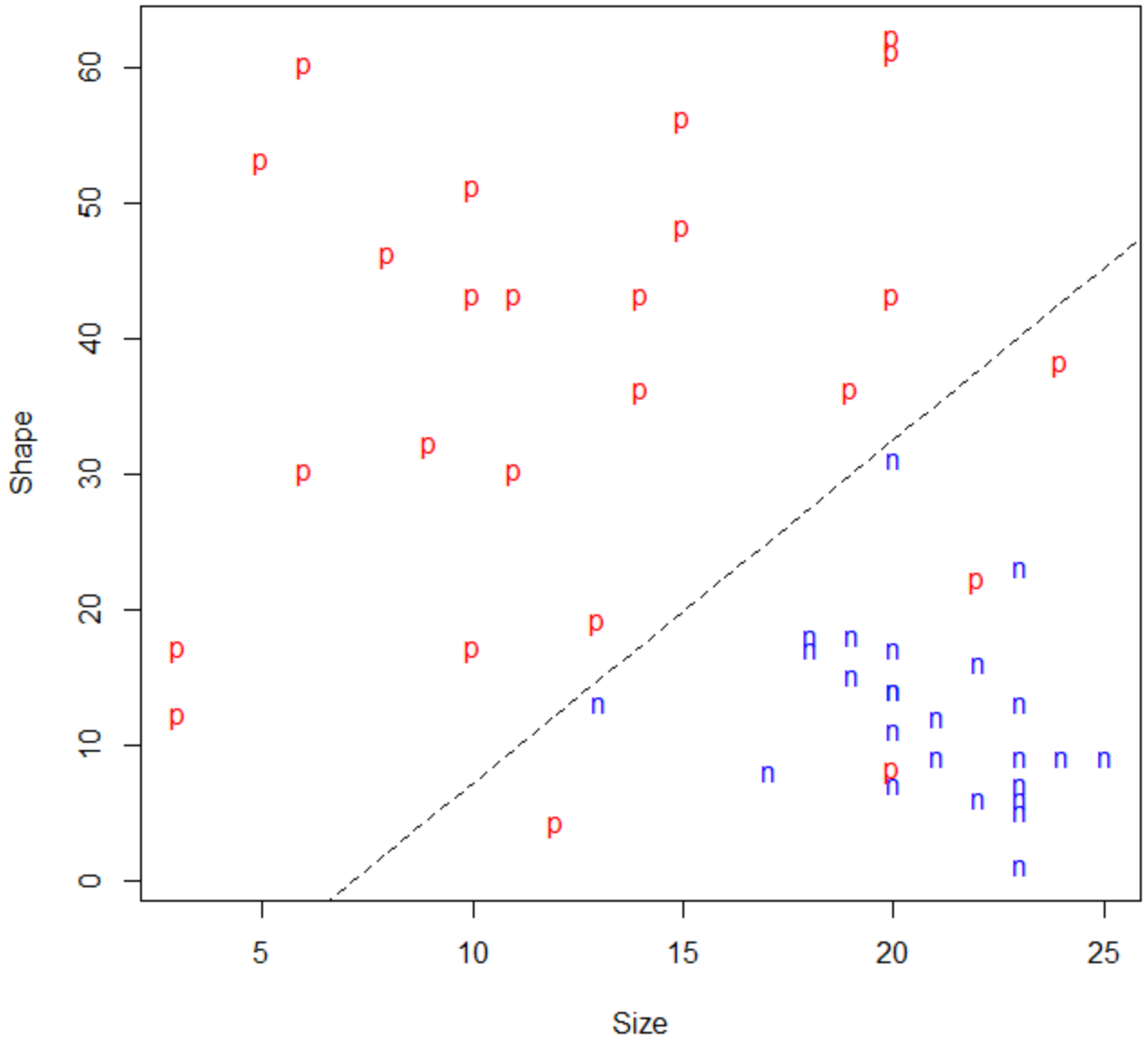
Example – Psychology Data

The score plot:



Example – Psychology Data

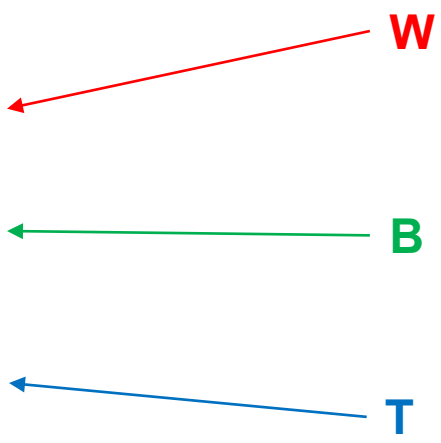
The canonical discrimination:



Example – Psychology Data

Constructing W, B, T:

```
> W_normal<-t(data.normal-xbar.normal)%*%(data.normal-xbar.normal)
> W_psychotic<-t(data.psychotic-xbar.psychotic)%*%(data.psychotic-xbar.psychotic)
>
> W<-W_normal+W_psychotic
> B<-25*(colMeans(data.normal)-colMeans(data.all))%*%
+         t(colMeans(data.normal)-colMeans(data.all))+
+    25*(colMeans(data.psychotic)-colMeans(data.all))%*%
+         t(colMeans(data.psychotic)-colMeans(data.all))
>
> T<-W+B
>
> W; B; T
        size  shape
size  1048.0  207.6
shape  207.6 7891.4
      size shape
[1,]   800 -2408
[2,] -2408  7248
       size shape
size   1848 -2200
shape -2200 15140
```



Example – Psychology Data

Same analysis using `lda` function:

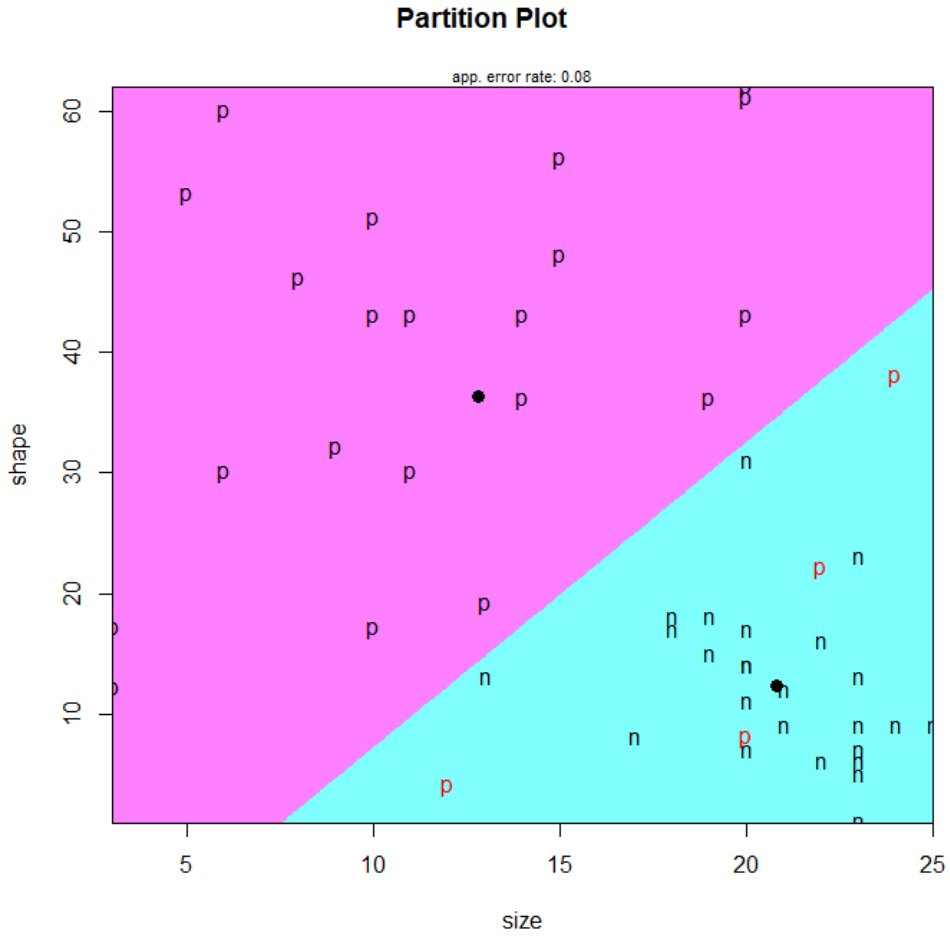
```
> (my.lda<-lda(group ~ size + shape, prior=c(1/2,1/2), data=psychotic) )
Call:
lda(group ~ size + shape, data = psychotic, prior = c(1/2, 1/2))

Prior probabilities of groups:
   normal psychotic
      0.5       0.5

Group means:
          size shape
normal    20.8 12.32
psychotic 12.8 36.40

Coefficients of linear discriminants:
           LD1
size  -0.15068
shape  0.05949

> d1/my.lda$scaling
          LD1
size  -0.1443
shape -0.1443
```

Constant: parallel vectors.
The slope of the discrimination line is the same.

Example – Psychology Data

Discrimination plot using `lda`:

```
library(klaR)
partimat(group ~ shape + size, prior=c(1/2,1/2), data=psychotic, method="lda",prec=1000)
```



Similar to slide 51!

Example – Psychology Data

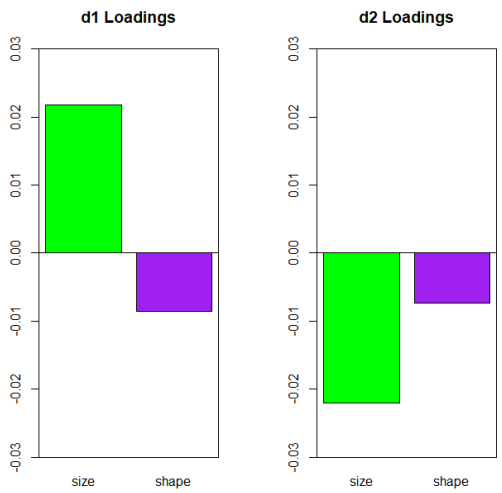
Confusion matrix:

```
> table(psychotic$group,predict(my.lda)[[1]],dnn=c("Nature","Classification"))
           Classification
Nature      normal psychotic
  normal        25         0
  psychotic      4        21
```

Error rate: (0+4)/(25+0+4+21)=0.08.

Sensitivity: 21/(21+4)=0.84.

Example – Psychology Data



- Test for additional information – does Shape contribute?

- Testing in the conditional distribution of Shape given Size.

  Test statistic (Theorem 20):

$$\Lambda_{shape|size} = \frac{T_{11}}{|T|} \cdot \frac{|W|}{W_{11}} = \frac{1848}{23136073} \cdot \frac{8227131}{1048} = 0.627$$

$$\frac{N-k-p+1}{k-1} \cdot \frac{1-\Lambda_{shape|size}}{\Lambda_{shape|size}} = \frac{50-2-2+1}{2-1} \cdot \frac{1-0.627}{0.627} = 27.95 \sim F(1,47)$$

$$p = P(F(1,47) > 27.95) = 0.000003$$

- **Shape contributes significantly (p=3e-6).**

Example – Psychology Data

Is LDA a good idea?

```
> W_normal/24;W_psychotic/24
         size  shape
size    6.917 -5.267
shape -5.267 40.893

        size  shape
size   36.75  13.92
shape 13.92 287.92
```

- Vastly different variances! Psychotic individuals are much more variable!

Taking that into account means Quadratic Discrimination analysis!

Example – Psychology Data

- Quadratic discrimination analysis using `qda` function:

```
> (my.qda<-qda(group ~ size + shape, prior=c(1/2,1/2), data=psychotic) )
Call:
qda(group ~ size + shape, data = psychotic, prior = c(1/2, 1/2))

Prior probabilities of groups:
   normal psychotic
      0.5       0.5

Group means:
           size shape
normal     20.8 12.32
psychotic 12.8 36.40
```

Example – Psychology Data

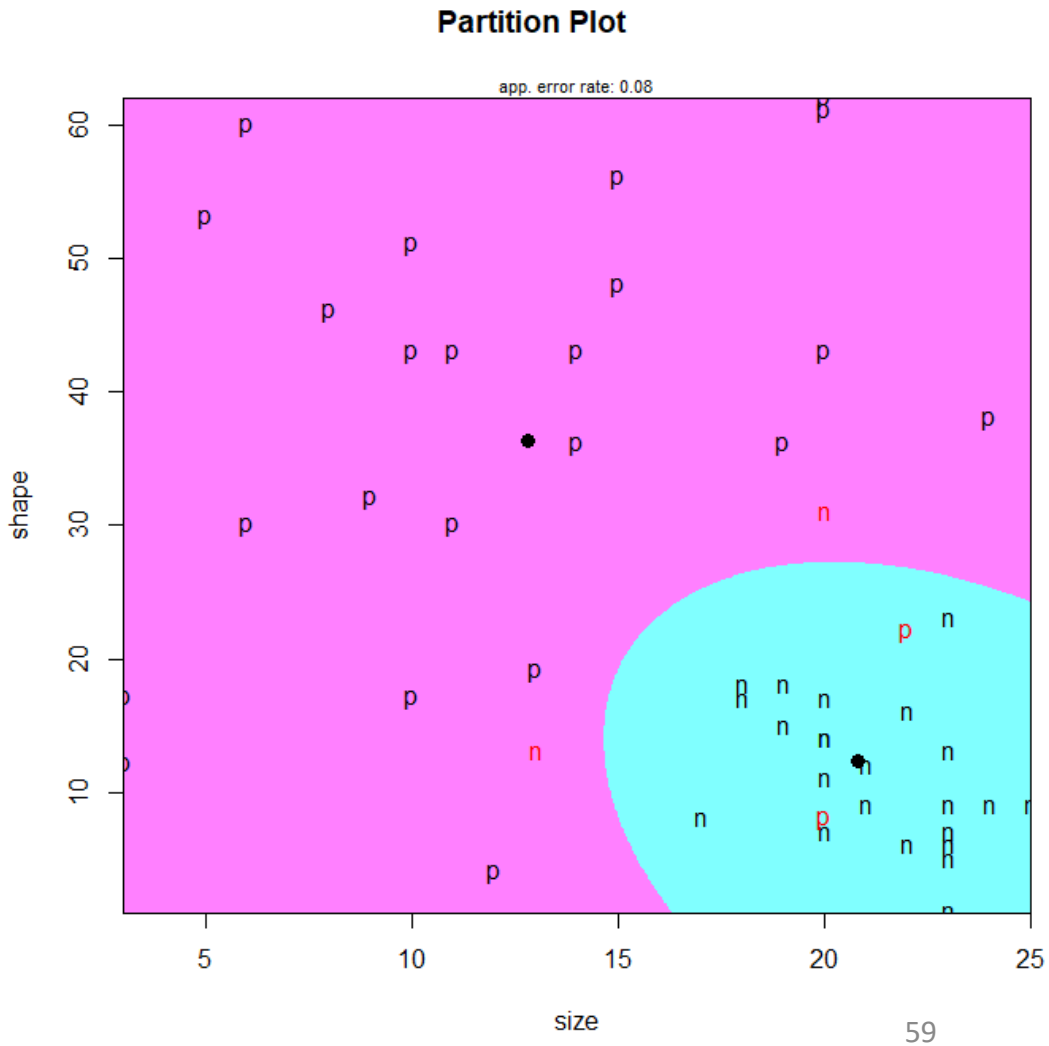
```
> partimat(group ~ shape + size, prior=c(1/2,1/2), data=psychotic,
method="qda",prec=1000)
```

- Confusion matrix, quadratic discrimination:

```
>table(psychotic$group,predict(my.qda)[[1]],
>dnn=c("Nature","Classification"))
          Classification
Nature      normal psychotic
  normal        23         2
  psychotic      2        23
```

Error rate: (2+2)/(23+2+2+23) =0.04 (unchanged)

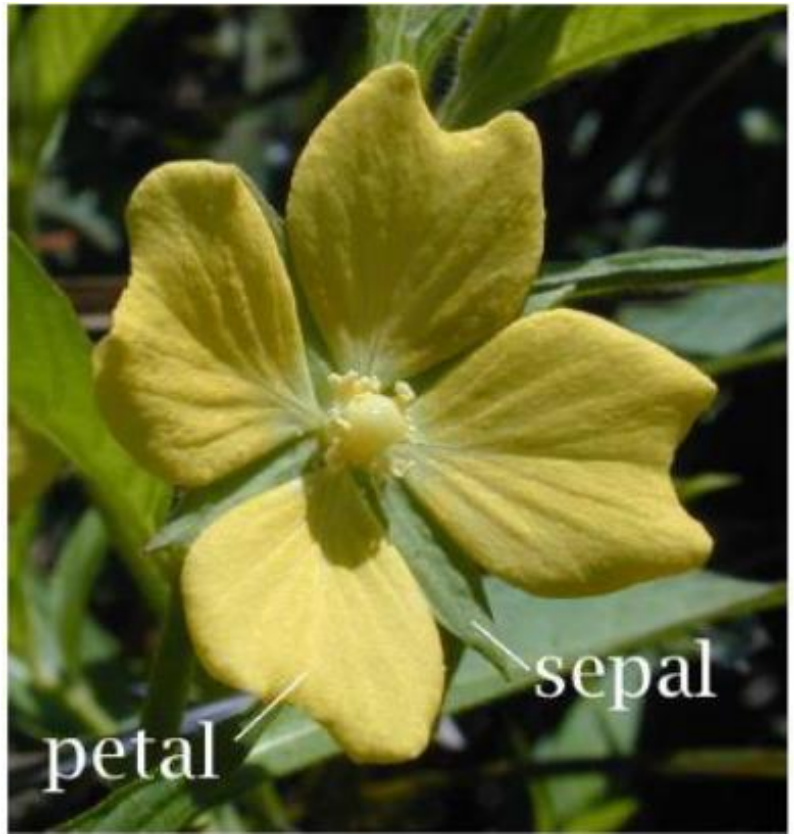
Sensitivity: 23/(2+23)=0.92 (increased).



Example – Iris Data

- Measumrents on 50 Iris Setosa, 50 Iris versicolor and 50 Iris virginica plants.
- Sepal length, sepal width, petal length and petal width measured in mm.
- Original hypothesis:

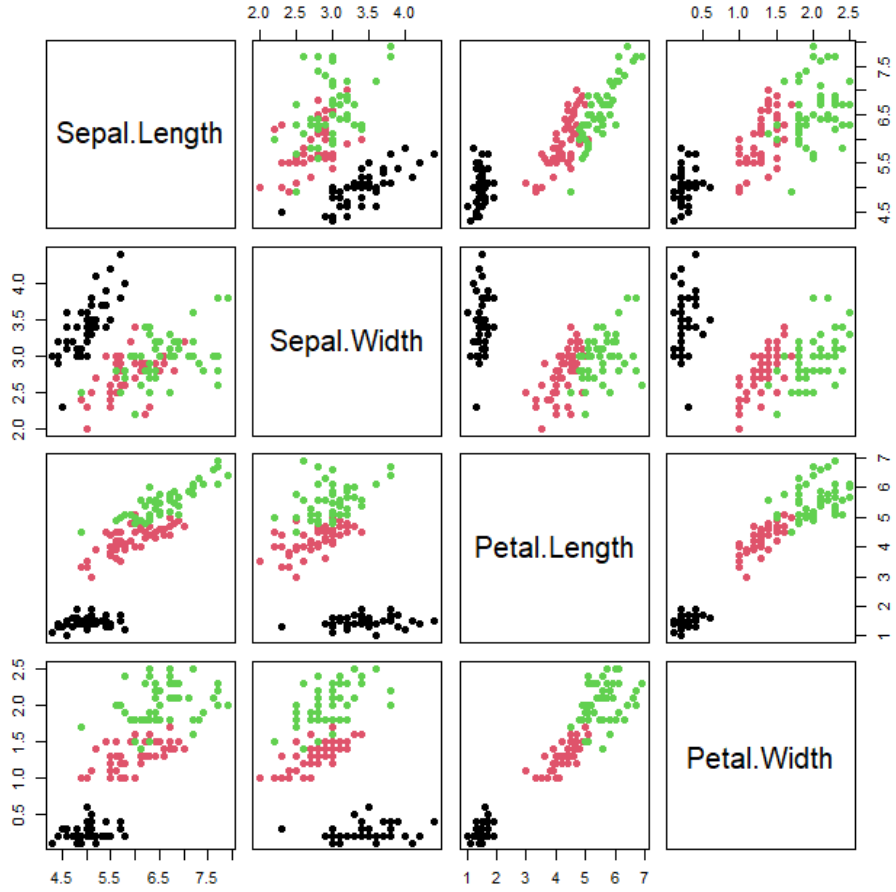
Versicolor is a hybrid of the two other species.



Example – Iris Data

- Built-in dataset:

```
> iris<-datasets::iris
> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length"
[2] "Petal.Width"  "Species"
> pairs(iris[,1:4], col=iris$Species, pch=19)
```



Example – Iris Data

- We want to construct decision rule that **separates species**.
- Method: canonical discriminant analysis.

- We will start by constructing the W's.

- Data management:

```
iris.all<-as.matrix(iris[,1:4])
iris.setosa<-as.matrix(iris[iris$Species=="setosa",1:4])
iris.versicolor<-as.matrix(iris[iris$Species=="versicolor",1:4])
iris.virginica<-as.matrix(iris[iris$Species=="virginica",1:4])

xbar.all<-matrix(rep(1,4*150),ncol=4)%*%diag(colMeans(iris.all))
xbar.setosa<-matrix(rep(1,4*50),ncol=4)%*%diag(colMeans(iris.setosa))
xbar.versicolor<-matrix(rep(1,4*50),ncol=4)%*%diag(colMeans(iris.versicolor))
xbar.virginica<-matrix(rep(1,4*50),ncol=4)%*%diag(colMeans(iris.virginica))
```

Example – Iris Data

- Constructing W's:

```
> W_setosa<-t(iris.setosa-xbar.setosa)%*%(iris.setosa-xbar.setosa)
W_versicolor<-t(iris.versicolor-xbar.versicolor)%*%(iris.versicolor-xbar.versicolor)
W_virginica<-t(iris.virginica-xbar.virginica)%*%(iris.virginica-xbar.virginica)
```

- Variance:

```
> W_setosa/49
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length      0.12425    0.099216     0.016355    0.010331
Sepal.Width       0.09922    0.143690     0.011698    0.009298
Petal.Length      0.01636    0.011698     0.030159    0.006069
Petal.Width       0.01033    0.009298     0.006069    0.011106
>
> W_versicolor/49
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length      0.26643     0.08518      0.18290     0.05578
Sepal.Width       0.08518     0.09847      0.08265     0.04120
Petal.Length      0.18290     0.08265      0.22082     0.07310
Petal.Width       0.05578     0.04120      0.07310     0.03911
>
> W_virginica/49
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length      0.40434     0.09376      0.30329     0.04909
Sepal.Width       0.09376     0.10400      0.07138     0.04763
Petal.Length      0.30329     0.07138      0.30459     0.04882
Petal.Width       0.04909     0.04763      0.04882     0.07543
```

- Again, vastly different variances.
- We will need QDA.

Example – Iris Data

Quadratic Discrimination functions:

$$S_i^Q(x) = -\frac{1}{2}\log\left(det\left(\hat{\Sigma}_i\right)\right) + < x, \hat{\mu}_i >_{\hat{\Sigma}_i^{-1}} - \frac{1}{2}\|\hat{\mu}_i\|^2_{\hat{\Sigma}_i^{-1}}, i \in \{setosa, versicolor, virginica\}$$

$$R_{setosa} = \{x \in \mathbb{R}^4 | S^Q_{setosa}(x) \geq \max(S^Q_{versicolor}(x), S^Q_{virginica}(x)\}$$
$$R_{versicolor} = \{x \in \mathbb{R}^4 | S^Q_{versicolor}(x) \geq \max(S^Q_{setosa}(x), S^Q_{virginica}(x)\}$$
$$R_{virginica} = \{x \in \mathbb{R}^4 | S^Q_{virginica}(x) \geq \max(S^Q_{setosa}(x), S^Q_{virginica}(x)\}$$

Posterior probabilities:

$$k^Q(\pi_i|x) = \frac{\exp\left(S_i^Q(x)\right)}{exp\left(S^Q_{setosa}(x) + S^Q_{versicolor}(x) + S^Q_{virginica}(x)\right)}, i \in \{setosa, versicolor, virginica\}$$

Example – Iris Data

Quadratic Discrimination functions:

$$\hat{\mu}_{setosa} = \begin{bmatrix} 5.006 \\ 3.428 \\ 1.462 \\ 0.246 \end{bmatrix}, \qquad \hat{\Sigma}_{setosa} = \begin{bmatrix} 0.124250 & 0.099216 & 0.016355 & 0.010331 \\ 0.099216 & 0.143690 & 0.011698 & 0.009298 \\ 0.016355 & 0.011698 & 0.030159 & 0.006069 \\ 0.01033 & 0.009298 & 0.006069 & 0.011106 \end{bmatrix}$$

$$\log\left(\det\left(\hat{\Sigma}_{setosa}\right)\right) = -13.07, \qquad \hat{\Sigma}_{setosa}^{-1} = \begin{bmatrix} 18.943 & -12.405 & -4.500 & -4.776 \\ -12.405 & 15.571 & 1.111 & -2.104 \\ -4.500 & 1.111 & 38.776 & -17.935 \\ -4.776 & -2.104 & -17.935 & 106.046 \end{bmatrix}$$

Example – Iris Data

Quadratic Discrimination functions:

$$\hat{\mu}_{setosa} = \begin{bmatrix} 5.006 \\ 3.428 \\ 1.462 \\ 0.246 \end{bmatrix}, \qquad \hat{\Sigma}_{setosa} = \begin{bmatrix} 0.124250 & 0.099216 & 0.016355 & 0.010331 \\ 0.099216 & 0.143690 & 0.011698 & 0.009298 \\ 0.016355 & 0.011698 & 0.030159 & 0.006069 \\ 0.01033 & 0.009298 & 0.006069 & 0.011106 \end{bmatrix}$$

$$\log\left(\det\left(\hat{\Sigma}_{setosa}\right)\right) = -13.07, \qquad \hat{\Sigma}_{setosa}^{-1} = \begin{bmatrix} 18.943 & -12.405 & -4.500 & -4.776 \\ -12.405 & 15.571 & 1.111 & -2.104 \\ -4.500 & 1.111 & 38.776 & -17.935 \\ -4.776 & -2.104 & -17.935 & 106.046 \end{bmatrix}$$

- **For example:**

```
S.setosa.Q<-function(x){-(log(det(Sigma.setosa))/2+
                          t(x)%*%solve(Sigma.setosa)%*% mu.setosa -
                          t(mu.setosa)%*%solve(Sigma.setosa)%*%mu.setosa/2}
```

Example – Iris Data

Quadratic Discrimination analysis:

```
my.qda <-qda(Species ~ Sepal.Length+Sepal.Width + Petal.Length+
                Petal.Width, prior=c(1,1,1)/3, data=iris)
```

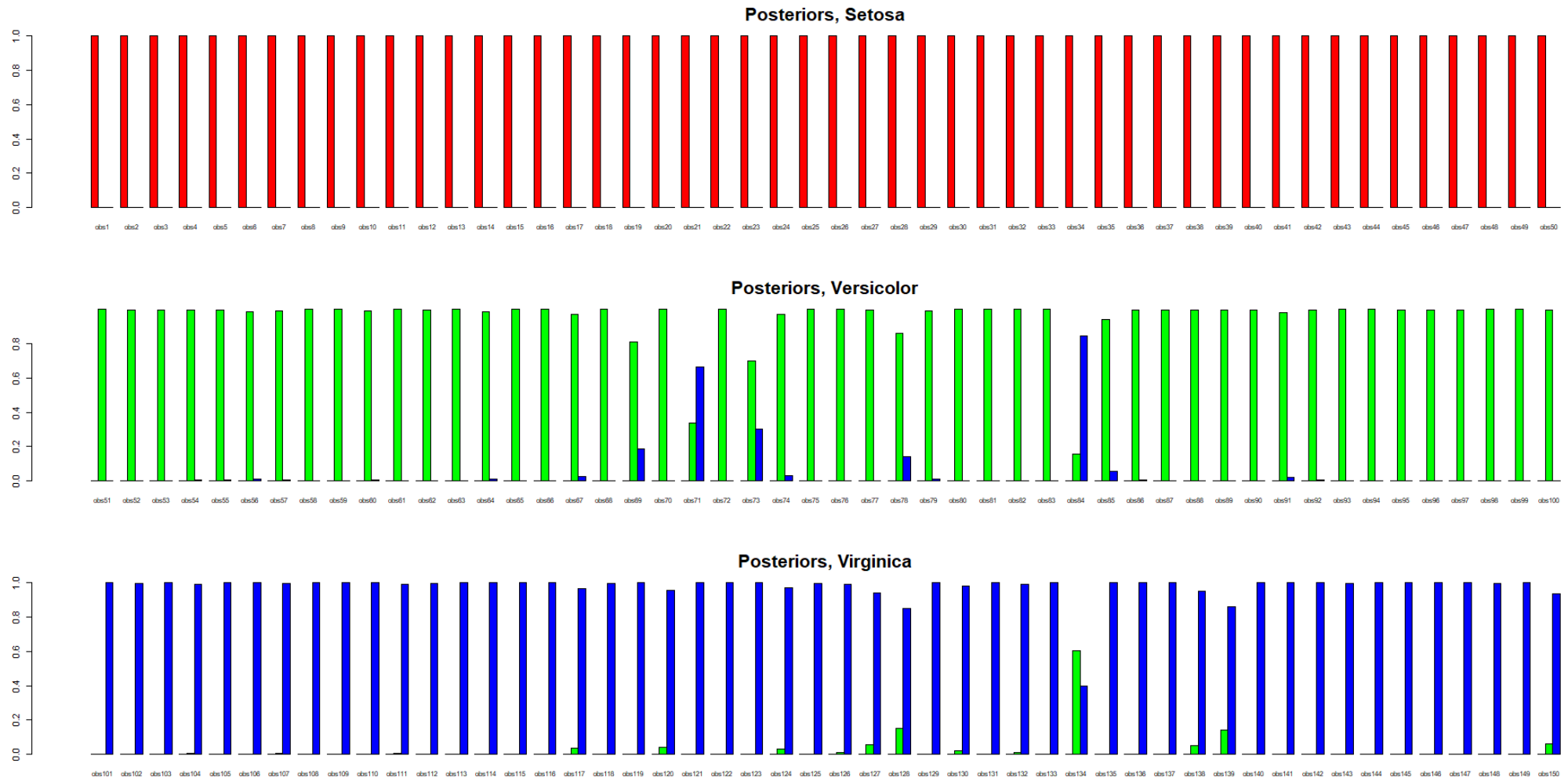
- ***Posteriors:***

```
> head(predict(my.qda)[[2]])
  setosa   versicolor    virginica
1      1 4.918517e-26 2.981541e-41
2      1 7.655808e-19 1.311032e-34
3      1 1.552279e-21 3.380440e-36
4      1 8.300396e-19 8.541858e-32
5      1 3.365614e-27 2.010147e-41
6      1 1.472533e-26 1.271928e-40
```

- ***Classification:***

```
> head(predict(my.qda)[[1]])
[1] setosa setosa setosa setosa setosa setosa
Levels: setosa versicolor virginica
```

Example – Iris Data

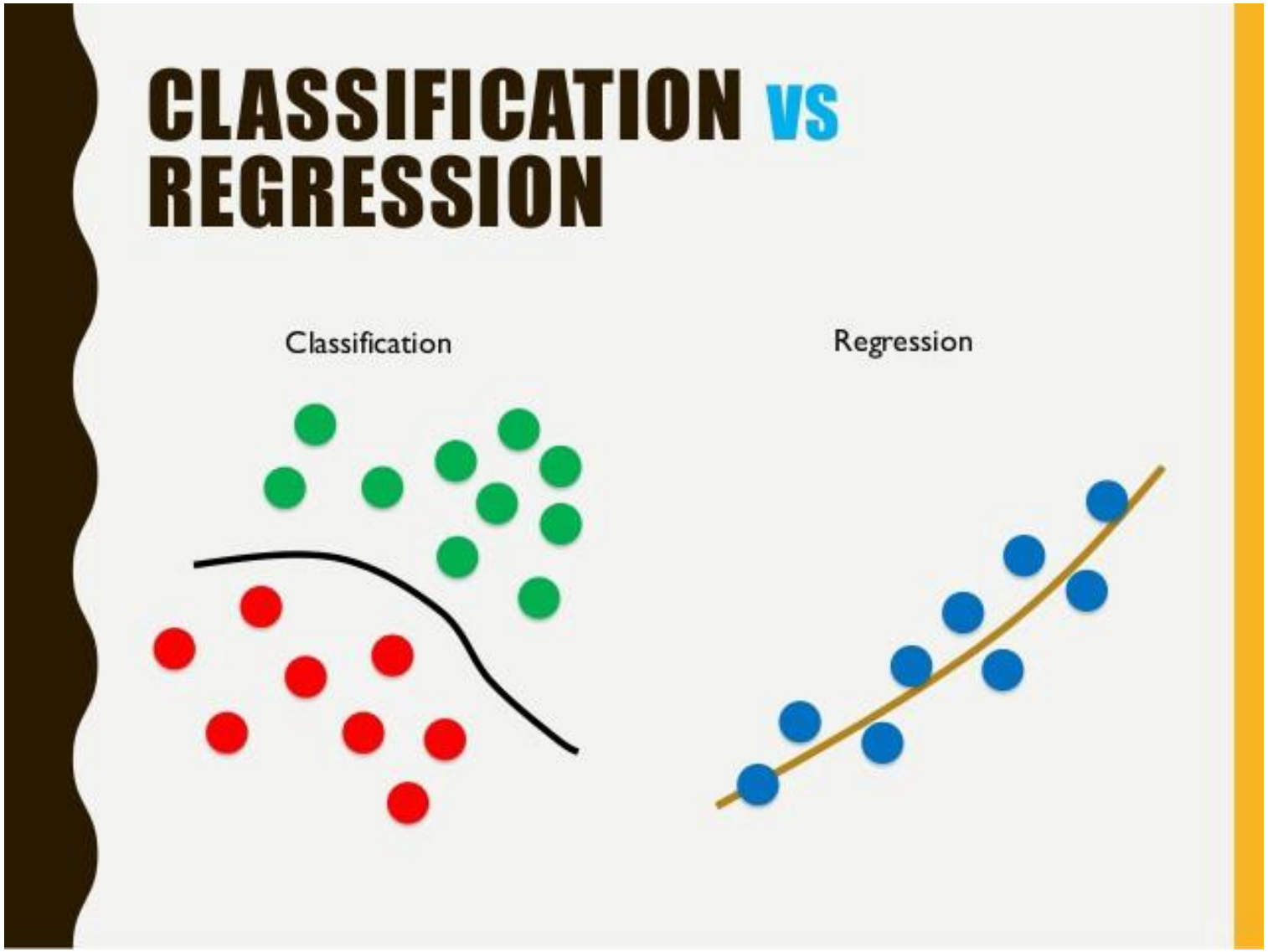


Example – Iris Data

Confusion matrix:

	Classification		
Nature	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

Error rate: 3/150=0.02



https://cdn-images-1.medium.com/max/638/1*4sixxtuD8unWceZ-yp9TgQ.jpeg

The General Linear Model - GLM

$$Y \sim N_n(\mu, \sigma^2 \Sigma)$$

$$\mu = X\theta$$

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix}$$

X is the design matrix; assumed known, and with full rank

Σ is assumed known

σ^2 is assumed unknown

The General Linear Model - GLM

- Example: $\Sigma = I_n$:

$$Y_i = \sum_{j=1}^{n} x_{ij}\,\theta_j + \varepsilon_i, i = 1, \dots, n, \qquad \varepsilon_1, \dots, \varepsilon_n \; iidsv, \varepsilon_1 \sim N(0, \sigma^2)$$

- Standard univariate linear regression.

The General Linear Model - GLM

- Take

$$M = \{X\theta \big| \theta \in \mathbb{R}^k\}$$

and let $M^{\perp}$ be the orthogonal complement with respect to $\|\cdot\|_{\Sigma^{-1}}$.

Since $\mathbb{R}^n = M \oplus M^{\perp}$, and vector $y \in \mathbb{R}^n$ may be written uniquely as

$$y = y_1 + y_2 \text{ with } y_1 \in M \text{ and } y_2 \in M^{\perp}$$

Note that

$$y_1 = P_M y = X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y$$

The General Linear Model - GLM

- Estimation:

$$L(\theta,\sigma^2) = k\cdot(\sigma^2)^{-\frac{n}{2}} exp\left(-\frac{1}{2\sigma^2}\|X\theta - Y\|^2_{\Sigma^{-1}}\right)$$

ie.

$$\ell(\theta,\sigma^2) = log\big(L(\theta,\sigma^2)\big) = k_1 - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\|X\theta - Y\|^2_{\Sigma^{-1}}$$

$$= k_1 - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\left\|X\theta - P_M Y - P_{M^\perp} Y\right\|^2_{\Sigma^{-1}}$$

$$= k_1 - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\left(\|X\theta - P_M Y\|^2_{\Sigma^{-1}} + \left\|P_{M^\perp} Y\right\|^2_{\Sigma^{-1}}\right)$$

$\boldsymbol{x\theta \in M}$:
$\boldsymbol{< x\theta, P_{M^\perp} Y >_{\Sigma^{-1}} = 0}$

The General Linear Model - GLM

$$\ell(\theta, \sigma^2) = k_1 - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2}\left(\|X\theta - P_M Y\|^2_{\Sigma^{-1}} - \left\|P_{M^\perp} Y\right\|^2_{\Sigma^{-1}}\right)$$

Largest in θ if $x\theta - P_M Y = 0$; ie. if

$$X\theta = X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

ie. (full rank of X)

$$\theta = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y =: \hat{\theta}$$

The General Linear Model - GLM

Mean and variance of $\hat{\theta}$:

$$E(\hat{\theta}) = E\left((X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y\right)$$
$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}E(Y)$$
$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X\theta = \theta$$

$$V(\hat{\theta}) = V\left((X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y\right)$$
$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}V(Y)\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}$$
$$= \sigma^2(X^T\Sigma^{-1}X)^{-1}$$

$\boldsymbol{V(AY) = AV(Y)A^T}$

The General Linear Model - GLM

Estimator for σ^2:

$$\ell\left(\hat{\theta}, \sigma^2\right) = k_1 - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \left\| P_{M^\perp} Y \right\|^2_{\Sigma^{-1}}$$

Largest if

$$-\frac{n}{\sigma^2} + \frac{1}{2\sigma^4} \left\| P_{M^\perp} Y \right\|^2_{\Sigma^{-1}} = 0,$$

ie.

$$\sigma^2 = \frac{1}{n} \left\| P_{M^\perp} Y \right\|^2_{\Sigma^{-1}} = \tilde{\sigma}^2.$$

The General Linear Model - GLM

- $\tilde{\sigma}^2$ is biased:

$$E(n\tilde{\sigma}^2) = E\left(tr\left((P_{M^\perp}Y)^T\Sigma^{-1}P_{M^\perp}Y\right)\right)$$
$$= E\left(tr\left((Y-x\theta)^T P_{M^\perp}\Sigma^{-1}P_{M^\perp}(Y-x\theta)\right)\right)$$
$$= tr\left(P_{M^\perp}\Sigma^{-1}P_{M^\perp}E\left((Y-x\theta)(Y-x\theta)^T\right)\right)$$
$$= tr\left(P_{M^\perp}\Sigma^{-1}P_{M^\perp}V(Y)\right)$$
$$= \sigma^2 tr(P_{M^\perp}\Sigma^{-1}P_{M^\perp}\Sigma)$$
$$= \sigma^2 tr(P_{M^\perp}\Sigma^{-1}(I-P_M)\Sigma)$$
$$= \sigma^2 tr(P_{M^\perp}(I_n+\mathbf{0})) = \sigma^2 tr(P_{M^\perp}) = \sigma^2(n-k)$$

$tr(AB) = tr(BA)$ and linearity of tr

- Unbiased estimator:

$$\hat{\sigma}^2 = \frac{1}{n-k}\left\|P_{M^\perp}Y\right\|^2_{\Sigma^{-1}} = \frac{1}{n-k}\left(Y-X\hat{\theta}\right)^T\Sigma^{-1}\left(Y-X\hat{\theta}\right)$$

The General Linear Model - GLM

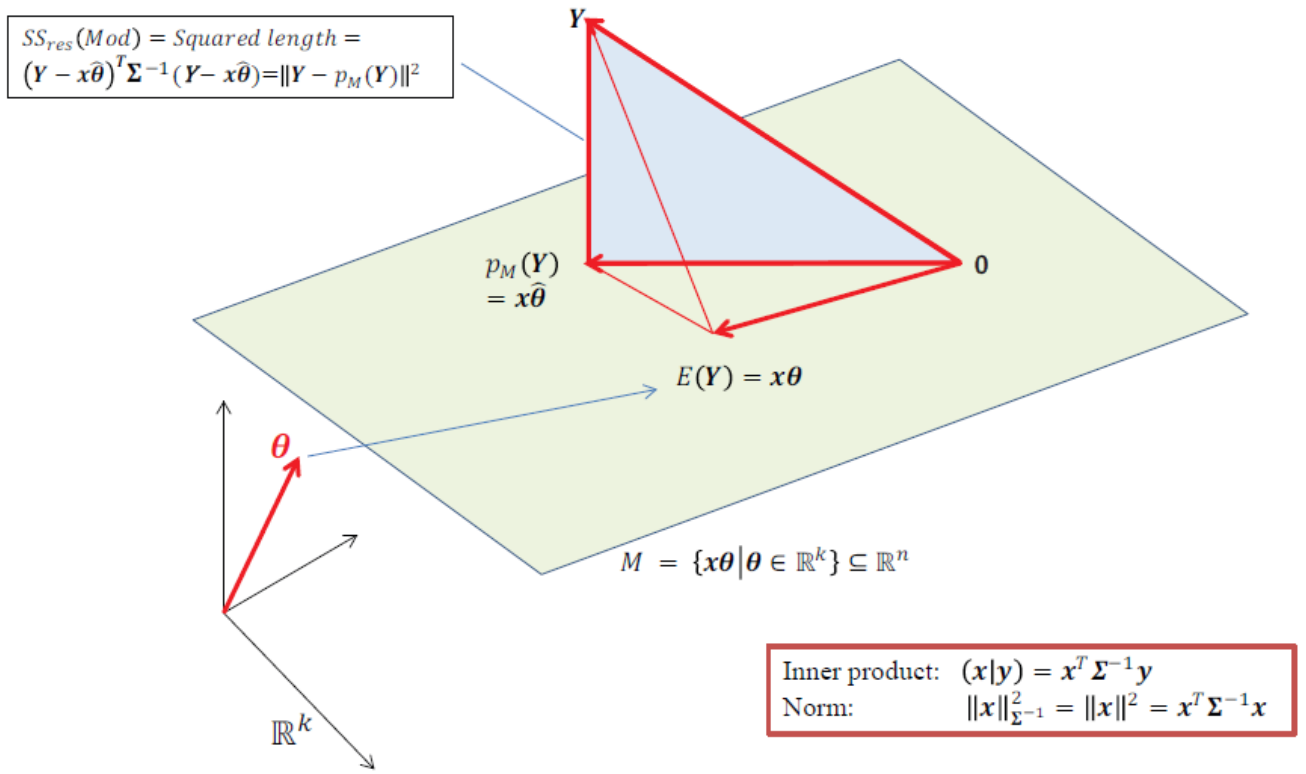


Figure 2.2 – Geometric sketch of the problem of estimation in the general linear model.

The General Linear Model - GLM

Theorem 2.3

Let $\mathbf{x}$ and $\boldsymbol{\theta}$ be given as in the preceding section and let $\boldsymbol{Y} \sim \mathrm{N}_n(\mathbf{x}\,\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite. Then the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is given by $\mathbf{x}\,\hat{\boldsymbol{\theta}}$ being the projection (with respect to $\boldsymbol{\Sigma}$) onto M, $\hat{\boldsymbol{\theta}}$ is a solution to the so-called ***normal equation(s)***

$$(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\hat{\boldsymbol{\theta}} = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}y.$$

If $\mathbf{x}$ has full rank k, then

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y},$$

and being a linear combination of normally distributed variables $\hat{\boldsymbol{\theta}}$ is also normally distributed with parameters

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\theta}}) &= \boldsymbol{\theta} \\ \mathrm{D}(\hat{\boldsymbol{\theta}}) &= \sigma^2(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}. \end{aligned}$$

It is especially noted that $\hat{\boldsymbol{\theta}}$ is an unbiased estimate of $\boldsymbol{\theta}$.

The General Linear Model - GLM

Theorem 2.5

Let the situation be as above. The maximum likelihood estimator of σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}\|^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}).$$

The unbiased estimator of σ^2 is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n - \mathrm{rk}(\mathbf{x})}\|\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}\|^2 \\ &= \frac{1}{n - \mathrm{rk}(\mathbf{x})}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})\end{aligned}$$

where $\mathbf{x}\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\mathrm{E}(\mathbf{Y})$. The following holds

$$\hat{\sigma}^2 \sim \sigma^2\chi^2(n - \mathrm{rk}(\mathbf{x}))/(n - \mathrm{rk}(\mathbf{x}))$$

and $\hat{\sigma}^2$ is independent of the maximum likelihood estimator of the expected value and is therefore independent of $\hat{\boldsymbol{\theta}}$.

GLM in R

- The `lm()` function.

```
synprod <- data.frame("Acontent"=c(1,0,0.5),
                      "Bcontent"=c(0,1,0.5),
                      "Qual"=c(90,30,75))

glmSyn <- lm(Qual ~ Acontent + Bcontent - 1, data=synprod)
# the "-1" in the line above removes the intercept
# If not specified an intercept is estimated by default
```

GLM in R

```
glmSyn <- lm(Qual ~ Acontent + Bcontent - 1, data=synprod)

> summary(glmSyn)

Call:
lm(formula = Qual ~ Acontent + Bcontent - 1, data = synprod)

Residuals:
 1  2  3
-5 -5 10

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
Acontent    95.00      11.18   8.497   0.0746 .
Bcontent    35.00      11.18   3.130   0.1968
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.25 on 1 degrees of freedom
Multiple R-squared:  0.9897,    Adjusted R-squared:  0.9692
F-statistic: 48.25 on 2 and 1 DF,  p-value: 0.1013
```

Exercises

- Exercise 6.4: ”Pen & paper” on CDF
- Exam 2011 Q. 3.3-3.6 : Classification
- Exam 2009 Problem 1 : Classification
- Exercise 3.1 and 3.2: GLM