What to do today ?

Part I. Introduction and Preparation
Part II. Inference under Multivariate Normal Distribution

Part III. Commonly-Used Multivariate Analysis Methods (Chp 8-9, 11-12)

III.1. Discrimination and Classification (Chp 11)

- **III.1.1 Introduction**
- **III.1.2 Two-group discriminant analysis**
- **III.1.3 Classification with two populations**
- *III.1.4 Discrimination and classification with several populations*

III.2. Principal Component Analysis (Chp 8)

III.3. Factor Analysis (Chp 9)

III.4. Clustering (Chp 12)

II.4 Multivariate Linear Regression (Chp 7.1-7)

- To explore how rvs $Y_1, \ldots, Y_m$ depend on $X_1, \ldots, X_k$.
  That is, to explore how r.v. $\mathbf{Y}$ depends on $X_1, \ldots, X_k$.

Assume

$$\begin{aligned} \mathbf{Y} &= [\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k] + \boldsymbol{\epsilon} \\ (\text{response} &= [\text{ linear function of } X_1, \ldots, X_k] + \text{error}) \end{aligned}$$

$E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$.
With n indpt obs on $\mathbf{Y}$, where $\mathbf{Y}_i$ is associated with the values $x_{1i}, \ldots, x_{ki}$ of $X_1, \ldots, X_k$ for $i = 1, \ldots, n$:

$$\begin{aligned} \mathbf{Y}_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{1k} + \boldsymbol{\epsilon}_1 \\ \mathbf{Y}_2 &= \beta_0 + \beta_1 x_{21} + \ldots + \beta_k x_{2k} + \boldsymbol{\epsilon}_2 \\ \vdots &\quad\quad \vdots \\ \mathbf{Y}_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \beta_k x_{nk} + \boldsymbol{\epsilon}_n \end{aligned}$$

where (1) $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n$ are indpt, and (2) $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$, $Var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}$.

$$\begin{pmatrix} \mathbf{Y}_1' \\ \mathbf{Y}_2' \\ \vdots \\ \mathbf{Y}_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_0' \\ \boldsymbol{\beta}_1' \\ \vdots \\ \boldsymbol{\beta}_k' \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1' \\ \boldsymbol{\epsilon}_2' \\ \vdots \\ \boldsymbol{\epsilon}_n' \end{pmatrix}$$

Multivariate Linear Regression Model

$$\mathbf{Y}_{n\times m} = \mathbf{X}_{n\times(k+1)}\boldsymbol{B}_{(k+1)\times m} + \boldsymbol{E}_{n\times m}$$

with (1) $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$, and (2) $Cov(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_l) = \boldsymbol{\Sigma}$ if $i = l$; $\mathbf{0}_{m\times m}$ if $i \neq l$.
The jth component $\mathbf{Y}_{(j)}$ follows

$$\mathbf{Y}_{(j)} = \mathbf{X}_{n\times(k+1)}\boldsymbol{\beta}_{(j)} + \boldsymbol{\epsilon}_{(j)} \quad j = 1, \dots, m$$

with $Var(\boldsymbol{\epsilon}_{(j)}) = \sigma_{jj}\mathbf{I}$.
Least Squares Estimator (LSE).

$$\hat{\boldsymbol{\beta}}_{(j)} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}_{(j)}$$

provided $\mathbf{X}$ has full rank.

$$\widehat{\boldsymbol{B}} = \left(\hat{\boldsymbol{\beta}}_{(1)}\hat{\boldsymbol{\beta}}_{(2)}\dots\hat{\boldsymbol{\beta}}_{(m)}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{Y}_{(1)}\mathbf{Y}_{(2)}\dots\mathbf{Y}_{(m)}\right)$$

That is

$$\widehat{\boldsymbol{B}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

- the fitted values: $\hat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{B}} = \mathbf{HY}$ with the "hat" matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$.
- the residuals: $\hat{\boldsymbol{E}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, and $\mathbf{X}'\hat{\boldsymbol{E}} = \mathbf{0}$
- $\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}$
- $\widehat{\boldsymbol{B}} = \left(\hat{\boldsymbol{\beta}}_{(1)} \ldots \hat{\boldsymbol{\beta}}_{(m)}\right)$ is unbiased: $E(\widehat{\boldsymbol{B}}) = \boldsymbol{B}$

- $Cov(\hat{\boldsymbol{\beta}}_{(j)}, \hat{\boldsymbol{\beta}}_{(l)}) = \sigma_{jl}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ for $j, l = 1, \ldots, k$.

- $E(\hat{\boldsymbol{E}}) = \mathbf{0}$ and $E(\hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}) = (n - [k+1])\boldsymbol{\Sigma}$; $\widehat{\boldsymbol{\Sigma}} = \frac{\hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}}{n-[k+1]}$.
- $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{E}}$ are uncorrelated.

With the model $\mathbf{Y} = \mathbf{X}\boldsymbol{B} + \boldsymbol{E}$, inferential procedures under the normality assumption on the error terms $\epsilon_i \sim MN(\mathbf{0}, \boldsymbol{\Sigma})$ iid for $i = 1, \ldots, n$:

- The LSE/MLE $\hat{\beta}_{(j)} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}_{(j)} \sim MN(\beta_{(j)}, \sigma_{jj}\left(\mathbf{X}'\mathbf{X}\right)^{-1})$.
- The MLE $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}/n$; an unbiased estimator $\tilde{\boldsymbol{\Sigma}} = \hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}/(n - [k+1])$;
- To test $H_0 : \boldsymbol{B}_{(2)} = \mathbf{0}$

  Likelihood ratio test:

$$\Lambda = \frac{\max_{\beta_{(1)}, \boldsymbol{\Sigma}} L(\beta_{(1)}, \boldsymbol{\Sigma})}{\max_{\beta, \boldsymbol{\Sigma}} L(\beta, \boldsymbol{\Sigma})} = \left(\frac{|\hat{\boldsymbol{\Sigma}}_1|}{|\hat{\boldsymbol{\Sigma}}|}\right)^{-n/2}$$

Wilks' lambda statistic: $\Lambda^{2/n} = \frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|}$

- To estimate $E(\mathbf{Y}_0|\mathbf{x}_0) = \beta_0 + \beta_1 x_{01} + \ldots + \beta_k x_{0k}$, or to predict for $\mathbf{Y}_0 = \boldsymbol{B}'\mathbf{x}_0 + \epsilon_0$

II.4.4C Multivariate Multiple Regression: inference

Provided the LSE $\hat{\boldsymbol{B}}$ using data with size n for the model $\mathbf{Y} = \mathbf{X}\boldsymbol{B} + \boldsymbol{E}$ under the normality assumption $\epsilon_i \sim MN(\mathbf{0}, \boldsymbol{\Sigma})$ iid:
if $\mathbf{Y}_0$ is the response when the indpendent variables are $x_{01}, \ldots, x_{0k}$,

- *to estimate* $E(\mathbf{Y}_0|\mathbf{x}_0) = \beta_0 + \beta_1 x_{01} + \ldots + \beta_k x_{0k}$, *which is* $E(\mathbf{Y}_0|\mathbf{x}_0) = \boldsymbol{B}'\mathbf{x}_0$
  - point estimate: $\hat{\boldsymbol{B}}'\mathbf{x}_0 \sim MN\big(E(\mathbf{Y}_0|\mathbf{x}_0), \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\boldsymbol{\Sigma}\big)$.
  - $100(1-\alpha)\%$ simultaneous confidence interval for $E(Y_{0j}) = \mathbf{x}_0'\boldsymbol{\beta}_{(j)}$

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}}_{(j)} \pm \sqrt{\frac{m(n-[k+1])}{n-k-m}F_{m,n-k-m}(\alpha)}\sqrt{[\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0]\frac{n}{n-[k+1]}\hat{\sigma}_{jj}}$$

- *to predict for* $\mathbf{Y}_0 = \boldsymbol{B}'\mathbf{x}_0 + \boldsymbol{\epsilon}_0$
  - point estimate: $\hat{\boldsymbol{B}}'\mathbf{x}_0 \sim MN\big(E(\mathbf{Y}_0|\mathbf{x}_0), \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\boldsymbol{\Sigma}\big)$.
  - $100(1-\alpha)\%$ prediction interval for Y_{0j}:

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}}_{(j)} \pm \sqrt{\frac{m(n-[k+1])}{n-k-m}F_{m,n-k-m}(\alpha)}\sqrt{[1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0]\frac{n}{n-[k+1]}\hat{\sigma}_{jj}}$$

II.4.4C Multivariate Multiple Regression: inference

Provided the LSE $\hat{\boldsymbol{B}}$ using data with size n for the model $\mathbf{Y} = \mathbf{X}\boldsymbol{B} + \boldsymbol{E}$ under the normality assumption $\boldsymbol{\epsilon}_i \sim MN(\mathbf{0}, \boldsymbol{\Sigma})$ iid:
if $\mathbf{Y}_0$ is the response when the indpendent variables are $x_{01}, \ldots, x_{0k}$,

- *to estimate* $E(\mathbf{Y}_0|\mathbf{x}_0) = \beta_0 + \beta_1 x_{01} + \ldots + \beta_k x_{0k}$, *which is* $E(\mathbf{Y}_0|\mathbf{x}_0) = \boldsymbol{B}^{'}\mathbf{x}_0$
  - point estimate: $\hat{\boldsymbol{B}}^{'}\mathbf{x}_0 \sim MN\big(E(\mathbf{Y}_0|\mathbf{x}_0), \mathbf{x}_0^{'}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{x}_0\boldsymbol{\Sigma}\big)$.
  - $100(1-\alpha)\%$ simultaneous confidence interval for $E(Y_{0j}) = \mathbf{x}_0^{'}\boldsymbol{\beta}_{(j)}$

$$\mathbf{x}_0^{'}\hat{\boldsymbol{\beta}}_{(j)} \pm \sqrt{\frac{m(n-[k+1])}{n-k-m}F_{m,n-k-m}(\alpha)}\sqrt{[\mathbf{x}_0^{'}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{x}_0]\frac{n}{n-[k+1]}\hat{\sigma}_{jj}}$$

- *to predict for* $\mathbf{Y}_0 = \boldsymbol{B}^{'}\mathbf{x}_0 + \boldsymbol{\epsilon}_0$
  - point estimate: $\hat{\boldsymbol{B}}^{'}\mathbf{x}_0 \sim MN\big(E(\mathbf{Y}_0|\mathbf{x}_0), \mathbf{x}_0^{'}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{x}_0\boldsymbol{\Sigma}\big)$.
  - $100(1-\alpha)\%$ prediction ellipsoid for $\mathbf{Y}_0$:

$$\Big\{\mathbf{y} : (\mathbf{y}-\hat{\boldsymbol{B}}^{'}\mathbf{x}_0)^{'}\Big(\frac{n}{n-[k+1]}\hat{\boldsymbol{\Sigma}}\Big)^{-1}(\mathbf{y}-\hat{\boldsymbol{B}}^{'}\mathbf{x}_0) \leq \frac{m(n-[k+1])}{n-k-m}F_{m,n-k-m}(\alpha)[1+\mathbf{x}_0^{'}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{x}_0$$

Example using Table 7.6 in the textbook (page 426)

Midterm Information

- **Time:** Mar 05 (in class), 50 minutes.
- Please arrive **10 minutes early**.

- Closed book.
- Non-communicating calculator permitted.
- One **A4 double-sided** cheatsheet permitted. Write your name and student number on it. The cheatsheet will be collected.

- Bring your **student ID**.

Midterm Coverage

Part I. Preparation

- **I.2** Matrix Algebra (Chp 2.1–2.4, Supplement 2A)
- **I.4** Multivariate Random Variables and Distributions
  - Random vectors and matrices (2.5)
  - Mean vectors and covariance matrices (2.6)
  - Descriptive multivariate analysis (Chp 1, 3)

Part II.1 Multivariate Normal Distribution (Chp 4.1 - 4.5)

- Definition and properties of $MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
- Properties of $\bar{\mathbf{X}}$ and $\mathbf{S}$

Midterm Coverage

II.2 Inferences on Mean Vector (Chp 5.1 - 5.5)

- Hotelling's T^2
- Likelihood ratio test
- Confidence regions and simultaneous comparisons
- Large sample inference on Population Mean

II.3 Comparisons of Several Mean Vectors (Chp 6.1 - 6.7)

- Comparing two populations
- MANOVA (one-way and two-way)

II.4 Multivariate Linear Regression (Chp 7)

- Classical linear regression (7.1–7.3)
- Inference for linear regression (7.4–7.5)
- Multivariate multiple regression (7.7)

Part III. Commonly-Used Multivariate Analysis Methods (Chp 8-11)

What to study?

- *III.1. Discriminant and Classification (Chp 11)*
- *III.2. Principal Component Analysis (Chp 8)*
- *III.3. Factor Analysis (Chp 9)*
- *III.4. Clustering (Chp 12)*

Why bother?

Data Mining and Machine Learning:

- Supervised learning (classification and regression)
- Clustering (unsupervised learning)
- Dimensionality reduction
- Structured prediction
- ... ...

III.1.1 Discrimination and Classification (Chp 11): Introduction

Discrimination (Separation Analysis): To describe (graphically or algebraically) the differential features of data from two or several *known* populations.

(Description of group separation.)

- e.g. to find "discriminants" whose numeric values separate as much as possible (e.g. biomarkers, patient's demographics etc.) of data from several known populations (e.g. progressive and non-progressive groups)

Classification (Allocation Analysis): To develop a rule to allocate data cases (e.g. patients) into two or more labeled classes.

(Allocation of observations into groups.)

- e.g. to allocate patients into the progressive or non-progressive groups

In practice, these two tasks often overlap.

III.1.2 Two-group discriminant analysis

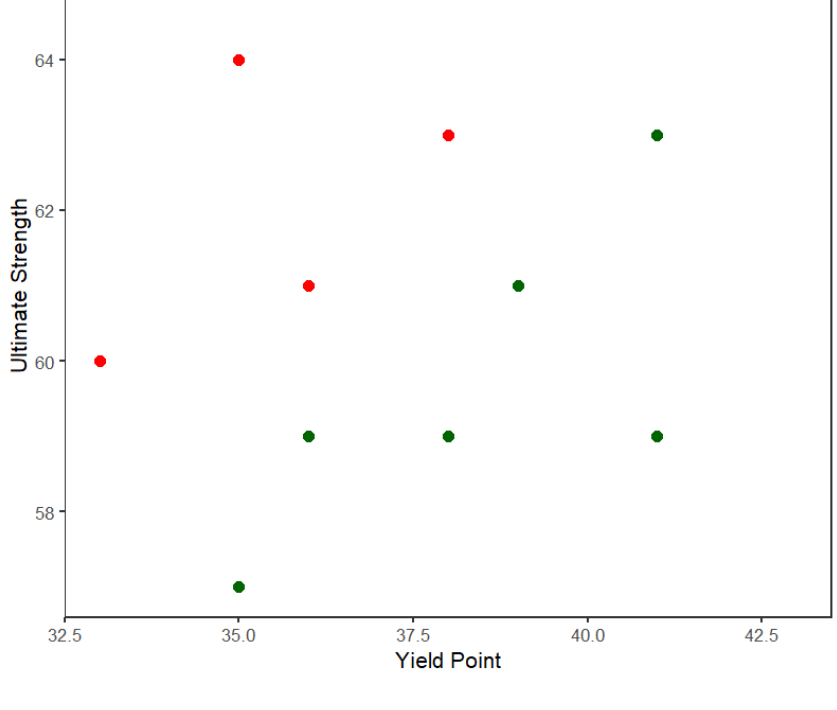
How to describe the difference of two groups?

- If they are from two populations?
  e.g. $\mathbf{X}_A \sim MN(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)$ and $\mathbf{X}_B \sim MN(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$

- If they are two sets of observations?

Example. Yield Point and Ultimate Strength of Steel Producted at Two Rolling Temperatures (5th Ed by Johnson and Wichern)

Temperature A		Temperature B	
x_{1A}	x_{2A}	x_{1B}	x_{2B}
33	60	35	57
36	61	36	59
35	64	38	59
38	63	39	61
40	65	41	63
		43	65
		41	59



Temperature A			Temperature B		
x_{1A}	x_{2A}	$\mathbf{a}'\mathbf{x}_A$	x_{1B}	x_{2B}	$\mathbf{a}'\mathbf{x}_B$
33	60	55.29	35	57	46.56
36	61	52.20	36	59	48.57
35	64	59.30	38	59	45.30
38	63	52.58	39	61	47.30
40	65	52.95	41	63	47.68
			43	65	48.05
			41	59	40.40

How to find the direction along which the largest difference between the two groups is shown?

III.1.2 Two-group discriminant analysis

Consider two groups of observations: $\mathbf{x}_{A1}, \ldots, \mathbf{x}_{Am}$; $\mathbf{x}_{B1}, \ldots, \mathbf{x}_{Bn}$
The difference between their projections on $\mathbf{b}$ is

$$diff_{AB}(\mathbf{b}) = \left[\frac{\bar{z}_A(\mathbf{b}) - \bar{z}_B(\mathbf{b})}{s_z}\right]^2 = \frac{\left[\mathbf{b}'(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B)\right]^2}{\mathbf{b}'\mathbf{S}_{pooled}\mathbf{b}}$$

- $\bar{z}_A(\mathbf{b})$ and $\bar{z}_B(\mathbf{b})$: the sample means of $\mathbf{b}'\mathbf{x}_{A1}, \ldots, \mathbf{b}'\mathbf{x}_{Am}$ and $\mathbf{b}'\mathbf{x}_{B1}, \ldots, \mathbf{b}'\mathbf{x}_{Bn}$;
- s_z^2: the projections' pooled sample variance.
- $\mathbf{S}_{pooled}$: the two sets of observations pooled sample variance,

$$\mathbf{S}_{pooled} = \frac{1}{n+m-2}\left[(m-1)\mathbf{S}_A + (n-1)\mathbf{S}_B\right]$$

Keep $\mathbf{b}'\mathbf{S}_{pooled}\mathbf{b}$ as a constant and maximize $diff_{AB}(\mathbf{b})$
$\Longrightarrow \mathbf{a} = \mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)$
$\Longrightarrow diff_{AB}(\mathbf{a}) = (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B) \propto T^2$-test statistic
Discriminant function: $\mathbf{a}'\mathbf{x} = (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)'\mathbf{S}_{pooled}^{-1}\mathbf{x}$.

III.1.3A Classification with two populations: with known distn

Consider two populations with sample space Ω: Population π_A with $f_A(\cdot)$; Population π_B with $f_B(\cdot)$

Classification rule: $\Omega = R_A \bigcup R_B$

- R_A: the set of all $\mathbf{x}$ for subjects classified into π_A
- R_B: the set of all $\mathbf{x}$ for subjects classified into π_B

Measures of classification accuracy:

- Conditional prbabilities:
  - $p(B|A) = P(\mathbf{x} \in R_B|\pi_A) = \int_{R_B} f_A(\mathbf{x})d\mathbf{x}$: Prob of classifying a subject into π_B when it's from π_A
  - $p(A|B) = P(\mathbf{x} \in R_A|\pi_B) = \int_{R_A} f_B(\mathbf{x})d\mathbf{x}$: Prob of classifying a subject into π_A when it's from π_B

Measures of classification accuracy: (cont'd)

- Marginal prbabilities of the accuracy: provided the prior prob $p_A = P(\pi_A)$ and $p_B = P(\pi_B)$,
  - $P(B|A)p_A$ = Prob of misclassifying a subject into π_B when it's from π_A
  - $P(A|B)p_B$ = Prob of classifying a subject into π_A when it's from π_B

- Total probability of misclassification (TPM): TPM=$p(A|B)p_B + p(B|A)p_A$.

- Expected cost of misclassification (ECM): ECM=$c(A|B)p(A|B)p_B + c(B|A)p(B|A)p_A$, provided
  - the cost of misclassifying a subject into π_A when it's from π_B is $c(A|B)$; the cost of misclassifying a subject into π_B when it's from π_A is $c(B|A)$; $c(A|A) = c(B|B) = 0$

How to find a classification rule R_A (or R_B) that minimizes ECM (or TPM)?

Optimal classification rule: R_A minimizes ECM if

$$R_A = \{\mathbf{x} : \frac{f_A(\mathbf{x})}{f_B(\mathbf{x})} \geq \frac{c(A|B)p_B}{c(B|A)p_A}\}.$$

Special cases:

- (a) If $p_A/p_B = 1$ (a subject from the two populations with the same probability),

$$R_A = \{\mathbf{x} : \frac{f_A(\mathbf{x})}{f_B(\mathbf{x})} \geq \frac{c(A|B)}{c(B|A)}\}.$$

- (b) If $c(A|B)/c(B|A) = 1$ (the costs of the two types of misclassification are equal),

$$R_A = \{\mathbf{x} : \frac{f_A(\mathbf{x})}{f_B(\mathbf{x})} \geq \frac{p_B}{p_A}\}.$$

- (c) If $c(A|B)/c(B|A) = p_A/p_B = 1$,

$$R_A = \{\mathbf{x} : \frac{f_A(\mathbf{x})}{f_B(\mathbf{x})} \geq 1\}.$$

Remarks:

- The optimal classification involves only the ratios of the costs and the prior probabilities.
- Can the classification rule be extended to more than two populations?

III.1.3B Classification with two populations: examples

Consider two populations with sample space Ω: Population π_A with $MN(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)$; Population π_B with $MN(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$;

- Case 1: $\boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_B = \boldsymbol{\Sigma}$
  - When $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$ and $\boldsymbol{\Sigma}$ are known, the optimal classification rule is

$$\begin{aligned} R_A &= \left\{\mathbf{x} : \frac{f_A(\mathbf{x})}{f_B(\mathbf{x})} \geq \frac{c(A|B)p_B}{c(B|A)p_A}\right\} \\ &= \left\{\mathbf{x} : (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)' \boldsymbol{\Sigma}^{-1}[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_A + \boldsymbol{\mu}_B)] \geq \log\left(\frac{c(A|B)p_B}{c(B|A)p_A}\right)\right\} \end{aligned}$$

  - When $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$ and $\boldsymbol{\Sigma}$ are unknown, the optimal classification rule is

$$R_A = \left\{\mathbf{x} : (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)' \mathbf{S}_{pooled}^{-1}[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_A + \bar{\mathbf{x}}_B)] \geq \log\left(\frac{c(A|B)p_B}{c(B|A)p_A}\right)\right\}.$$

    $\bar{\mathbf{x}}_A, \bar{\mathbf{x}}_B$ and $\mathbf{S}_{pooled}$ are the sample means and pooled sample variance.

How is R_A related to the discriminant function?

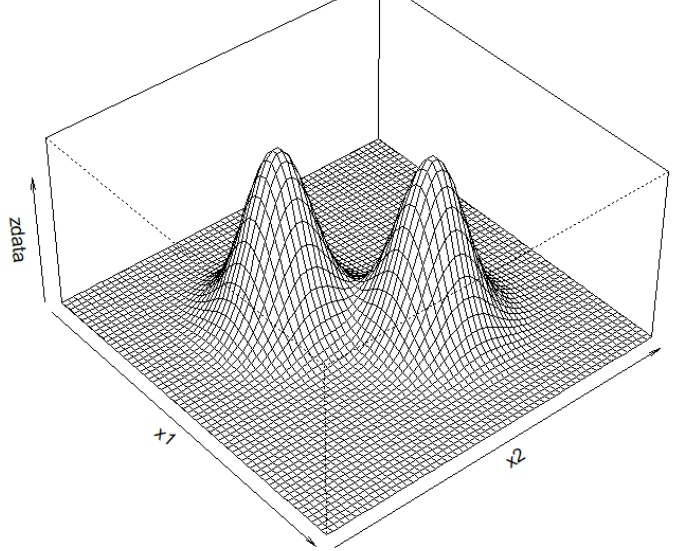
```
> library(mvtnorm)
> zxy <- cbind(x=rep((-30:30)/5, rep(61, 61)), y = rep((-30:30)/5, 61))
> muA <- c(-1.5,-1.5)
> muB <- c(1.5,1.5)
> Sigma <- rbind(c(1, 0), c(0, 1))
> a = solve(Sigma)%*%(muA-muB)
> a
     [,1]
[1,]   -3
[2,]   -3
> t(a)%*%(muA+muB)/2
     [,1]
[1,]    0
```

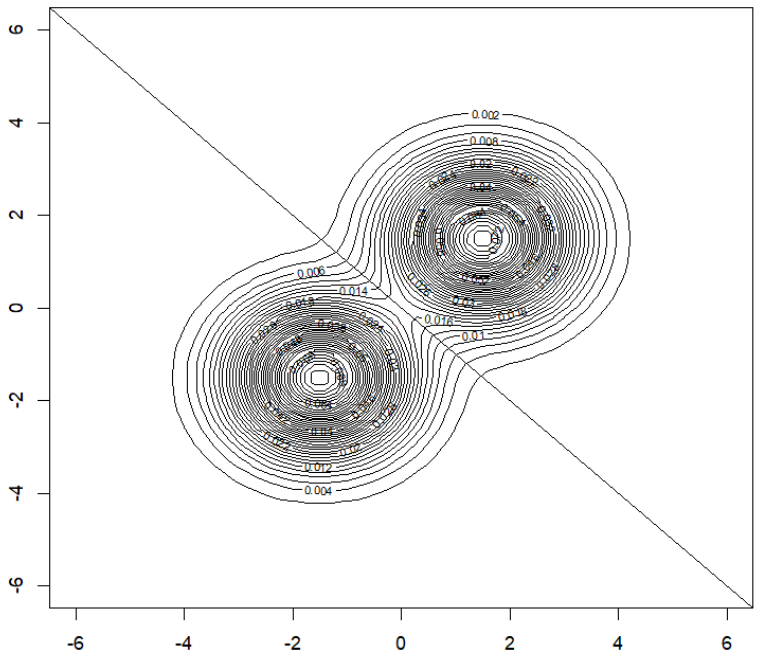
With $\mathbf{a} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)$,

$$R_A = \left\{\mathbf{x} : \mathbf{a}^{'}[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_A + \boldsymbol{\mu}_B)] \geq \log\left(\frac{c(A|B)p_B}{c(B|A)p_A}\right)\right\}$$

$\Longrightarrow$ When $c(A|B) = c(B|A)$ and $p_A = p_B$,
$R_A = \left\{\mathbf{x} : \mathbf{a}^{'}[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_A + \boldsymbol{\mu}_B)] \geq 0\right\} = \left\{\mathbf{x} : x_1 + x_2 \leq 0\right\}$

$\Longrightarrow$ With $\Delta = \sqrt{(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)^{'}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)}$, $p(B|A) = \Phi(-\frac{1}{2}\Delta)$ and $p(A|B) = \Phi(-\frac{1}{2}\Delta)$: if $\Delta = 0$, $p(A|B) = p(B|A) = 1/2$.


zdata
x1
x2


6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6
0.002
0.008
0.006
0.014
0.01
0.012
0.004

- Case 2: $\boldsymbol{\Sigma}_A \neq \boldsymbol{\Sigma}_B$
  - When $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$ and $\boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_B$ are known, the optimal classification rule is

$$R_A = \{\mathbf{x} : \frac{f_A(\mathbf{x})}{f_B(\mathbf{x})} \geq \frac{c(A|B)p_B}{c(B|A)p_A}\}$$

$$= \{\mathbf{x} : (\boldsymbol{\mu}_A^{'}\boldsymbol{\Sigma}_A^{-1} - \boldsymbol{\mu}_B^{'}\boldsymbol{\Sigma}_B^{-1})\mathbf{x} - \frac{1}{2}\mathbf{x}^{'}(\boldsymbol{\Sigma}_A^{-1} - \boldsymbol{\Sigma}_B^{-1})\mathbf{x} - k \geq \log\left(\frac{c(A|B)p_B}{c(B|A)p_A}\right)\}$$

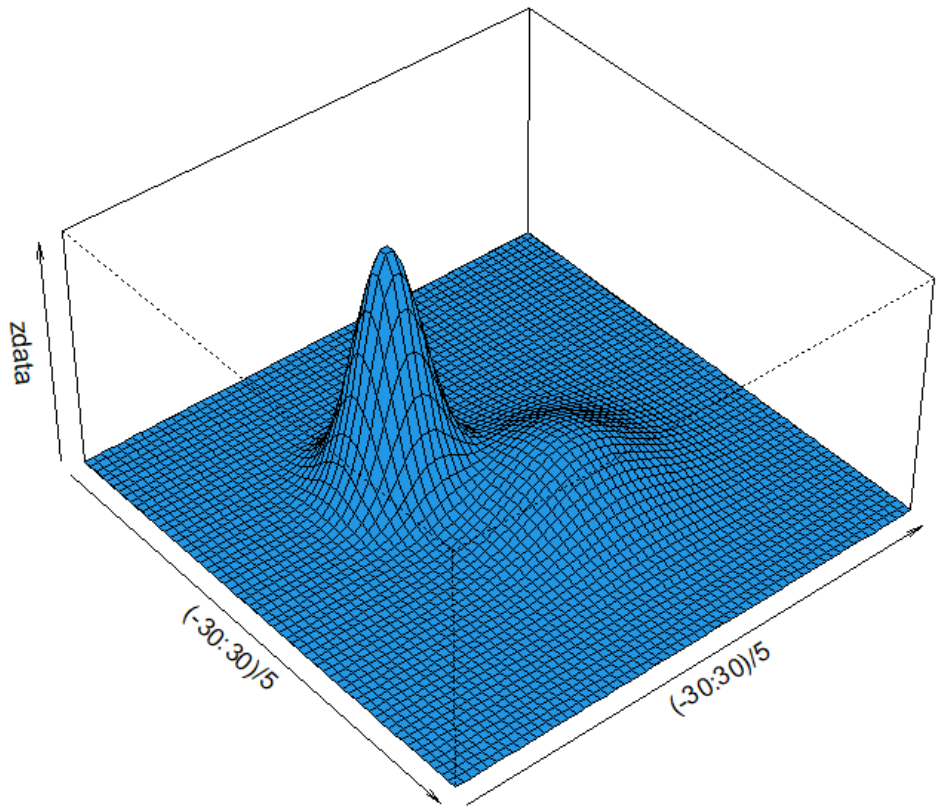
$$\text{constant } k = \tfrac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_A|}{|\boldsymbol{\Sigma}_B|}\right) + \tfrac{1}{2}(\boldsymbol{\mu}_A^{'}\boldsymbol{\Sigma}_A^{-1}\boldsymbol{\mu}_A - \boldsymbol{\mu}_B^{'}\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\mu}_B)$$

  - When $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$ and $\boldsymbol{\Sigma}_A, \boldsymbol{\Sigma}_B$ are unknown, use $\bar{\mathbf{x}}_A, \bar{\mathbf{x}}_B$ and $\mathbf{S}_A, \mathbf{S}_B$ to approximate them in R_A above.

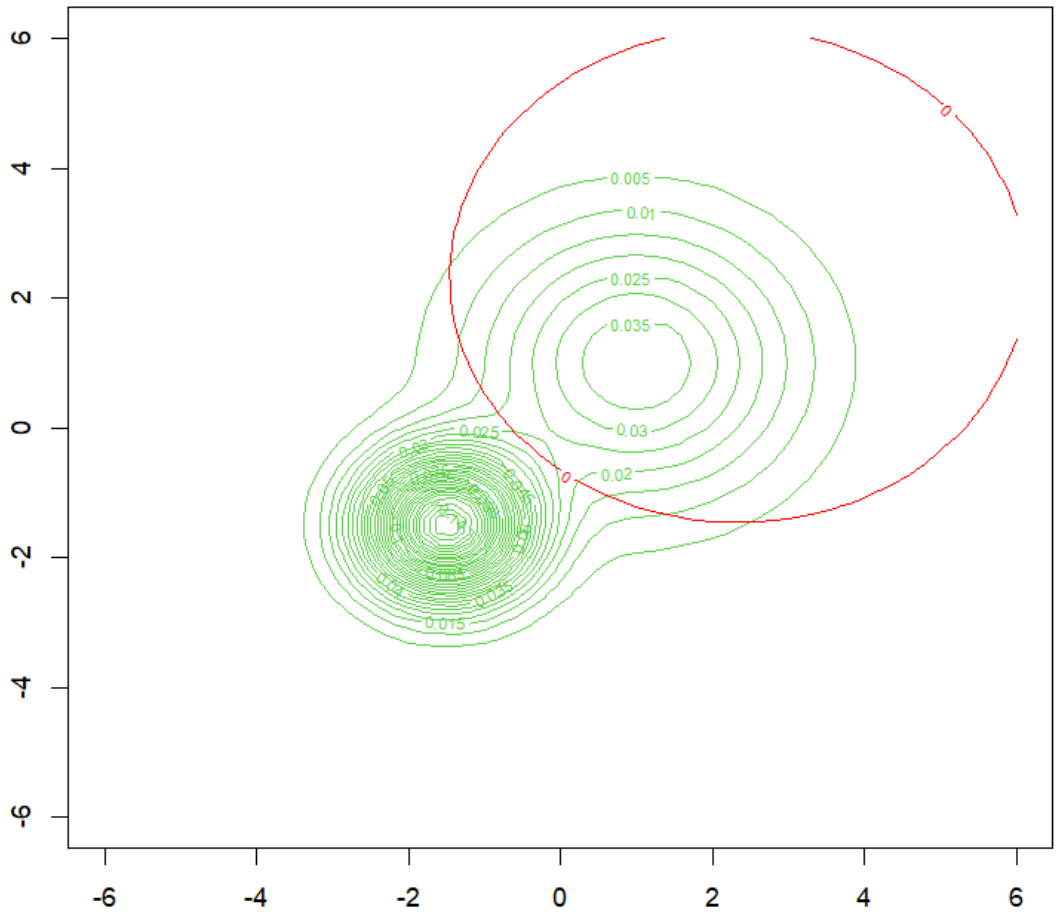
Remarks:

- R_A is determined by a quadratic curve instead of a straight line.
- The classification rule may be more sensitive to the normal assumption.

```
> library(mvtnorm)
> mu1 <- c(-1.5, -1.5)
> mu2 <- c(1, 1)
> Sigma1 <- rbind(c(.5, 0), c(0, .5))
> Sigma2 <- rbind(c(2, 0), c(0, 2))
> zdata <- matrix(.5*dmvnorm(zxy, mean = mu1, sigma = Sigma1)
+ + .5*dmvnorm(zxy, mean = mu2, sigma = Sigma2), ncol = 61, byrow = T)
> persp((-30:30)/5, (-30:30)/5, zdata, theta = 50, phi = 40, r = 10, expand = .5,
+ ltheta = 50,
```



```
> f <- function(x){
+ k = log(det(Sigma1)/det(Sigma2))/2 +(t(mu1)%*%solve(Sigma1)%*%mu1-t(mu2)%*%solve(Sigma2)%*%mu2)/2 -t(x)%
+  + (t(mu1)%*%solve(Sigma1)-t(mu2)%*%solve(Sigma2))%*%x - k
+  }
> rr <- matrix(apply(zxy,1,f),ncol=61,byrow=T)
> contour((-30:30)/5, (-30:30)/5, zdata, nlevels = 30,col=3)
> contour((-30:30)/5, (-30:30)/5,rr,nlevels=0,add=T,col="red")
```



III.1.3C Classification with two populations: evaluating classification

Actual error rate (AER). Given a sample classification rule R_A,

$$AER = p_B \int_{R_A} f_B(\mathbf{x})d\mathbf{x} + p_A \int_{R_B} f_A(\mathbf{x})d\mathbf{x} +$$

Problem: $f_A(\cdot)$ and $f_B(\cdot)$ are unknown.

Apparent error rate (APER): $\frac{n_{AM}+n_{BM}}{n_A+n_B}$
n_A, n_B: num of subjects in Populations A,B; n_{AM}, n_{BM}: num of subjects in Populations A,B and misclassified.

- Pros: easy to obtain and no need of parametric assumptions;
- Cons: may underestimate AER – the data used to build the classification function are also used to evaluate it, bias is smaller for larger training samples.

III.1.3C Classification with two populations: evaluating classification

Cross validation:

- (a) Randomly split data in a sample into g groups
- (b) Set aside one of the g groups as a validation sample
  - build the classification rule from the other $g-1$ groups (the training sample)
  - classify the data in the validation sample and record the results
- (c) Repeat the previous step g times

Holdout procedure (jackknife procedure)

- Cross-validation with $g = n$ groups, each with one obs.
- Estimate AER by $\frac{n_{AM}^{(H)}+n_{BM}^{(H)}}{n_A+n_B}$: $n_{AM}^{(H)}, n_{BM}^{(H)}$ are the number of holdout subjects in Populations A, B that are misclassified to Populations B,A.

What will we study next?

- *Part I. Introduction and Preparation*
- *Part II. Inference under Multivariate Normal Distribution (Chp 4-7)*
- **Part III. Commonly-Used Multivariate Analysis Methods (Chp 8-9; 11-12)**
  - **III.1. Discriminant Analysis and Classification (Chp 11)**
  - *III.2. Principal Component Analysis (Chp 8)*
  - *III.3. Factor Analysis (Chp 9)*
  - *III.4. Clustering (Chp 12)*