CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Discrimination and Classification: Applications to Discriminant Analysis

AUTHOR: Shiying Wang

SUPERVISOR: Dr. Natalia Stepanova

DATE: May 1st, 2020

Abstract

The project is focused on discrimination and classification with two or more populations. The goal is to gain some basic knowledge of the classical theory of classification and discrimination for two and more populations, in normal and non-normal setups. In the project, the notions of separation and classification, as well as methods of evaluating the classification functions and Fisher's method for discriminating among several populations are introduced and discussed. These methods are then used to solve selected theoretical and practical problems posed after Chapter 11 in the book of Johnson and Wichern [4]. To make the solutions more illustrative, MINITAB and R are used.

Contents

Page

1 Introduction **3**

2 Separation and Classification for Two (or More) Populations **3**

2.1 Basic Definitions 3

2.2 Two Examples of Discriminating Between Two Multivariate Populations 6

3 Classification with Two Multivariate Normal Populations **9**

3.1 The Estimated Minimum ECM Rule for Two Normal Population 10

3.2 An allocation rule based on Fisher's discriminant function 11

3.3 An Example of Allocation Based on Fisher's Discriminant Function 13

4 Evaluating Classification Functions **15**

5 Classification with Several Populations **17**

5.1 The Minimum Expected Cost of Misclassification Method 17

5.2 Minimum ECM Classification Rule with Equal Misclassification Costs 18

5.3 Classification with Several Normal Populations (Unequal $\mathbf{\Sigma}_i$) 19

5.4 An Example of Finding The Quadratic Discriminator 21

6 Fisher's Method for Discriminating among Several Populations **22**

6.1 Fisher's Sample Linear Discriminants 23

6.2 Two Examples of Fisher's Sample Discriminants for Two (or More) Populations . 25

7 Conclusion **38**

8 Appendix **39**

1 Introduction

Discrimination and classification are multivariate techniques that are based on a multivariate observation $\mathbf{X}$. The goal of discrimination is to describe the differential features of objects (observations) that can separate the known collections (populations). The goal of classification is to allocate a new object (observation) to previously defined groups.
In practice, when we want to discriminate the known populations (observations), we first need to allocate them. Conversely, a discriminator will be needed to allocate the objects (observations), so the goals of discrimination and classification are frequently overlapped (see pp. 376–377 in [2]). For example, in the university, the grades that students get for final exam could be classified as *pass* or *fail*. The determination can be made on the basis of exam scores by professor.

2 Separation and Classification for Two (or More) Populations

2.1 Basic Definitions

The idea of the separation and classification is to separate two (or more) classes of objects or to assign a new object into two (or more) labeled classes π_1 and π_2 (or π_1, π_2,..., π_k). We shall first focus on two-class classification.

Suppose that the observed values of $\mathbf{X}^\top = (X_1, X_2, \ldots, X_p)$ from population π_1 can be described by probability density functions $f_1(\mathbf{x})$ and the observed values from population π_2 can be described by probability density functions $f_2(\mathbf{x})$. The set of all possible sample outcomes $\mathbf{x} = (x_1, \ldots, x_n)^\top$ is divided into two regions R_1 and R_2 such that if a new object is allocated to π_1, it belongs to R_1, and if a new object is allocated to π_2, it belongs to R_2.

It is clear that classification rules cannot be completely precise; there might be some cases that we cannot distinguish between the features of the measured population because of the following possible conditions that might occur in practice (see [3]): (1) incomplete knowledge of future performance; (2) perfect information requires destroying the object; (3) unavailable or expensive information.

Then, it is possible that the errors occur when we try to classify the measured objects into the incorrect population, for example, classifying a π_2 object as belonging to π_1 or a π_1 object as belonging to π_2. However, it is assured that a "good" classification procedure should have as small possibilities of misclassification as possible and minimize the cost of classification.

Let $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ be the probability density functions associated with the $p \times 1$ random vector $\mathbf{X}$ from the populations π_1 and π_2, respectively. An object with associated measurement $\mathbf{x}$ *must* be assigned to either π_1 or π_2. Let Ω be the sample space – the collection of all possible

realizations $\mathbf{x}$ of $\mathbf{X}$. Typically, Ω is either $\mathbb{R}^n$ or a subset of $\mathbb{R}^n$ for some integer n. Let R_1 be the set of $\mathbf{x}$-values for which we classify objects as belonging to π_1 and $R_2 = \Omega \backslash R_1$ be the remaining $\mathbf{x}$-values for which we classify objects as belonging to π_2. R_1 and R_2 are mutually exclusive since any $\mathbf{x}$ must be assigned to one of the two populations (see Section 11.1 of [4]).

The conditional probability of classifying an object as π_2 when it is actually from π_1 is

$$P(2|1) := P(X \in R_2|\pi_1) = \int_{R_2=\Omega\backslash R_1} f_1(\mathbf{x})d\mathbf{x}.$$

Similarly, the conditional probability of classifying an object as π_1 when it is, infact, from π_2 is

$$P(1|2) := P(X \in R_1|\pi_2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x}.$$

Let p_1 and p_2 be the *prior* probability of π_1 and π_2 respectively. We have $p_1 + p_2 = 1$. Let $P(\pi_1)$ be the probability that observation comes from π_1 and $P(\pi_2)$ be the probability that observation comes from π_2. Then, the overall probabilities of correctly or incorrectly classifying objects can be derived as follows (see pp. 580–581 in [4]):

$$\begin{gathered} P(\text{observation is correctly classified as } \pi_1) \\ = P(\text{observation comes from } \pi_1 \text{ and is correctly classified as } \pi_1) \\ = P(\mathbf{X} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1, \end{gathered}$$

$$\begin{gathered} P(\text{observation is misclassified as } \pi_1) \\ = P(\text{observation comes from } \pi_2 \text{ and is misclassified as } \pi_1) \\ = P(\mathbf{X} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2, \end{gathered}$$

$$\begin{gathered} P(\text{observation is correctly classified as } \pi_2) \\ = P(\text{observation comes from } \pi_2 \text{ and is correctly classified as } \pi_2) \\ = P(\mathbf{X} \in R_2|\pi_2)P(\pi_2) = P(2|2)p_2, \end{gathered}$$

$$\begin{gathered} P(\text{observation is misclassified as } \pi_2) \\ = P(\text{observation comes from } \pi_1 \text{ and is misclassified as } \pi_2) \\ = P(\mathbf{X} \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1. \end{gathered}$$

The costs of misclassification can be defined by a *cost matrix*:

		Classify as: π_1	π_2
True population:	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

Consequently, the average, or *expected cost of misclassification* (ECM) is

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

Fact 1 (Result 11.1 in [4]). The regions R_1 and R_2 that minimize the ECM are known to be defined as follows:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right), \tag{2}$$

$$(density\ ratio) \geq (cost\ ratio)(prior\ probability\ ratio)$$

$$R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right).$$

$$(density\ ratio) < (cost\ ratio)(prior\ probability\ ratio)$$

Special Cases of Minimum Expected Cost Regions

a) If $p_2/p_1 = 1$ (equal prior probability), then

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

b) If $c(1|2)/c(2|1) = 1$ (equal misclassification costs), then

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1}, \quad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1}. \tag{3}$$

c) If $p_2/p_1 = c(1|2)/c(2|1) = 1$ or $p_2/p_1 = 1/(c(1|2)/c(2|1))$ (equal prior probability and equal misclassification costs), then

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1, \quad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1. \tag{4}$$

2.2 Two Examples of Discriminating Between Two Multivariate Populations

Exercise 11.1 in [4]. A researcher wants to determine a procedure for discriminating between two multivariate populations. The researcher has enough data available to estimate the density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ associated with populations π_1 and π_2, respectively. Let $c(2|1) = 50$ (this is the cost of assigning items as π_2, given that π_1 is true) and $c(1|2) = 100$. In addition, it is known that about 20% of all possible items (for which the measurements $\mathbf{x}$ can be recorded) belong to π_2.

(a) Give the minimum ECM rule (in general form) for assigning a new item to one of the two populations.

(b) Measurements recorded on a new item yield the density values $f_1(\mathbf{x}) = 0.3$ and $f_2(\mathbf{x}) = 0.5$. Given the preceding information, assign this item to population π_1 or population π_2.

Solution: By assumption, $p_2 = 0.2$ and $p_1 = 0.8$.

(a) By the definition of the ECM (1), we have

$$\begin{aligned}\text{ECM} &= c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \\ &= 40P(2|1) + 20P(1|2).\end{aligned}$$

$$\begin{aligned}R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &\geq \frac{c(1|2)}{c(2|1)}\left(\frac{p_2}{p_1}\right) \\ &= \frac{100}{50} \times \frac{0.2}{0.8} = 0.5,\end{aligned}$$

that is,

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 0.5,$$

and hence

$$R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 0.5.$$

(b) Since

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{0.3}{0.5} = 0.6 > 0.5,$$

according to part (a), we assign this item to population π_1.

Exercise 11.7 in [4]. Let $f_1(x) = (1-|x|)$ for $|x| \leq 1$ and $f_2(x) = (1-|x-0.5|)$ for $-.5 \leq x \leq 1.5$.

(a) Sketch the two densities.

(b) Identify the classification regions when $p_1 = p_2$ and $c(1|2) = c(2|1)$.

(c) Identify the classification regions when $p_1 = .2$ and $c(1|2) = c(2|1)$.

Solution:

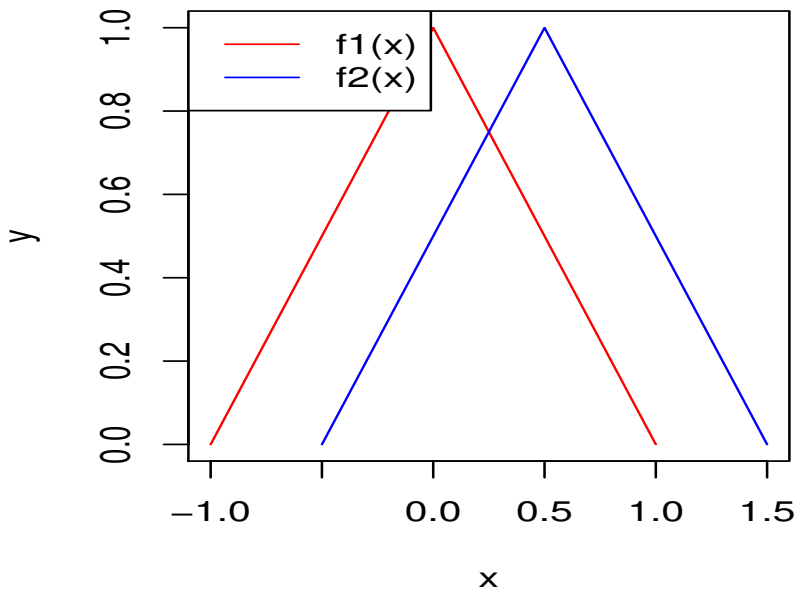


Figure 1: The graphs of $f_1(x)$ and $f_2(x)$ in Exercise 11.7

(a)

(b) By assumption, $p_1 = p_2$ and $c(1|2) = c(2|1)$. Therefore, using (4), we obtain

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq 1,$$

that is,

$$f_1(x) \geq f_2(x).$$

From this (see Figure 1),

$$-0.5 \leq x \leq 0.25,$$

and hence

$$R_2 : \frac{f_1(x)}{f_2(x)} < 1,$$

giving (see Figure 1)

$$0.25 < x \leq 1.$$

Thus, $R_1 : -0.5 \leq x \leq 0.25$ and $R_2 : 0.25 < x \leq 1$.

(c) By assumption, $p_1 = 0.2$ and $c(1|2) = c(2|1)$, that is, $p_2 = 0.8$. By (3), we have

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1}.$$

We consider the above inequality for those $x \in [-0.5, 1]$ for which $f_1(x) > 0$ and $f_1(x) > 0$. That is, the region R_1 is given by

$$\frac{f_1(x)}{f_2(x)} \geq 4,$$

or equivalently,

$$\frac{1 - |x|}{1 - |x - 0.5|} \geq 4,$$

and hence,

$$(4|x - 0.5| - |x|) \geq 3.$$

Case 1: for $-0.5 \leq x \leq 0$, the above inequality gives

$$4(0.5 - x) + x \geq 3,$$

$$-0.5 \leq x \leq -1/3,$$

Case 2: for $0 < x \leq 0.5$, we obtain

$$4(0.5 - x) - x \geq 3,$$

$$x \leq -1/5,$$

Case 3: for $0.5 < x \leq 1$, we have

$$4(x - 0.5) - x \geq 3,$$

$$x \geq 5/3.$$

In conclusion, $R_1 : -0.5 \leq x \leq -1/3$ and $R_2 : -1/3 < x \leq 1$.

3 Classification with Two Multivariate Normal Populations

Assuming that $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are multivariate normal densities, the corresponding mean vectors are $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and the common covariance matrix is $\boldsymbol{\Sigma}$. The density of a multivariate normal $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ distribution is

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right], \quad i = 1, 2.$$

If the population parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known, then the minimum ECM regions become

$$R_1 : \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right),$$

$$R_2 : \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right).$$

Taking the natural logarithms on the left-hand side, we obtain the following:

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2).$$

Consequently, the allocation rule that minimizes the ECM is as follows: allocate $\mathbf{x}_0$ to π_1 if

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right], \tag{5}$$

allocate $\mathbf{x}_0$ to π_2 otherwise (see Section 11.3 in [4]).

Suppose that we have n_1 observations of random vector $\mathbf{X}^\top = (X_1, X_2, \ldots, X_p)$ from π_1 and n_2 measurements of this quantity from π_2, with $n_1 + n_2 - 2 \geq p$. That is, n_1 observations $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1}$ are from population π_1, and n_2 observations $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2}$ are from population π_2. Then the respective data matrices are (see [1] and [5])

$$\underset{(n_1\times p)}{\mathbb{X}_1} = \begin{pmatrix} \mathbf{X}_{11}^\top \\ \mathbf{X}_{12}^\top \\ \ldots \\ \mathbf{X}_{1n_1}^\top \end{pmatrix}, \quad \underset{(n_2\times p)}{\mathbb{X}_2} = \begin{pmatrix} \mathbf{X}_{21}^\top \\ \mathbf{X}_{22}^\top \\ \ldots \\ \mathbf{X}_{2n_2}^\top \end{pmatrix}.$$

Let $\mathbf{X}_{ij}$ be the element of $\mathbb{X}_i$ with $i = 1, 2$ and $j = 1, 2, \ldots, n_i$. The sample mean vectors and

covariance matrices are defined as follows:

$$\bar{\mathbf{X}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{X}_{1j}, \quad \mathbf{S}_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{X}_{1j}-\bar{\mathbf{X}}_1)(\mathbf{X}_{1j}-\bar{\mathbf{X}}_1)^\top,$$

$$\bar{\mathbf{X}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{X}_{2j}, \quad \mathbf{S}_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\mathbf{X}_{2j}-\bar{\mathbf{X}}_2)(\mathbf{X}_{2j}-\bar{\mathbf{X}}_2)^\top.$$

We assume the parent populations have the same covariance matrix $\boldsymbol{\Sigma}$, and the sample covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are combined (pooled) to derive a single unbiased estimate of $\boldsymbol{\Sigma}$. Namely,

$$\mathbf{S}_{\text{pooled}} = \left[\frac{n_1-1}{(n_1-1)+(n_2-1)}\right]\mathbf{S}_1 + \left[\frac{n_2-1}{(n_1-1)+(n_2-1)}\right]\mathbf{S}_2$$

is an unbiased estimate of $\boldsymbol{\Sigma}$ if the data matrices $\mathbb{X}_1$ and $\mathbb{X}_2$ contain random samples from the population π_1 and π_2, respectively. Substituting $\bar{\mathbf{X}}_1$ for $\boldsymbol{\mu}_1$, $\bar{\mathbf{X}}_2$ for $\boldsymbol{\mu}_2$, and $\mathbf{S}_{\text{pooled}}$ for $\boldsymbol{\Sigma}$ in (5) gives the classification rule, called the *estimated minimum ECM rule*.

3.1 The Estimated Minimum ECM Rule for Two Normal Population

The estimated minimum ECM rule prescribes to allocate $\mathbf{x}_0$ to π_1 if

$$(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)^\top\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)^\top\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{X}}_1+\bar{\mathbf{X}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right], \tag{6}$$

and to allocate $\mathbf{x}_0$ to π_2 otherwise.

If in (6) we have

$$\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = 1,$$

then $\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] = 0$, and the estimated minimum ECM rule for two normal populations amounts to comparing the scalar variable (see p. 586 of [4])

$$\hat{\mathbf{y}} = (\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)^\top\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} =: \hat{\mathbf{a}}^\top\mathbf{x} \tag{7}$$

evaluated at $\mathbf{x}_0$, with the number

$$\hat{\mathbf{m}} := \frac{1}{2}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)^\top\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{X}}_1+\bar{\mathbf{X}}_2) = \frac{1}{2}(\bar{\mathbf{y}}_1+\bar{\mathbf{y}}_2), \tag{8}$$

where

$$\bar{\mathbf{y}}_1 = (\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)^\top\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{X}}_1 = \hat{\mathbf{a}}^\top\bar{\mathbf{X}}_1$$

and

$$\bar{\mathbf{y}}_2 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\top \mathbf{S}_{\text{pooled}}^{-1} \bar{\mathbf{X}}_2 = \hat{\mathbf{a}}^\top \bar{\mathbf{X}}_2.$$

The following approach to the separation problem is due to Fisher [2]. This approach does not assume that the two populations are normal.

3.2 An allocation rule based on Fisher's discriminant function

Allocate $\mathbf{x}_0$ to π_1 if

$$\hat{\mathbf{y}}_0 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\top \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x}_0 \geq \hat{\mathbf{m}} = \frac{1}{2}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\top \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2) \tag{9}$$

or

$$\hat{\mathbf{y}}_0 - \hat{\mathbf{m}} \geq 0,$$

allocate $\mathbf{x}_0$ to π_2 if

$$\hat{\mathbf{y}}_0 \leq \hat{\mathbf{m}}$$

or

$$\hat{\mathbf{y}}_0 - \hat{\mathbf{m}} \leq 0.$$

Fisher's linear discriminant rule as given in (9) was developed under the assumption of a common covariance matrix. In the case of two normal population with different covariance matrices, substituting normal densities into (2) gives (see Section 11.3 in [4])

$$R_1 : -\frac{1}{2}\mathbf{X}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{X} + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1})\mathbf{X} - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right],$$

$$R_2 : -\frac{1}{2}\mathbf{X}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{X} + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1})\mathbf{X} - k < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right],$$

where

$$k = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2). \tag{10}$$

Proof: The regions of minimum ECM depend on the ratio of the densities, $f_1(\mathbf{x})/f_2(\mathbf{x})$, or equivalently, the natural logarithm of the density ratio

$$\ln[f_1(\mathbf{x})/f_2(\mathbf{x})] = \ln f_1(\mathbf{x}) - \ln f_2(\mathbf{x}).$$

The density of a multivariate normal $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ distribution is

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right], \quad i = 1, 2.$$

Hence,

$$\begin{aligned}
\ln f_1(\mathbf{x}) - \ln f_2(\mathbf{x}) &= -\ln((2\pi)^{p/2}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_1| - \frac{(\mathbf{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}{2} \\
&\quad + \ln((2\pi)^{p/2}) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_2| + \frac{(\mathbf{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-2}(\mathbf{x}-\boldsymbol{\mu}_2)}{2} \\
&= -\frac{1}{2}(\ln|\boldsymbol{\Sigma}_1| - \ln|\boldsymbol{\Sigma}_2|) - \frac{1}{2}(\mathbf{x}^\top\boldsymbol{\Sigma}_1^{-1}\mathbf{x} - \mathbf{x}^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\mathbf{x} + \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 \\
&\quad + \mathbf{x}^\top\boldsymbol{\Sigma}_2^{-1}\mathbf{x} - \mathbf{x}^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\mathbf{x} + \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) \\
&= -\frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) - \frac{1}{2}\mathbf{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + \frac{1}{2}(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\mathbf{x} \\
&\quad - \frac{1}{2}(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) + \frac{1}{2}(\mathbf{x}^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{x}^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) \\
&= -\frac{1}{2}\mathbf{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + \frac{1}{2}(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\mathbf{x} \\
&\quad + \frac{1}{2}(\mathbf{x}^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{x}^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) - k, \qquad (11)
\end{aligned}$$

where k is defined in (10).

Since

$$\frac{1}{2}(\mathbf{x}^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{x}^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)^\top = \frac{1}{2}(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\mathbf{x} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\mathbf{x}) = \frac{1}{2}(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\mathbf{x},$$

we have

$$\frac{1}{2}(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\mathbf{x} + \frac{1}{2}(\mathbf{x}^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{x}^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) = (\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\mathbf{x}. \qquad (12)$$

Combining (11) and (12) with (2) yields

$$R_1 : -\frac{1}{2}\mathbf{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right],$$

$$R_2 : -\frac{1}{2}\mathbf{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right],$$

where k is given by (10).

3.3 An Example of Allocation Based on Fisher's Discriminant Function

Exercise 11.10 in [4]. Suppose that $n_1 = 11$ and $n_2 = 12$ observations are made on two random variables $\mathbf{X}_1$ and $\mathbf{X}_2$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are assumed to have a bivariate normal distribution with a common covariance matrix $\boldsymbol{\Sigma}$, but possibly different mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ for the two samples. The sample mean vectors and pooled covariance matrix are

$$\bar{\mathbf{X}}_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \bar{\mathbf{X}}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \mathbf{S}_{\text{pooled}} = \begin{pmatrix} 7.3 & -1.1 \\ -1.1 & 4.8 \end{pmatrix}.$$

(a) Test for the difference in population mean vectors using Hotelling's two-sample T^2-statistic. Let $\alpha = 0.10$.

(b) Construct Fisher's (sample) linear discriminant function. (See relations (7) and (9)).

(c) Assign the observation $\mathbf{x}_0^\top = (0,1)$ to either population π_1 or π_2. Assume equal costs and equal prior probabilities.

Solution: First, we recall the following result (see Result 6.2 in [4]): If $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$ is a random sample of size n_1 from $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_1}$ is an independent random sample of size n_2 from $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, then the Hotelling's two-sample test statistics for testing the hypothesis $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0$ is given by

$$T^2 := [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta}_0]^\top \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{\text{pooled}} \right]^{-1} [\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta}_0]$$

where $\bar{\mathbf{X}}_2$ and $\bar{\mathbf{X}}_2$ are as in (5) and (6), and, under H_0, it is distributed as follows:

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

Consequently,

$$P_{H_0} \left[(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta}_0)^\top \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{\text{pooled}} \right]^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta}_0) \leq c^2 \right] = 1 - \alpha,$$

where

$$c^2 = \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1+n_2-p-1}(1 - \alpha)$$

and $F_{p,n_1+n_2-p-1}(1-\alpha)$ is the upper α percentage point of an F-distribution with p and $n_1 + n_2 - p - 1$ degrees of freedom.

(a) The hypotheses for the test of difference in population mean vectors are:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

Then Hotelling's two-sample T^2-statistic is given by

$$T^2 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\top \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{\text{pooled}} \right]^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2).$$

By assumption, $\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 = \begin{pmatrix} -3 \\ -2 \end{pmatrix}$, $n_1 = 11$, $n_2 = 12$, $\mathbf{S}_{\text{pooled}} = \begin{pmatrix} 7.3 & -1.1 \\ -1.1 & 4.8 \end{pmatrix}$ and $p = 2$. Plugging these values and vectors into the expression for T^2, we obtain that the observed value of T^2 is equal to

$$\begin{aligned} T^2_{calculated} &= \begin{pmatrix} -3 \\ -2 \end{pmatrix}^\top \left(\frac{1}{11} + \frac{1}{12} \begin{pmatrix} 7.3 & -1.1 \\ -1.1 & 4.8 \end{pmatrix} \right)^{-1} \begin{pmatrix} -3 \\ -2 \end{pmatrix} \\ &= \begin{pmatrix} -3, & -2 \end{pmatrix} \begin{pmatrix} 1.2719 & -0.1917 \\ -0.1917 & 0.8363 \end{pmatrix}^{-1} \begin{pmatrix} -3 \\ -2 \end{pmatrix} \\ &= \begin{pmatrix} -2.8163, & -3.0369 \end{pmatrix} \begin{pmatrix} -3 \\ -2 \end{pmatrix} \\ &= 14.5225. \end{aligned}$$

Under the null hypothesis,

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

thus,

$$\frac{(23 - 2 - 1)}{2 \times 21} T^2 \sim F_{2,20},$$

that is,

$$\frac{10}{21} T^2 \sim F_{2,20},$$

Since $\alpha = 0.1$, using R, we get that the 90^{th} percentile of the F distribution with (2,20) degrees of freedom equals 2.5893. Since $\frac{10}{21} \times 14.5225 > 2.5893$, we reject the null hypothesis and accept the alternative hypothesis.

(b) By assumption, $\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathbf{S}^{-1}_{\text{pooled}} = \frac{1}{33.83} \begin{pmatrix} 4.8 & 1.1 \\ 1.1 & 7.3 \end{pmatrix}$. According to (7), the estimated minimum ECM rule for two normal population amounts to comparing the scalar

variable

$$\hat{\mathbf{Y}} = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_1)^\top \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x}_0 =: \hat{\mathbf{a}}^\top \mathbf{x}_0$$

with $\hat{\mathbf{m}}$ defined in (8). Plugging $\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2$ and $\mathbf{S}_{pooled}^{-1}$ into the above formula, we get

$$\hat{\mathbf{y}} = \frac{1}{33.83} \begin{pmatrix} -3, & 2 \end{pmatrix} \begin{pmatrix} 4.8 & 1.1 \\ 1.1 & 7.3 \end{pmatrix} \mathbf{x}_0,$$

that is,

$$\hat{\mathbf{y}} = \begin{pmatrix} -\frac{1660}{3383}, & -\frac{1790}{3383} \end{pmatrix} \mathbf{x}_0.$$

An allocation rule based on Fisher's discriminant function (9) is: allocate $\mathbf{x}_0$ to π_1 if

$$\hat{\mathbf{y}}_0 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\top \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x}_0 \geq \hat{\mathbf{m}} = \frac{1}{2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\top \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2)$$

or

$$\hat{\mathbf{y}}_0 - \hat{\mathbf{m}} \geq 0,$$

and allocate $\mathbf{x}_0$ to π_2 otherwise.
In this problem, we have

$$\hat{\mathbf{m}} = \frac{1}{2} \begin{pmatrix} -\frac{1660}{3383}, & -\frac{1790}{3383} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = -\frac{830}{3383} = -0.245.$$

(c) For the observation $\mathbf{x}_0^\top = (0, 1)$, we have

$$\hat{\mathbf{y}}_0 = -\frac{1790}{3383} = -0.529 < \hat{\mathbf{m}},$$

so we allocate $\mathbf{x}_0$ to π_2.

4 Evaluating Classification Functions

One important way of judging the performance of any classification procedure is to calculate the misclassification probabilities. The total probability of misclassification (TPM) is

$$\text{TPM} = p_1 \int_{R_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}.$$

The smallest value of this quantity obtained by a judicious choice of R_1 and R_2 is called the optimum error rate (OER):

$$\text{OER} = p_1 \int_{R_2} f_1(\mathbf{x})d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x},$$

where R_1 and R_2 are determined by (3). That is, the OER is the error rate for the minimum TPM classification rule.

The performance of sample classification functions can be evaluated by calculating the actual error rate (AER)

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

where $\hat{R}_1$ and $\hat{R}_2$ represent the classification regions determined by samples of sizes n_1 and n_2, respectively. The AER indicates how the sample classification function will perform in future samples.

The apparent error rate (APER) is defined as the fraction of observations in the training sample that are misclassified by the sample classification function. It can be calculated from the *confusion matrix* which shows actual versus predicted group membership. For n_1 observations from π_1 and n_2 observations from π_2, the *confusion matrix* has the form

		Predicted membership		
		π_1	π_2	
Actual membership	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

where

$$n_{1C} = \text{number of } \pi_1 \text{ items correctly classified as } \pi_1 \text{ items},$$

$$n_{1M} = \text{number of } \pi_1 \text{ items misclassified as } \pi_2 \text{ items},$$

$$n_{2C} = \text{number of } \pi_2 \text{ items correctly classified as } \pi_2 \text{ items},$$

$$n_{2M} = \text{number of } \pi_2 \text{ items misclassified as } \pi_1 \text{ items}.$$

The apparent error rate is then

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2},$$

which is recognized as the proportion of items in the training set that are misclassified (see Section 11.4 in [4]). The apparent error rate can be any number between 0 and 1. The smaller the apparent error rate is, the better the classification will be.

5 Classification with Several Populations

5.1 The Minimum Expected Cost of Misclassification Method

Let $f_i(\mathbf{x})$ be the density associated with population π_i, $i = 1, 2, \ldots, g$. We define

$$\begin{aligned} p_i &:= \text{the prior probability of population } \pi_i, \qquad i = 1, 2, \ldots, g, \\ c(k|i) &:= \text{the cost of allocating an item to } \pi_k \text{ when, in fact, it belongs to } \pi_i \\ &\quad \text{for } k, i = 1, 2, \ldots, g. \end{aligned}$$

For $k = i$, $c(i|i) = 0$.

Let R_k be the set of $\mathbf{x}$'s classified as π_k and

$$P(k|i) = P(\text{classifying item as } \pi_k|\pi_i) = \int_{R_k} f_i(\mathbf{x})d\mathbf{x},$$

for $k, i = 1, 2, \ldots, g$ with $P(i|i) = 1 - \sum_{k=1k\neq i}^{g} P(k|i)$. The conditional expected cost of misclassifying an $\mathbf{x}$ from π_1 into π_2, or π_3,..., or π_g is

$$\begin{aligned} \text{ECM}(1) &= P(2|1)c(2|1) + P(3|1)c(3|1) + \ldots + P(g|1)c(g|1) \\ &= \sum_{k=2}^{g} P(k|1)c(k|1). \end{aligned}$$

We can obtain the conditional expected costs of misclassification ECM(2),..., ECM(g) in a similar way. Multiplying each conditional ECM by its prior probability and summing everything gives the overall ECM:

$$\begin{aligned} \text{ECM} &= p_1\text{ECM}(1) + p_2\text{ECM}(2) + \ldots + p_g\text{ECM}(g) \\ &= p_1\left(\sum_{k=2}^{g} P(k|1)c(k|1)\right) + p_2\left(\sum_{k=1,\ k\neq 2} P(k|2)c(k|2)\right) \\ &\quad + \ldots + p_g\left(\sum_{k=1}^{g-1} P(k|g)c(k|g)\right) \\ &= \sum_{i=1}^{g} p_i\left(\sum_{k=1,\ k\neq i}^{g} P(k|i)c(k|i)\right). \end{aligned}$$

In order to determine an optimal classification procedure, we need to find the partition $R_1, \ldots, R_g$ of the sample space Ω for which the value of ECM is minimal(see Section 11.5 in [4]).

Fact 2 (Result 11.5 in [4]): The classification regions that minimize the ECM are defined

by allocating $\mathbf{x}$ to that population π_k, $k = 1, 2, \ldots, g$, for which

$$\sum_{i=1,\ i\neq k}^{g} p_i f_i(\mathbf{x}) c(k|i)$$

is smallest. If a tie occurs, $\mathbf{x}$ can be assigned to any of the tied populations.

Suppose all the misclassification costs are equal, in which case the minimum expected cost of misclassification rule is the minimum total probability of misclassification rule. (As a particular case, we set all the misclassification costs equal to 1.) Using Result 11.5, we would allocate $\mathbf{x}$ to that population π_k, $k = 1, 2, \ldots, g$, for which

$$\sum_{i=1,\ i\neq k}^{g} p_i f_i(\mathbf{x})$$

is smallest. This sum will be smallest when the omitted term $p_k f_k(\mathbf{x})$ is largest. Hence, in case of equal misclassification costs, we arrive at the following rule, called the *minimum ECM classification rule with equal misclassification costs*.

5.2 Minimum ECM Classification Rule with Equal Misclassification Costs

When the misclassification costs are the same, we allocate $\mathbf{x}_0$ to π_k if

$$p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}) \quad \text{for all } i \neq k,$$

or, equivalently, allocate $\mathbf{x}_0$ to π_k if

$$\ln\left(p_k f_k(\mathbf{x})\right) > \ln\left(p_i f_i(\mathbf{x})\right) \quad \text{for all } i \neq k. \tag{13}$$

The classification rule above is identical to the one that maximizes the "posterior" probability $P(\pi_k|\mathbf{x}) = P(\mathbf{x}$ comes from π_k given that $\mathbf{x}$ was observed), where

$$P(\pi_k|\mathbf{x}) = \frac{p_k f_k(\mathbf{x})}{\sum_{i=1}^{g} p_i f_i(\mathbf{x})} = \frac{(\text{prior}) \times (\text{likelihood})}{\sum[(\text{prior}) \times (\text{likelihood})]} \quad \text{for } k = 1, 2, \ldots, g,$$

see Section 11.5 in [4] for more details.

We shall now consider how this classification rule could be applied in case of two normal populations with different covariance matrices.

5.3 Classification with Several Normal Populations (Unequal $\boldsymbol{\Sigma}_i$)

We now assume that $f_i(\mathbf{x})$ are multivariate normal densities, the corresponding mean vectors are $\boldsymbol{\mu}_i$ and covariance matrices are $\boldsymbol{\Sigma}_i$. If, further, $c(i|i) = 0, c(k|i) = 1, k \neq i$ (or, equivalently, the misclassification costs are all equal), then (13) becomes: allocate $\mathbf{x}$ to π_k if

$$\begin{aligned} \ln(p_k f_k(\mathbf{x})) &= \ln p_k - \left(\frac{p}{2}\right)\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) \\ &= \max_i \ln(p_i f_i(\mathbf{x}))\,. \end{aligned}$$

Since $(p/2)\ln(2\pi)$ is constant, it can be ignored and we therefore define the quadratic discrimination score for the ith population to be

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln p_i, \qquad i = 1, 2, \ldots, g.$$

Using discriminant scores, we arrive at the following

Minimum Total Probability of Misclassification (TPM) Rule for Normal Populations with Unequal $\boldsymbol{\Sigma}_i$

Allocate $\mathbf{x}$ to π_k if

$$\text{the quadratic score } d_k^Q(\mathbf{x}) = \text{the largest of } d_1^Q(\mathbf{x}), d_2^Q(\mathbf{x}), \ldots, d_g^Q(\mathbf{x}).$$

In practice, the $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are unknown, but a training set of correctly classified observations is often available for the construction of estimates. The relevant sample quantities for population π_i are

$$\begin{aligned} \bar{\mathbf{x}}_i &= \text{sample mean vector,} \\ \mathbf{S}_i &= \text{sample covariance matrix,} \\ n_i &= \text{sample size.} \end{aligned}$$

The estimator of the quadratic discrimination score $\hat{d}_i^Q(\mathbf{x})$ is then

$$\hat{d}_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^\top \mathbf{S}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i) + \ln p_i, \qquad i = 1, 2, \ldots, g, \tag{14}$$

and the classification rule based on the sample is as follows (see Section 11.5 in [4]):

Estimated Minimum TPM Rule for Several Normal Populations with Unequal $\boldsymbol{\Sigma}_i$

Allocate $\mathbf{x}$ to π_k if

$$\text{the quadratic score } \hat{d}_k^Q(\mathbf{x}) = \text{the largest of } \hat{d}_1^Q(\mathbf{x}), \hat{d}_2^Q(\mathbf{x}), \ldots, \hat{d}_g^Q(\mathbf{x}). \tag{15}$$

If the population covariance matrices $\boldsymbol{\Sigma}_i$ are equal, that is, $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ for $i = 1, 2, \ldots, g$, the discriminant score becomes

$$\hat{d}_k^Q(\mathbf{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i, \qquad i = 1, 2, \ldots, g.$$

Since $c_i = \ln p_i - \frac{1}{2}\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$ is constant, we can ignore them for allocative purposes and then define the linear discriminant score

$$d_i(\mathbf{x}) = \boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \qquad \text{for } i = 1, 2, \ldots, g. \tag{16}$$

An estimator $\hat{d}_i(\mathbf{x})$ of the linear discriminant score $d_i(\mathbf{x})$ is based on the pooled estimate of $\boldsymbol{\Sigma}$, which is defined by

$$\mathbf{S}_{\text{pooled}} = \frac{1}{n_1 + n_2 + \ldots + n_g - g}((n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \ldots + (n_g - 1)\mathbf{S}_g),$$

and is given by

$$\hat{d}_i(\mathbf{x}) = \bar{\mathbf{x}}_i^\top\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \frac{1}{2}\bar{\mathbf{x}}_i^\top\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_i + \ln p_i \qquad \text{for } i = 1, 2, \ldots, g. \tag{17}$$

Consequently, we have the following rule (see Section 11.5 in [4]):

Estimated Minimum TPM Rule for Equal-Covariance Normal Populations

Allocate $\mathbf{x}$ to π_k if

$$\text{the linear discriminant score } \hat{d}_k(\mathbf{x}) = \text{the largest of } \hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x}), \ldots, \hat{d}_g(\mathbf{x}). \tag{18}$$

When we try to compare two linear discriminant scores at a time, we can obtain that the condition that $d_k(\mathbf{x})$ as in (16) is the largest linear discriminant score among $d_1(\mathbf{x}), d_2(\mathbf{x}), \ldots, d_g(\mathbf{x})$ is equivalent to

$$d_k(\mathbf{x}) - d_i(\mathbf{x}) = (\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_i) + \ln\left(\frac{p_k}{p_i}\right) \geq 0,$$

for all $i = 1, 2, \ldots, g$. From this, noting that $\ln(p_k/p_i) = -\ln(p_i/p_k)$, we can obtain the alternative classification rule: allocate $\mathbf{x}$ to π_k if

$$(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_i) \geq \ln\left(\frac{p_i}{p_k}\right),$$

for all $i = 1, 2, \ldots, g$ such that $i \neq k$.

To get the sample version of the classification rule above, we can substitute $\bar{\mathbf{x}}_i$ for $\boldsymbol{\mu}_i$ and $\mathbf{S}_{\text{pooled}}$ for $\boldsymbol{\Sigma}$, and obtain the following rule: allocate $\mathbf{x}$ to π_k if

$$\hat{d}_{ki}(\mathbf{x}) = (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_i)^\top \mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_i)^\top \mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_k + \bar{\mathbf{x}}_i) \geq \ln\left(\frac{p_i}{p_k}\right), \tag{19}$$

for all $i = 1, 2, \ldots, g$ such that $i \neq k$.

5.4 An Example of Finding The Quadratic Discriminator

Exercise 11.5 in [4]. Suppose $\mathbf{x}$ comes from one of two p-variate normal populations: either normal with mean $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$ (population π_1) or normal with mean $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$ (population π_2). If the respective density functions are denoted by $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, find the expression for the quadratic discrimininator

$$Q := \ln\left[\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right]. \tag{20}$$

If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, Q becomes

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

Solution: The density of a multivariate normal $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ distribution is

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right], \quad i = 1, 2.$$

Hence, the quadratic discrimininator in (20) takes the form

$$\begin{aligned}
Q = \ln\left[\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right] &= \ln\left[\frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right)}{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right)}\right] \\
&= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) \\
&= -\frac{1}{2}(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \\
&\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2) \\
&= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).
\end{aligned}$$

6 Fisher's Method for Discriminating among Several Populations

We shall now discuss an extension of Fisher's discriminant method for two populations to the case of several populations. To use Fisher's method, we need to find a reasonable representation of the populations that involve only a few linear combinations of the observations, such as $\mathbf{a}_1^\top \mathbf{x}$, $\mathbf{a}_2^\top \mathbf{x}$, and $\mathbf{a}_3^\top \mathbf{x}$. Now, it is not necessary to assume that the g populations are multivariate normal. However, we do assume that $p \times p$ population covariance matrices are equal and of full rank . That is, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_g = \mathbf{\Sigma}$ and $rank(\mathbf{\Sigma}) = p$.

Let $\bar{\boldsymbol{\mu}}$ denote the mean vector of the combined populations and let $\mathbf{B}_{\boldsymbol{\mu}}$ be the between groups sum of cross products, that is,

$$\mathbf{B}_{\boldsymbol{\mu}} = \sum_{i=1}^{g} (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^\top, \qquad \text{where} \quad \bar{\boldsymbol{\mu}} = \frac{1}{g}\sum_{i=1}^{g} \boldsymbol{\mu}_i.$$

We consider the linear combination

$$\mathbf{Y} = \mathbf{a}^\top \mathbf{X}$$

which has expected value

$$E(\mathbf{Y}) = \mathbf{a}^\top E(\mathbf{X}|\pi_i) = \mathbf{a}^\top \boldsymbol{\mu}_i, \qquad \text{for population } \pi_i, \quad i = 1, 2, \ldots, g,$$

and variance

$$\text{Var}(\mathbf{Y}) = \mathbf{a}^\top \text{Cov}(\mathbf{X})\mathbf{a} = \mathbf{a}^\top \mathbf{\Sigma} \mathbf{a}, \qquad \text{for all populations } \pi_i, \quad i = 1, 2, \ldots, g.$$

Consequently, the expected value $\boldsymbol{\mu}_{iY} = \mathbf{a}^\top \boldsymbol{\mu}_i$ changes as the population from which $\mathbf{X}$ is selected changes. We first define the overall mean

$$\begin{aligned} \bar{\boldsymbol{\mu}}_{\mathbf{Y}} &= \frac{1}{g}\sum_{i=1}^{g} \boldsymbol{\mu}_{i\mathbf{Y}} = \frac{1}{g}\sum_{i=1}^{g} \mathbf{a}^\top \boldsymbol{\mu}_i = \mathbf{a}^\top \left(\frac{1}{g}\sum_{i=1}^{g} \boldsymbol{\mu}_i \right) \\ &= \mathbf{a}^\top \bar{\boldsymbol{\mu}}, \end{aligned}$$

and form the ratio

$$\begin{aligned} \frac{\sum_{i=1}^{g} (\boldsymbol{\mu}_{i\mathbf{Y}} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}})^2}{\boldsymbol{\sigma}_{\mathbf{Y}}^2} &= \frac{\sum_{i=1}^{g} (\mathbf{a}^\top \boldsymbol{\mu}_i - \mathbf{a}^\top \bar{\boldsymbol{\mu}})^2}{\mathbf{a}^\top \mathbf{\Sigma} \mathbf{a}} \\ &= \frac{\mathbf{a}^\top \left(\sum_{i=1}^{g} (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^\top \right) \mathbf{a}}{\mathbf{a}^\top \mathbf{\Sigma} \mathbf{a}}, \end{aligned}$$

or

$$\frac{\sum_{i=1}^{g}(\boldsymbol{\mu}_{i\mathbf{Y}} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}})^2}{\boldsymbol{\sigma}_{\mathbf{Y}}^2} = \frac{\mathbf{a}^\top \mathbf{B}_{\boldsymbol{\mu}} \mathbf{a}}{\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}},$$

where $\boldsymbol{\sigma}_{\mathbf{Y}}^2$ is the variance of $\mathbf{Y}$.

Typically, $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_i, i = 1, 2, \ldots, g$, are unavailable, but the training set consists of correctly classified observations. Suppose the training set consists of a random sample of size n_i from population π_i, $i = 1, 2, \ldots, g$. Denote the $n_i \times p$ data set, from population π_i, by $\mathbb{X}_i$ and its jth row by $\mathbf{X}_{ij}^\top$. Consider the sample mean vectors defined by

$$\bar{\mathbf{X}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij},$$

and the covariance matrices denoted by $\mathbf{S}_i$, $i = 1, 2, \ldots, g$. We then define the "overall average" vector

$$\bar{\mathbf{X}} = \frac{1}{g} \sum_{i=1}^{g} \bar{\mathbf{X}}_i,$$

which is the $p \times 1$ vector average of the individual sample averages. The sample between groups matrix $\mathbf{B}$ is defined by

$$\mathbf{B} = \sum_{i=1}^{g} (\bar{\mathbf{X}}_i - \bar{\mathbf{X}})(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})^\top.$$

Also, a natural estimator of $\boldsymbol{\Sigma}$ is based on the sample within groups matrix

$$\mathbf{W} = \sum_{i=1}^{g} (n_i - 1)\mathbf{S}_i = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\bar{\mathbf{X}}_{ij} - \bar{\mathbf{X}}_i)(\bar{\mathbf{X}}_{ij} - \bar{\mathbf{X}}_i)^\top,$$

and is equal to

$$\mathbf{S}_{\text{pooled}} = \mathbf{W}/(n_1 + n_2 + \ldots + n_g - g).$$

We note that $\mathbf{W}$ is the constant $(n_1 + n_2 + \cdots + n_g - g)$ times $\mathbf{S}_{\text{pooled}}$, so the same $\hat{\mathbf{a}}$ that maximizes $\hat{\mathbf{a}}^\top \mathbf{B} \hat{\mathbf{a}} / \hat{\mathbf{a}}^\top \mathbf{S}_{\text{pooled}} \hat{\mathbf{a}}$ also maximizes $\hat{\mathbf{a}}^\top \mathbf{B} \hat{\mathbf{a}} / \hat{\mathbf{a}}^\top \mathbf{W} \hat{\mathbf{a}}$. Moreover, we can present the optimizing $\hat{\mathbf{a}}$ in the form of eigenvectors $\hat{\mathbf{e}}_i$ of $\mathbf{W}^{-1}\mathbf{B}$, because if $\mathbf{W}^{-1}\mathbf{B}\hat{\mathbf{e}} = \hat{\lambda}\hat{\mathbf{e}}$, then $\mathbf{S}_{\text{pooled}}^{-1}\mathbf{B}\hat{\mathbf{e}} = \hat{\lambda}(n_1 + n_2 + \ldots + n_g - g)\hat{\mathbf{e}}$, where $\hat{\lambda}$ is the nonzero eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ corresponding to the eigenvalue $\hat{\mathbf{e}}$ (See pp.621–623 in [4])

6.1 Fisher's Sample Linear Discriminants

Let $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_s > 0$ denote the $s \leq \min(g-1, p)$ nonzero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ and let $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_s$ be the corresponding eigenvectors (scaled so that $\hat{\mathbf{e}}_k^\top \mathbf{S}_{\text{pooled}} \hat{\mathbf{e}}_k = 1$, $k = 1, \ldots, s$). Then (see

Exercise 11.21 in [4]) the vector of coefficients $\hat{\mathbf{a}}$ that maximizes the ratio

$$\frac{\hat{\mathbf{a}}^\top \mathbf{B}\hat{\mathbf{a}}}{\hat{\mathbf{a}}^\top \mathbf{W}\hat{\mathbf{a}}} = \frac{\hat{\mathbf{a}}^\top \left(\sum_{i=1}^{g}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^\top\right)\hat{\mathbf{a}}}{\hat{\mathbf{a}}^\top \left[\sum_{i=1}^{g}(n_i-1)\mathbf{S}_i = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_i)(\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_i)^\top\right]\hat{\mathbf{a}}}$$

is given by $\hat{\mathbf{a}}_1 = \hat{\mathbf{e}}_1$. The linear combination $\hat{\mathbf{a}}_1^\top \mathbf{x}$ is called the *sample first discriminant*. The choice $\hat{\mathbf{a}}_2 = \hat{\mathbf{e}}_2$ produces the *sample second discriminant*, $\hat{\mathbf{a}}_2^\top \mathbf{x}$, and continuing, we obtain $\hat{\mathbf{a}}_k^\top \mathbf{x} = \hat{\mathbf{e}}_k^\top \mathbf{x}$, the *sample kth discriminant*, $k \leq s$. For more details, see Section 11.6 in [4].

Fisher's discriminants are used to obtain a low-dimensional representation of the data that separates populations as much as possible. In spite of the fact that they were obtained for the purpose of separation, the discriminants also provide the basis for a classification rule.

Consider the separatory measure given by (see relation (11-68) in [4])

$$\triangle_{\mathbf{S}}^2 = \sum_{i=1}^{g}(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}), \tag{21}$$

where

$$\bar{\boldsymbol{\mu}} = \frac{1}{g}\sum_{i=1}^{g}\boldsymbol{\mu}_i$$

and $(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})$ is the squared statistical (Mahalanobis) distance from the ith population mean $\boldsymbol{\mu}_i$ to the centroid $\bar{\boldsymbol{\mu}}$.

Exercise 11.22 in [4]. Show that $\triangle_{\mathbf{S}}^2 = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \lambda_1 + \lambda_2 + \ldots + \lambda_s$, where λ_1, λ_2, $\ldots$, λ_s are the nonzero eigenvalues of $\boldsymbol{\Sigma}^{-1}\mathbf{B}_{\boldsymbol{\mu}}$ (or $\boldsymbol{\Sigma}^{-1/2}\mathbf{B}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}^{-1/2}$) and $\triangle_{\mathbf{S}}^2$ is given by (21). Also, show that $\lambda_1 + \lambda_2 + \ldots + \lambda_r$ is the resulting separation when only the first r discriminants $Y_1, Y_2, \ldots, Y_r$ are used.

Solution: Let $\mathbf{P}$ be the orthogonal matrix whose ith row $\mathbf{e}_i^\top$ is the eigenvector of $\boldsymbol{\Sigma}^{-1/2}\mathbf{B}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}^{-1/2}$ corresponding to the ith largest eigenvalue, $i = 1, 2, \ldots, p$. Consider

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_s \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} \mathbf{e}_1^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{X} \\ \vdots \\ \mathbf{e}_s^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{X} \\ \vdots \\ \mathbf{e}_p^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{X} \end{pmatrix} = \mathbf{P}\boldsymbol{\Sigma}^{-1/2}\mathbf{X}.$$

We have $\boldsymbol{\mu}_{i\mathbf{Y}} = E(\mathbf{Y}|\pi_i) = \mathbf{P}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_i$ and $\bar{\boldsymbol{\mu}}_{\mathbf{Y}} = \mathbf{P}\boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{\mu}}$, so

$$\begin{aligned}(\boldsymbol{\mu}_{i\mathbf{Y}} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}})^\top(\boldsymbol{\mu}_{i\mathbf{Y}} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}}) &= (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{P}^\top\mathbf{P}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}) \\ &= (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}).\end{aligned}$$

Therefore, $\triangle^2_{\mathbf{S}} = \sum_{i=1}^{g}(\boldsymbol{\mu}_{i\mathbf{Y}} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}})^\top(\boldsymbol{\mu}_{i\mathbf{Y}} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}})$. Using Y_1, we have

$$\begin{aligned}\sum_{i=1}^{g}(\boldsymbol{\mu}_{iY_1} - \bar{\boldsymbol{\mu}}_{Y_1})^2 &= \sum_{i=1}^{g}\mathbf{e}_1{}^\top\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_1 \\ &= \mathbf{e}_1{}^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{B}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_1 = \lambda_1,\end{aligned}$$

because $\mathbf{e}_1$ has eigenvalue λ_1. Similarly, Y_2 produces

$$\sum_{i=1}^{g}(\boldsymbol{\mu}_{iY_2} - \bar{\boldsymbol{\mu}}_{Y_2})^2 = \mathbf{e}_2{}^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{B}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_2 = \lambda_2,$$

and Y_p produces

$$\sum_{i=1}^{g}(\boldsymbol{\mu}_{iY_p} - \bar{\boldsymbol{\mu}}_{Y_p})^2 = \mathbf{e}_p{}^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{B}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_p = \lambda_p.$$

Thus,

$$\begin{aligned}\triangle^2_{\mathbf{S}} &= \sum_{i=1}^{g}(\boldsymbol{\mu}_{i\mathbf{Y}} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}})^\top(\boldsymbol{\mu}_{i\mathbf{Y}} - \bar{\boldsymbol{\mu}}_{\mathbf{Y}}) \\ &= \sum_{i=1}^{g}(\boldsymbol{\mu}_{iY_1} - \bar{\boldsymbol{\mu}}_{Y_1})^2 + \sum_{i=1}^{g}(\boldsymbol{\mu}_{iY_2} - \bar{\boldsymbol{\mu}}_{Y_2})^2 + \ldots + \sum_{i=1}^{g}(\boldsymbol{\mu}_{iY_p} - \bar{\boldsymbol{\mu}}_{Y_p})^2 \\ &= \lambda_1 + \lambda_2 + \ldots + \lambda_p = \lambda_1 + \lambda_2 + \ldots + \lambda_s,\end{aligned}$$

where we used the fact that $\lambda_{s+1} = \ldots = \lambda_p = 0$. If only the first r discriminants are used, their contribution to $\triangle^2_{\mathbf{S}}$ is $\lambda_1 + \lambda_2 + \ldots + \lambda_r$.

6.2 Two Examples of Fisher's Sample Discriminants for Two (or More) Populations

Exercise 11.23 in [4]. Consider the data given in Table 1 (see Table 1.6 in [4]).

(a) Check the marginal distributions of the x_i's in both the multiple-sclerosis (MS) group and non-multiple-sclerosis (NMS) group for normality by graphing the corresponding observations as normal probability plots. Suggest appropriate data transformations if the normality assumption is suspect.

(b) Assume that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. Construct Fisher's linear discriminant function. Do all the variables in the discriminant function appear to be important? Discuss your answer. Develop a classification rule assuming equal prior probabilities and equal costs of misclassification.

(c) Using the results in (b), calculate the apparent error rate.

Table 1: **Multiple-Sclerosis Data**

Non-Multiple-Sclerosis Group Data					
Subject Number	$\mathbf{x}_1$ (Age)	$\mathbf{x}_2$ $(S1L + S1R)$	$\mathbf{x}_3$ $\lvert S1L - S1R\rvert$	$\mathbf{x}_4$ $(S2L + S2R)$	$\mathbf{x}_5$ $\lvert S2L - S2R\rvert$
1	18	152	1.6	198.4	0
2	19	138	0.4	180.8	1.6
3	20	144	0	186.4	0.8
4	20	143.6	3.2	194.8	0
5	20	148.8	0	217.6	0
6	21	141.6	0.8	181.6	0.8
7	21	136	1.6	180	0.8
8	21	137.6	1.6	185.6	3.2
9	22	140.4	3.2	182	3.2
10	22	137.2	0	181.8	0.2
11	22	125.4	1	169.2	0
12	22	142.4	4.8	185.6	0
13	22	150.4	0	214.4	3.2
14	22	145.6	1.6	203.6	5.2
15	23	147.2	3.2	196.8	1.6
16	23	139.2	1.6	179.2	0
17	24	169.6	0	204.8	0
18	24	139.2	1.6	176	3.2
19	24	153.6	0	212	0.8
20	25	146.8	0	194.8	3.2
21	25	139.2	1.6	198.4	3.2
22	25	136	1.6	181.6	2.4
23	26	138.8	1.6	191.6	0
24	26	150.4	0	205.2	0.4
25	26	139	1.4	178.6	0.2
26	27	133.8	0.2	180.8	0
27	27	139	1.8	190.4	1.6
28	28	136	1.6	193.2	3.6
29	28	146.4	0.8	195.6	2.8
30	29	145.2	4.8	194.2	3.8
31	29	146.4	0.8	208.2	0.2
32	29	138	2.8	181.2	0.4
33	30	148.8	1.6	196.4	1.6
34	31	137.2	0	184	0
35	31	147.2	0	197.6	0.8
36	32	144	0	185.8	0.2
37	32	156	0	192.8	2.4
38	34	137	0.2	182.4	0

39	35	143.2	2.4	184	1.6
40	36	141.6	0.8	187.2	1.6
41	37	152	1.6	189.2	2.8
42	39	157.4	3.4	227	2.6
43	40	141.4	0.6	209.2	1.6
44	42	156	2.4	195.2	3.2
45	43	150.4	1.6	180	0.8
46	43	142.4	1.6	188.8	0
47	46	158	2	192	3.2
48	48	130	3.6	190	0.4
49	49	152.2	1.4	200	4.8
50	49	150	3.2	206.6	2.2
51	50	146.4	2.4	191.6	2.8
52	54	146	1.2	203.2	1.6
53	55	140.8	0	184	1.6
54	56	140.4	0.4	203.2	1.6
55	56	155.8	3	187.8	2.6
56	56	141.6	0.8	196.8	1.6
57	57	144.8	0.8	188	0.8
58	57	146.8	3.2	191.6	0
59	59	176.8	2.4	232.8	0.8
60	60	171	1.8	202	3.6
61	60	163.2	0	224	0
62	60	171.6	1.2	213.8	3.4
63	60	146.4	4	203.2	4.8
64	62	146.8	3.6	201.6	3.2
65	67	154.4	2.4	205.2	6
66	69	171.2	1.6	210.4	0.8
67	73	157.2	0.4	204.8	0
68	74	175.2	5.6	235.6	0.4
69	79	155	1.4	204.4	0

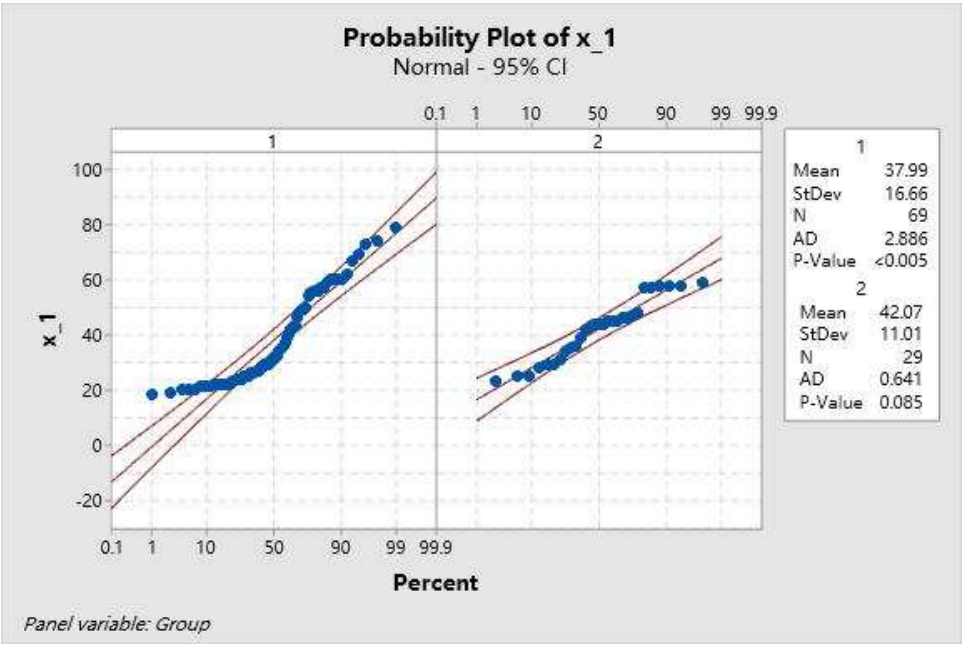
Multiple-Sclerosis Group Data

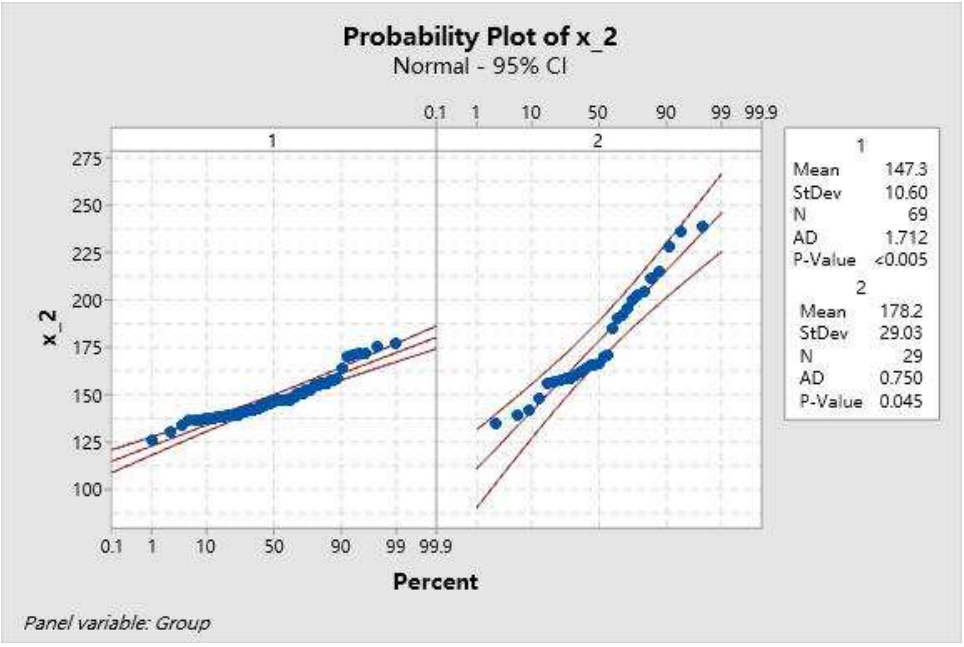
Subject Number	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
1	23	148	0.8	205.4	0.6
2	25	195.2	3.2	262.8	0.4
3	25	158	8	209.8	12.2
4	28	134.4	0	198.4	3.2
5	29	190.2	14.2	243.8	10.6
6	29	160.4	18.4	222.8	31.2
7	31	227.8	90.2	270.2	83
8	34	211	3	250.8	5.2
9	35	203.8	12.8	254.4	11.2
10	36	141.2	6.8	194.4	21.6
11	39	157.4	3.4	227	2.6

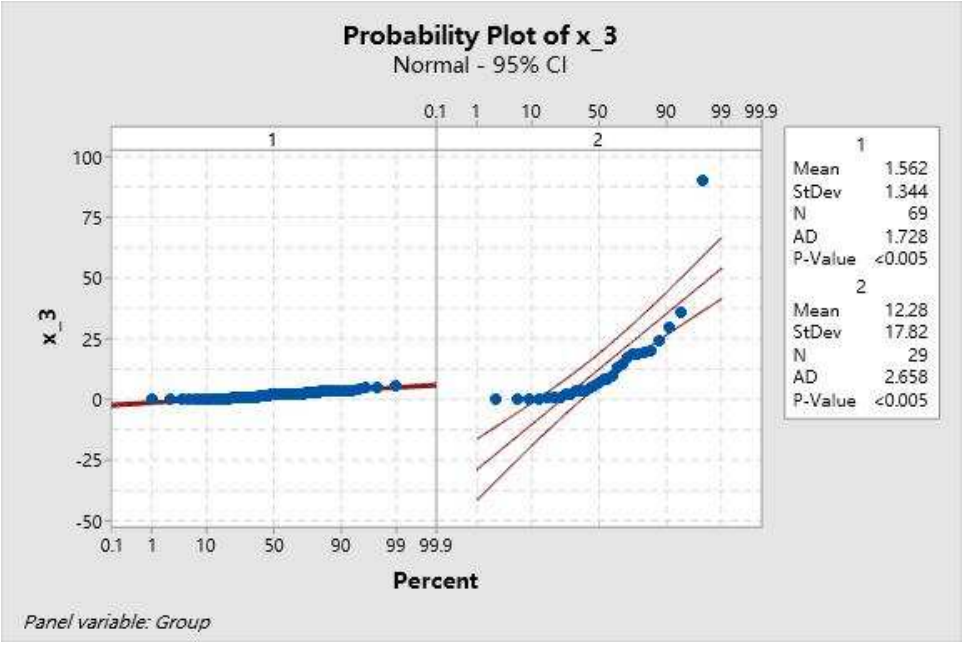
12	42	166.4	0	226	0
13	43	191.8	35.4	243.6	40.8
14	44	156.8	0	203.2	0
15	44	202.8	29.3	246.4	24.8
16	44	165.2	18.4	254	46.4
17	45	162	5.6	224.4	8.8
18	45	138.4	0.8	176.8	4
19	45	158.4	1.6	214.4	0
20	46	155.4	1.8	201.2	6
21	46	214.8	9.2	290.6	0.6
22	47	185	19	274.4	7.6
23	48	236	20	328	0
24	57	170.8	24	228.4	33.6
25	57	165.6	16.8	229.2	15.6
26	58	238.4	8	304.4	6
27	58	164	0.8	216.8	0.8
28	58	169.8	0	219.2	1.6
29	59	199.8	4.6	250.3	1

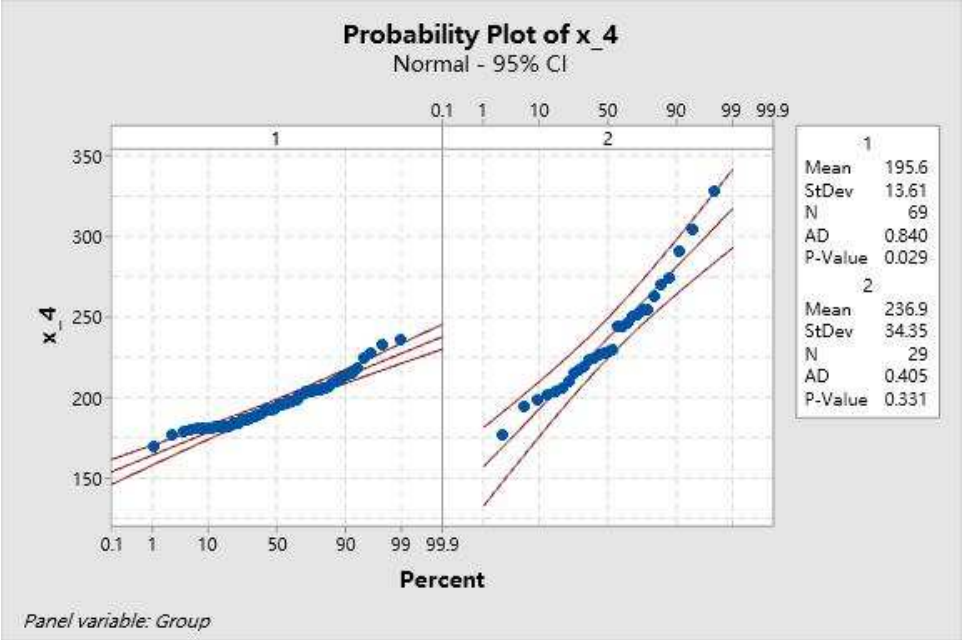
Solution:

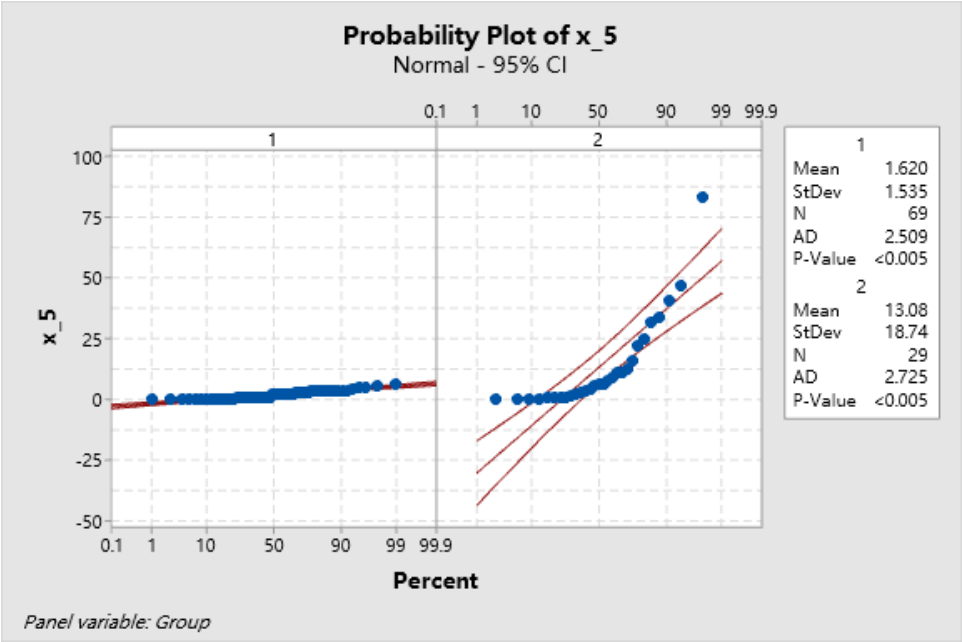
(a) The normal probability plots for each of the variables $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5$ are the following:




Probability Plot of x_2
Normal - 95% CI
0.1 1 10 50 90 99 99.9
1
2
275
250
225
200
175
150
125
100
x_2
0.1 1 10 50 90 99 99.9
Percent
1
Mean 147.3
StDev 10.60
N 69
AD 1.712
P-Value <0.005
2
Mean 178.2
StDev 29.03
N 29
AD 0.750
P-Value 0.045
Panel variable: Group


Probability Plot of x_3
Normal - 95% CI
0.1 1 10 50 90 99 99.9
1
2
100
75
50
25
0
-25
-50
x_3
0.1 1 10 50 90 99 99.9
Percent
1
Mean 1.562
StDev 1.344
N 69
AD 1.728
P-Value <0.005
2
Mean 12.28
StDev 17.82
N 29
AD 2.658
P-Value <0.005
Panel variable: Group





We can see from the above that variables $\mathbf{x}_2$ and $\mathbf{x}_4$ appear to be normal. We can use the transformations $\ln(\mathbf{x}_1)$, $\ln(\mathbf{x}_3 + 1)$ and $\ln(\mathbf{x}_5 + 1)$ for variables $\mathbf{x}_1$, $\mathbf{x}_3$ and $\mathbf{x}_5$, respectively, to bring them closer to the normality assumption since these three variables look like being exponentially distributed. For $\mathbf{x}_3$ and $\mathbf{x}_5$, the asymptotes seem to have one unit downward shift comparing to $\mathbf{x}_1$. Therefore, for logarithmic function, we have to shift $\ln(\mathbf{x}_3)$ and $\ln(\mathbf{x}_5)$ to the right.

Linear Discriminant Function for Groups

	1	2
Constant	-56.871	-80.100
x_1	-0.080	-0.103
x_2	0.333	0.367
x_3	-1.499	-1.709
x_4	0.350	0.434
x_5	1.076	1.330

(b) From the output using Minitab, we obtain that the linear discriminant function is

$$\begin{aligned}\hat{\mathbf{y}} &= [(-0.079)-(-0.103)]\mathbf{x}_1+[(0.330)-(0.364)]\mathbf{x}_2+[(-1.498)-(-1.708)]\mathbf{x}_3\\ &\quad +[(0.351)-(0.435)]\mathbf{x}_4+[(1.076)-(1.329)]\mathbf{x}_5+[(-56.833)-(-80.067)]\\ &= 0.024\mathbf{x}_1-0.034\mathbf{x}_2+0.21\mathbf{x}_3-0.084\mathbf{x}_4-0.253\mathbf{x}_5-23.234,\end{aligned}$$

and $\hat{\mathbf{m}} = -23.234$. The classification rule is therefore: allocate $\mathbf{x}_0$ to π_1 (NMS group) if

$$\hat{\mathbf{a}}^\top\mathbf{x}_0-\hat{\mathbf{m}} = 0.024\mathbf{x}_1-0.034\mathbf{x}_2+0.21\mathbf{x}_3-0.084\mathbf{x}_4-0.253\mathbf{x}_5-23.234\geq 0,$$

and allocate $\mathbf{x}_0$ to π_2 (MS group) otherwise.
We can see from the above that each variable has a non-zero coefficient in the discriminant function so all the variables appear to be important.

(c) From the output using Minitab, we have the confusion matrix in the form

Summary of Classification

	True Group	
Put into Group	1	2
1	66	7
2	3	22
Total N	69	29
N correct	66	22
Proportion	0.957	0.759

Correct Classifications

N	Correct	Proportion
98	88	0.898

Summary of Misclassified Observations

Observation	True Group	Pred Group	Group	Squared Distance	Probability
13**	1	2	1	4.740	0.440
			2	4.255	0.560
42**	1	2	1	4.087	0.431
			2	3.531	0.569
59**	1	2	1	4.238	0.328
			2	2.804	0.672
70**	2	1	1	1.928	0.809
			2	4.819	0.191
73**	2	1	1	3.316	0.857
			2	6.899	0.143
83**	2	1	1	0.6149	0.859
			2	4.2238	0.141
87**	2	1	1	1.629	0.979
			2	9.294	0.021
88**	2	1	1	0.9227	0.763
			2	3.2580	0.237
89**	2	1	1	1.793	0.716
			2	3.645	0.284
96**	2	1	1	2.383	0.670
			2	3.798	0.330

Hence, we obtain

$$\begin{aligned} \text{APER} &= \frac{n_{1M} + n_{2M}}{n_1 + n_2} \\ &= \frac{3+7}{69+29} \\ &= 0.102. \end{aligned}$$

Exercise 11.27 in [4]. The data in Table 2 contain observations on $\mathbf{x}_2$ = sepal width and $\mathbf{x}_4$ = petal width for samples from three species of iris. There are $n_1 = n_2 = n_3 = 50$ observations in each sample.

(a) Plot the data in the $(\mathbf{x}_2, \mathbf{x}_4)$ variable space. Do the observations for the three groups appear to be bivariate normal?

(b) Assume that the samples are from bivariate normal populations with a common covariance matrix. Test the hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$ versus H_1 : at least one $\boldsymbol{\mu}_i$ is different from the others at the $\alpha = 0.05$ significance level. Is the assumption of a common covariance matrix reasonable in this case? Explain.

(c) Assuming that the populations are bivariate normal, construct the quadratic discriminate scores $\hat{d}_i^{Q}(\mathbf{x})$ given by (14) with $p_1 = p_2 = p_3 = 1/3$. Using the rule given by (15), classify the new observation $\mathbf{x}_0^\top = (3.5,\ 1.75)$ into population π_1, π_2, or π_3.

(d) Assume that the covariance matrices $\mathbf{\Sigma}_i$ are the same for all three bivariate normal populations. Construct the linear discriminate score $\hat{d}_i(\mathbf{x})$ given by (17), and use it to assign $\mathbf{x}_0^\top = (3.5,\ 1.75)$ to one of the populations π_1, $i = 1, 2, 3$ according to (18). Take $p_1 = p_2 = p_3 = 1/3$. Compare the results in parts (c) and (d). Which approach do you prefer? Explain.

(e) Assuming equal covariance matrices and bivariate normal populations and supposing that $p_1 = p_2 = p_3 = 1/3$, allocate $\mathbf{x}_0^\top = (3.5,\ 1.75)$ to π_1, π_2, or π_3 using rule (19). Compare the result with that in part (d). Delineate the classification regions $\hat{R}_1$, $\hat{R}_2$ and $\hat{R}_3$ on your graph from part (a) determined by the linear functions $\hat{d}_{ki}(\mathbf{x}_0)$ in (19).

(f) Using the linear discriminant scores from part (d), classify the sample observations. Calculate the APER.

Table 2: **Data on Irises**

$\pi_1 : Irissetosa$				$\pi_2 : Irisversicolor$				$\pi_3 : Irisvirginica$			
Sepal length $\mathbf{x}_1$	**Sepal width** $\mathbf{x}_2$	**Sepal length** $\mathbf{x}_3$	**Sepal width** $\mathbf{x}_4$	**Sepal length** $\mathbf{x}_1$	**Sepal width** $\mathbf{x}_2$	**Sepal length** $\mathbf{x}_3$	**Sepal width** $\mathbf{x}_4$	**Sepal length** $\mathbf{x}_1$	**Sepal width** $\mathbf{x}_2$	**Sepal length** $\mathbf{x}_3$	**Sepal width** $\mathbf{x}_4$
5.1	3.5	1.4	0.2	7	3.2	4.7	1.4	6.3	3.3	6	2.5
4.9	3	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4	1.3	6.3	2.9	5.6	1.8
5	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5	3.4	1.5	0.2	4.9	2.4	3.3	1	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5	2	3.5	1	6.5	3.2	5.1	2
4.8	3.4	1.6	0.2	5.9	3	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3	1.4	0.1	6	2.2	4	1	6.8	3	5.5	2.1
4.3	3	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5	2
5.8	4	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3	4.5	1.5	6.5	3	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3

5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6	2.2	5	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4	1.3	5.6	2.8	4.9	2
4.6	3.6	1	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5	3	1.6	0.2	6.6	3	4.4	1.4	7.2	3.2	6	1.8
5	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3	5	1.7	6.1	3	4.9	1.8
5.2	3.4	1.4	0.2	6	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1	7.2	3	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1	7.9	3.8	6.4	2
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3	4.5	1.5	6.1	2.6	5.6	1.4
5	3.2	1.2	0.2	6	3.4	4.5	1.6	7.7	3	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3	1.3	0.2	5.6	3	4.1	1.3	6	3	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4	1.3	6.9	3.1	5.4	2.1
5	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4	1.2	5.8	2.7	5.1	1.9
5	3.5	1.6	0.6	5	2.3	3.3	1	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3	1.4	0.3	5.7	3	4.2	1.2	6.7	3	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3	5.2	2
5.3	3.7	1.5	0.2	5.1	2.5	3	1.1	6.2	3.4	5.4	2.3
5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3	5.1	1.8

Solution:

(a) From the figure above by using Minitab, with X-data being $\mathbf{x}_2$ and Y-data being $\mathbf{x}_4$, we can see that all the points from each group fall in an elliptical form, that is, the three groups appear to be bivariate normal.

(b) The output is given above. Since 0.05 is greater than p-value (it is close to 0) shown above, we reject the null hypothesis. The assumption of a common covariance matrix is not reasonable because from the plot in part (a), we can see that the points of the group 1 are in a different ellipse from the points of groups 2 and 3.

(c) By assumption, we have $p_1 = p_2 = p_3 = 1/3$ and $\mathbf{x}_0^\top = (3.5,\ 1.75)$. Using R, we obtain

$$\mathbf{S}_1 = \begin{pmatrix} 0.1437 & 0.0093 \\ 0.0093 & 0.0111 \end{pmatrix}, \quad \mathbf{S}_2 = \begin{pmatrix} 0.0985 & 0.0412 \\ 0.0412 & 0.0391 \end{pmatrix}, \quad \mathbf{S}_3 = \begin{pmatrix} 0.1437 & 0.0093 \\ 0.0093 & 0.0111 \end{pmatrix},$$

and also

$$\bar{\mathbf{X}}_1 = \begin{pmatrix} 3.428 \\ 0.246 \end{pmatrix}, \quad \bar{\mathbf{X}}_2 = \begin{pmatrix} 2.77 \\ 1.326 \end{pmatrix}, \quad \bar{\mathbf{X}}_3 = \begin{pmatrix} 2.974 \\ 2.026 \end{pmatrix}.$$

According to (14) and the classification rule (15), the estimates of the quadratic discrimination scores are

$$\begin{aligned} \hat{d}_1^Q(\mathbf{x}_0) &= -\frac{1}{2}\ln|\mathbf{S}_1| - \frac{1}{2}(\mathbf{x}_0 - \bar{\mathbf{x}}_1)^\top \mathbf{S}_1^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1) + \ln p_1 = -104.9361, \\ \hat{d}_2^Q(\mathbf{x}_0) &= -\frac{1}{2}\ln|\mathbf{S}_2| - \frac{1}{2}(\mathbf{x}_0 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_2^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_2) + \ln p_2 = -1.0554, \\ \hat{d}_3^Q(\mathbf{x}_0) &= -\frac{1}{2}\ln|\mathbf{S}_3| - \frac{1}{2}(\mathbf{x}_0 - \bar{\mathbf{x}}_3)^\top \mathbf{S}_3^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_3) + \ln p_3 = -2.3244, \end{aligned}$$

and we allocate $\mathbf{x}_0$ to population 2 (Irisversicolor) since $\hat{d}_2^Q(\mathbf{x}_0)$ is the greatest among the three discrimination scores.

(d) Using part (c), we get

$$\mathbf{S}_{\text{pooled}} = \frac{\mathbf{S}_1 + \mathbf{S}_2 + \mathbf{S}_3}{3} = \begin{pmatrix} 0.1731 & 0.0491 \\ 0.0491 & 0.0628 \end{pmatrix}.$$

According to (17) and the classification rule (18), the estimated linear discriminant scores are

$$\begin{aligned} \hat{d}_1(\mathbf{x}_0) &= \bar{\mathbf{x}}_1^\top \mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}\bar{\mathbf{x}}_1^\top \mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_1 + \ln p_1 = 17.6437, \\ \hat{d}_2(\mathbf{x}_0) &= \bar{\mathbf{x}}_2^\top \mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}\bar{\mathbf{x}}_2^\top \mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_2 + \ln p_2 = 38.1425, \\ \hat{d}_3(\mathbf{x}_0) &= \bar{\mathbf{x}}_3^\top \mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}\bar{\mathbf{x}}_3^\top \mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_3 + \ln p_3 = 37.5175. \end{aligned}$$

Therefore we allocate $\mathbf{x}_0$ to population 2 (Irisversicolor) since $\hat{d}_2(\mathbf{x})$ is the greatest among the three discrimination scores. We get the same result in part (c) and (d) but I prefer using the quadratic discrimination score since there is no need to assume the covariance matrices are the same for all bivariate normal population.

(e) Recall the classification rule (19), according to which we allocate $\mathbf{x}_0$ to π_k if

$$(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_i)^\top \mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_i)^\top \mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_k + \bar{\mathbf{x}}_i) \geq \ln\left(\frac{p_i}{p_k}\right) = 0, \tag{22}$$

for $i = 1, 2, 3$ such that $i \neq k$.

Case 1 (checking the conditions for allocating $\mathbf{x}_0$ to π_1):

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = -20.4988 < 0,$$

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_3)^\top \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_3)^\top \mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_3) = -19.8738 < 0.$$

Case 2 (checking the conditions for allocating $\mathbf{x}_0$ to π_2):

$$(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)^\top \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)^\top \mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_2 + \bar{\mathbf{x}}_1) = 20.4988 > 0,$$

$$(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_3)^\top \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_3)^\top \mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_2 + \bar{\mathbf{x}}_3) = 0.625 > 0.$$

Case 3 (checking the conditions for allocating $\mathbf{x}_0$ to π_3):

$$(\bar{\mathbf{x}}_3 - \bar{\mathbf{x}}_1)^\top \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_3 - \bar{\mathbf{x}}_1)^\top \mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_3 + \bar{\mathbf{x}}_1) = 19.8738 > 0,$$

$$(\bar{\mathbf{x}}_3 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_3 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_3 + \bar{\mathbf{x}}_2) = -0.625 < 0.$$

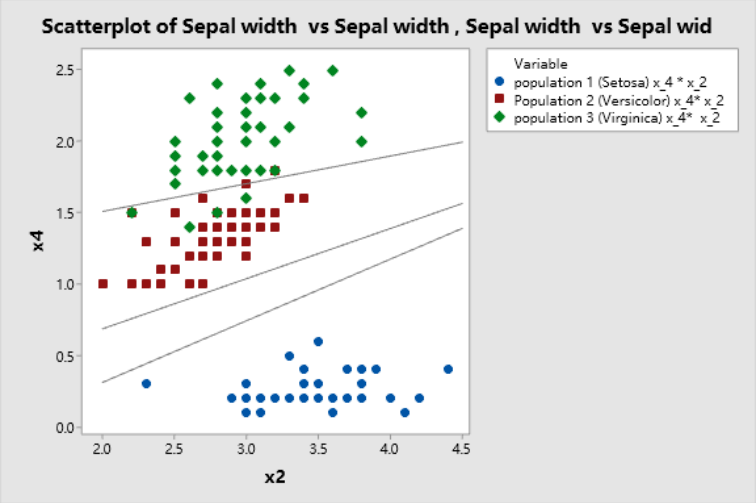
We therefore allocate $\mathbf{x}_0$ to π_2 and it is the same conclusion as we get in part (d). In order to sketch the regions, we have to calculate $\hat{d}_{12}(\mathbf{x}_0)$, $\hat{d}_{23}(\mathbf{x}_0)$ and $\hat{d}_{13}(\mathbf{x}_0)$. We have

$$\hat{d}_{12}(\mathbf{x}_0) = 11.1399\mathbf{x}_2 - 25.8982\mathbf{x}_4 - 14.1664,$$

$$\hat{d}_{13}(\mathbf{x}_0) = 13.6826\mathbf{x}_2 - 39.0308\mathbf{x}_4 + 0.5409,$$

$$\hat{d}_{23}(\mathbf{x}_0) = 2.5427\mathbf{x}_2 - 13.1325\mathbf{x}_4 + 14.7073.$$

The classification regions are shown on the graph below:



Summary of Classification

	True Group		
Put into Group	1	2	3
1	50	0	0
2	0	48	1
3	0	2	49
Total N	50	50	50
N correct	50	48	49
Proportion	1.000	0.960	0.980

Summary of Misclassified Observations

Observation	True Group	Pred Group	Group	Squared Distance	Probability
71**	2	3	1	469.688	0.000
			2	-2.360	0.336
			3	-3.723	0.664
84**	2	3	1	515.644	0.000
			2	-2.785	0.154
			3	-6.187	0.846
134**	3	2	1	501.643	0.000
			2	-5.495	0.605
			3	-4.642	0.395

(f) Hence, the apparent error rate is

$$\begin{aligned} \text{APER} &= \frac{n_{1M} + n_{2M} + n_{3M}}{n_1 + n_2 + n_3} \\ &= \frac{0 + 2 + 1}{50 + 50 + 50} \\ &= 0.02. \end{aligned}$$

7 Conclusion

The project aims at presenting, analytically and numerically, some basic methods of the classification and discrimination analysis that have not been covered by the statistical courses offered by the School of Mathematics and Statistics. The theory of classification is different from the estimation and hypothesis testing theories. In the course of writing this project, we have learnt various interesting methods and tools of the classification theory and have also practiced with applying them to real-life data.

For instance, in the last example related to Fisher's method for discriminating among several populations, we have used different ways to allocate the given variables to one of four predefined groups under the assumption of bivariate normally distributed populations. The graphics obtained by using MINITAB allow us to visually examine multivariate data. The obtained APER of 0.02 indicates that the misclassification rate is small. From the output in MINITAB, there are 3 observations misclassified and they are all from groups 2 and 3. This indicates that the whole group 1 is correctly classified.

Lastly, classification and discrimination might be useful in various practical applications in many fields of industry, medicine and science.

References

[1] Anderson,T. W. *An Introduction to Multivariate Statistical Analysis*, 3d ed. New York: John Wiley, 2003.

[2] Fisher, R. A. The Statistical Utilization of Multiple Measurements. *Annals of Eugenics*, **8** (1938), 376–386.

[3] Kendall, M. G. *Multivariate Analysis*. New York: Hafner Press, 1975.

[4] Johnson, R. A. and Wichern, D. W. *Applied Multivariate Analysis*, 6th ed. New Jersey: Pearson Prentice Hall, 2019.

[5] Wald, A.On a Statistical Problem Arising in the Classification of an Individual into One of Two Groups (1944). *Annals of Mathematical Statistics*, **15**, 145–162.

8 Appendix

1. **Code for example in 11.2**

```
x1<- seq(-1,1,0.05)
y1<-1-abs(x1)
x2<- seq(-0.5,1.5,0.05)
y2<-1-abs(x2-0.5)
plot(x1,y1,type="l", col="red",xlim = c(-1,1.5),ylim =
c(0,1),xlab="x",ylab="y")
lines(x2,y2,col="blue")
legend("topleft", legend = c("f1(x)","f2(x)"),
       col = c ("red","blue"),lty = 1:1)
```

2. **Code for calculation in the example 2 of 11.6**

```
x <-matrix(c(3.5,1.75),nrow=2,ncol=1,byrow=TRUE)
x0<-matrix(c(a,b),nrow=2,ncol=1,byrow=TRUE)
s1 <-matrix(c(0.1437,0.0093,0.0093,0.0111),nrow=2,ncol=2,byrow=TRUE)
s2 <-matrix(c(0.0985,0.0412,0.0412,0.0391),nrow=2,ncol=2,byrow=TRUE)
s3 <-matrix(c(0.104,0.0476,0.0476,0.0754),nrow=2,ncol=2,byrow=TRUE)
x1 <-matrix(c(3.4280,0.246),nrow=2,ncol=1,byrow=TRUE)
x2 <-matrix(c(2.77,1.326),nrow=2,ncol=1,byrow=TRUE)
x3 <-matrix(c(2.974,2.026),nrow=2,ncol=1,byrow=TRUE)

d1x<- -(1/2)*log(det(s1))-(1/2)*(t(x-x1)%*%solve(s1)%*%(x-x1))+ log(1/3)
d2x<- -(1/2)*log(det(s2))-(1/2)*(t(x-x2)%*%solve(s2)%*%(x-x2))+ log(1/3)
d3x<- -(1/2)*log(det(s3))-(1/2)*(t(x-x3)%*%solve(s3)%*%(x-x3))+ log(1/3)


Spooled<- (1/2)*(s1+s2+s3)
d11x<-t(x1)%*%solve(Spooled)%*%x-(1/2)*t(x1)\\
%*%solve(Spooled)%*%x1+log(1/3)
d12x<-t(x2)%*%solve(Spooled)%*%x-(1/2)*t(x2)\\
%*%solve(Spooled)%*%x2+log(1/3)
d13x<-t(x3)%*%solve(Spooled)%*%x-(1/2)*t(x3)\\
```

```
%*%solve(Spooled)%*%x3+log(1/3)


t(x2-x1)%*%solve(Spooled)%*%x-(1/2)*t(x2-x1)%*%solve(Spooled)%*%(x2+x1)
t(x2-x3)%*%solve(Spooled)%*%x-(1/2)*t(x2-x3)%*%solve(Spooled)%*%(x2+x3)

t(x1-x2)%*%solve(Spooled)%*%x-(1/2)*t(x1-x2)%*%solve(Spooled)%*%(x1+x2)
t(x1-x3)%*%solve(Spooled)%*%x-(1/2)*t(x1-x3)%*%solve(Spooled)%*%(x1+x3)

t(x3-x1)%*%solve(Spooled)%*%x-(1/2)*t(x3-x1)%*%solve(Spooled)%*%(x3+x1)
t(x3-x2)%*%solve(Spooled)%*%x-(1/2)*t(x3-x2)%*%solve(Spooled)%*%(x3+x2)
```