Discriminant Analysis

INTRODUCTION

Discrimination and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups. Discriminant analysis is rather exploratory in nature. As a separative procedure, it is often employed on a one-time basis in order to investigate observed differences when causal relationships are not well understood. Classification procedures are less exploratory in the sense that they lead to well-defined rules, which can be used for assigning new objects. Classification ordinarily requires more problem structure than discrimination does.

Thus, the immediate goals of discrimination and classification, respectively, are as follows:

Goal 1. To describe, either graphically (in three or fewer dimensions) or algebraically, the differential features of objects (observations) from several known collections (populations). We try to find "discriminants" whose numerical values are such that the collections are separated as much as possible.

Goal 2. To sort objects (observations) into two or tore labeled classes. The emphasis is on deriving a rule that can be used to optimally assign new objects to the labeled classes.

We shall follow convention and use the term discrimination to refer to Goal 1. We shall refer to the second goal as classification or allocation. A function that separates objects may sometimes serve as an allocator, and. conversely, a rule that allocates objects may suggest a discriminatory procedure. In practice, Goals 1 and 2 frequently overlap, and the distinction between separation and allocation becomes blurred.

SEPARATION AND CLASSIFICATION FOR TWO POPULATIONS

To fix ideas, let us list situations in which one may be interested in (1) separating two classes of objects or (2) assigning a new object to one of two classes (or both).

It is convenient to label the classes π_1, and π_2.

The objects are ordinarily separated or classified on the basis of measurements on, for instance, p associated random variables = [X_1, X_2, ..., X_p].

The observed values of X differ to some extent from one class to the other.' We can think of the totality of values from the first class as being the population of x values for $\pi\pi_1$, and those from the second class as the population of X values for π_2.

These two populations can then be described by probability density functions f(x) and f2(x), and consequently, we can talk of assigning observations to populations or objects to classes interchangeably.

For example, that objects (consumers) are to be separated into two labeled classes (“purchases” and “laggards” on the basis of observed values of presumably relevant variables (education, income, and so forth).

In the terminology of observation and population, we want to identify an observation of the form x'= [x_1, (education), x_2(income), x_3 (family size), x_4 (amount of brand switching)] as population π_1, purchasers, or population π_2, laggards.

Allocation or classification rules are usually developed from "learning" samples. Measured characteristics of randomly selected objects known to come from each of the two populations are examined for differences. Essentially, the set of all possible sample outcomes is divided into two regions, R_1, and R_2, such that if a new observation falls in R_1, it is allocated to population π_1, and if it falls in R_2, we allocate it to population π_2. Thus, one set of observed values favors π_1, while the other set of values tavors π_2.

You may wonder at this point how it is we know that some observations belong to a particular population, but we are unsure about others.

1.Incomplete knowledge of future performance.

Example: A medical school applications office might want to classify an applicant as likely to become M.D. or unlikely to become M.D. on the basis of test scores and other college records. Here the actual determination can be made only at the end of several years of training.

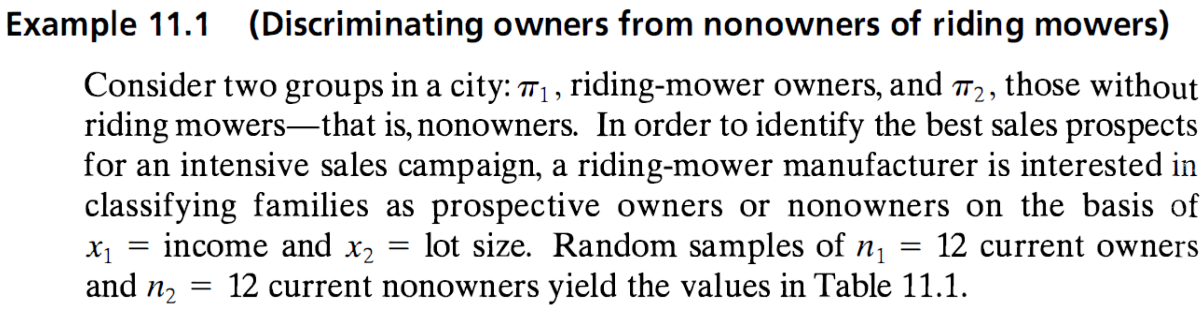
2."Perfect" information requires destroying the object.

Example: The lifetime of a calculator battery is determined by using it until it fails, and the strength of a piece of lumber is obtained by loading it until it breaks. Failed products cannot be sold. One would like to classify products as good or bad (not meeting specifications) on the basis of certain preliminary measurements.

3. Unavailable or expensive information.

Example: It is assumed that certain of the Federalist Papers were written by James Madison or Alexander Hamilton because they signed them. Others of the Papers, however, were unsigned and it is of interest to determine which of the two men wrote the unsigned Papers. Clearly, we cannot ask them. Word frequencies and sentence lengths may help classify the disputed Papers.

It should be clear from these examples that classification rules cannot usually provide an error-free method of assignment. This is because there may not be a clear distinction between the measured characteristics of the populations; that is, the groups may overlap. It is then possible, for example, to incorrectly classify a π_2, objects as belonging to π_1, or a π_1 object as belonging to π_2.



Example 11.1 (Discriminating owners from nonowners of riding mowers)

Consider two groups in a city: π_1, riding-mower owners, and π_2, those without riding mowers—that is, nonowners. In order to identify the best sales prospects for an intensive sales campaign, a riding-mower manufacturer is interested in classifying families as prospective owners or nonowners on the basis of x_1 = income and x_2 = lot size. Random samples of $n_1 = 12$ current owners and $n_2 = 12$ current nonowners yield the values in Table 11.1.

TABLE 11.1

π_1: Riding-mower owners		π_2: Nonowners	
x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)	x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)
60.0	18.4	75.0	19.6
85.5	16.8	52.8	20.8
64.8	21.6	64.8	17.2
61.5	20.8	43.2	20.4
87.0	23.6	84.0	17.6
110.1	19.2	49.2	17.6
108.0	17.6	59.4	16.0
82.8	22.4	66.0	18.4
69.0	20.0	47.4	16.4
93.0	20.8	33.0	18.8
51.0	22.0	51.0	14.0
81.0	20.0	63.0	14.8

These data are plotted in Figure 11.1. We see that riding-mower owners tend to have larger incomes and bigger lots than nonowners, although income seems to be a better "discriminator" than lot size. On the other hand, there is some overlap between the two groups. If, for example, we were to allocate those values of (x_1, x_2) that fall into region R_1 (as determined by the solid line in the figure) to π_1, mower owners, and those (x_1, x_2) values which fall into R_2 to π_2, nonowners, we would make some mistakes. Some riding-mower owners would be incorrectly classified as nonowners and, conversely, some nonowners as owners. The idea is to create a rule (regions R_1 and R_2) that minimizes the chances of making these mistakes.

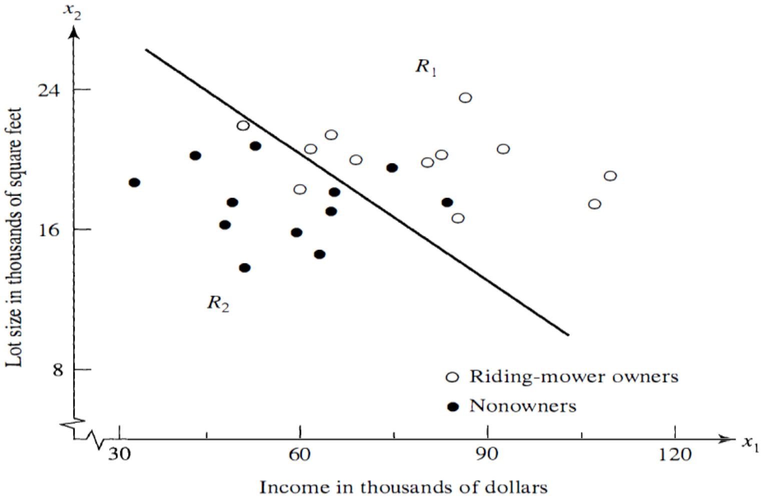


Figure 11.1 Income and lot size for riding-mower owners and nonowners.

The population of each class should be nearly equal to avoid misclassification. When the population is not equal, then the classification overwhelmingly favours the class with higher population.

Another aspect of classification is cost. Suppose that classifying a π_1 object as belonging to π_2 represents a more serious error than classifying a π_2 object as belonging to π_1. Then one should be cautious about making the former assignment. As an example, failing to diagnose a potentially fatal illness is substantially more "costly" than concluding that the disease is present when, in fact, it is not. An optimal classification procedure should, whenever possible, account for the costs associated with misclassification.

Let $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ be the probability density functions associated with the $p \times 1$ vector random variable $\mathbf{X}$ for the populations π_1 and π_2, respectively. An object with associated measurements $\mathbf{x}$ *must* be assigned to either π_1 or π_2. Let Ω be the sample space—that is, the collection of all possible observations $\mathbf{x}$. Let R_1 be that set of $\mathbf{x}$ values for which we classify objects as π_1 and $R_2 = \Omega - R_1$ be the remaining $\mathbf{x}$ values for which we classify objects as π_2. Since every object must be assigned to one and only one of the two populations, the sets R_1 and R_2 are mutually exclusive and exhaustive. For $p = 2$, we might have a case like the one pictured in Figure 11.2.

The conditional probability, $P(2 \mid 1)$, of classifying an object as π_2 when, in fact, it is from π_1 is

$$P(2 \mid 1) = P(\mathbf{X} \in R_2 \mid \pi_1) = \int_{R_2 = \Omega - R_1} f_1(\mathbf{x})\, d\mathbf{x} \qquad (11\text{-}1)$$

Similarly, the conditional probability, $P(2 \mid 1)$, of classifying an object as π_1 when it is really from π_2 is

$$P(1 \mid 2) = P(\mathbf{X} \in R_1 \mid \pi_2) = \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} \qquad (11\text{-}2)$$

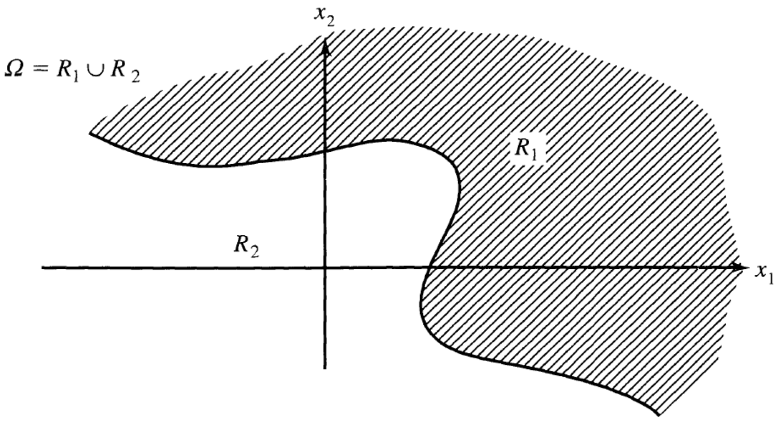


Figure 11.2 Classification regions for two populations.

The integral sign in (11-1) represents the volume formed by the density function $f_1(\mathbf{x})$ over the region R_2. Similarly, the integral sign in (11-2) represents the volume formed by $f_2(\mathbf{x})$ over the region R_1. This is illustrated in Figure 11.3 for the univariate case, $p = 1$.

Let p_1 be the *prior* probability of π_1 and p_2 be the *prior* probability of π_2, where $p_1 + p_2 = 1$. Then the overall probabilities of correctly or incorrectly classifying objects can be derived as the product of the prior and conditional classification probabilities:

$$\begin{aligned}
P(\text{observation is correctly classified as } \pi_1) &= P(\text{observation comes from } \pi_1 \\
&\qquad \text{and is correctly classified as } \pi_1) \\
&= P(\mathbf{X} \in R_1 \mid \pi_1)P(\pi_1) = P(1 \mid 1)p_1 \\
P(\text{observation is misclassified as } \pi_1) &= P(\text{observation comes from } \pi_2 \\
&\qquad \text{and is misclassified as } \pi_1) \\
&= P(\mathbf{X} \in R_1 \mid \pi_2)P(\pi_2) = P(1 \mid 2)p_2 \\
P(\text{observation is correctly classified as } \pi_2) &= P(\text{observation comes from } \pi_2 \\
&\qquad \text{and is correctly classified as } \pi_2) \\
&= P(\mathbf{X} \in R_2 \mid \pi_2)P(\pi_2) = P(2 \mid 2)p_2
\end{aligned}$$

$$\begin{aligned}
P(\text{observation is misclassified as } \pi_2) &= P(\text{observation comes from } \pi_1 \\
&\qquad \text{and is misclassified as } \pi_2) \\
&= P(\mathbf{X} \in R_2 \mid \pi_1)P(\pi_1) = P(2 \mid 1)p_1
\end{aligned} \tag{11-3}$$

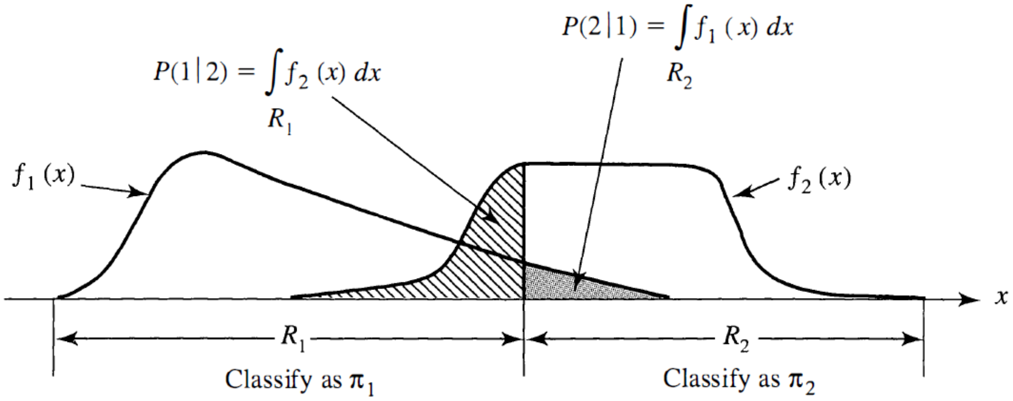


Figure 11.3 Misclassification probabilities for hypothetical classification regions when $p = 1$.

Classification schemes are often evaluated in terms of their misclassification probabilities (see Section 11.4), but this ignores misclassification cost. For example even a seemingly small probability such as $.06 = P(2 \mid 1)$ may be too large if the cost of making an incorrect assignment to π_2 is extremely high. A rule that ignores costs may cause problems.

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

$$\begin{array}{ll|cc} & & \multicolumn{2}{c}{\text{Classify as:}} \\ & & \pi_1 & \pi_2 \\ \hline \text{True population:} & \pi_1 & 0 & c(2 \mid 1) \\ & \pi_2 & c(1 \mid 2) & 0 \end{array} \tag{11-4}$$

The costs are (1) zero for correct classification, (2) $c(1 \mid 2)$ when an observation from π_2 is incorrectly classified as π_1, and (3) $c(2 \mid 1)$ when a π_1 observation is incorrectly classified as π_2.

For any rule, the average, or *expected cost of misclassification* (ECM) is provided by multiplying the off-diagonal entries in (11-4) by their probabilities of occurrence, obtained from (11-3). Consequently,

$$\text{ECM} = c(2 \mid 1)P(2 \mid 1)p_1 + c(1 \mid 2)P(1 \mid 2)p_2 \tag{11-5}$$

A reasonable classification rule should have an ECM as small, or nearly as small, as possible.

Suppose, then, that we have n_1 observations of the multivariate random variable $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ from π_1 and n_2 measurements of this quantity from π_2, with $n_1 + n_2 - 2 \geq p$. Then the respective data matrices are

$$\begin{gathered} \underset{(n_1 \times p)}{\mathbf{X}_1} = \begin{bmatrix} \mathbf{x}'_{11} \\ \mathbf{x}'_{12} \\ \vdots \\ \mathbf{x}'_{1n_1} \end{bmatrix} \\ \underset{(n_2 \times p)}{\mathbf{X}_2} = \begin{bmatrix} \mathbf{x}'_{21} \\ \mathbf{x}'_{22} \\ \vdots \\ \mathbf{x}'_{2n_2} \end{bmatrix} \end{gathered} \tag{11-15}$$

From these data matrices, the sample mean vectors and covariance matrices are determined by

$$\underset{(p\times 1)}{\bar{\mathbf{x}}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j}, \qquad \underset{(p\times p)}{\mathbf{S}_1} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)'$$
$$\underset{(p\times 1)}{\bar{\mathbf{x}}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{x}_{2j}, \qquad \underset{(p\times p)}{\mathbf{S}_2} = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)' \qquad (11\text{-}16)$$

Since it is assumed that the parent populations have the same covariance matrix $\boldsymbol{\Sigma}$, the sample covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are combined (pooled) to derive a single, unbiased estimate of $\boldsymbol{\Sigma}$. In particular, the weighted average

$$\mathbf{S}_{\text{pooled}} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_2 \qquad (11\text{-}17)$$

is an unbiased estimate of $\boldsymbol{\Sigma}$ if the data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ contain *random* samples from the populations π_1 and π_2, respectively.
the "sample" classification rule:

THE ESTIMATED MINIMUM ECM RULE FOR TWO NORMAL POPULATIONS

Allocate $\mathbf{x}_0$ to π_1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln\left[\left(\frac{c(1\,|\,2)}{c(2\,|\,1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (11\text{-}18)$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

If, in (11-18),

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

then $\ln(1) = 0$, and the estimated minimum ECM rule for two normal populations amounts to comparing the scalar variable

$$\hat{y} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} = \hat{\mathbf{a}}'\mathbf{x} \qquad (11\text{-}19)$$

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

$$\begin{aligned}\hat{m} &= \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\\ &= \frac{1}{2}(\bar{y}_1 + \bar{y}_2) \qquad (11\text{-}20)\end{aligned}$$

where

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

and

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

That is, the estimated minimum ECM rule for two normal populations is tantamount to creating two *univariate* populations for the y values by taking an appropriate linear combination of the observations from populations π_1 and π_2 and then assigning a new observation $\mathbf{x}_0$ to π_1 or π_2, depending upon whether $\hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0$ falls to the right or left of the midpoint $\hat{m}$ between the two univariate means $\bar{y}_1$ and $\bar{y}_2$.

Example 11.3 (Classification with two normal populations—common Σ and equal costs)

This example is adapted from a study [4] concerned with the detection of hemophilia A carriers. (See also Exercise 11.32.)

To construct a procedure for detecting potential hemophilia A carriers, blood samples were assayed for two groups of women and measurements on the two variables,

$$X_1 = \log_{10}(\text{AHF activity})$$
$$X_2 = \log_{10}(\text{AHF-like antigen})$$

recorded. ("AHF" denotes antihemophilic factor.) The first group of $n_1 = 30$ women were selected from a population of women who did not carry the hemophilia gene. This group was called the *normal* group. The second group of $n_2 = 22$ women was selected from known hemophilia A carriers (daughters of hemophiliacs, mothers with more than one hemophilic son, and mothers with one hemophilic son and other hemophilic relatives). This group was called the *obligatory carriers*. The pairs of observations (x_1, x_2) for the two groups are plotted in Figure 11.4. Also shown are estimated contours containing 50% and 95% of the probability for bivariate normal distributions centered at $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$, respectively. Their common covariance matrix was taken as the pooled sample covariance matrix $\mathbf{S}_{\text{pooled}}$. In this example, bivariate normal distributions seem to fit the data fairly well.

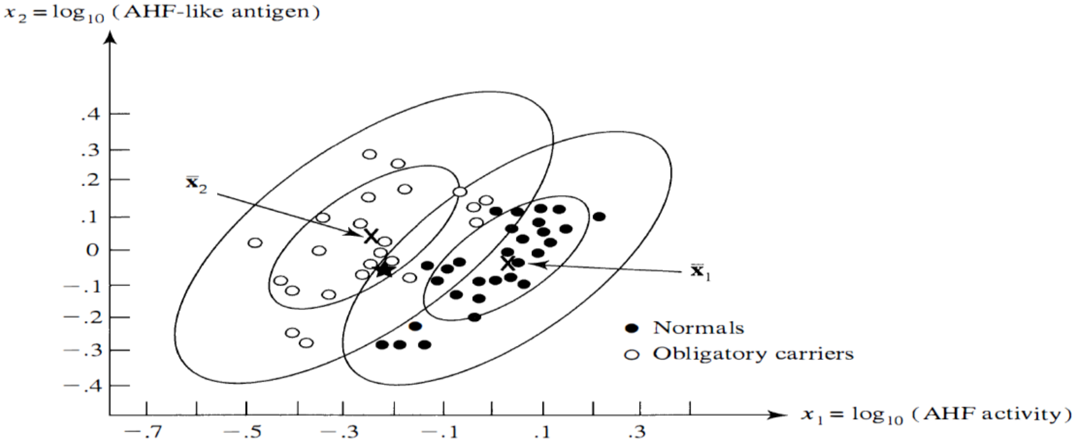


Figure 11.4 Scatter plots of [$\log_{10}$(AHF activity), $\log_{10}$(AHF-like antigen)] for the normal group and obligatory hemophilia A carriers.

The investigators (see [4]) provide the information

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} -.0065 \\ -.0390 \end{bmatrix}, \qquad \bar{\mathbf{x}}_2 = \begin{bmatrix} -.2483 \\ .0262 \end{bmatrix}$$

and

$$\mathbf{S}_{\text{pooled}}^{-1} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Therefore, the equal costs and equal priors discriminant function [see (11-19)] is

$$\begin{aligned} \hat{y} &= \hat{\mathbf{a}}'\mathbf{x} = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} \\ &= [.2418 \quad -.0652] \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 37.61x_1 - 28.92x_2 \end{aligned}$$

Moreover,

$$\bar{y}_1 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_1 = [37.61 \quad -28.92] \begin{bmatrix} -.0065 \\ -.0390 \end{bmatrix} = .88$$

$$\bar{y}_2 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_2 = [37.61 \quad -28.92] \begin{bmatrix} -.2483 \\ .0262 \end{bmatrix} = -10.10$$

and the midpoint between these means [see (11-20)] is

$$\hat{m} = \tfrac{1}{2}(\bar{y}_1 + \bar{y}_2) = \tfrac{1}{2}(.88 - 10.10) = -4.61$$

Measurements of AHF activity and AHF-like antigen on a woman who may be a hemophilia A carrier give $x_1 = -.210$ and $x_2 = -.044$. Should this woman be classified as π_1 (normal) or π_2 (obligatory carrier)?

Using (11-18) with equal costs and equal priors so that $\ln(1) = 0$, we obtain

$$\text{Allocate } \mathbf{x}_0 \text{ to } \pi_1 \text{ if } \hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 \geq \hat{m} = -4.61$$

$$\text{Allocate } \mathbf{x}_0 \text{ to } \pi_2 \text{ if } \hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 < \hat{m} = -4.61$$

where $\mathbf{x}'_0 = [-.210, -.044]$. Since

$$\hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 = [37.61 \quad -28.92] \begin{bmatrix} -.210 \\ -.044 \end{bmatrix} = -6.62 < -4.61$$

we classify the woman as π_2, an obligatory carrier. The new observation is indicated by a star in Figure 11.4. We see that it falls within the estimated .50 probability contour of population π_2 and about on the estimated .95 probability contour of population π_1. Thus, the classification is not clear cut.

AN ALLOCATION RULE BASED ON FISHER'S DISCRIMINANT FUNCTION[8]

Allocate $\mathbf{x}_0$ to π_1 if

$$\begin{aligned} \hat{y}_0 &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 \\ &\geq \hat{m} = \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \end{aligned}$$

or (11-35)

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

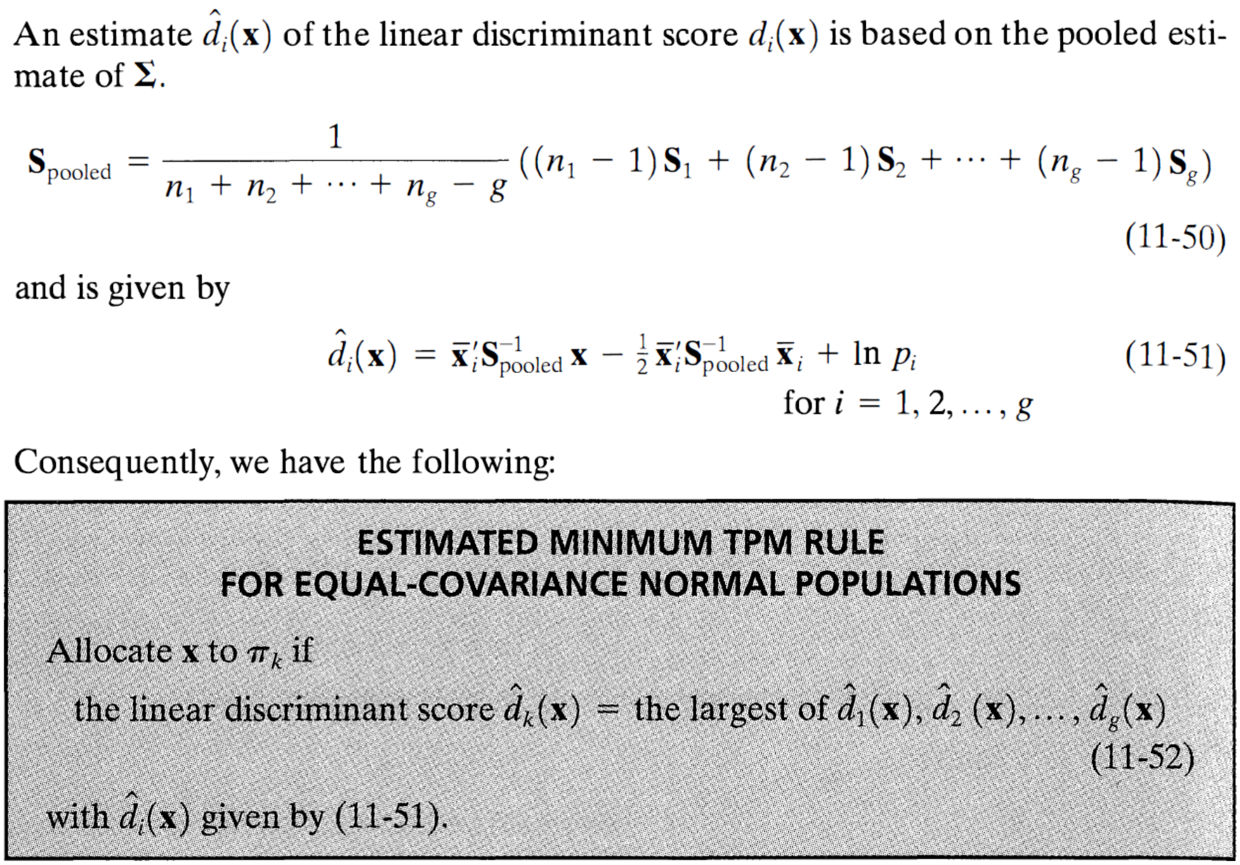
Allocate $\mathbf{x}_0$ to π_2 if

$$\hat{y}_0 < \hat{m}$$

or

$$\hat{y}_0 - \hat{m} < 0$$

Fisher discriminant Formulae 1:



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

$$\mathbf{S}_{\text{pooled}} = \frac{1}{n_1 + n_2 + \cdots + n_g - g}\left((n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \cdots + (n_g - 1)\mathbf{S}_g\right) \tag{11-50}$$

and is given by

$$\hat{d}_i(\mathbf{x}) = \bar{\mathbf{x}}_i'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \tfrac{1}{2}\bar{\mathbf{x}}_i'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_i + \ln p_i \qquad \text{for } i = 1, 2, \ldots, g \tag{11-51}$$

Consequently, we have the following:

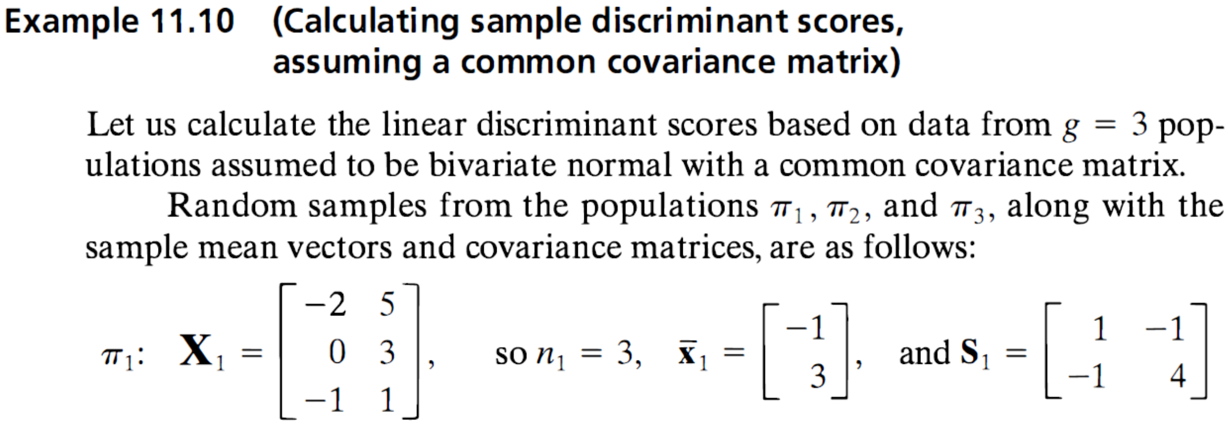
ESTIMATED MINIMUM TPM RULE
FOR EQUAL-COVARIANCE NORMAL POPULATIONS

Allocate $\mathbf{x}$ to π_k if

the linear discriminant score $\hat{d}_k(\mathbf{x})$ = the largest of $\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x}), \ldots, \hat{d}_g(\mathbf{x})$ (11-52)

with $\hat{d}_i(\mathbf{x})$ given by (11-51).

Problem 1:



Example 11.10 (Calculating sample discriminant scores, assuming a common covariance matrix)

Let us calculate the linear discriminant scores based on data from $g = 3$ populations assumed to be bivariate normal with a common covariance matrix.

Random samples from the populations π_1, π_2, and π_3, along with the sample mean vectors and covariance matrices, are as follows:

$$\pi_1\!: \quad \mathbf{X}_1 = \begin{bmatrix} -2 & 5 \\ 0 & 3 \\ -1 & 1 \end{bmatrix}, \qquad \text{so } n_1 = 3, \quad \bar{\mathbf{x}}_1 = \begin{bmatrix} -1 \\ 3 \end{bmatrix}, \quad \text{and } \mathbf{S}_1 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

$$\pi_2: \quad \mathbf{X}_2 = \begin{bmatrix} 0 & 6 \\ 2 & 4 \\ 1 & 2 \end{bmatrix}, \qquad \text{so } n_2 = 3, \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \qquad \text{and } \mathbf{S}_2 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

$$\pi_3: \quad \mathbf{X}_3 = \begin{bmatrix} 1 & -2 \\ 0 & 0 \\ -1 & -4 \end{bmatrix}, \quad \text{so } n_3 = 3, \quad \bar{\mathbf{x}}_3 = \begin{bmatrix} 0 \\ -2 \end{bmatrix}, \quad \text{and } \mathbf{S}_3 = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

Given that $p_1 = p_2 = .25$ and $p_3 = .50$, let us classify the observation $\mathbf{x}_0' = [x_{01}, x_{02}] = [-2 \quad -1]$ according to (11-52). From (11-50),

$$\mathbf{S}_{\text{pooled}} = \frac{3-1}{9-3}\begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} + \frac{3-1}{9-3}\begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix} + \frac{3-1}{9-3}\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

$$= \frac{2}{6}\begin{bmatrix} 1+1+1 & -1-1+1 \\ -1-1+1 & 4+4+4 \end{bmatrix} = \begin{bmatrix} 1 & -\frac{1}{3} \\ -\frac{1}{3} & 4 \end{bmatrix}$$

so

$$\mathbf{S}_{\text{pooled}}^{-1} = \frac{9}{35}\begin{bmatrix} 4 & \frac{1}{3} \\ \frac{1}{3} & 1 \end{bmatrix} = \frac{1}{35}\begin{bmatrix} 36 & 3 \\ 3 & 9 \end{bmatrix}$$

Next,

$$\bar{\mathbf{x}}_1'\mathbf{S}_{\text{pooled}}^{-1} = [-1 \quad 3]\frac{1}{35}\begin{bmatrix} 36 & 3 \\ 3 & 9 \end{bmatrix} = \frac{1}{35}[-27 \quad 24]$$

and

$$\bar{\mathbf{x}}_1'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_1 = \frac{1}{35}[-27 \quad 24]\begin{bmatrix} -1 \\ 3 \end{bmatrix} = \frac{99}{35}$$

so

$$\begin{aligned}\hat{d}_1(\mathbf{x}_0) &= \ln p_1 + \bar{\mathbf{x}}_1'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}\bar{\mathbf{x}}_1'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_1 \\ &= \ln(.25) + \left(\frac{-27}{35}\right)x_{01} + \left(\frac{24}{35}\right)x_{02} - \frac{1}{2}\left(\frac{99}{35}\right)\end{aligned}$$

Notice the linear form of $\hat{d}_1(\mathbf{x}_0) = \text{constant} + (\text{constant})\,x_{01} + (\text{constant})\,x_{02}$. In a similar manner,

$$\bar{\mathbf{x}}_2'\mathbf{S}_{\text{pooled}}^{-1} = [1 \quad 4]\,\frac{1}{35}\begin{bmatrix}36 & 3\\ 3 & 9\end{bmatrix} = \frac{1}{35}[48 \quad 39]$$

$$\bar{\mathbf{x}}_2'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_2 = \frac{1}{35}[48 \quad 39]\begin{bmatrix}1\\ 4\end{bmatrix} = \frac{204}{35}$$

and

$$\hat{d}_2(\mathbf{x}_0) = \ln(.25) + \left(\frac{48}{35}\right)x_{01} + \left(\frac{39}{35}\right)x_{02} - \frac{1}{2}\left(\frac{204}{35}\right)$$

Finally,

$$\bar{\mathbf{x}}_3'\mathbf{S}_{\text{pooled}}^{-1} = [0 \quad -2]\,\frac{1}{35}\begin{bmatrix}36 & 3\\ 3 & 9\end{bmatrix} = \frac{1}{35}[-6 \quad -18]$$

$$\bar{\mathbf{x}}_3'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_3 = \frac{1}{35}[-6 \quad -18]\begin{bmatrix}0\\ -2\end{bmatrix} = \frac{36}{35}$$

and

$$\hat{d}_3(\mathbf{x}_0) = \ln(.50) + \left(\frac{-6}{35}\right)x_{01} + \left(\frac{-18}{35}\right)x_{02} - \frac{1}{2}\left(\frac{36}{35}\right)$$

Substituting the numerical values $x_{01} = -2$ and $x_{02} = -1$ gives

$$\hat{d}_1(\mathbf{x}_0) = -1.386 + \left(\frac{-27}{35}\right)(-2) + \left(\frac{24}{35}\right)(-1) - \frac{99}{70} = -1.943$$

$$\hat{d}_2(\mathbf{x}_0) = -1.386 + \left(\frac{48}{35}\right)(-2) + \left(\frac{39}{35}\right)(-1) - \frac{204}{70} = -8.158$$

$$\hat{d}_3(\mathbf{x}_0) = -.693 + \left(\frac{-6}{35}\right)(-2) + \left(\frac{-18}{35}\right)(-1) - \frac{36}{70} = -.350$$

Since $\hat{d}_3(\mathbf{x}_0) = -.350$ is the largest discriminant score, we allocate $\mathbf{x}_0$ to π_3. ■

FISHER'S DISCRIMINANT FUNCTION—SEPARATION OF POPULATIONS

Fisher [9] actually arrived at the linear classification statistic (11-19) using an entirely different argument from the one in Section 11.3. Fisher's idea was to transform the multivariate observations $\mathbf{x}$ to univariate observations y such that the y's derived from populations π_1 and π_2 were separated as much as possible. Fisher suggested taking linear combinations of $\mathbf{x}$ to create y's because they are simple enough functions of the $\mathbf{x}$ to be handled easily. Fisher's approach does not assume that the populations are normal. It does, however, implicitly assume that the population covariance matrices are equal, because a pooled estimate of the common covariance matrix is used.

A fixed linear combination of the $\mathbf{x}$'s takes the values $y_{11}, y_{12}, \ldots, y_{1n_1}$ for the observations from the first population and the values $y_{21}, y_{22}, \ldots, y_{2n_2}$ for the observations from the second population. The separation of these two sets of univariate y's is assessed in terms of the difference between $\bar{y}_1$ and $\bar{y}_2$ expressed in standard deviation units. That is,

$$\text{separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \qquad \text{where } s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

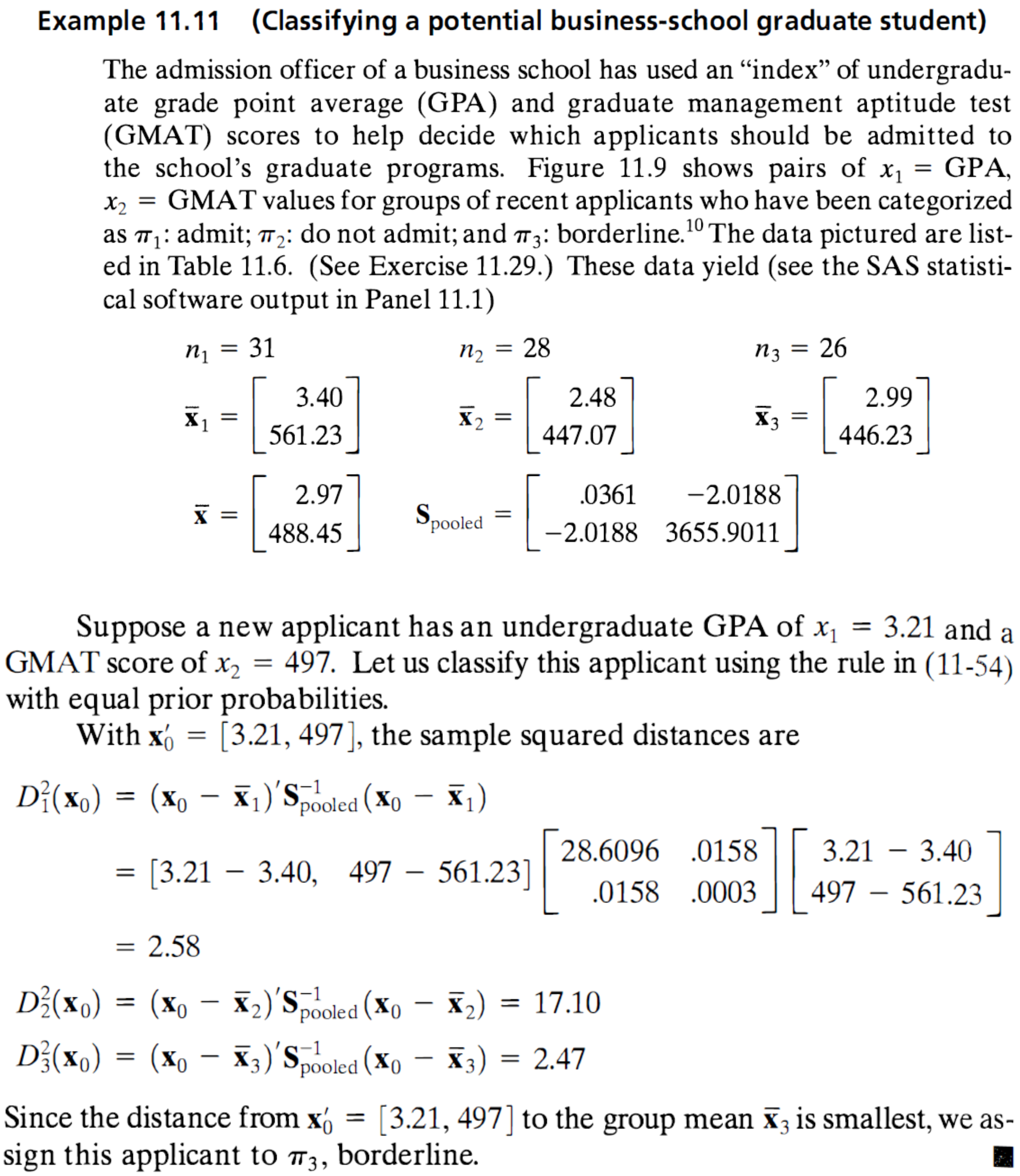
is the pooled estimate of the variance. The objective is to select the linear combination of the $\mathbf{x}$ to achieve maximum separation of the sample means $\bar{y}_1$ and $\bar{y}_2$.

Result 11.4. The linear combination $\hat{y} = \hat{\mathbf{a}}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}$ maximizes the ratio

$$\frac{\left(\begin{array}{c}\text{Squared distance}\\ \text{between sample means of } y\end{array}\right)}{(\text{Sample variance of } y)} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{\mathbf{a}}'\bar{\mathbf{x}}_1 - \hat{\mathbf{a}}'\bar{\mathbf{x}}_2)^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}} = \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}} \qquad (11\text{-}33)$$

over all possible coefficient vectors $\hat{\mathbf{a}}$ where $\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. The maximum of the ratio (11-33) is $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

Problem 2:



Example 11.11 (Classifying a potential business-school graduate student)

The admission officer of a business school has used an "index" of undergraduate grade point average (GPA) and graduate management aptitude test (GMAT) scores to help decide which applicants should be admitted to the school's graduate programs. Figure 11.9 shows pairs of x_1 = GPA, x_2 = GMAT values for groups of recent applicants who have been categorized as π_1: admit; π_2: do not admit; and π_3: borderline.[10] The data pictured are listed in Table 11.6. (See Exercise 11.29.) These data yield (see the SAS statistical software output in Panel 11.1)

$$n_1 = 31 \qquad n_2 = 28 \qquad n_3 = 26$$

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 3.40 \\ 561.23 \end{bmatrix} \qquad \bar{\mathbf{x}}_2 = \begin{bmatrix} 2.48 \\ 447.07 \end{bmatrix} \qquad \bar{\mathbf{x}}_3 = \begin{bmatrix} 2.99 \\ 446.23 \end{bmatrix}$$

$$\bar{\mathbf{x}} = \begin{bmatrix} 2.97 \\ 488.45 \end{bmatrix} \qquad \mathbf{S}_{\text{pooled}} = \begin{bmatrix} .0361 & -2.0188 \\ -2.0188 & 3655.9011 \end{bmatrix}$$

Suppose a new applicant has an undergraduate GPA of $x_1 = 3.21$ and a GMAT score of $x_2 = 497$. Let us classify this applicant using the rule in (11-54) with equal prior probabilities.

With $\mathbf{x}_0' = [3.21, 497]$, the sample squared distances are

$$\begin{aligned} D_1^2(\mathbf{x}_0) &= (\mathbf{x}_0 - \bar{\mathbf{x}}_1)'\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1) \\ &= [3.21 - 3.40, \quad 497 - 561.23] \begin{bmatrix} 28.6096 & .0158 \\ .0158 & .0003 \end{bmatrix} \begin{bmatrix} 3.21 - 3.40 \\ 497 - 561.23 \end{bmatrix} \\ &= 2.58 \\ D_2^2(\mathbf{x}_0) &= (\mathbf{x}_0 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_2) = 17.10 \\ D_3^2(\mathbf{x}_0) &= (\mathbf{x}_0 - \bar{\mathbf{x}}_3)'\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_3) = 2.47 \end{aligned}$$

Since the distance from $\mathbf{x}_0' = [3.21, 497]$ to the group mean $\bar{\mathbf{x}}_3$ is smallest, we assign this applicant to π_3, borderline. ■

Misclassification calculation

Approach 1:

There is a measure of performance that does not depend on the form of the parent populations and that can be calculated for *any* classification procedure. This measure, called the *apparent error rate* (APER), is defined as the fraction of observations in the *training* sample that are misclassified by the sample classification function.

The apparent error rate can be easily 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

(11-29)

where

n_{1C} = number of π_1 items correctly classified as π_1 items
n_{1M} = number of π_1 items misclassified as π_2 items
n_{2C} = number of π_2 items correctly classified
n_{2M} = number of π_2 items misclassified

The apparent error rate is then

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \tag{11-30}$$

which is recognized as the *proportion* of items in the training set that are misclassified.

Example 11.5 (Calculating the apparent error rate)

Consider the classification regions R_1 and R_2 shown in Figure 11.1 for the riding-mower data. In this case, observations northeast of the solid line are classified as π_1, mower owners; observations southwest of the solid line are classified as π_2, nonowners. Notice that some observations are misclassified. The confusion matrix is

		Predicted membership		
		π_1: riding-mower owners	π_2: nonowners	
Actual membership	π_1: riding-mower owners	$n_{1C} = 10$	$n_{1M} = 2$	$n_1 = 12$
	π_2: nonowners	$n_{2M} = 2$	$n_{2C} = 10$	$n_2 = 12$

The apparent error rate, expressed as a percentage, is

$$\text{APER} = \left(\frac{2 + 2}{12 + 12}\right)100\% = \left(\frac{4}{24}\right)100\% = 16.7\%$$

■

Approach 2:

A second approach that seems to work well is called Lachenbruch's "holdout" procedure[6] (see also Lachenbruch and Mickey [22]):

1. Start with the π_1 group of observations. Omit one observation from this group, and develop a classification function based on the remaining $n_1 - 1$, n_2 observations.

2. Classify the "holdout" observation, using the function constructed in Step 1.
3. Repeat Steps 1 and 2 until all of the π_1 observations are classified. Let $n_{1M}^{(H)}$ be the number of holdout (H) observations misclassified in this group.
4. Repeat Steps 1 through 3 for the π_2 observations. Let $n_{2M}^{(H)}$ be the number of holdout observations misclassified in this group.

Estimates $\hat{P}(2 \mid 1)$ and $\hat{P}(1 \mid 2)$ of the conditional misclassification probabilities in (11-1) and (11-2) are then given by

$$\hat{P}(2 \mid 1) = \frac{n_{1M}^{(H)}}{n_1}$$

$$\hat{P}(1 \mid 2) = \frac{n_{2M}^{(H)}}{n_2} \qquad (11\text{-}31)$$

and the total proportion misclassified, $(n_{1M}^{(H)} + n_{2M}^{(H)})/(n_1 + n_2)$, is, for moderate samples, a nearly unbiased estimate of the *expected* actual error rate, $E(\text{AER})$.

$$\boxed{\hat{E}(\text{AER}) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} \qquad (11\text{-}32)}$$

Example 11.6 (Calculating an estimate of the error rate using the holdout procedure)

We shall illustrate Lachenbruch's holdout procedure and the calculation of error rate estimates for the equal costs and equal priors version of (11-18). Consider the following data matrices and descriptive statistics. (We shall assume that the $n_1 = n_2 = 3$ bivariate observations were selected randomly from two populations π_1 and π_2 with a common covariance matrix.)

$$\mathbf{X}_1 = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix}; \qquad \bar{\mathbf{x}}_1 = \begin{bmatrix} 3 \\ 10 \end{bmatrix}, \qquad 2\mathbf{S}_1 = \begin{bmatrix} 2 & -2 \\ -2 & 8 \end{bmatrix}$$

$$\mathbf{X}_2 = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}; \qquad \bar{\mathbf{x}}_2 = \begin{bmatrix} 4 \\ 7 \end{bmatrix}, \qquad 2\mathbf{S}_2 = \begin{bmatrix} 2 & -2 \\ -2 & 8 \end{bmatrix}$$

The pooled covariance matrix is

$$\mathbf{S}_{\text{pooled}} = \frac{1}{4}(2\mathbf{S}_1 + 2\mathbf{S}_2) = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

Using $\mathbf{S}_{\text{pooled}}$, the rest of the data, and Rule (11-18) with equal costs and equal priors, we may classify the sample observations. You may then verify (see Exercise 11.19) that the confusion matrix is

		Classify as:	
		π_1	π_2
True population:	π_1	2	1
	π_2	1	2

and consequently,

$$\text{APER(apparent error rate)} = \frac{2}{6} = .33$$

Holding out the first observation $\mathbf{x}'_H = [2, 12]$ from $\mathbf{X}_1$, we calculate

$$\mathbf{X}_{1H} = \begin{bmatrix} 4 & 10 \\ 3 & 8 \end{bmatrix}; \qquad \bar{\mathbf{x}}_{1H} = \begin{bmatrix} 3.5 \\ 9 \end{bmatrix}; \quad \text{and} \quad 1\mathbf{S}_{1H} = \begin{bmatrix} .5 & 1 \\ 1 & 2 \end{bmatrix}$$

The new pooled covariance matrix, $\mathbf{S}_{H,\text{pooled}}$, is

$$\mathbf{S}_{H,\text{pooled}} = \frac{1}{3}[1\mathbf{S}_{1H} + 2\mathbf{S}_2] = \frac{1}{3}\begin{bmatrix} 2.5 & -1 \\ -1 & 10 \end{bmatrix}$$

with inverse′

$$\mathbf{S}^{-1}_{H,\text{pooled}} = \frac{1}{8}\begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix}$$

It is computationally quicker to classify the holdout observation $\mathbf{x}_{1H}$ on the basis of its squared distances from the group means $\bar{\mathbf{x}}_{1H}$ and $\bar{\mathbf{x}}_2$. This procedure is equivalent to computing the value of the linear function $\hat{y} = \hat{\mathbf{a}}'_H \mathbf{x}_H = (\bar{\mathbf{x}}_{1H} - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}_{H,\text{pooled}}\mathbf{x}_H$ and comparing it to the midpoint $\hat{m}_H = \frac{1}{2}(\bar{\mathbf{x}}_{1H} - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}_{H,\text{pooled}}(\bar{\mathbf{x}}_{1H} + \bar{\mathbf{x}}_2)$. [See (11-19) and (11-20).]

Thus with $\mathbf{x}'_H = [2, 12]$ we have

$$\text{Squared distance from } \bar{\mathbf{x}}_{1H} = (\mathbf{x}_H - \bar{\mathbf{x}}_{1H})'\mathbf{S}^{-1}_{H,\text{pooled}}(\mathbf{x}_H - \bar{\mathbf{x}}_{1H})$$

$$= [2 - 3.5 \quad 12 - 9]\,\frac{1}{8}\begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix}\begin{bmatrix} 2 - 3.5 \\ 12 - 9 \end{bmatrix} = 4.5$$

$$\text{Squared distance from } \bar{\mathbf{x}}_2 = (\mathbf{x}_H - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}_{H,\text{pooled}}(\mathbf{x}_H - \bar{\mathbf{x}}_2)$$

$$= [2 - 4 \quad 12 - 7]\,\frac{1}{8}\begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix}\begin{bmatrix} 2 - 4 \\ 12 - 7 \end{bmatrix} = 10.3$$

Since the distance from $\mathbf{x}_H$ to $\bar{\mathbf{x}}_{1H}$ is smaller than the distance from $\mathbf{x}_H$ to $\bar{\mathbf{x}}_2$, we classify $\mathbf{x}_H$ as a π_1 observation. In this case, the classification is correct.

If $\mathbf{x}'_H = [4, 10]$ is withheld, $\bar{\mathbf{x}}_{1H}$ and $\mathbf{S}^{-1}_{H,\text{pooled}}$ become

$$\bar{\mathbf{x}}_{1H} = \begin{bmatrix} 2.5 \\ 10 \end{bmatrix} \quad \text{and} \quad \mathbf{S}^{-1}_{H,\text{pooled}} = \frac{1}{8}\begin{bmatrix} 16 & 4 \\ 4 & 2.5 \end{bmatrix}$$

We find that

$$(\mathbf{x}_H - \bar{\mathbf{x}}_{1H})'\mathbf{S}^{-1}_{H,\text{pooled}}(\mathbf{x}_H - \bar{\mathbf{x}}_{1H}) = [4 - 2.5 \quad 10 - 10]\,\frac{1}{8}\begin{bmatrix} 16 & 4 \\ 4 & 2.5 \end{bmatrix}\begin{bmatrix} 4 - 2.5 \\ 10 - 10 \end{bmatrix}$$
$$= 4.5$$

$$(\mathbf{x}_H - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}_{H,\text{pooled}}(\mathbf{x}_H - \bar{\mathbf{x}}_2) = [4 - 4 \quad 10 - 7]\,\frac{1}{8}\begin{bmatrix} 16 & 4 \\ 4 & 2.5 \end{bmatrix}\begin{bmatrix} 4 - 4 \\ 10 - 7 \end{bmatrix}$$
$$= 2.8$$

and consequently, we would incorrectly assign $\mathbf{x}'_H = [4, 10]$ to π_2. Holding out $\mathbf{x}'_H = [3, 8]$ leads to incorrectly assigning this observation to π_2 as well. Thus, $n^{(H)}_{1M} = 2$.

Turning to the second group, suppose $\mathbf{x}'_H = [5, 7]$ is withheld. Then

$$\mathbf{X}_{2H} = \begin{bmatrix} 3 & 9 \\ 4 & 5 \end{bmatrix}; \qquad \bar{\mathbf{x}}_{2H} = \begin{bmatrix} 3.5 \\ 7 \end{bmatrix}; \quad \text{and} \quad 1\mathbf{S}_{2H} = \begin{bmatrix} .5 & -2 \\ -2 & 8 \end{bmatrix}$$

The new pooled covariance matrix is

$$\mathbf{S}_{H,\text{pooled}} = \frac{1}{3}[2\mathbf{S}_1 + 1\mathbf{S}_{2H}] = \frac{1}{3}\begin{bmatrix} 2.5 & -4 \\ -4 & 16 \end{bmatrix}$$

with inverse

$$\mathbf{S}_{H,\text{pooled}}^{-1} = \frac{3}{24}\begin{bmatrix} 16 & 4 \\ 4 & 2.5 \end{bmatrix}$$

We find that

$$(\mathbf{x}_H - \overline{\mathbf{x}}_1)'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \overline{\mathbf{x}}_1) = [5-3 \quad 7-10]\frac{3}{24}\begin{bmatrix} 16 & 4 \\ 4 & 2.5 \end{bmatrix}\begin{bmatrix} 5-3 \\ 7-10 \end{bmatrix}$$
$$= 4.8$$

$$(\mathbf{x}_H - \overline{\mathbf{x}}_{2H})'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \overline{\mathbf{x}}_{2H}) = [5-3.5 \quad 7-7]\frac{3}{24}\begin{bmatrix} 16 & 4 \\ 4 & 2.5 \end{bmatrix}\begin{bmatrix} 5-3.5 \\ 7-7 \end{bmatrix}$$
$$= 4.5$$

and $\mathbf{x}'_H = [5, 7]$ is correctly assigned to π_2.

When $\mathbf{x}'_H = [3, 9]$ is withheld,

$$(\mathbf{x}_H - \overline{\mathbf{x}}_1)'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \overline{\mathbf{x}}_1) = [3-3 \quad 9-10]\frac{3}{24}\begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix}\begin{bmatrix} 3-3 \\ 9-10 \end{bmatrix}$$
$$= .3$$

$$(\mathbf{x}_H - \overline{\mathbf{x}}_{2H})'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \overline{\mathbf{x}}_{2H}) = [3-4.5 \quad 9-6]\frac{3}{24}\begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix}\begin{bmatrix} 3-4.5 \\ 9-6 \end{bmatrix}$$
$$= 4.5$$

and $\mathbf{x}'_H = [3, 9]$ is incorrectly assigned to π_1. Finally, withholding $\mathbf{x}'_H = [4, 5]$ leads to correctly classifying this observation as π_2. Thus, $n_{2M}^{(H)} = 1$.

An estimate of the expected actual error rate is provided by

$$\hat{E}(\text{AER}) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} = \frac{2+1}{3+3} = .5$$

Hence, we see that the apparent error rate APER = .33 is an optimistic measure of performance. Of course, in practice, sample sizes are larger than those we have considered here, and the difference between APER and $\hat{E}(\text{AER})$ may not be as large. ■