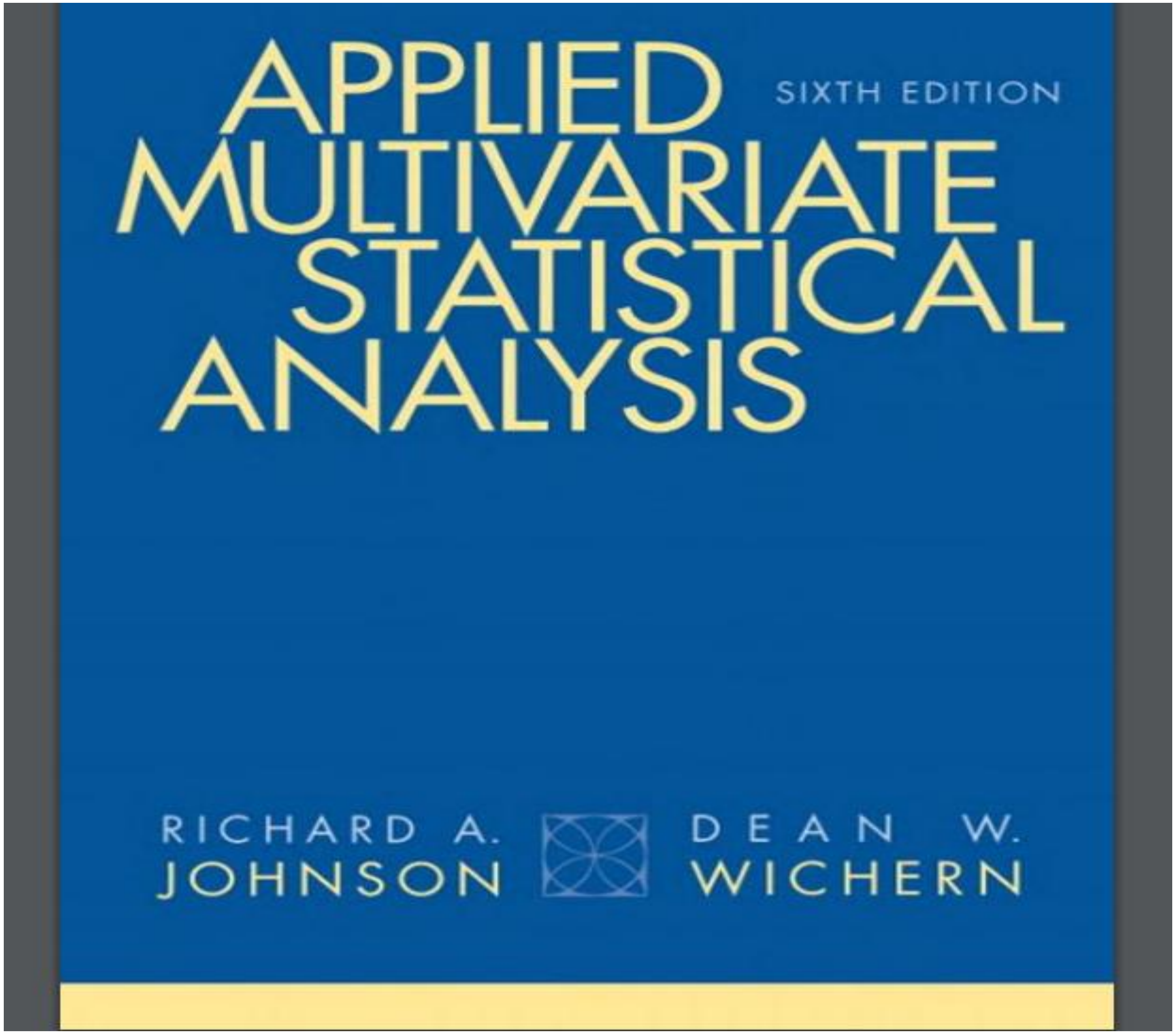
# Advanced Multivariate Methods



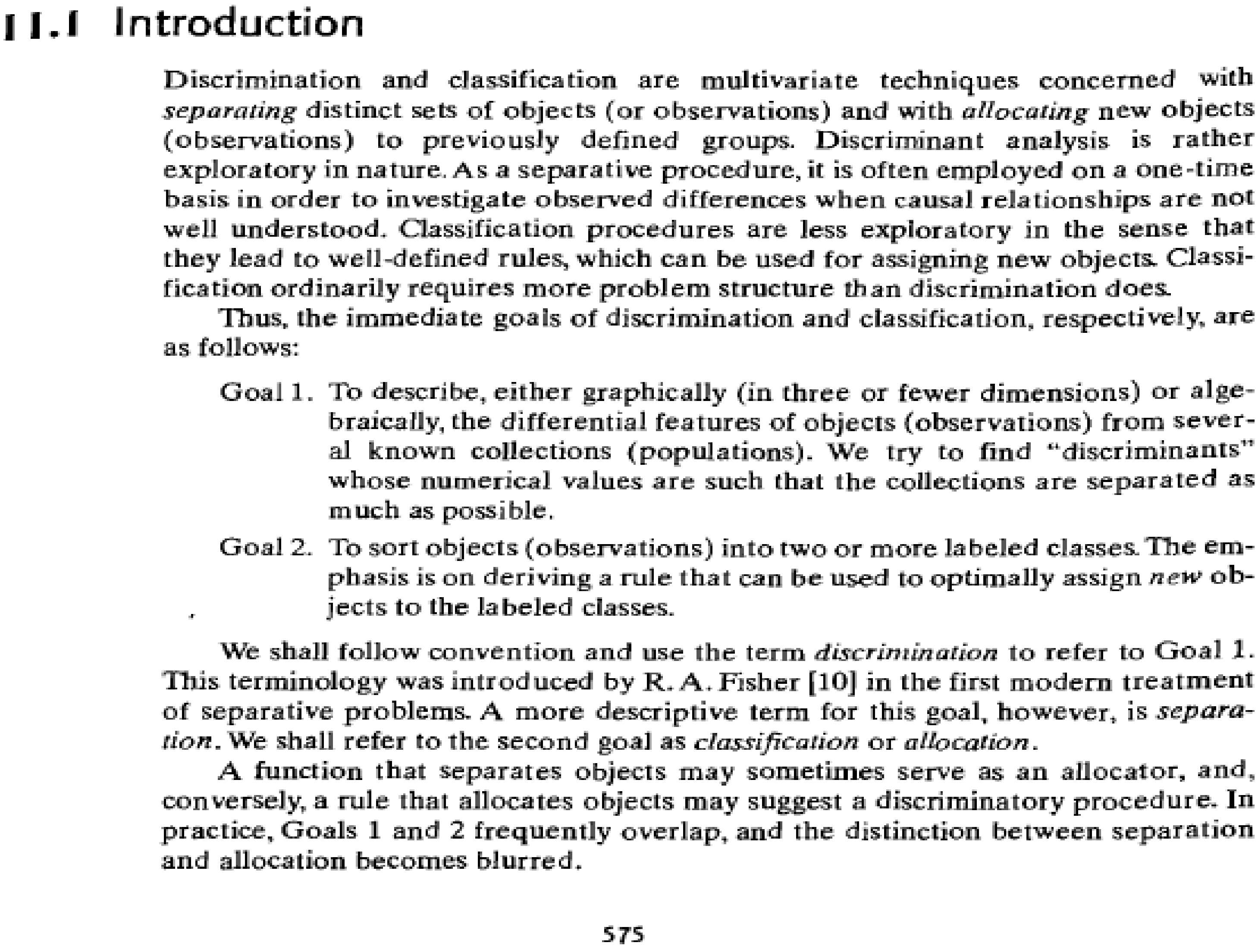
APPLIED MULTIVARIATE STATISTICAL ANALYSIS

SIXTH EDITION

RICHARD A. JOHNSON

DEAN W. WICHERN

# Discrimination and Classification



## 11.1 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:

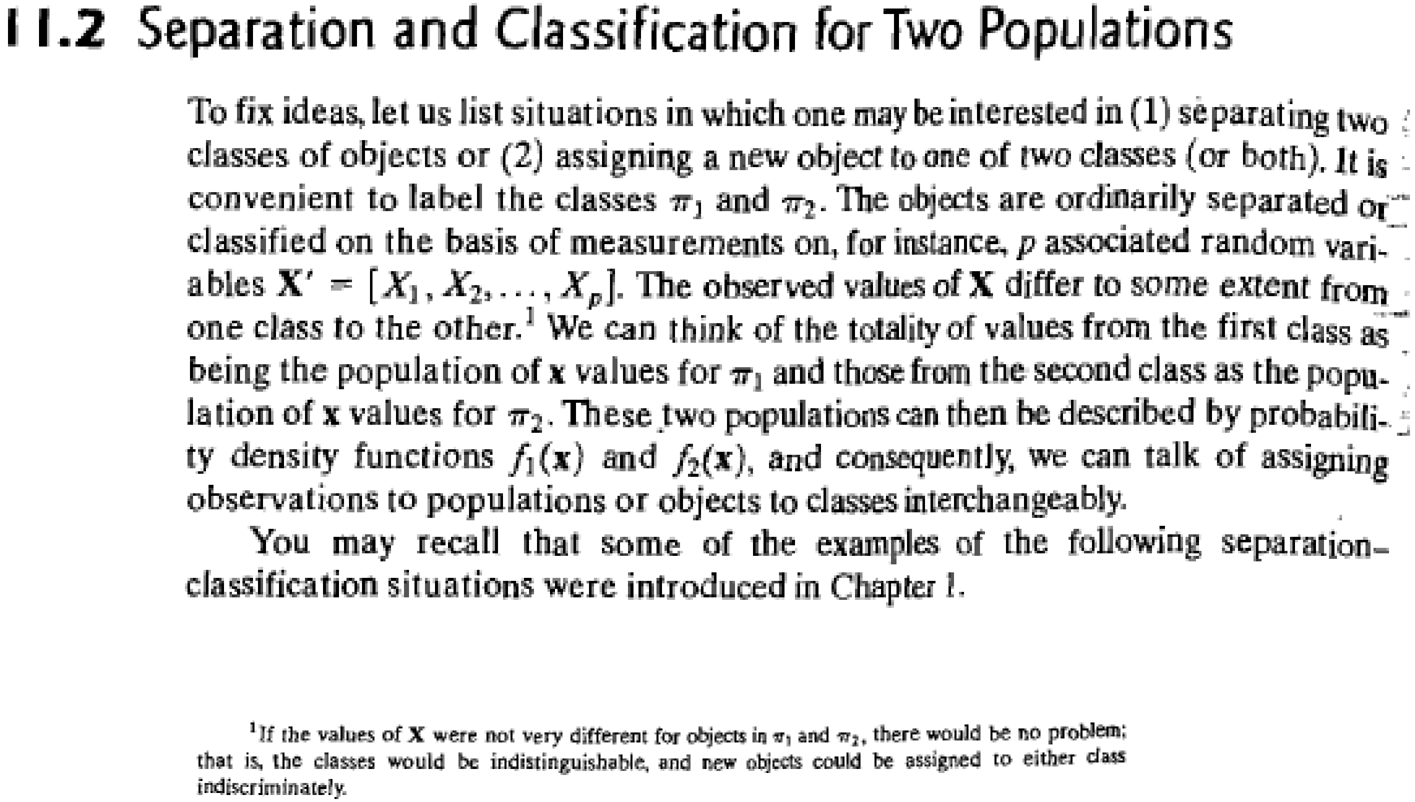
Goal 1. To describe, either graphically (in three or fewer dimensions) or algebraically, the differential features of objects (observations) from several known collections (populations). We try to find "discriminants" whose numerical values are such that the collections are separated as much as possible.

Goal 2. To sort objects (observations) into two or more 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. This terminology was introduced by R. A. Fisher [10] in the first modern treatment of separative problems. A more descriptive term for this goal, however, is *separation*. 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



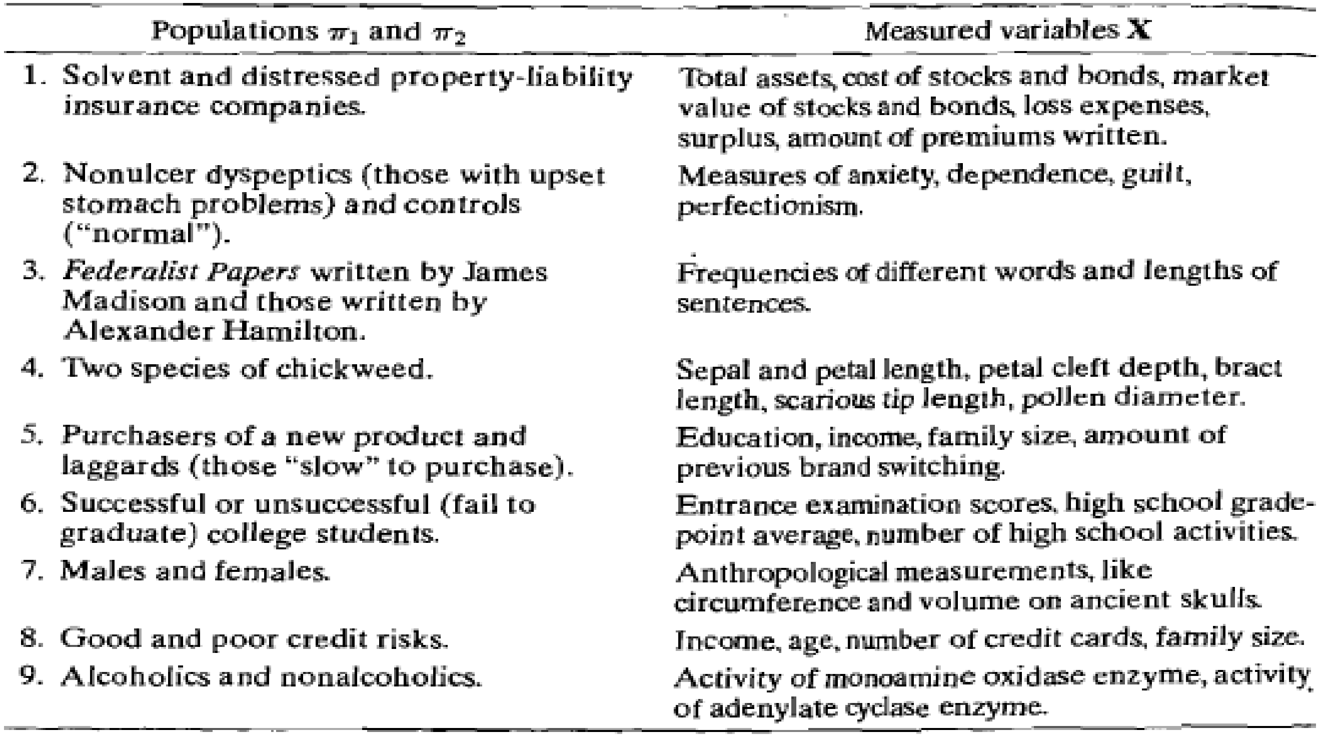
## 11.2 Separation and Classification for Two Populations

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 $\pi_1$ and $\pi_2$. The objects are ordinarily separated or classified on the basis of measurements on, for instance, $p$ associated random variables $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$. The observed values of $\mathbf{X}$ differ to some extent from one class to the other.[1] We can think of the totality of values from the first class as being the population of $\mathbf{x}$ values for $\pi_1$ and those from the second class as the population of $\mathbf{x}$ values for $\pi_2$. These two populations can then be described by probability density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, and consequently, we can talk of assigning observations to populations or objects to classes interchangeably.

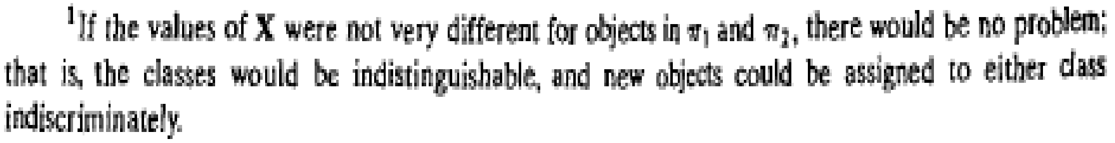
You may recall that some of the examples of the following separation-classification situations were introduced in Chapter 1.

[1] If the values of $\mathbf{X}$ were not very different for objects in $\pi_1$ and $\pi_2$, there would be no problem; that is, the classes would be indistinguishable, and new objects could be assigned to either class indiscriminately.

# Two Populations and Corresponding Measured Variables

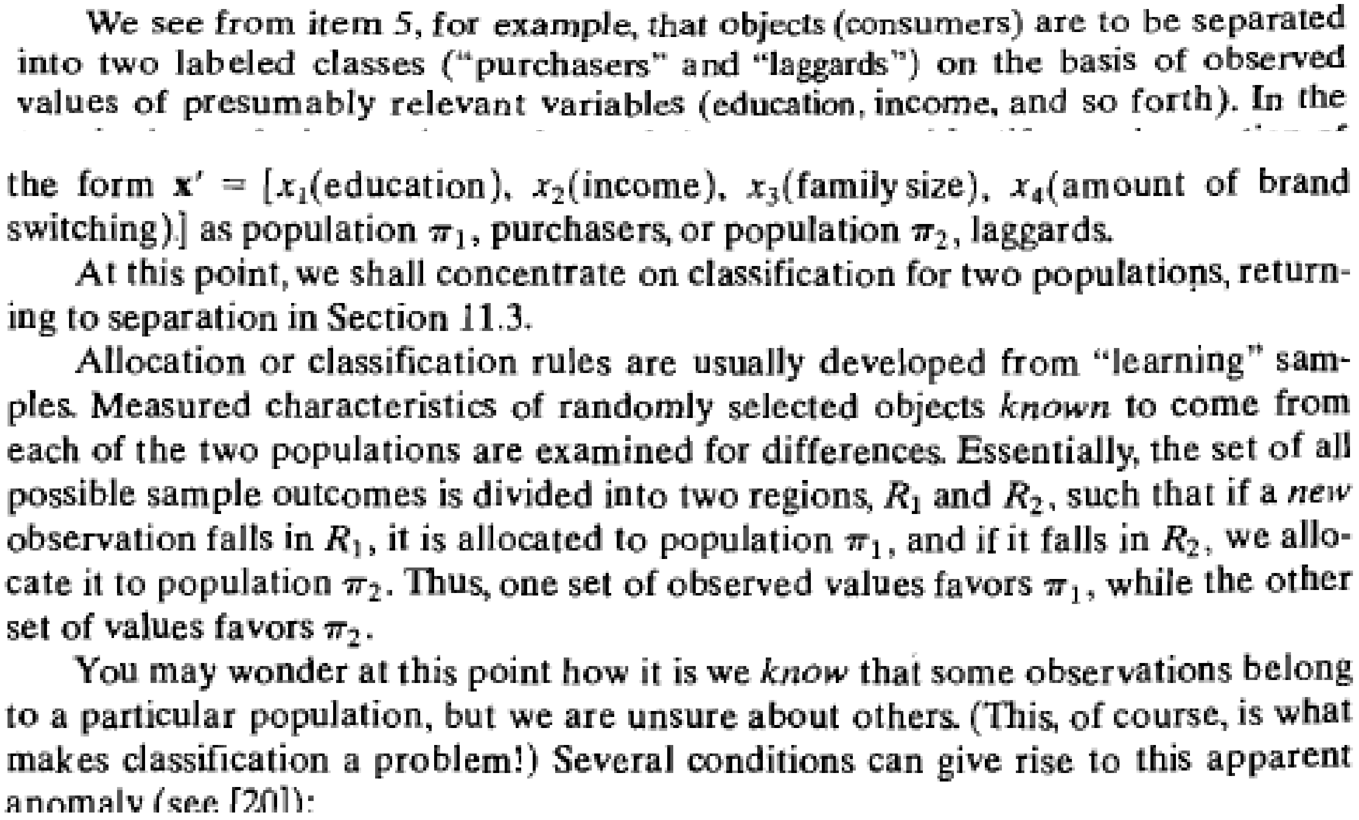


| Populations $\pi_1$ and $\pi_2$ | Measured variables **X** |
|---|---|
| 1. Solvent and distressed property-liability insurance companies. | Total assets, cost of stocks and bonds, market value of stocks and bonds, loss expenses, surplus, amount of premiums written. |
| 2. Nonulcer dyspeptics (those with upset stomach problems) and controls ("normal"). | Measures of anxiety, dependence, guilt, perfectionism. |
| 3. *Federalist Papers* written by James Madison and those written by Alexander Hamilton. | Frequencies of different words and lengths of sentences. |
| 4. Two species of chickweed. | Sepal and petal length, petal cleft depth, bract length, scarious tip length, pollen diameter. |
| 5. Purchasers of a new product and laggards (those "slow" to purchase). | Education, income, family size, amount of previous brand switching. |
| 6. Successful or unsuccessful (fail to graduate) college students. | Entrance examination scores, high school grade-point average, number of high school activities. |
| 7. Males and females. | Anthropological measurements, like circumference and volume on ancient skulls. |
| 8. Good and poor credit risks. | Income, age, number of credit cards, family size. |
| 9. Alcoholics and nonalcoholics. | Activity of monoamine oxidase enzyme, activity of adenylate cyclase enzyme. |



[1] If the values of **X** were not very different for objects in $\pi_1$ and $\pi_2$, there would be no problem; that is, the classes would be indistinguishable, and new objects could be assigned to either class indiscriminately.

# Predicting Group Membership: Objects (i.e. Consumers or Human Subjects) Separated into 2 Classes: Purchasers and Laggards Based on Values of the Variables



We see from item 5, for example, that objects (consumers) are to be separated into two labeled classes ("purchasers" and "laggards") on the basis of observed values of presumably relevant variables (education, income, and so forth). In the

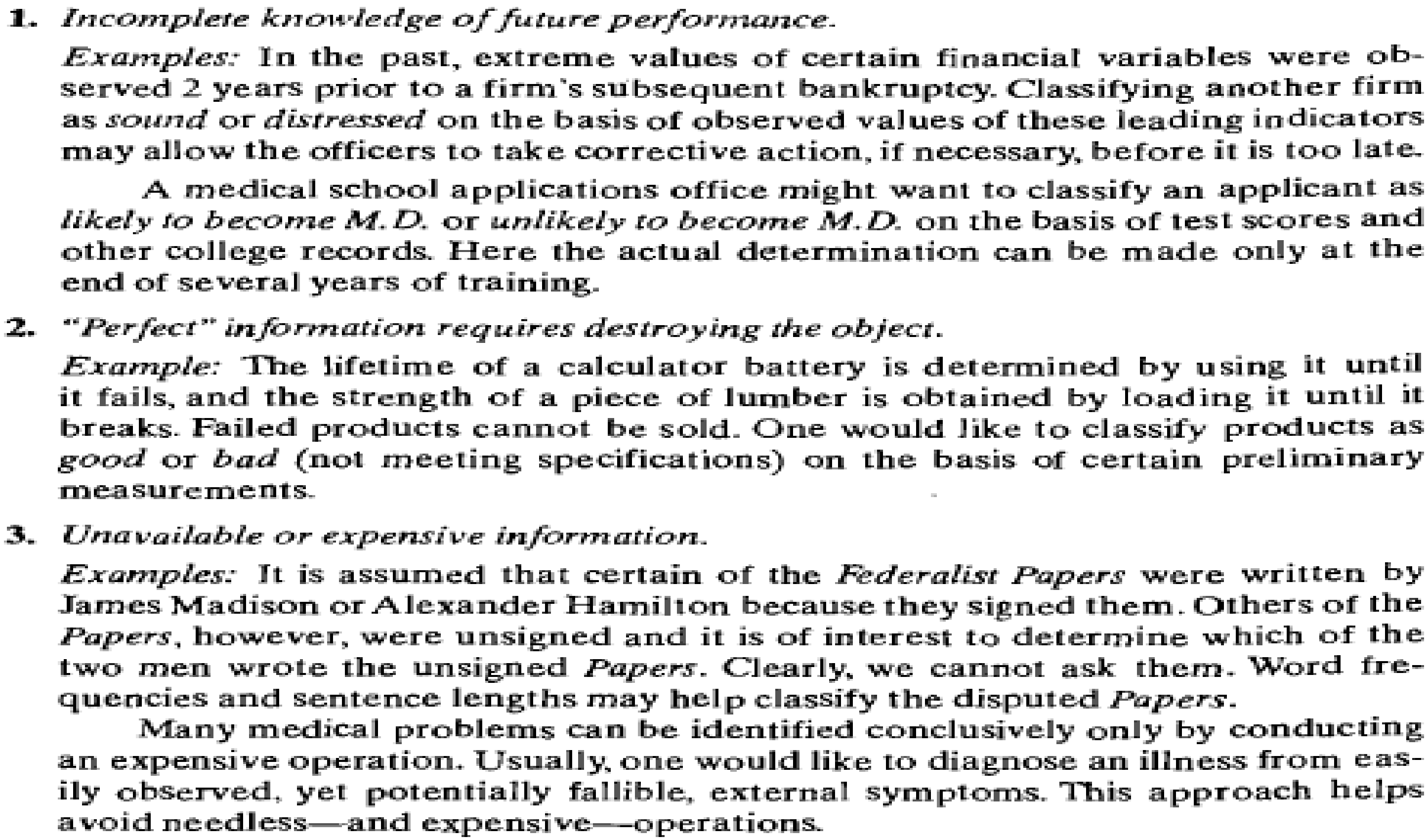
the form $\mathbf{x}' = [x_1(\text{education}), x_2(\text{income}), x_3(\text{family size}), x_4(\text{amount of brand switching})]$ as population $\pi_1$, purchasers, or population $\pi_2$, laggards.

At this point, we shall concentrate on classification for two populations, returning to separation in Section 11.3.

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 $\pi_1$, and if it falls in $R_2$, we allocate it to population $\pi_2$. Thus, one set of observed values favors $\pi_1$, while the other set of values favors $\pi_2$.

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. (This, of course, is what makes classification a problem!) Several conditions can give rise to this apparent anomaly (see [20]):

# Criterion for Classification or Discrimination by Case



1. *Incomplete knowledge of future performance.*

   *Examples:* In the past, extreme values of certain financial variables were observed 2 years prior to a firm's subsequent bankruptcy. Classifying another firm as *sound* or *distressed* on the basis of observed values of these leading indicators may allow the officers to take corrective action, if necessary, before it is too late.

   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.

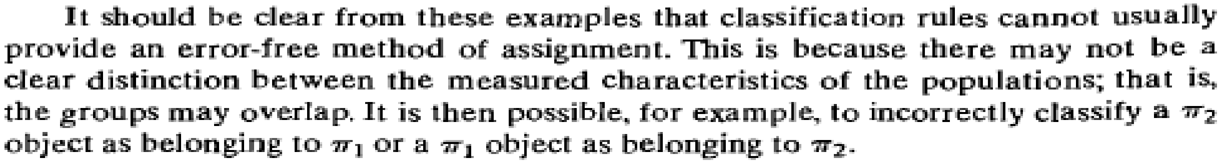
2. *"Perfect" information requires destroying the object.*

   *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.*

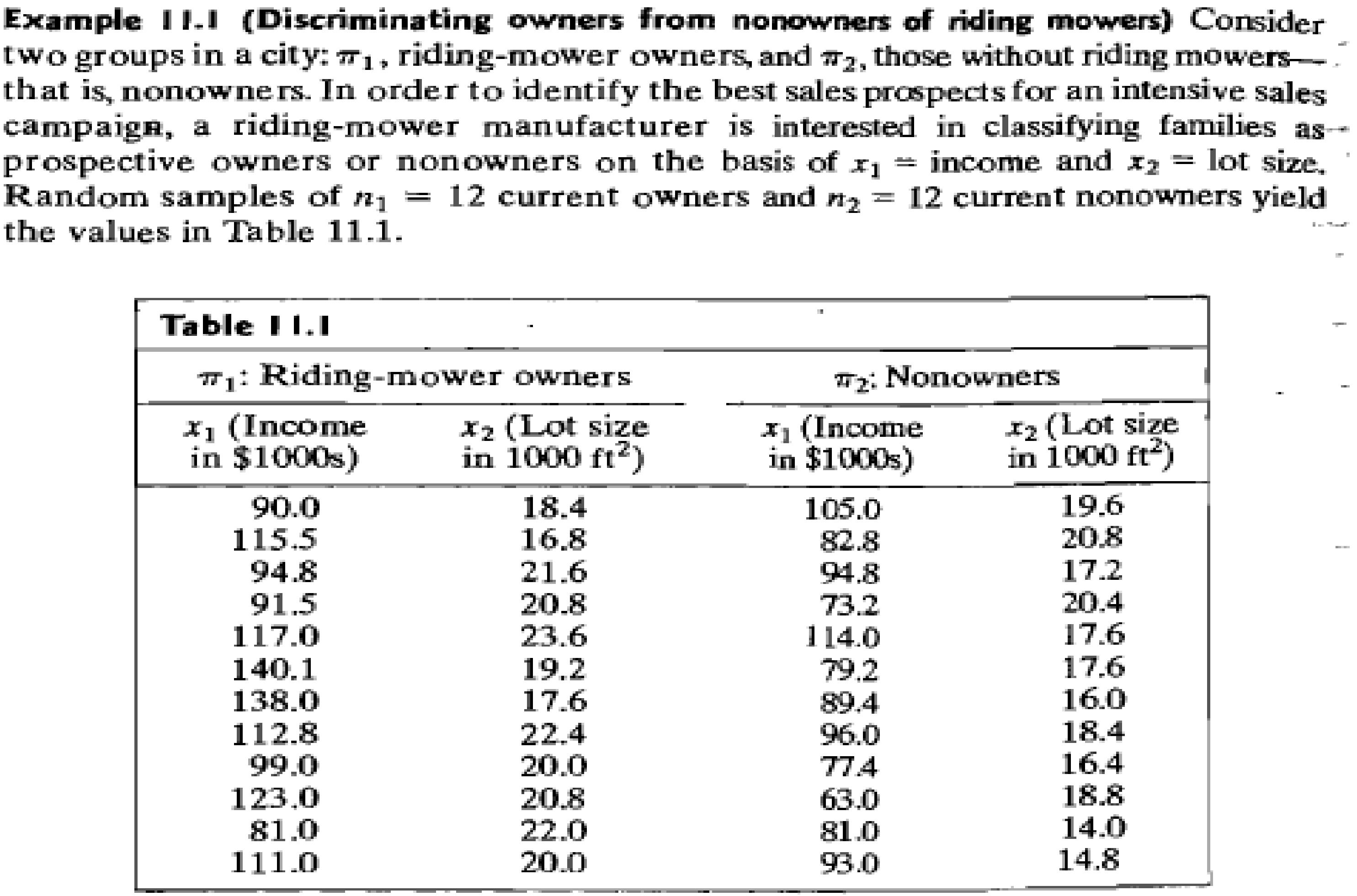
   *Examples:* It is assumed that certain of the *Federalist Papers* were written by James Madison or Alexander Hamilton because they signed them. Others of the *Papers*, however, were unsigned and it is of interest to determine which of the two men wrote the unsigned *Papers*. Clearly, we cannot ask them.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

   Many medical problems can be identified conclusively only by conducting an expensive operation. Usually, one would like to diagnose an illness from easily observed, yet potentially fallible, external symptoms. This approach helps avoid needless—and expensive—operations.



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 $\pi_2$ object as belonging to $\pi_1$ or a $\pi_1$ object as belonging to $\pi_2$.

# Example: Discrimination Owners from Non owners of Riding Mowers

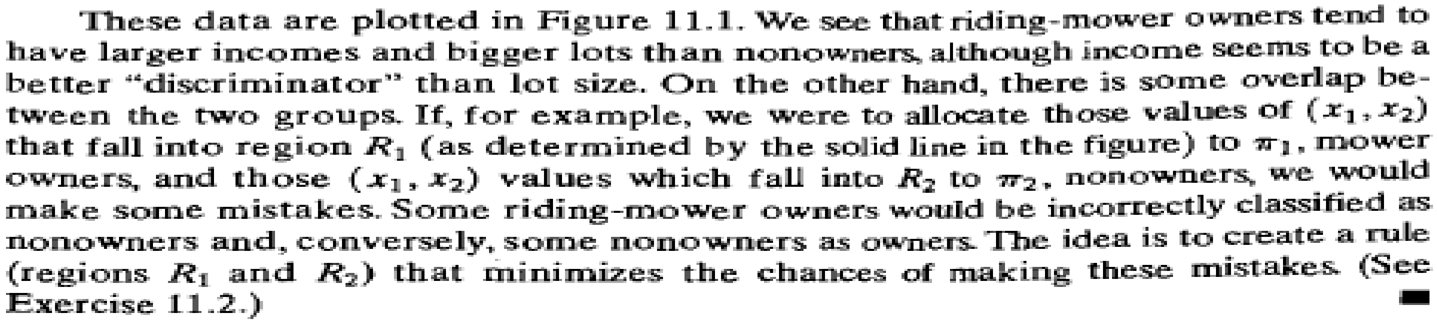


**Example 11.1 (Discriminating owners from nonowners of riding mowers)** Consider two groups in a city: $\pi_1$, riding-mower owners, and $\pi_2$, those without riding mowers—that is, nonowners. In order to identify the best sales prospects for an intensive sales campaign, a riding-mower manufacturer 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**

| $\pi_1$: Riding-mower owners | | $\pi_2$: 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$^2$) |
| 90.0 | 18.4 | 105.0 | 19.6 |
| 115.5 | 16.8 | 82.8 | 20.8 |
| 94.8 | 21.6 | 94.8 | 17.2 |
| 91.5 | 20.8 | 73.2 | 20.4 |
| 117.0 | 23.6 | 114.0 | 17.6 |
| 140.1 | 19.2 | 79.2 | 17.6 |
| 138.0 | 17.6 | 89.4 | 16.0 |
| 112.8 | 22.4 | 96.0 | 18.4 |
| 99.0 | 20.0 | 77.4 | 16.4 |
| 123.0 | 20.8 | 63.0 | 18.8 |
| 81.0 | 22.0 | 81.0 | 14.0 |
| 111.0 | 20.0 | 93.0 | 14.8 |

# Graphical Illustration: Distinguishing Owners from Non Owners Riding Mowers



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 $\pi_1$, mower owners, and those $(x_1, x_2)$ values which fall into $R_2$ to $\pi_2$, 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. (See Exercise 11.2.) ■

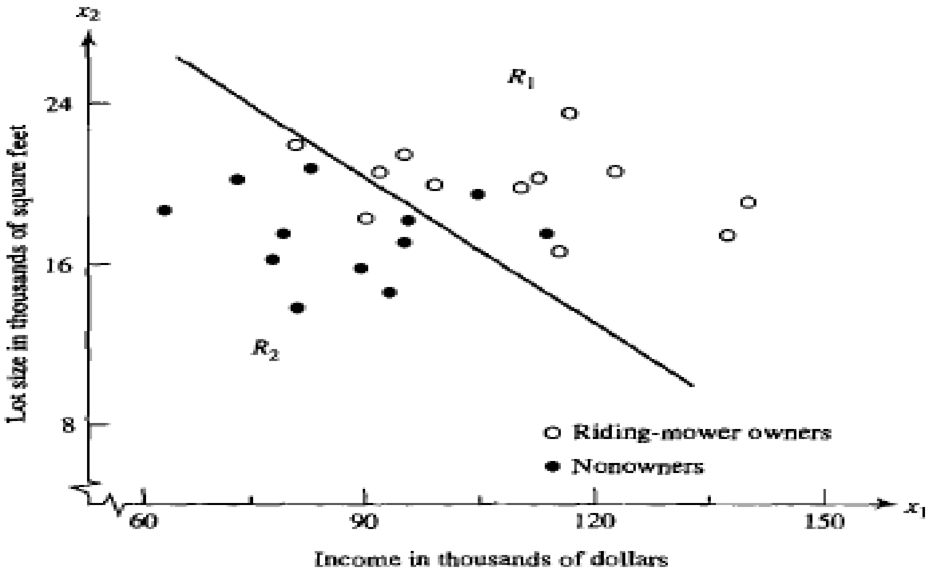
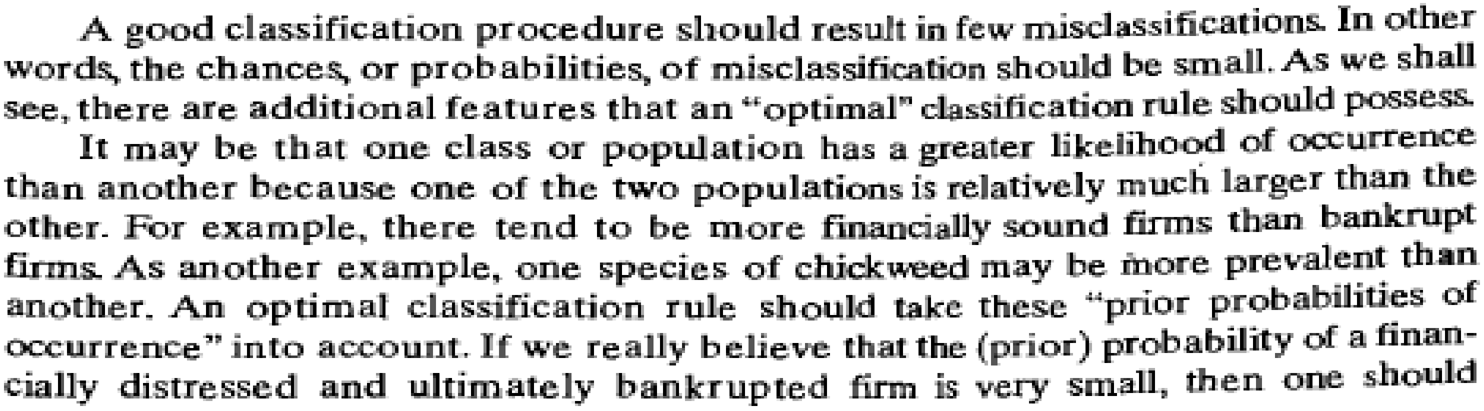


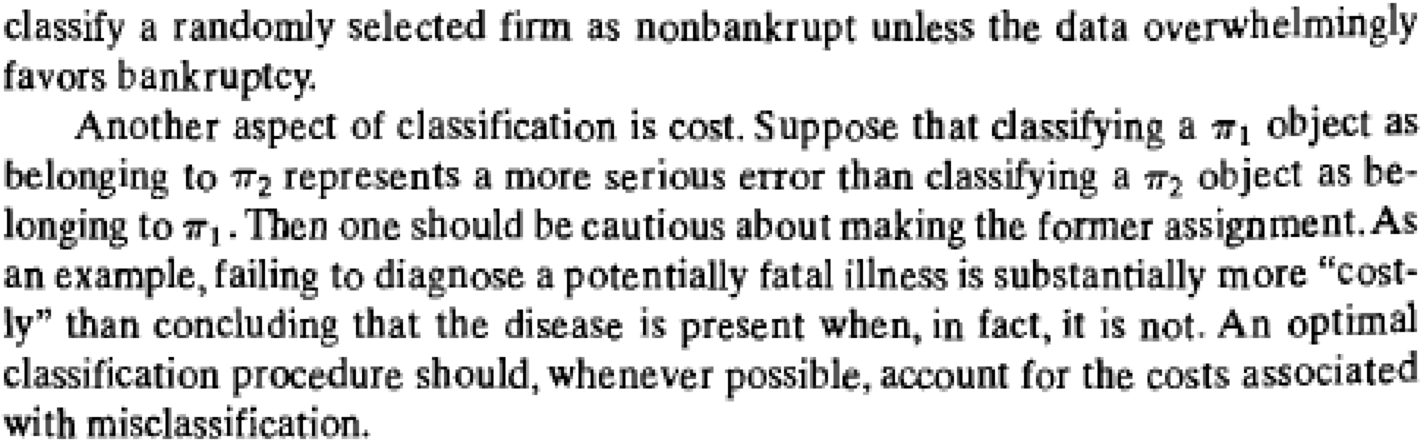
**Figure 11.1** Income and lot size for riding-mower owners and nonowners.

# Classification: Distribution of Cases or Objects Must Be Fairly Equal Sample Size or n



A good classification procedure should result in few misclassifications. In other words, the chances, or probabilities, of misclassification should be small. As we shall see, there are additional features that an "optimal" classification rule should possess.

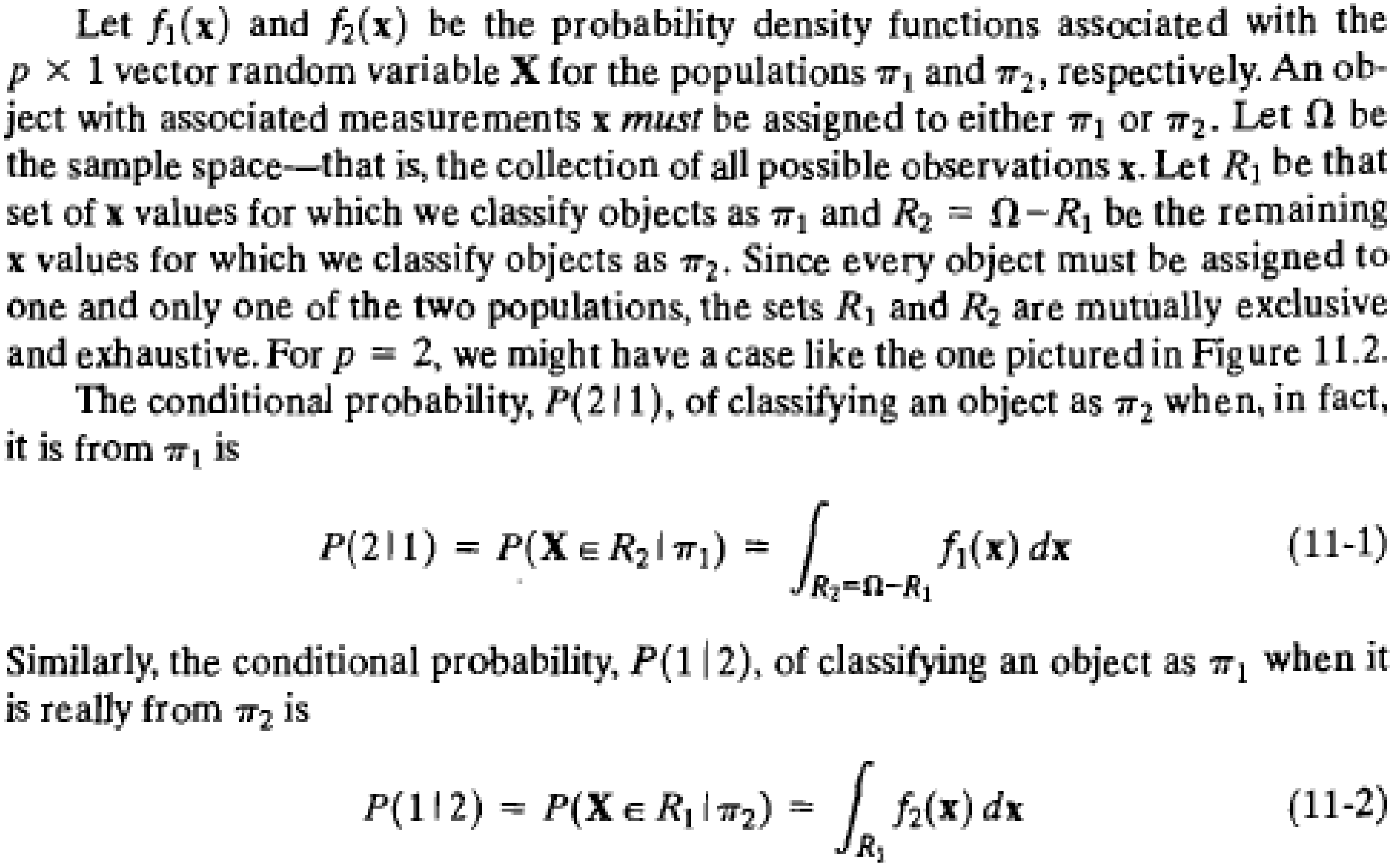
It may be that one class or population has a greater likelihood of occurrence than another because one of the two populations is relatively much larger than the other. For example, there tend to be more financially sound firms than bankrupt firms. As another example, one species of chickweed may be more prevalent than another. An optimal classification rule should take these "prior probabilities of occurrence" into account. If we really believe that the (prior) probability of a financially distressed and ultimately bankrupted firm is very small, then one should



classify a randomly selected firm as nonbankrupt unless the data overwhelmingly favors bankruptcy.

Another aspect of classification is cost. Suppose that classifying a $\pi_1$ object as belonging to $\pi_2$ represents a more serious error than classifying a $\pi_2$ object as belonging to $\pi_1$. 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.

# Underlying Normal Distribution Needed for Conditional Probability of Classifying an Object



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 $\pi_1$ and $\pi_2$, respectively. An object with associated measurements $\mathbf{x}$ *must* be assigned to either $\pi_1$ or $\pi_2$. Let $\Omega$ 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 $\pi_1$ and $R_2 = \Omega - R_1$ be the remaining $\mathbf{x}$ values for which we classify objects as $\pi_2$. 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, $P(2|1)$, of classifying an object as $\pi_2$ when, in fact, it is from $\pi_1$ is

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

Similarly, the conditional probability, $P(1|2)$, of classifying an object as $\pi_1$ when it is really from $\pi_2$ is

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

# Graph of Classification Regions for Two Populations

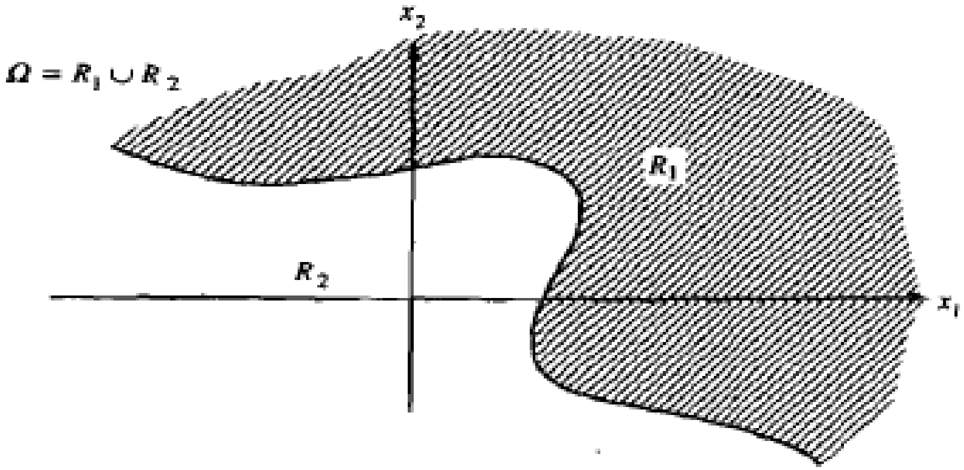
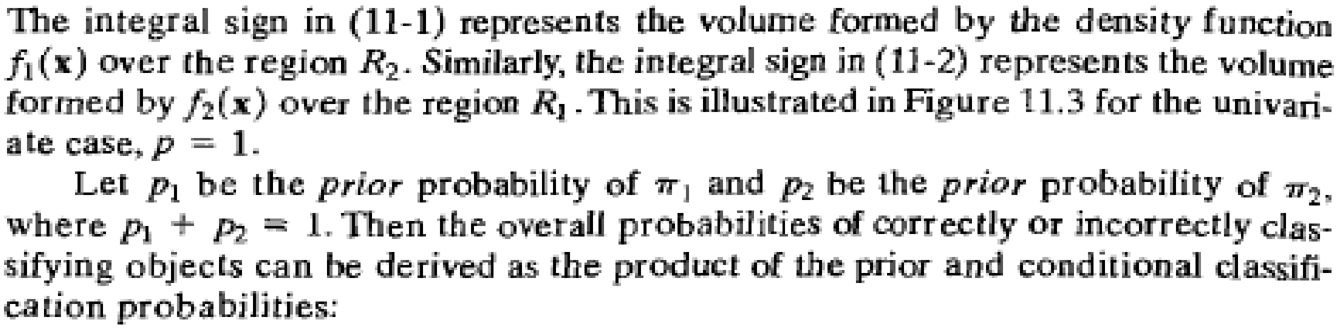


**Figure 11.2** Classification regions for two populations.



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 $\pi_1$ and $p_2$ be the *prior* probability of $\pi_2$, 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:

# Graph Misclassification Probabilities for Hypothetical Regions – i.e. Overlapping Distributions

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

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

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

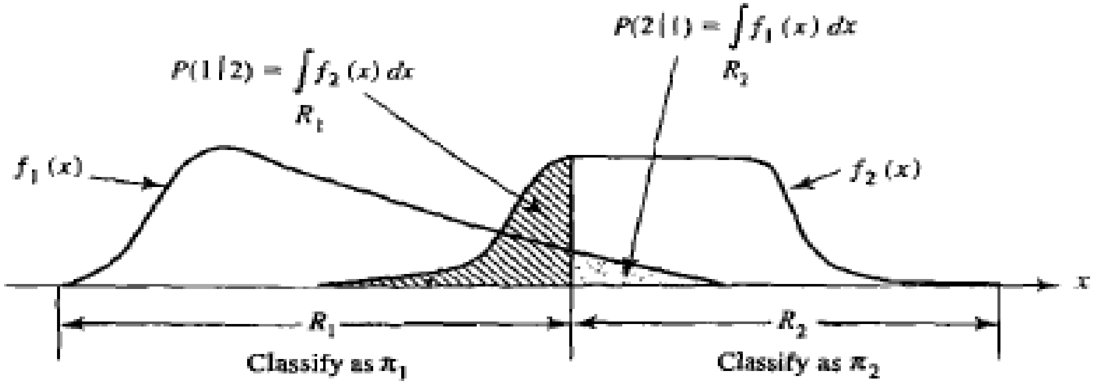
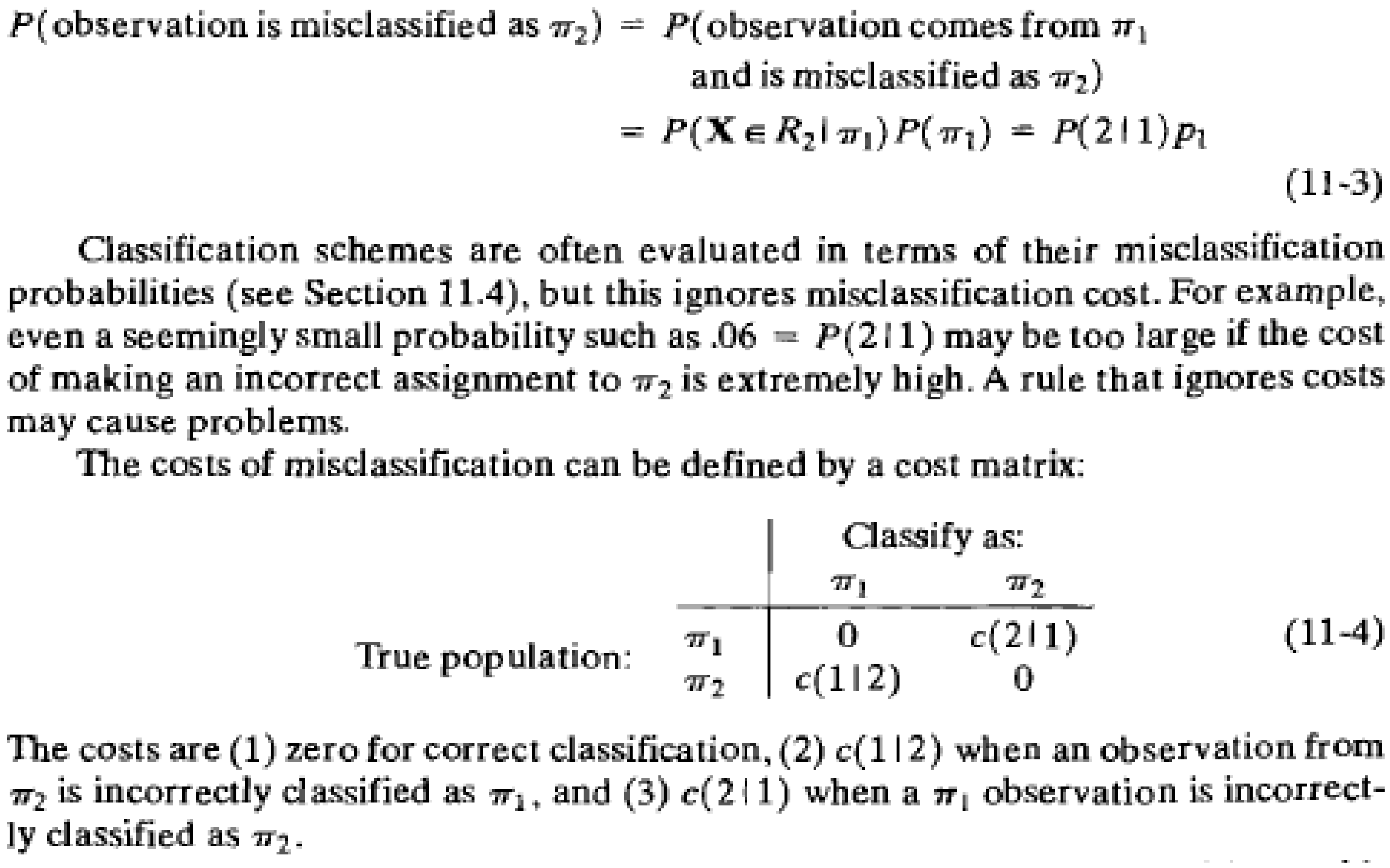


**Figure 11.3** Misclassification probabilities for hypothetical classification regions when $p = 1$.

# Classification Matrix as 2 x2 Table



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

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 $= P(2|1)$ may be too large if the cost of making an incorrect assignment to $\pi_2$ is extremely high. A rule that ignores costs may cause problems.

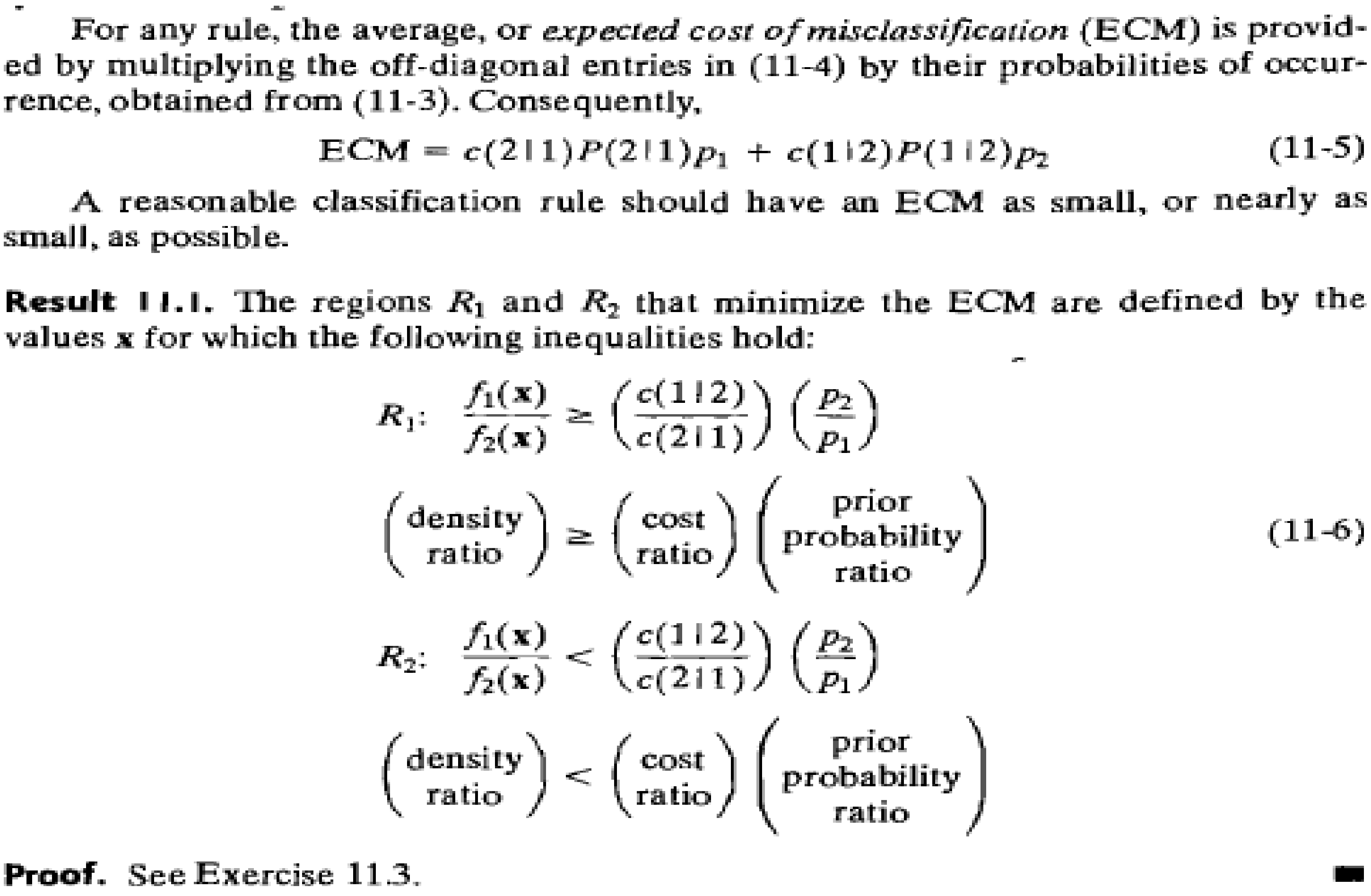
The costs of misclassification can be defined by a cost matrix:

| | | Classify as: | |
|---|---|---|---|
| | | $\pi_1$ | $\pi_2$ |
| True population: | $\pi_1$ | 0 | $c(2 \mid 1)$ |
| | $\pi_2$ | $c(1 \mid 2)$ | 0 |

(11-4)

The costs are (1) zero for correct classification, (2) $c(1|2)$ when an observation from $\pi_2$ is incorrectly classified as $\pi_1$, and (3) $c(2|1)$ when a $\pi_1$ observation is incorrectly classified as $\pi_2$.

# Expected Cost of Misclassification (ECM): Minimizing Overlap of Groups



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,

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

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

**Result 11.1.** The regions $R_1$ and $R_2$ that minimize the ECM are defined by the values $\mathbf{x}$ for which the following inequalities hold:

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

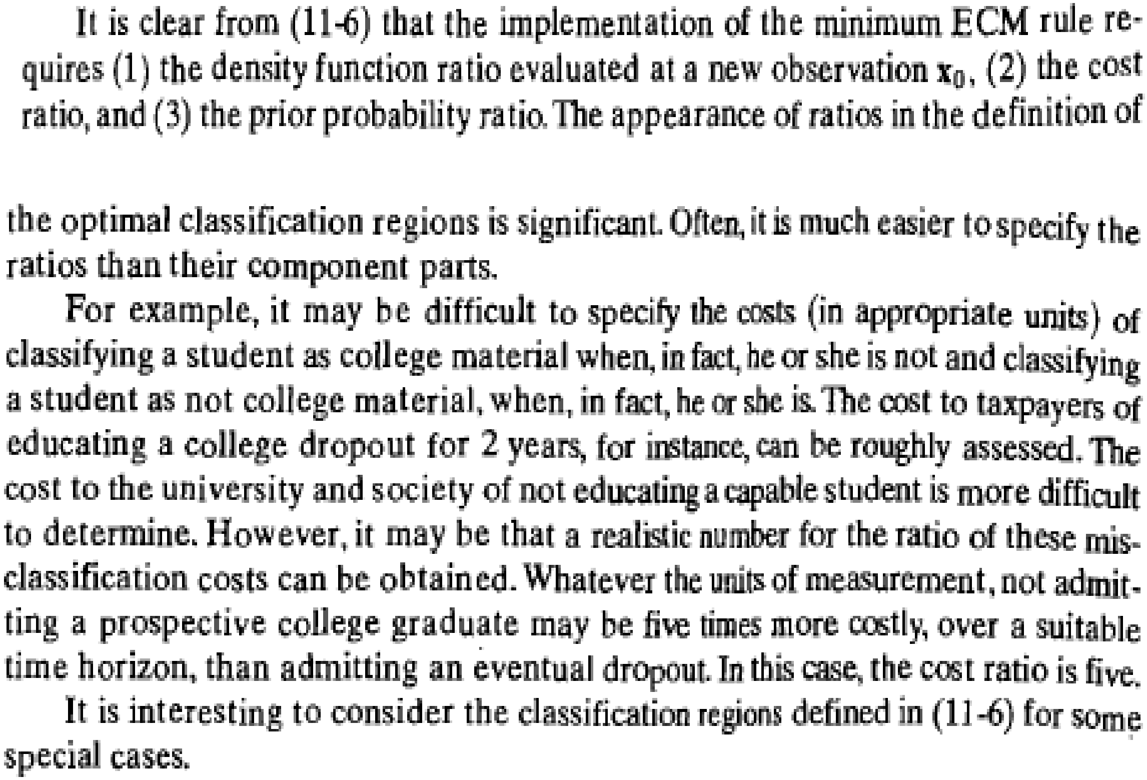
$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} \geq \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix} \tag{11-6}$$

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

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} < \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix}$$

**Proof.** See Exercise 11.3. ■

# Misclassification and Cost to the Funding Source (Private Corporation, Public Taxpayers, etc.)



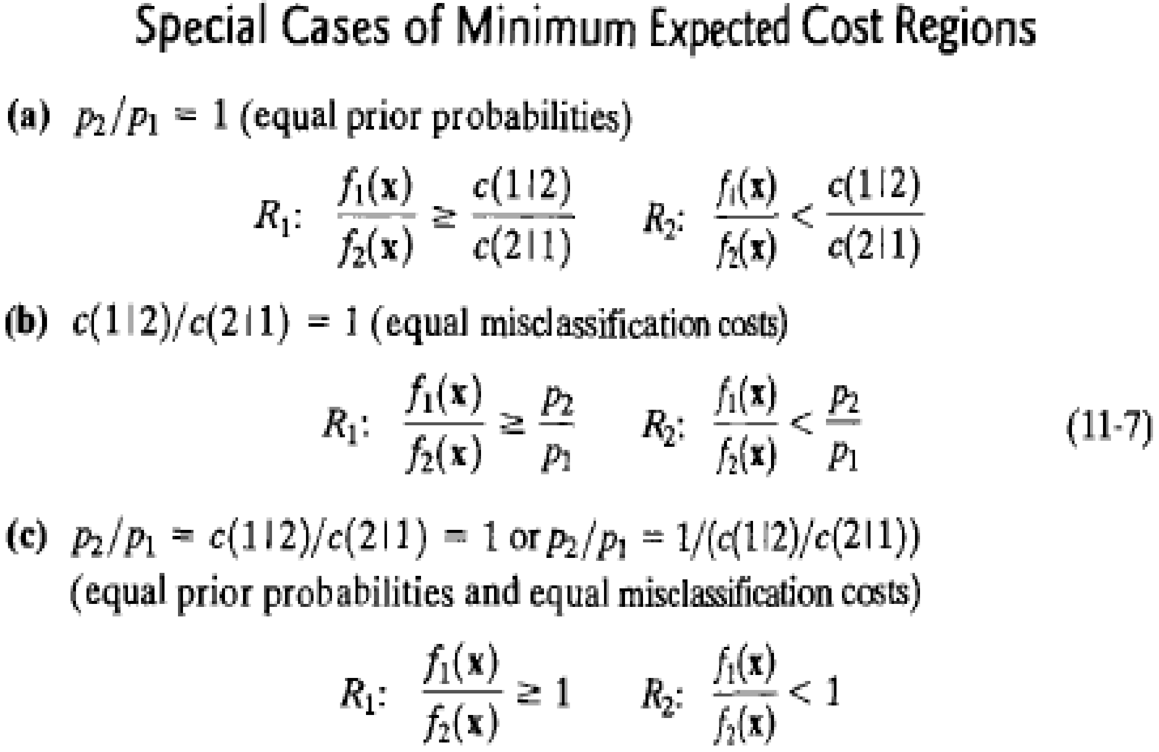
It is clear from (11-6) that the implementation of the minimum ECM rule requires (1) the density function ratio evaluated at a new observation $\mathbf{x}_0$, (2) the cost ratio, and (3) the prior probability ratio. The appearance of ratios in the definition of

the optimal classification regions is significant. Often, it is much easier to specify the ratios than their component parts.

For example, it may be difficult to specify the costs (in appropriate units) of classifying a student as college material when, in fact, he or she is not and classifying a student as not college material, when, in fact, he or she is. The cost to taxpayers of educating a college dropout for 2 years, for instance, can be roughly assessed. The cost to the university and society of not educating a capable student is more difficult to determine. However, it may be that a realistic number for the ratio of these misclassification costs can be obtained. Whatever the units of measurement, not admitting a prospective college graduate may be five times more costly, over a suitable time horizon, than admitting an eventual dropout. In this case, the cost ratio is five.

It is interesting to consider the classification regions defined in (11-6) for some special cases.

# Special Cases of Minimum Expected Cost Regions



Special Cases of Minimum Expected Cost Regions

**(a)** $p_2/p_1 = 1$ (equal prior probabilities)

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

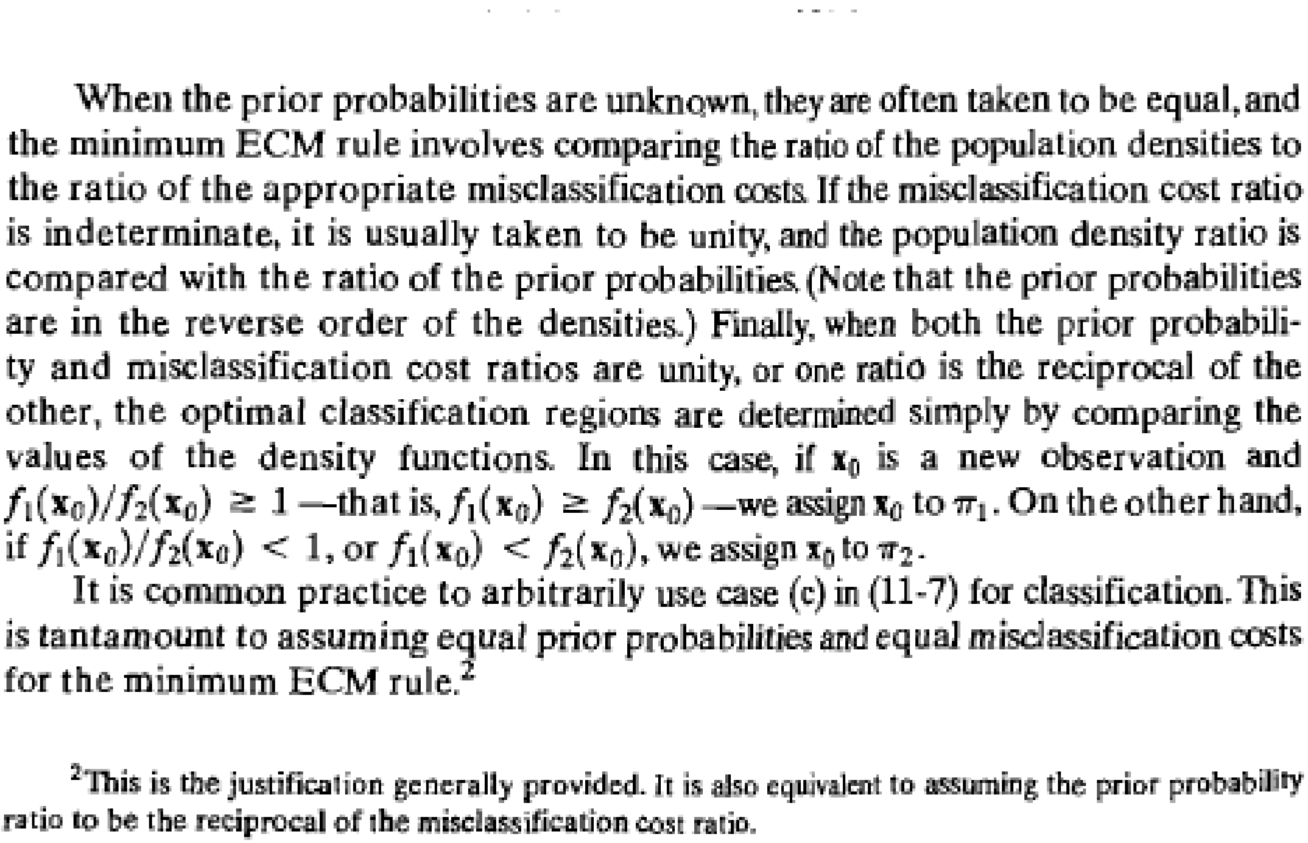
**(b)** $c(1|2)/c(2|1) = 1$ (equal misclassification costs)

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \qquad R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1} \tag{11-7}$$

**(c)** $p_2/p_1 = c(1|2)/c(2|1) = 1$ or $p_2/p_1 = 1/(c(1|2)/c(2|1))$
(equal prior probabilities and equal misclassification costs)

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \qquad R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

# When Both Prior Probabilities and Misclassification Cost Ratios Are Unity, Optimal Classification Regions Determined By Comparing Values of Density Functions

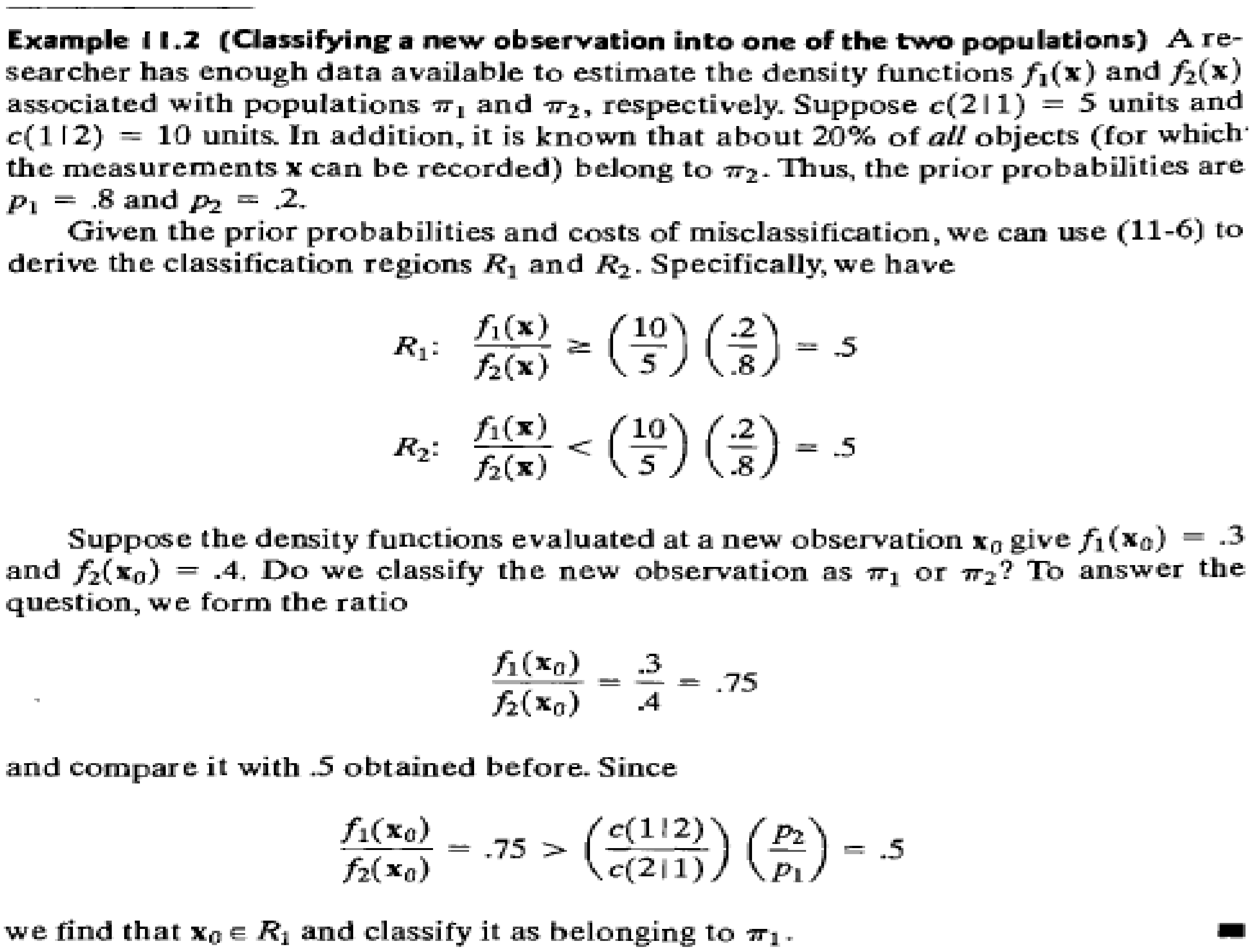


When the prior probabilities are unknown, they are often taken to be equal, and the minimum ECM rule involves comparing the ratio of the population densities to the ratio of the appropriate misclassification costs. If the misclassification cost ratio is indeterminate, it is usually taken to be unity, and the population density ratio is compared with the ratio of the prior probabilities. (Note that the prior probabilities are in the reverse order of the densities.) Finally, when both the prior probability and misclassification cost ratios are unity, or one ratio is the reciprocal of the other, the optimal classification regions are determined simply by comparing the values of the density functions. In this case, if $\mathbf{x}_0$ is a new observation and $f_1(\mathbf{x}_0)/f_2(\mathbf{x}_0) \geq 1$ —that is, $f_1(\mathbf{x}_0) \geq f_2(\mathbf{x}_0)$ —we assign $\mathbf{x}_0$ to $\pi_1$. On the other hand, if $f_1(\mathbf{x}_0)/f_2(\mathbf{x}_0) < 1$, or $f_1(\mathbf{x}_0) < f_2(\mathbf{x}_0)$, we assign $\mathbf{x}_0$ to $\pi_2$.

It is common practice to arbitrarily use case (c) in (11-7) for classification. This is tantamount to assuming equal prior probabilities and equal misclassification costs for the minimum ECM rule.[2]

[2] This is the justification generally provided. It is also equivalent to assuming the prior probability ratio to be the reciprocal of the misclassification cost ratio.

# Classifying New Observation into One of Two Populations



**Example 11.2 (Classifying a new observation into one of the two populations)** A researcher has enough data available to estimate the density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ associated with populations $\pi_1$ and $\pi_2$, respectively. Suppose $c(2|1) = 5$ units and $c(1|2) = 10$ units. In addition, it is known that about 20% of *all* objects (for which the measurements $\mathbf{x}$ can be recorded) belong to $\pi_2$. Thus, the prior probabilities are $p_1 = .8$ and $p_2 = .2$.

Given the prior probabilities and costs of misclassification, we can use (11-6) to derive the classification regions $R_1$ and $R_2$. Specifically, we have

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

Suppose the density functions evaluated at a new observation $\mathbf{x}_0$ give $f_1(\mathbf{x}_0) = .3$ and $f_2(\mathbf{x}_0) = .4$. Do we classify the new observation as $\pi_1$ or $\pi_2$? To answer the question, we form the ratio

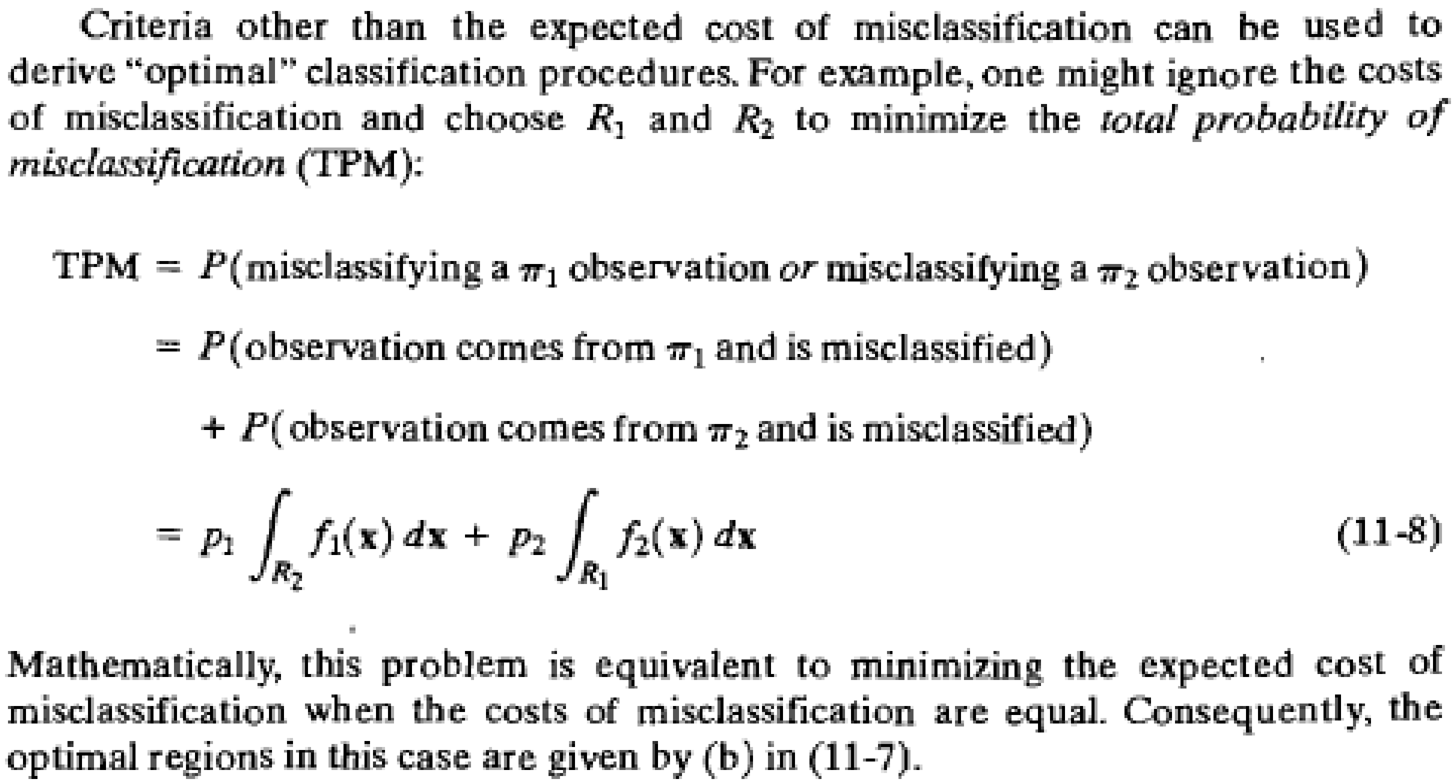
$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = \frac{.3}{.4} = .75$$

and compare it with .5 obtained before. Since

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = .75 > \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = .5$$

we find that $\mathbf{x}_0 \in R_1$ and classify it as belonging to $\pi_1$. ■

# Total Probability of Misclassification

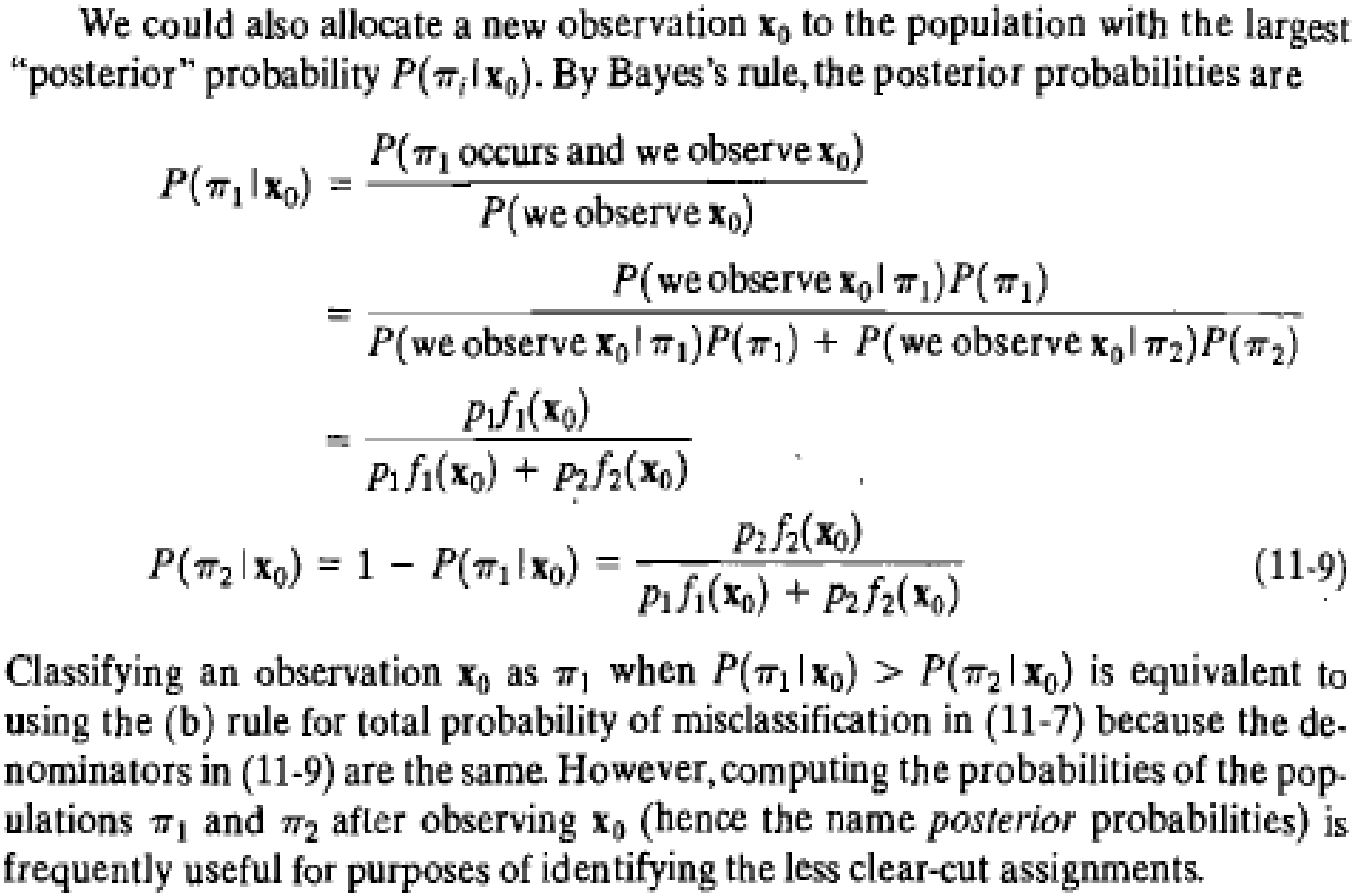


Criteria other than the expected cost of misclassification can be used to derive "optimal" classification procedures. For example, one might ignore the costs of misclassification and choose $R_1$ and $R_2$ to minimize the *total probability of misclassification* (TPM):

$$\begin{aligned}\text{TPM} &= P(\text{misclassifying a } \pi_1 \text{ observation } \textit{or} \text{ misclassifying a } \pi_2 \text{ observation})\\ &= P(\text{observation comes from } \pi_1 \text{ and is misclassified})\\ &\quad + P(\text{observation comes from } \pi_2 \text{ and is misclassified})\\ &= p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} \qquad (11\text{-}8)\end{aligned}$$

Mathematically, this problem is equivalent to minimizing the expected cost of misclassification when the costs of misclassification are equal. Consequently, the optimal regions in this case are given by (b) in (11-7).

# Appropriating an Object or Case to the Population with Largest Posterior Probability Use Bayes’ Rule



We could also allocate a new observation $\mathbf{x}_0$ to the population with the largest "posterior" probability $P(\pi_i | \mathbf{x}_0)$. By Bayes's rule, the posterior probabilities are

$$P(\pi_1 | \mathbf{x}_0) = \frac{P(\pi_1 \text{ occurs and we observe } \mathbf{x}_0)}{P(\text{we observe } \mathbf{x}_0)}$$

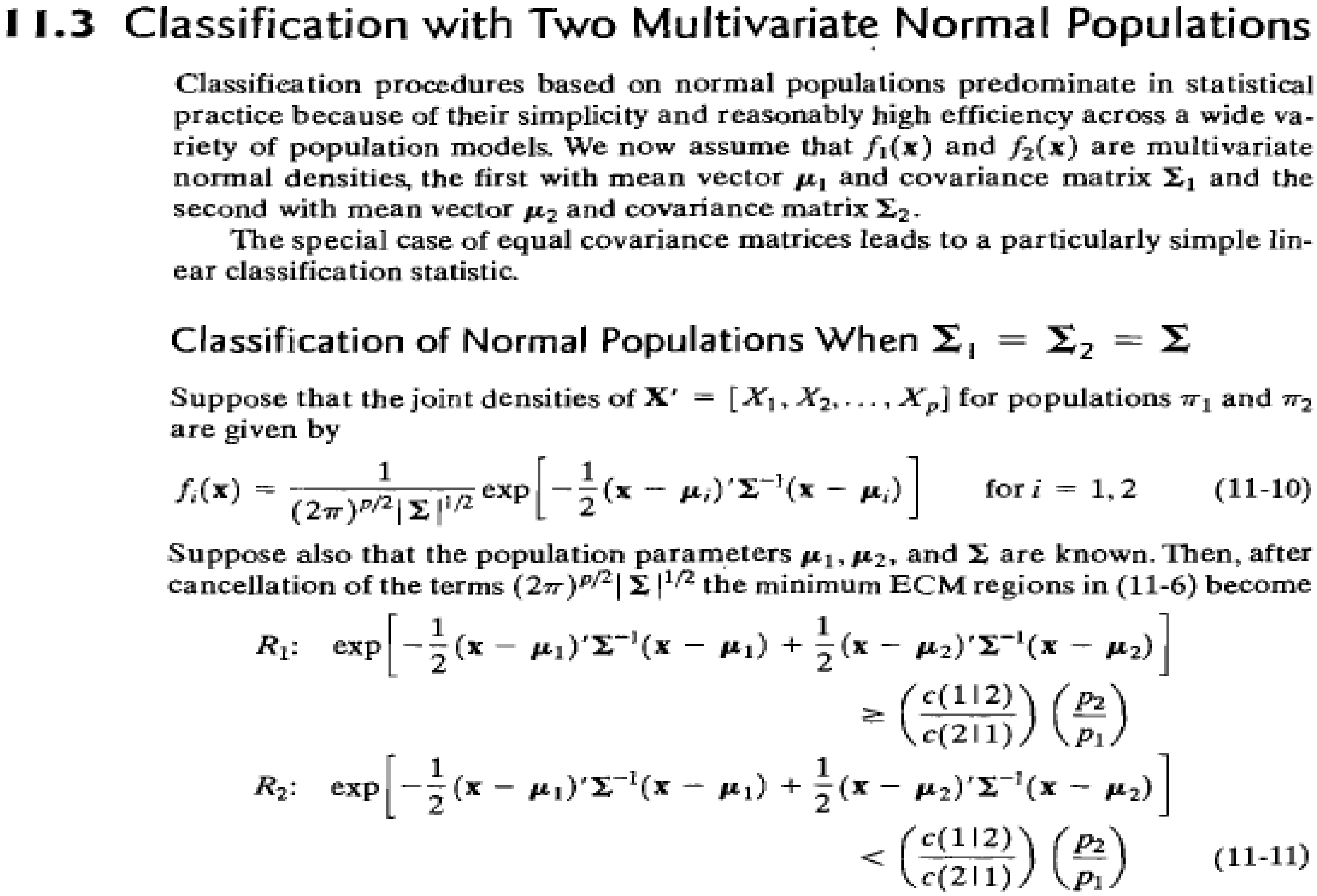
$$= \frac{P(\text{we observe } \mathbf{x}_0 | \pi_1) P(\pi_1)}{P(\text{we observe } \mathbf{x}_0 | \pi_1) P(\pi_1) + P(\text{we observe } \mathbf{x}_0 | \pi_2) P(\pi_2)}$$

$$= \frac{p_1 f_1(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}$$

$$P(\pi_2 | \mathbf{x}_0) = 1 - P(\pi_1 | \mathbf{x}_0) = \frac{p_2 f_2(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)} \tag{11-9}$$

Classifying an observation $\mathbf{x}_0$ as $\pi_1$ when $P(\pi_1 | \mathbf{x}_0) > P(\pi_2 | \mathbf{x}_0)$ is equivalent to using the (b) rule for total probability of misclassification in (11-7) because the denominators in (11-9) are the same. However, computing the probabilities of the populations $\pi_1$ and $\pi_2$ after observing $\mathbf{x}_0$ (hence the name *posterior* probabilities) is frequently useful for purposes of identifying the less clear-cut assignments.

# Classification with 2 Multivariate Normal Populations



## 11.3 Classification with Two Multivariate Normal Populations

Classification procedures based on normal populations predominate in statistical practice because of their simplicity and reasonably high efficiency across a wide variety of population models. We now assume that $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are multivariate normal densities, the first with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$ and the second with mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$.

The special case of equal covariance matrices leads to a particularly simple linear classification statistic.

### Classification of Normal Populations When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

Suppose that the joint densities of $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ for populations $\pi_1$ and $\pi_2$ are given by

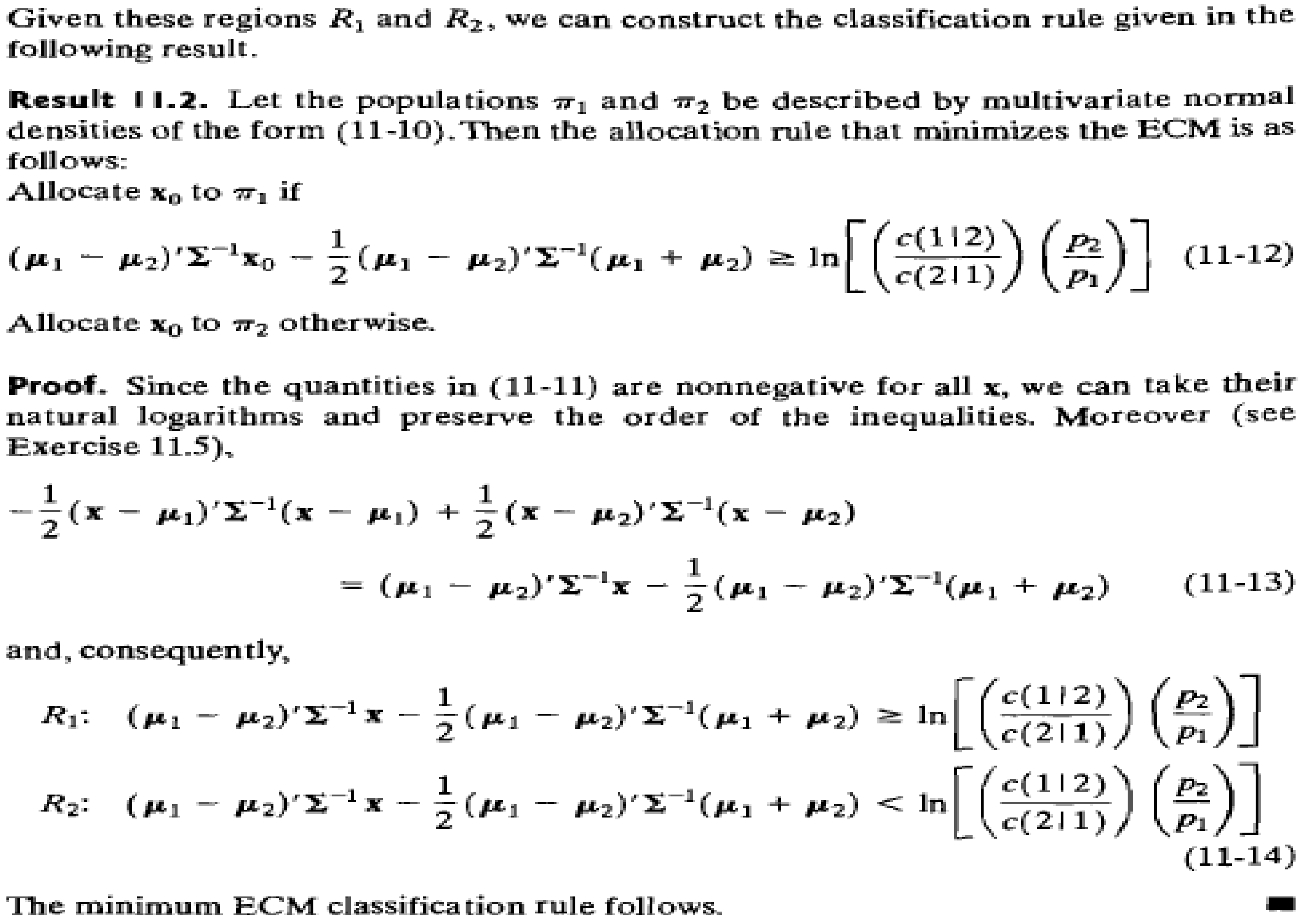
$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right] \qquad \text{for } i = 1, 2 \qquad (11\text{-}10)$$

Suppose also that the population parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are known. Then, after cancellation of the terms $(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}$ the minimum ECM regions in (11-6) become

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

$$R_2: \quad \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right] < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \qquad (11\text{-}11)$$

# Classification with 2 Multivariate Normal Populations: Equal ∑



Given these regions $R_1$ and $R_2$, we can construct the classification rule given in the following result.

**Result 11.2.** Let the populations $\pi_1$ and $\pi_2$ be described by multivariate normal densities of the form (11-10). Then the allocation rule that minimizes the ECM is as follows:

Allocate $\mathbf{x}_0$ to $\pi_1$ if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (11\text{-}12)$$

Allocate $\mathbf{x}_0$ to $\pi_2$ otherwise.

**Proof.** Since the quantities in (11-11) are nonnegative for all $\mathbf{x}$, we can take their natural logarithms and preserve the order of the inequalities. Moreover (see Exercise 11.5),

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

$$= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \qquad (11\text{-}13)$$

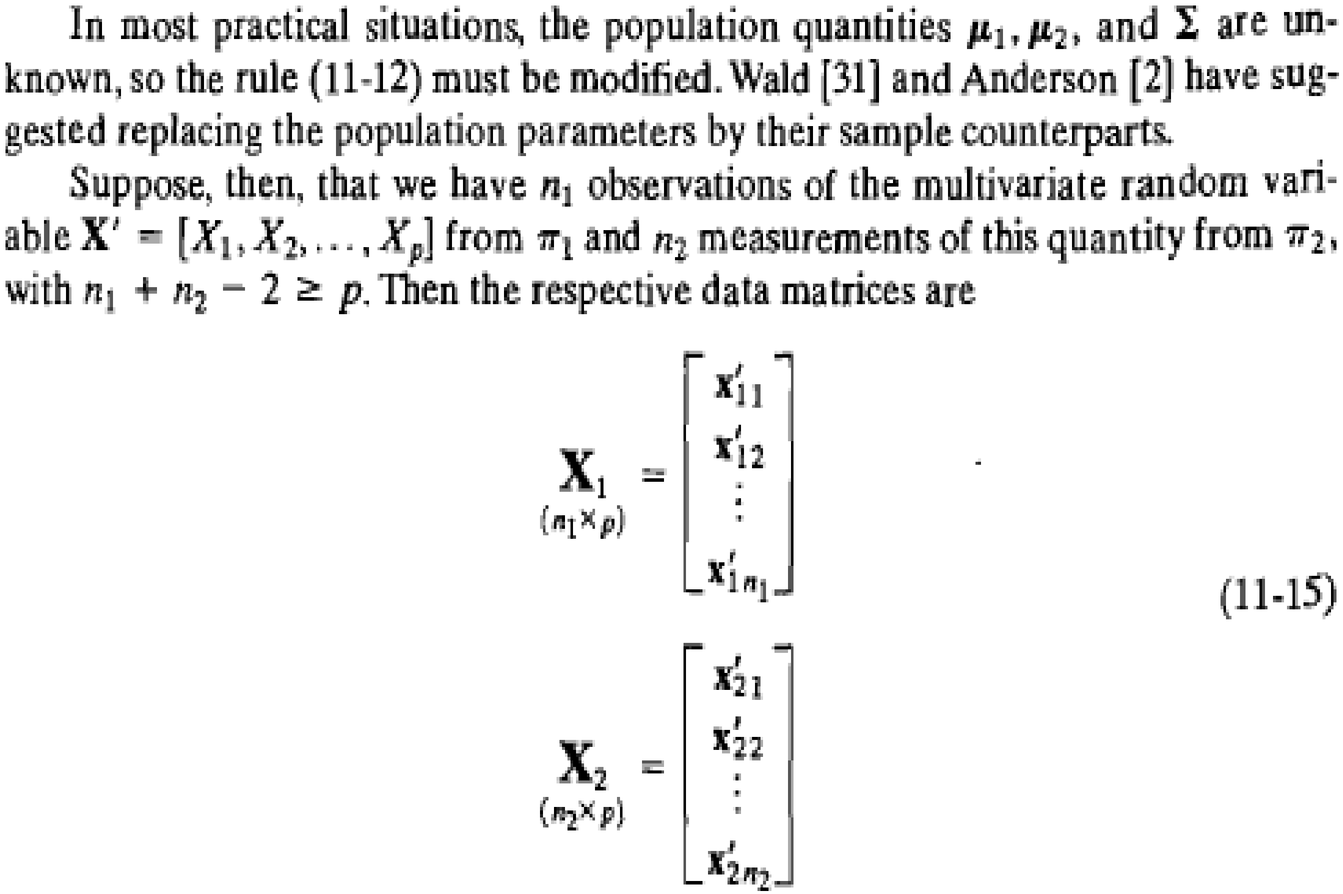
and, consequently,

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

$$R_2\text{:}\quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (11\text{-}14)$$

The minimum ECM classification rule follows. ■

# Matrices of the Two Populations Expressed as $X_1$ and $X_2$



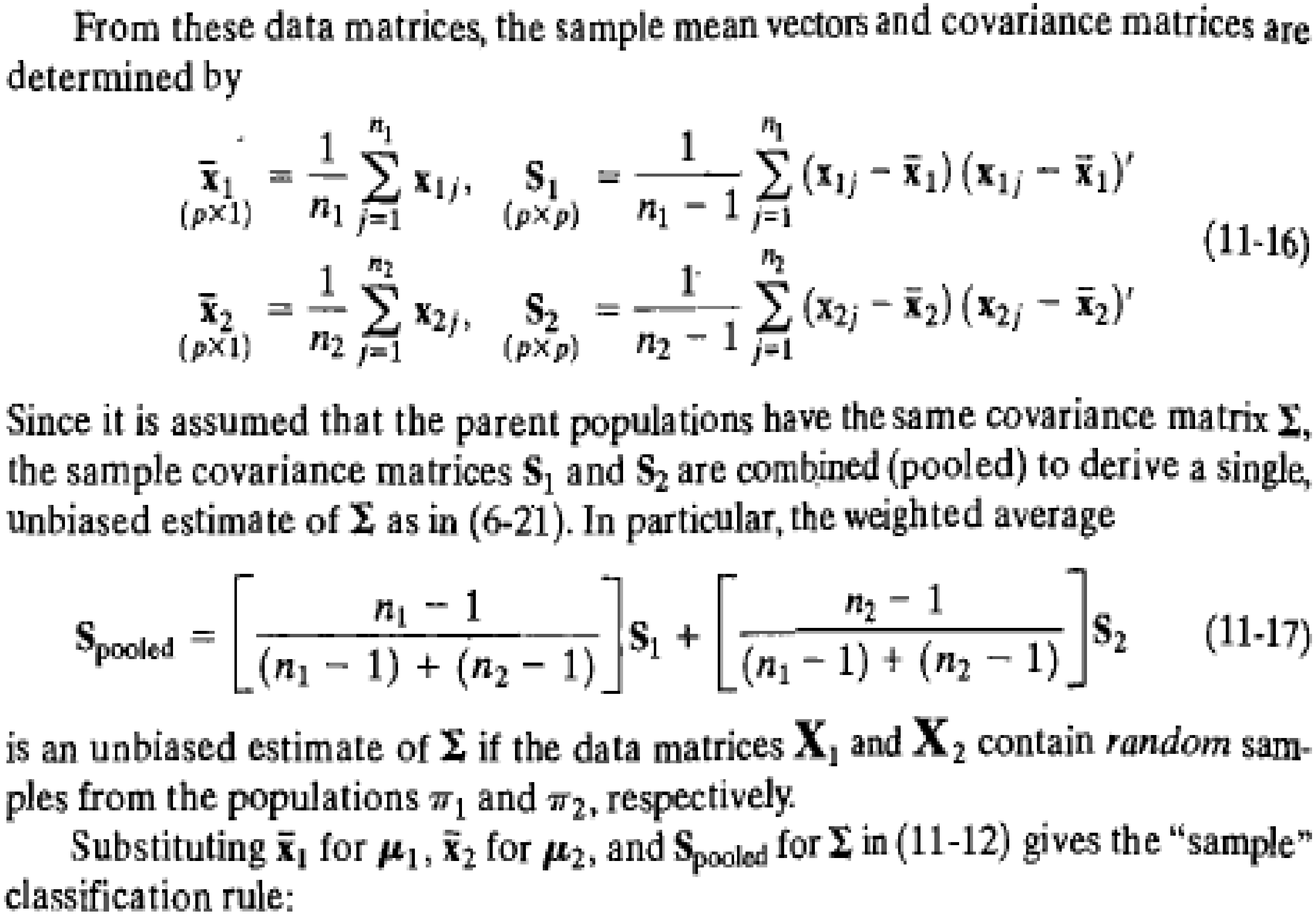
In most practical situations, the population quantities $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are unknown, so the rule (11-12) must be modified. Wald [31] and Anderson [2] have suggested replacing the population parameters by their sample counterparts.

Suppose, then, that we have $n_1$ observations of the multivariate random variable $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ from $\pi_1$ and $n_2$ measurements of this quantity from $\pi_2$, with $n_1 + n_2 - 2 \geq p$. Then the respective data matrices are

$$\underset{(n_1 \times p)}{\mathbf{X}_1} = \begin{bmatrix} \mathbf{x}'_{11} \\ \mathbf{x}'_{12} \\ \vdots \\ \mathbf{x}'_{1n_1} \end{bmatrix}$$

$$\underset{(n_2 \times p)}{\mathbf{X}_2} = \begin{bmatrix} \mathbf{x}'_{21} \\ \mathbf{x}'_{22} \\ \vdots \\ \mathbf{x}'_{2n_2} \end{bmatrix} \qquad (11\text{-}15)$$

# Sample Covariance Matrices S Pooled or Combined to Derive Single Unbiased Estimate ∑



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

$$\underset{(p\times 1)}{\bar{\mathbf{x}}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j}, \quad \underset{(p\times p)}{\mathbf{S}_1} = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)'$$

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

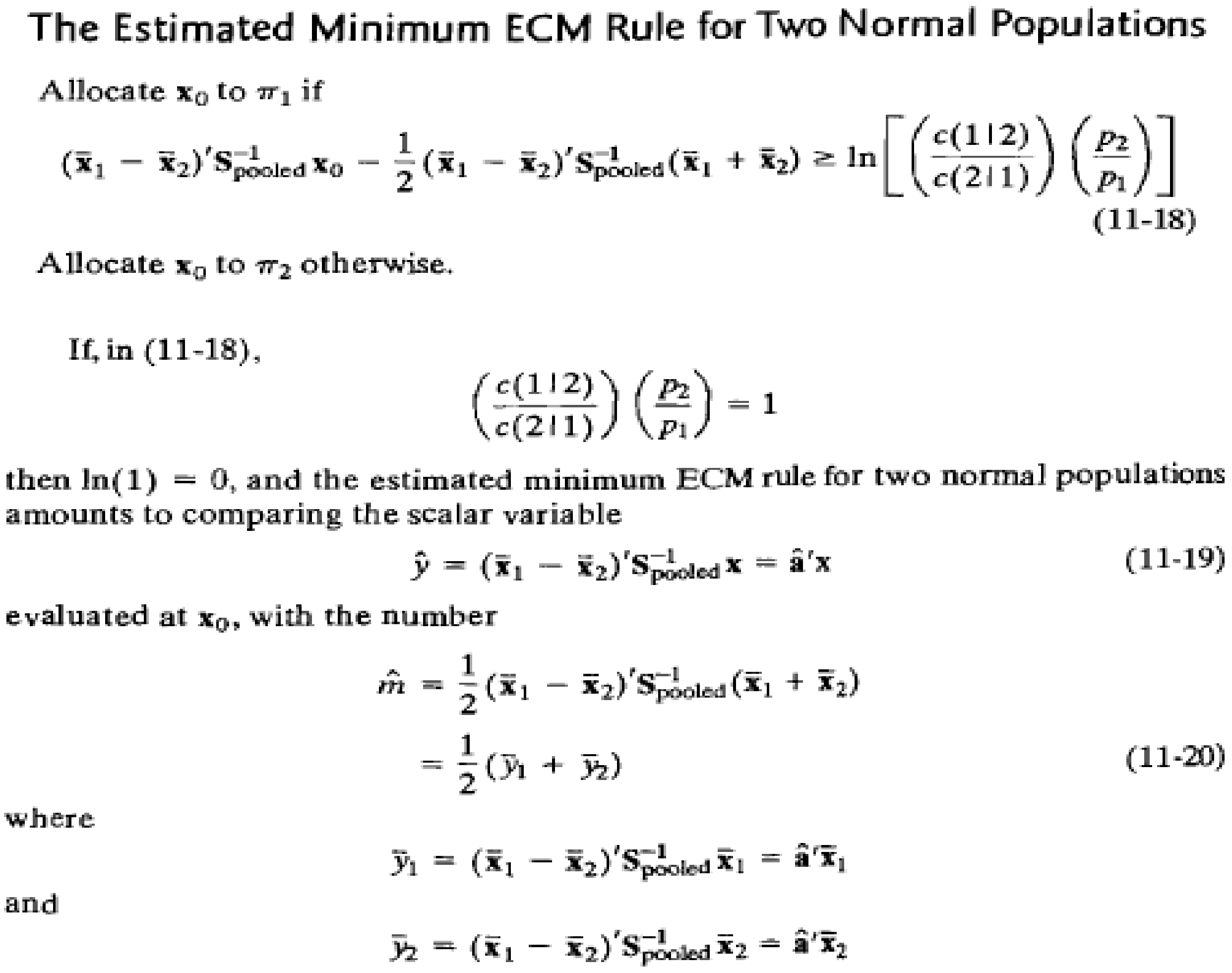
Since it is assumed that the parent populations have the same covariance matrix $\mathbf{\Sigma}$, the sample covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are combined (pooled) to derive a single, unbiased estimate of $\mathbf{\Sigma}$ as in (6-21). In particular, the weighted average

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

is an unbiased estimate of $\mathbf{\Sigma}$ if the data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ contain *random* samples from the populations $\pi_1$ and $\pi_2$, respectively.

Substituting $\bar{\mathbf{x}}_1$ for $\boldsymbol{\mu}_1$, $\bar{\mathbf{x}}_2$ for $\boldsymbol{\mu}_2$, and $\mathbf{S}_{\text{pooled}}$ for $\mathbf{\Sigma}$ in (11-12) gives the "sample" classification rule:

# Expected Cost of Misclassification (ECM) Rule for 2 Normal Populations



### The Estimated Minimum ECM Rule for Two Normal Populations

Allocate $\mathbf{x}_0$ to $\pi_1$ if

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

Allocate $\mathbf{x}_0$ to $\pi_2$ 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

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

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

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

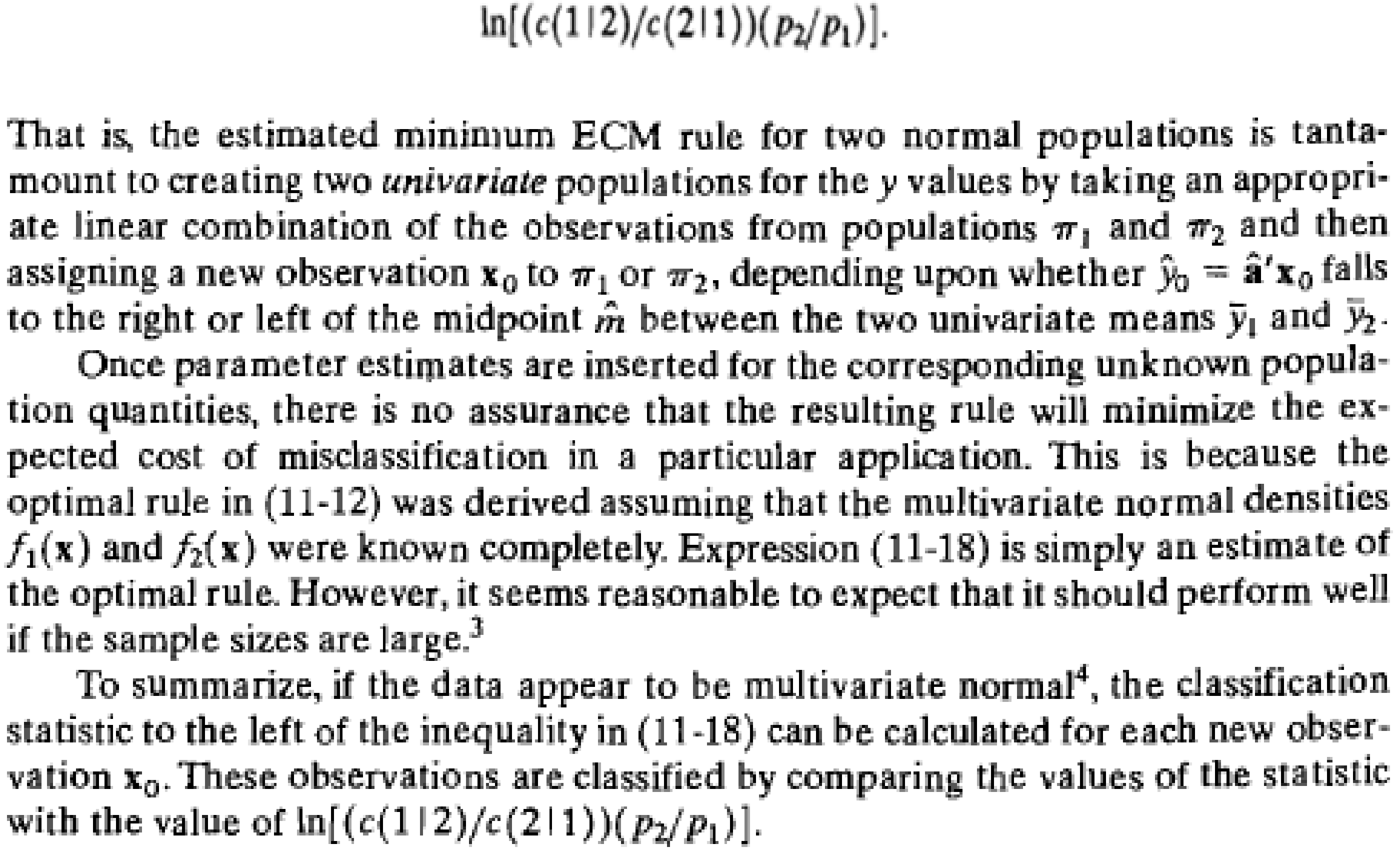
where

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

and

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

# Summary: Data Multivariate Normal, Classification Statistic Calculated for Each New Observation or Case, and then the Cases Classified by Comparing values of the Statistic With the Value of



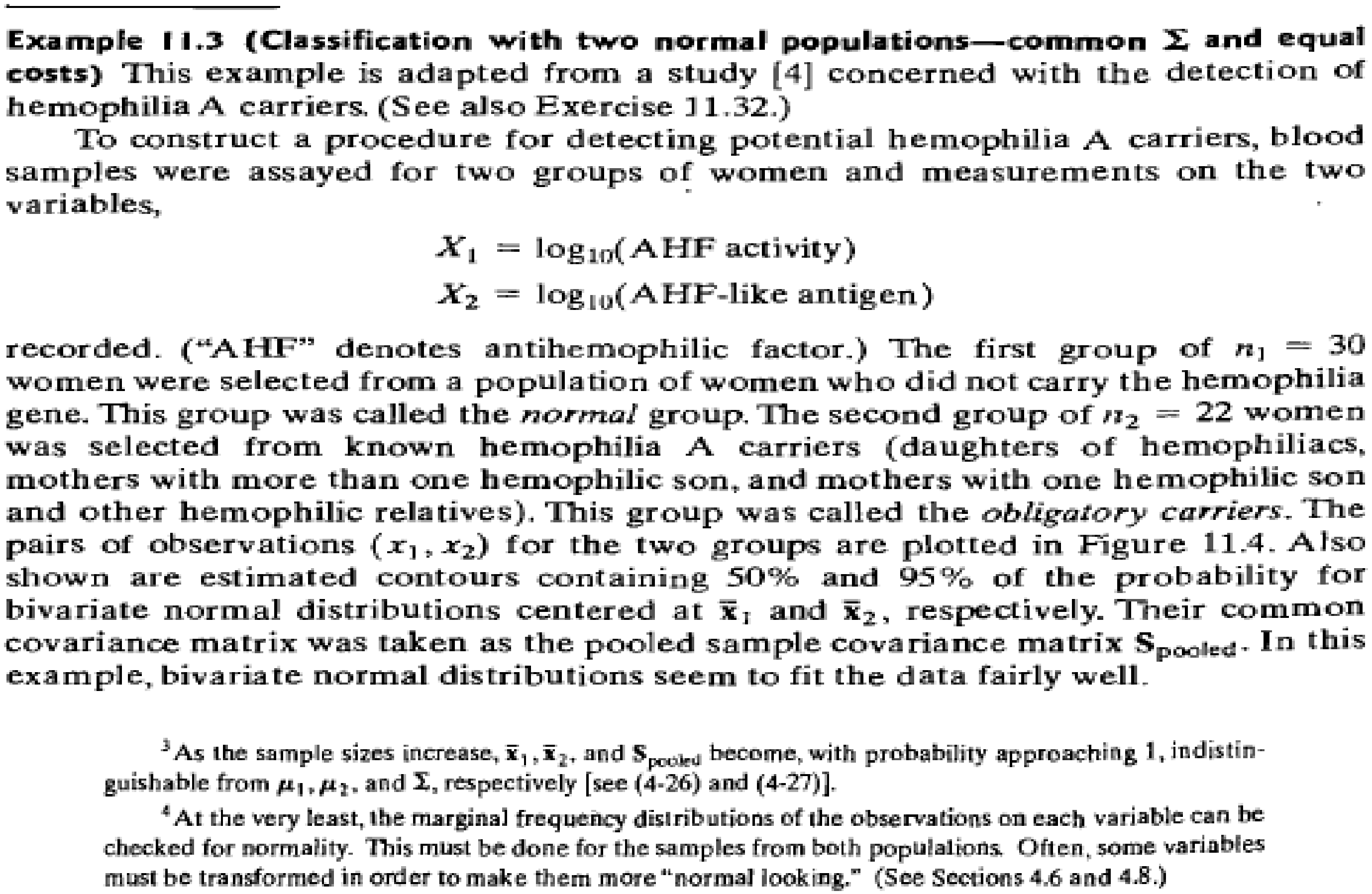
$$\ln[(c(1|2)/c(2|1))(p_2/p_1)].$$

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 $\pi_1$ and $\pi_2$ and then assigning a new observation $\mathbf{x}_0$ to $\pi_1$ or $\pi_2$, 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 $\bar{y}_2$.

Once parameter estimates are inserted for the corresponding unknown population quantities, there is no assurance that the resulting rule will minimize the expected cost of misclassification in a particular application. This is because the optimal rule in (11-12) was derived assuming that the multivariate normal densities $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ were known completely. Expression (11-18) is simply an estimate of the optimal rule. However, it seems reasonable to expect that it should perform well if the sample sizes are large.[3]

To summarize, if the data appear to be multivariate normal[4], the classification statistic to the left of the inequality in (11-18) can be calculated for each new observation $\mathbf{x}_0$. These observations are classified by comparing the values of the statistic with the value of $\ln[(c(1|2)/c(2|1))(p_2/p_1)]$.

# Classification with 2 Normal Populations – Common ∑ and Equal Costs



**Example 11.3 (Classification with two normal populations—common $\Sigma$ 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.

[3] As the sample sizes increase, $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2$, and $\mathbf{S}_{\text{pooled}}$ become, with probability approaching 1, indistinguishable from $\mu_1, \mu_2$, and $\Sigma$, respectively [see (4-26) and (4-27)].

[4] At the very least, the marginal frequency distributions of the observations on each variable can be checked for normality. This must be done for the samples from both populations. Often, some variables must be transformed in order to make them more "normal looking." (See Sections 4.6 and 4.8.)

# Anti-Hemophiliac or Normal vs. Obligatory Hemophilia A Carriers

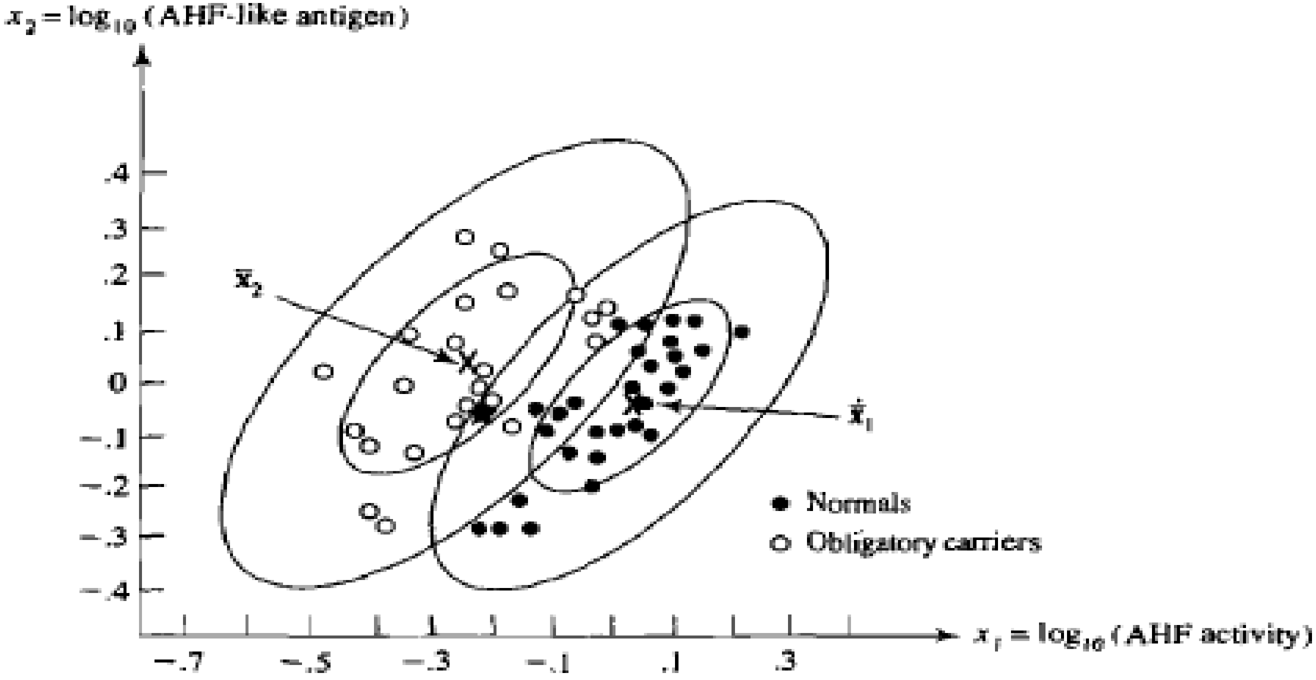
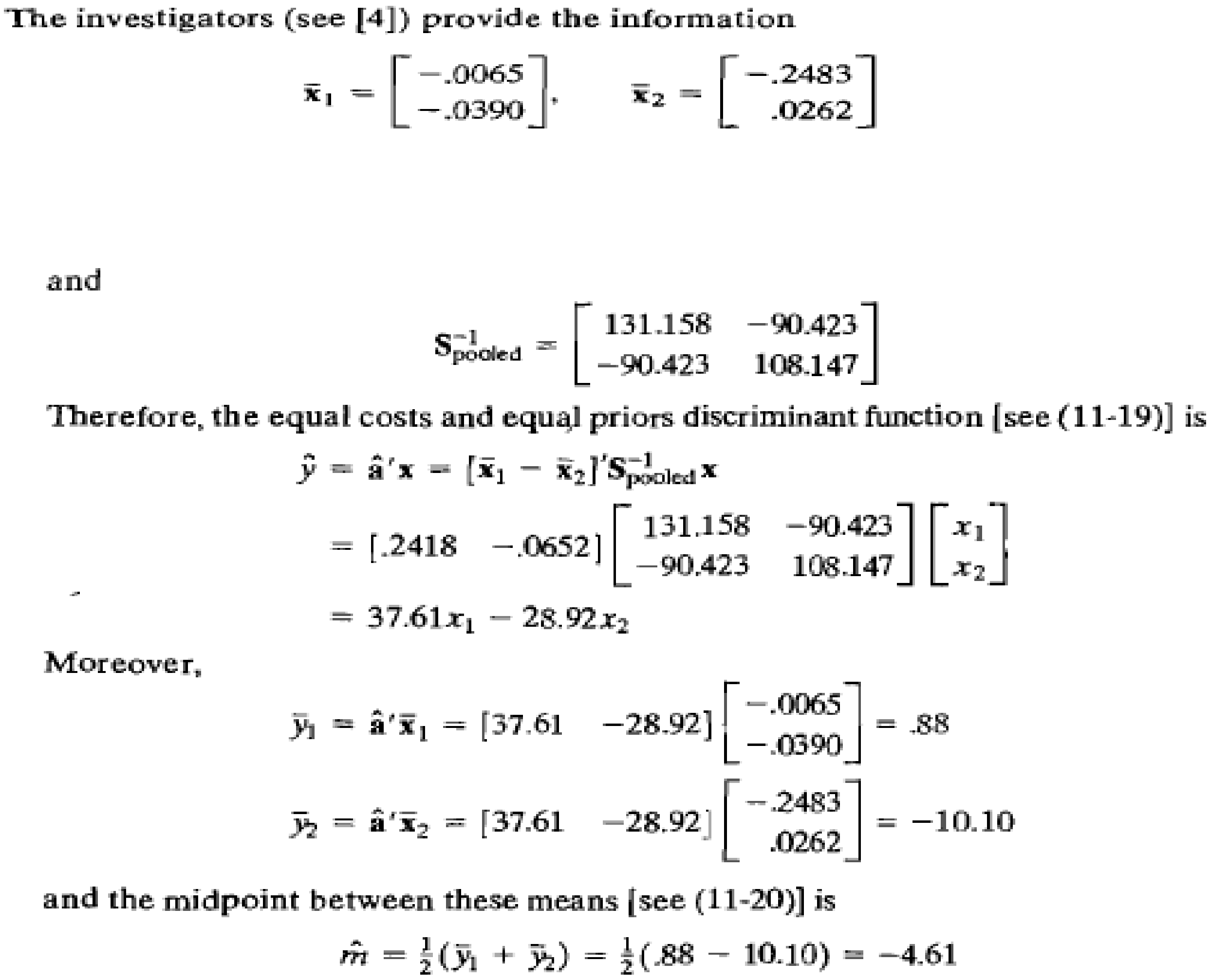


**Figure 11.4** Scatter plots of [$\log_{10}$(AHF activity), $\log_{10}$(AHF-like antigen)] for the normal group and obligatory hemophilia A carriers.

# Classification with 2 Normal Populations – Common ∑ and Equal Costs: Hemophiliac vs. Non-Hemophiliac



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,

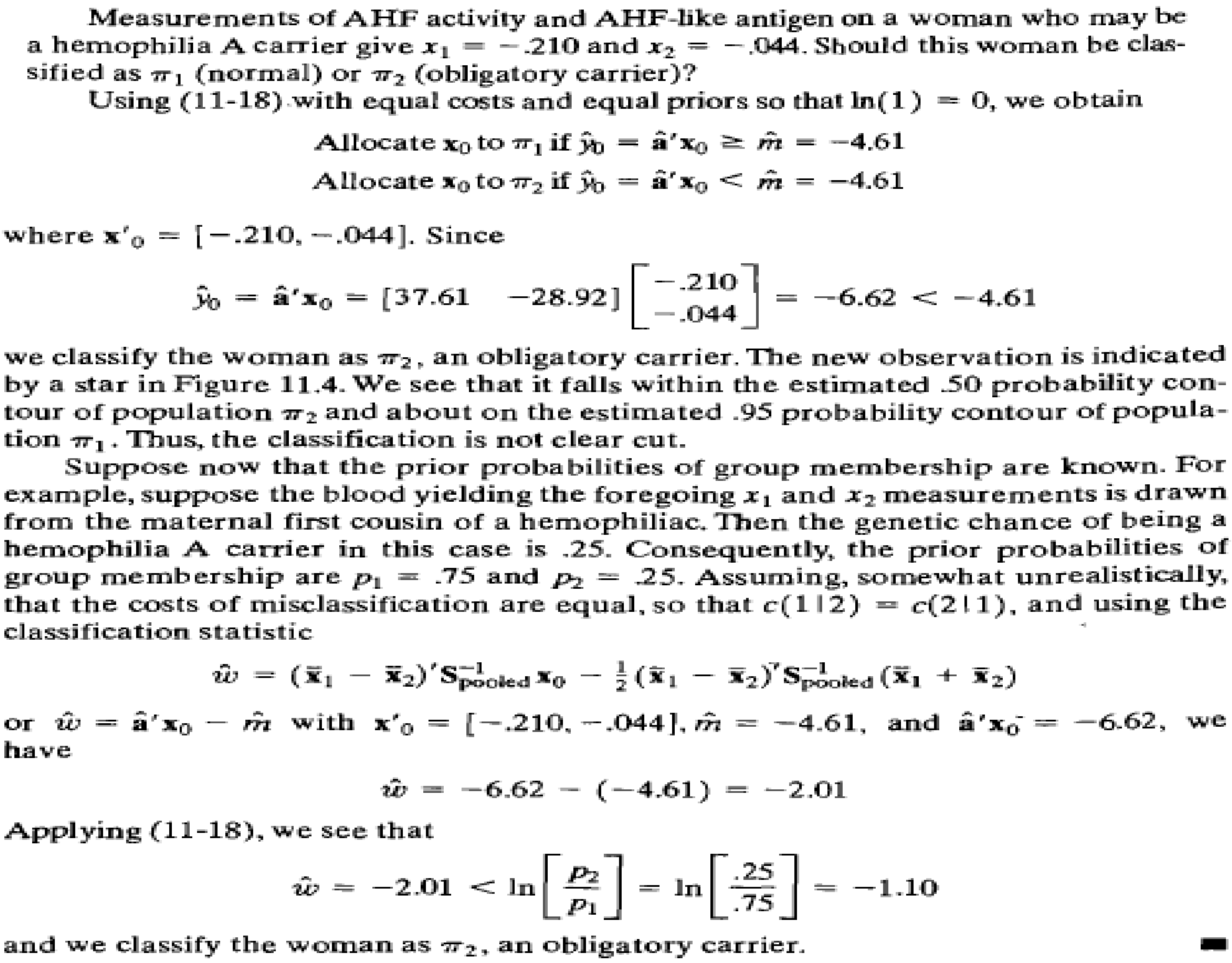
$$\bar{y}_1 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_1 = [37.61 \quad -28.92] \begin{bmatrix} -.0065 \\ -.0390 \end{bmatrix} = .88$$

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

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

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

# Classification of Specific Case



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 $\pi_1$ (normal) or $\pi_2$ (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 $\pi_2$, 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 $\pi_2$ and about on the estimated .95 probability contour of population $\pi_1$. Thus, the classification is not clear cut.

Suppose now that the prior probabilities of group membership are known. For example, suppose the blood yielding the foregoing $x_1$ and $x_2$ measurements is drawn from the maternal first cousin of a hemophiliac. Then the genetic chance of being a hemophilia A carrier in this case is .25. Consequently, the prior probabilities of group membership are $p_1 = .75$ and $p_2 = .25$. Assuming, somewhat unrealistically, that the costs of misclassification are equal, so that $c(1|2) = c(2|1)$, and using the classification statistic

$$\hat{w} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

or $\hat{w} = \hat{\mathbf{a}}'\mathbf{x}_0 - \hat{m}$ with $\mathbf{x}'_0 = [-.210, -.044]$, $\hat{m} = -4.61$, and $\hat{\mathbf{a}}'\mathbf{x}_0 = -6.62$, we have

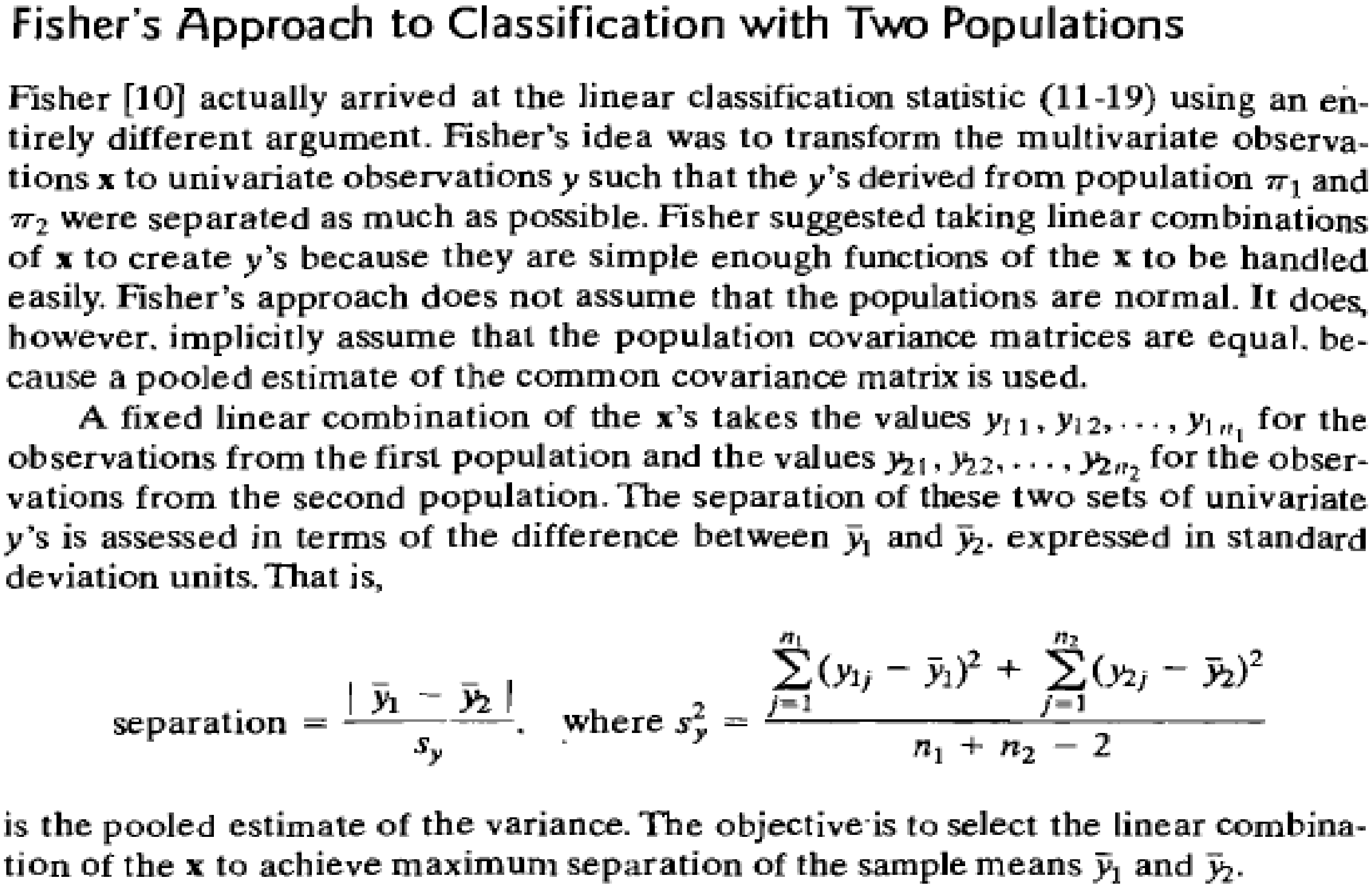
$$\hat{w} = -6.62 - (-4.61) = -2.01$$

Applying (11-18), we see that

$$\hat{w} = -2.01 < \ln\left[\frac{p_2}{p_1}\right] = \ln\left[\frac{.25}{.75}\right] = -1.10$$

and we classify the woman as $\pi_2$, an obligatory carrier. ■

# Fisher’s Approach to Classification With Two Populations



## Fisher's Approach to Classification with Two Populations

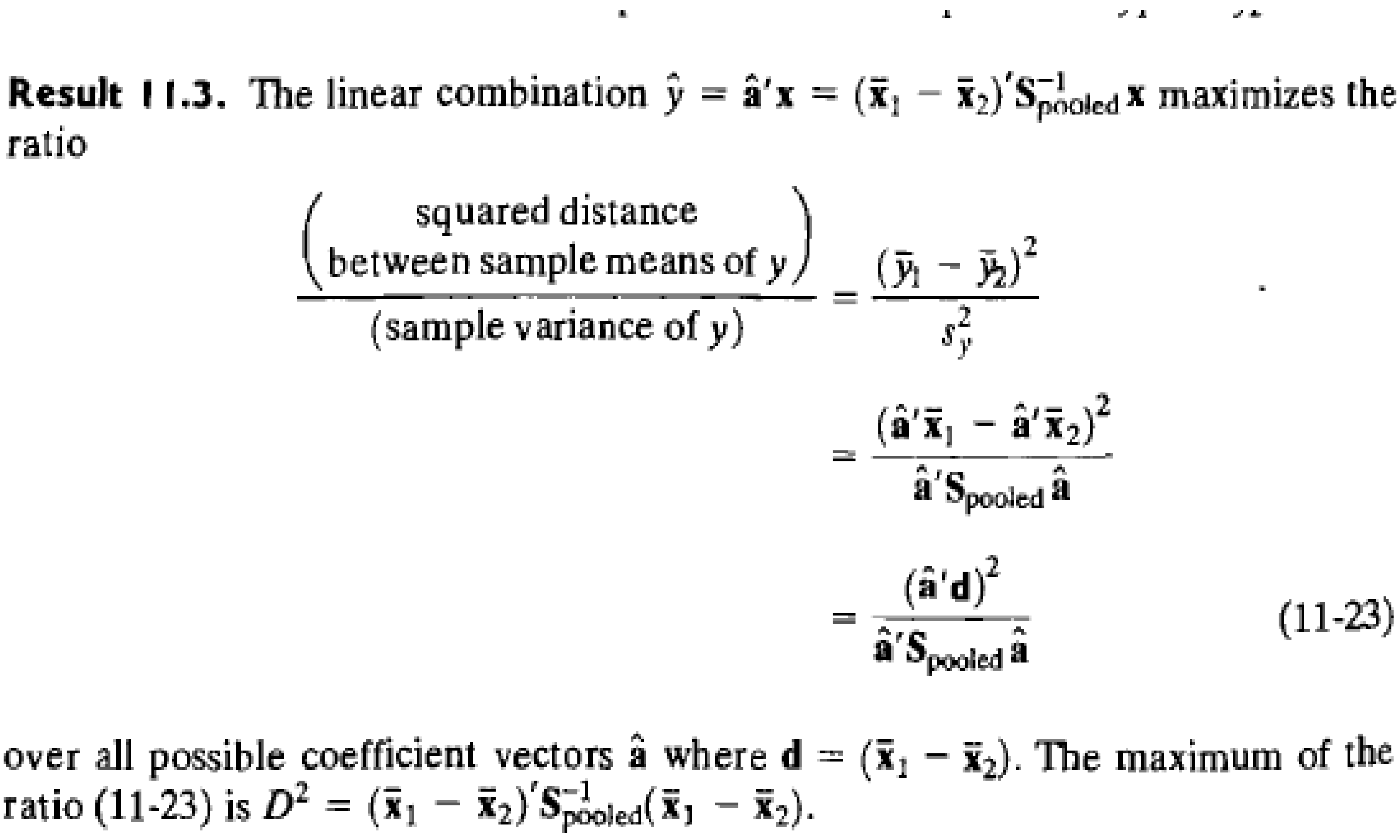
Fisher [10] actually arrived at the linear classification statistic (11-19) using an entirely different argument. Fisher's idea was to transform the multivariate observations $\mathbf{x}$ to univariate observations $y$ such that the $y$'s derived from population $\pi_1$ and $\pi_2$ were separated as 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,

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

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 $\bar{y}_2$.

# Linear Combination Maximizes the Ratio Squared Distance Between Sample Means of y Divided by Sample Variance y



**Result 11.3.** 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

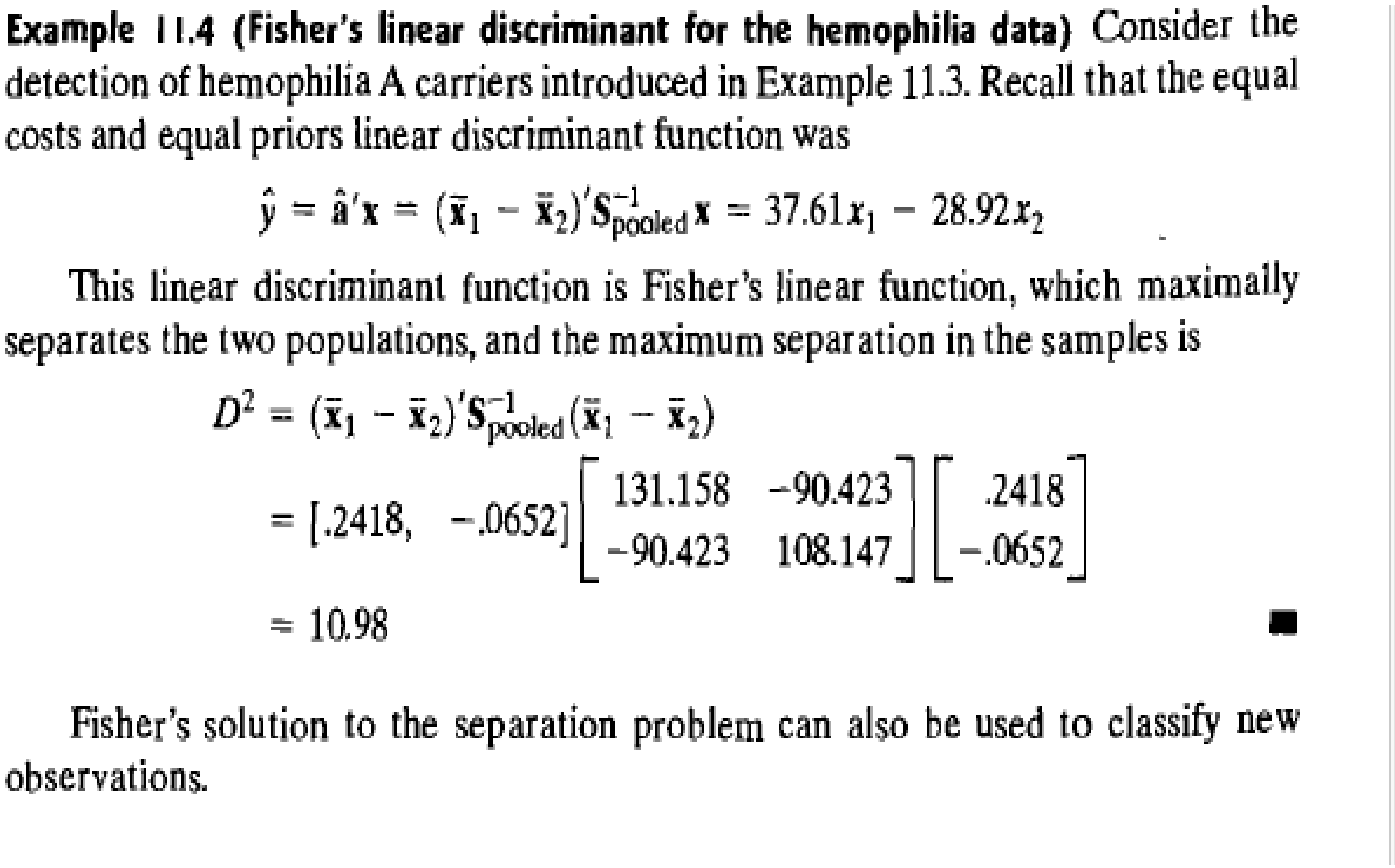
$$\frac{\begin{pmatrix}\text{squared distance} \\ \text{between sample means of } y\end{pmatrix}}{(\text{sample variance of } y)} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

$$= \frac{(\hat{\mathbf{a}}'\bar{\mathbf{x}}_1 - \hat{\mathbf{a}}'\bar{\mathbf{x}}_2)^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}}$$

$$= \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}} \qquad (11\text{-}23)$$

over all possible coefficient vectors $\hat{\mathbf{a}}$ where $\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. The maximum of the ratio (11-23) is $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

# Example: Fisher's Linear Discriminant for Hemophilia Data



**Example 11.4 (Fisher's linear discriminant for the hemophilia data)** Consider the detection of hemophilia A carriers introduced in Example 11.3. Recall that the equal costs and equal priors linear discriminant function was

$$\hat{y} = \hat{\mathbf{a}}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} = 37.61x_1 - 28.92x_2$$

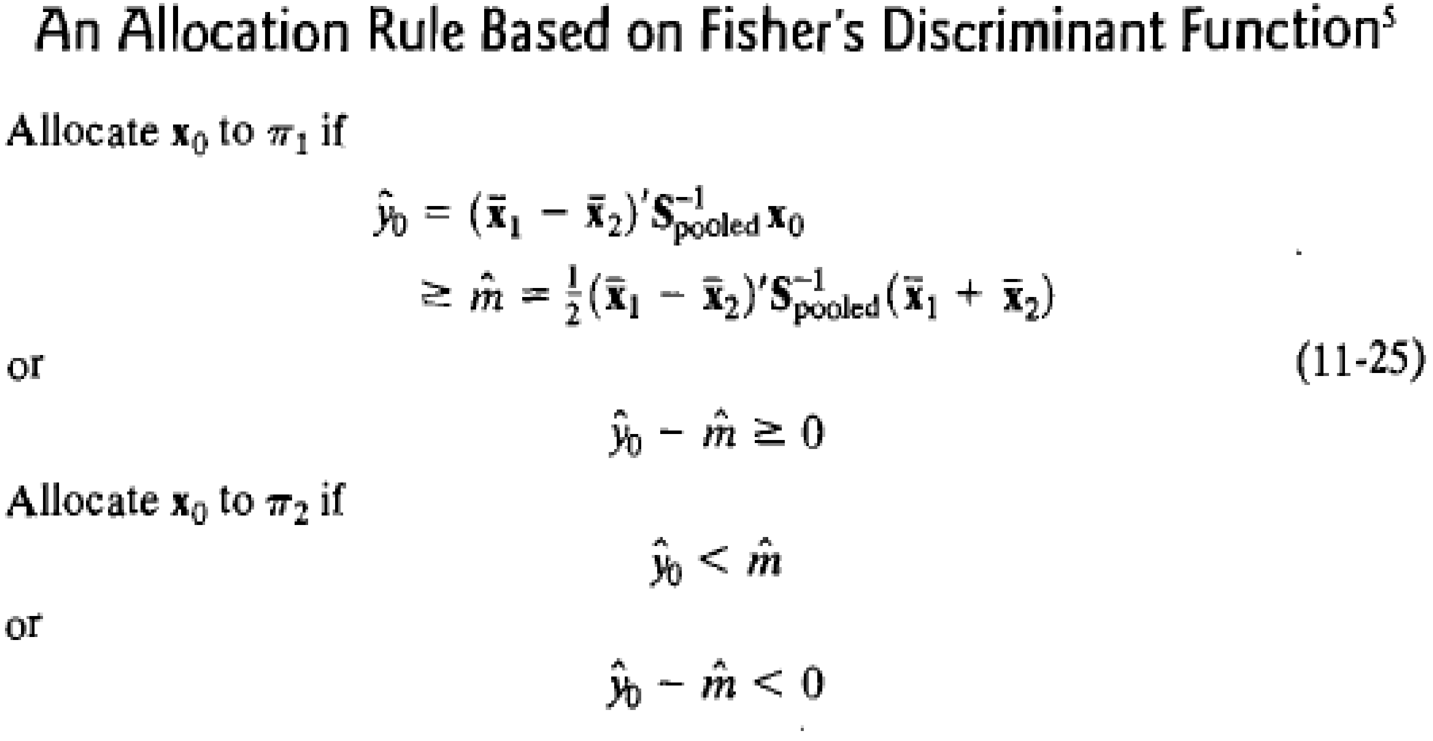
This linear discriminant function is Fisher's linear function, which maximally separates the two populations, and the maximum separation in the samples is

$$\begin{aligned} D^2 &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\ &= [.2418, \quad -.0652]\begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}\begin{bmatrix} .2418 \\ -.0652 \end{bmatrix} \\ &= 10.98 \end{aligned}$$

■

Fisher's solution to the separation problem can also be used to classify new observations.

# Allocation Rule Based on Fisher's Discriminant Function



## An Allocation Rule Based on Fisher's Discriminant Function[5]

Allocate $\mathbf{x}_0$ to $\pi_1$ 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 $\qquad$ (11-25)

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

Allocate $\mathbf{x}_0$ to $\pi_2$ if

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

or

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

# Graph: Classification Using Fisher's Procedure for Two Populations

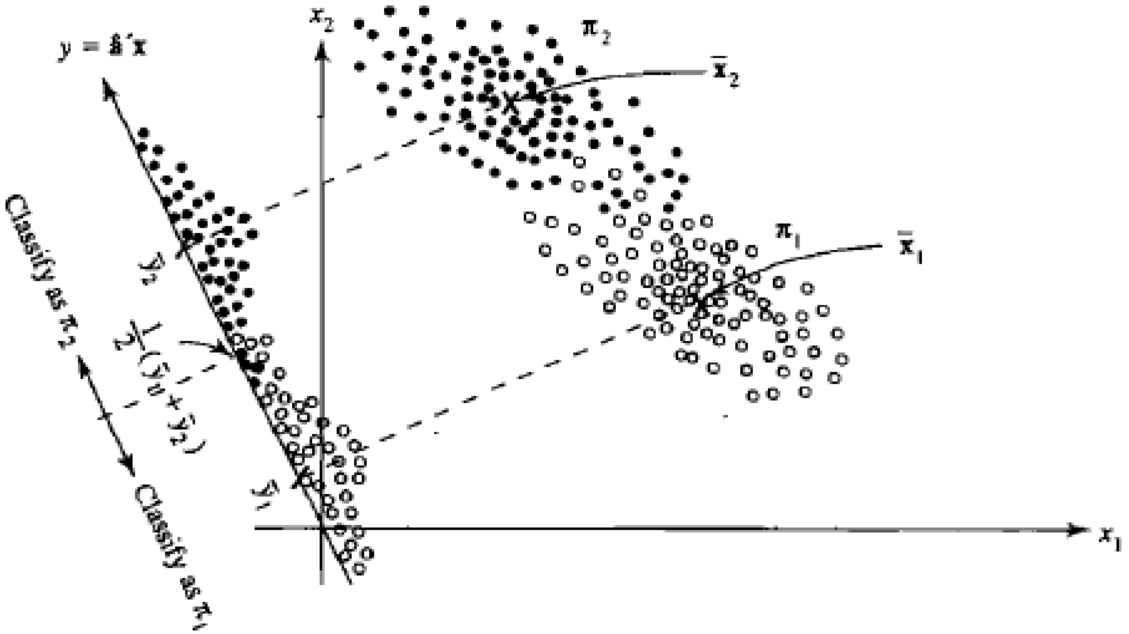
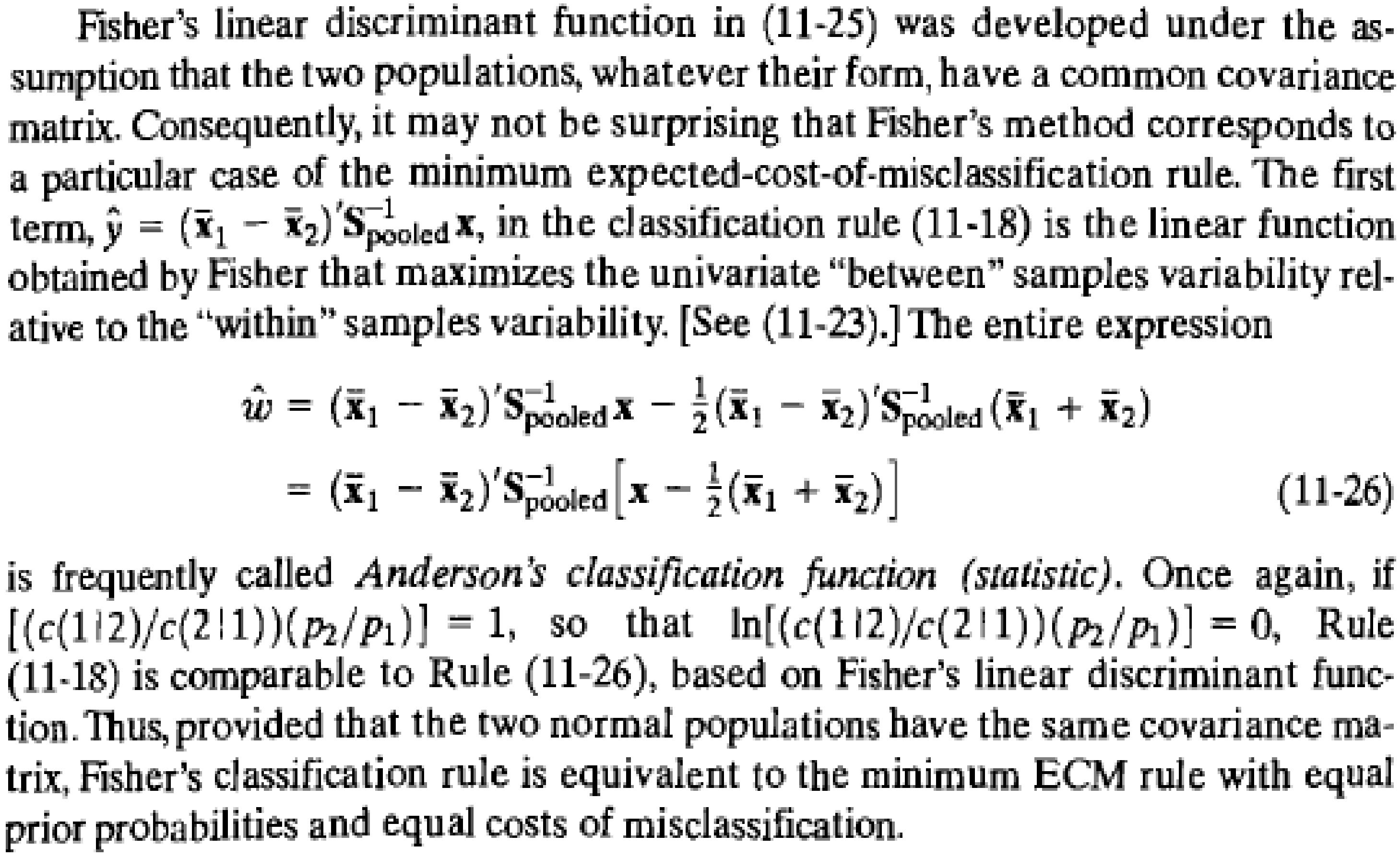


**Figure 11.5** A pictorial representation of Fisher's procedure for two populations with $p = 2$.

The procedure (11-23) is illustrated, schematically, for $p = 2$ in Figure 11.5. All points in the scatter plots are projected onto a line in the direction $\hat{\mathbf{a}}$, and this direction is varied until the samples are maximally separated.

# Fisher's Linear Discriminant Function: Common Covariance Matrix

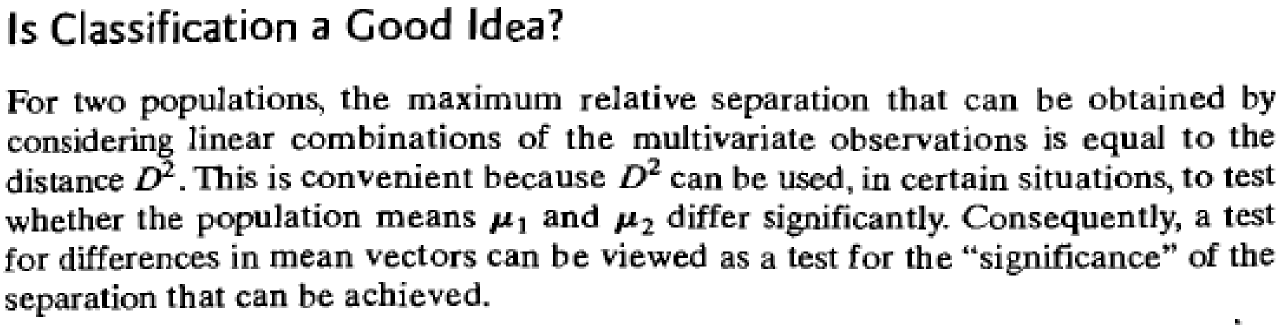


Fisher's linear discriminant function in (11-25) was developed under the assumption that the two populations, whatever their form, have a common covariance matrix. Consequently, it may not be surprising that Fisher's method corresponds to a particular case of the minimum expected-cost-of-misclassification rule. The first term, $\hat{y} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}$, in the classification rule (11-18) is the linear function obtained by Fisher that maximizes the univariate "between" samples variability relative to the "within" samples variability. [See (11-23).] The entire expression

$$\begin{aligned}\hat{w} &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\\ &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\left[\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right] \qquad (11\text{-}26)\end{aligned}$$

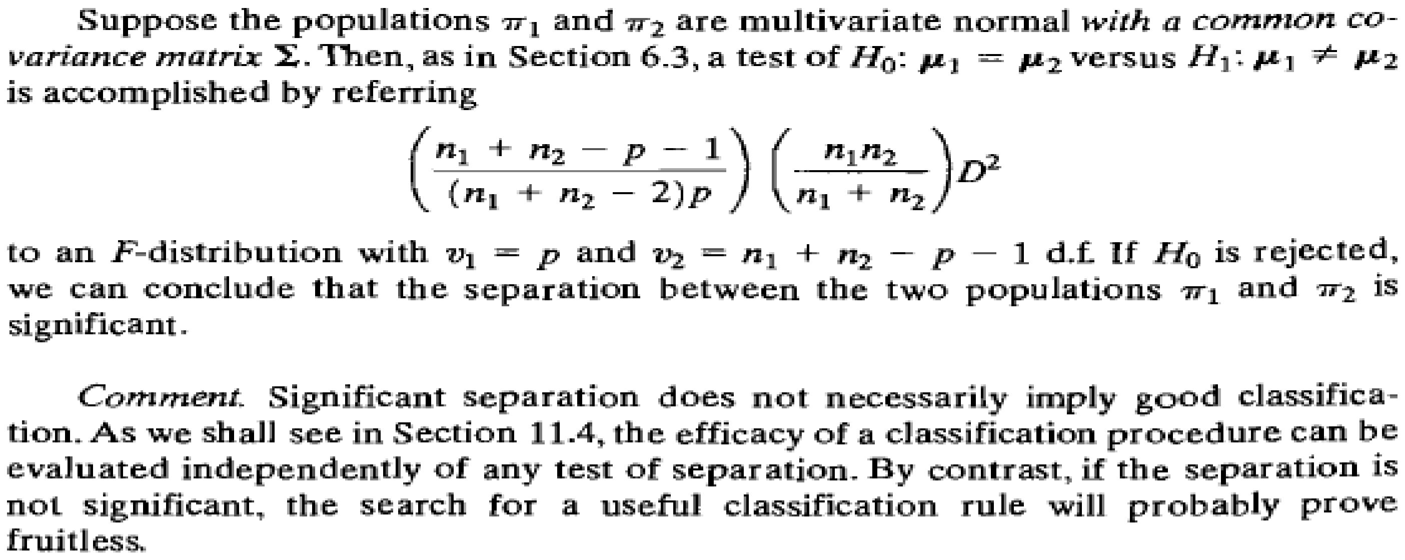
is frequently called *Anderson's classification function (statistic)*. Once again, if $[(c(1|2)/c(2|1))(p_2/p_1)] = 1$, so that $\ln[(c(1|2)/c(2|1))(p_2/p_1)] = 0$, Rule (11-18) is comparable to Rule (11-26), based on Fisher's linear discriminant function. Thus, provided that the two normal populations have the same covariance matrix, Fisher's classification rule is equivalent to the minimum ECM rule with equal prior probabilities and equal costs of misclassification.

# Value of Separation of Data – Is Classification a Good Idea



## Is Classification a Good Idea?

For two populations, the maximum relative separation that can be obtained by considering linear combinations of the multivariate observations is equal to the distance $D^2$. This is convenient because $D^2$ can be used, in certain situations, to test whether the population means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ differ significantly. Consequently, a test for differences in mean vectors can be viewed as a test for the "significance" of the separation that can be achieved.



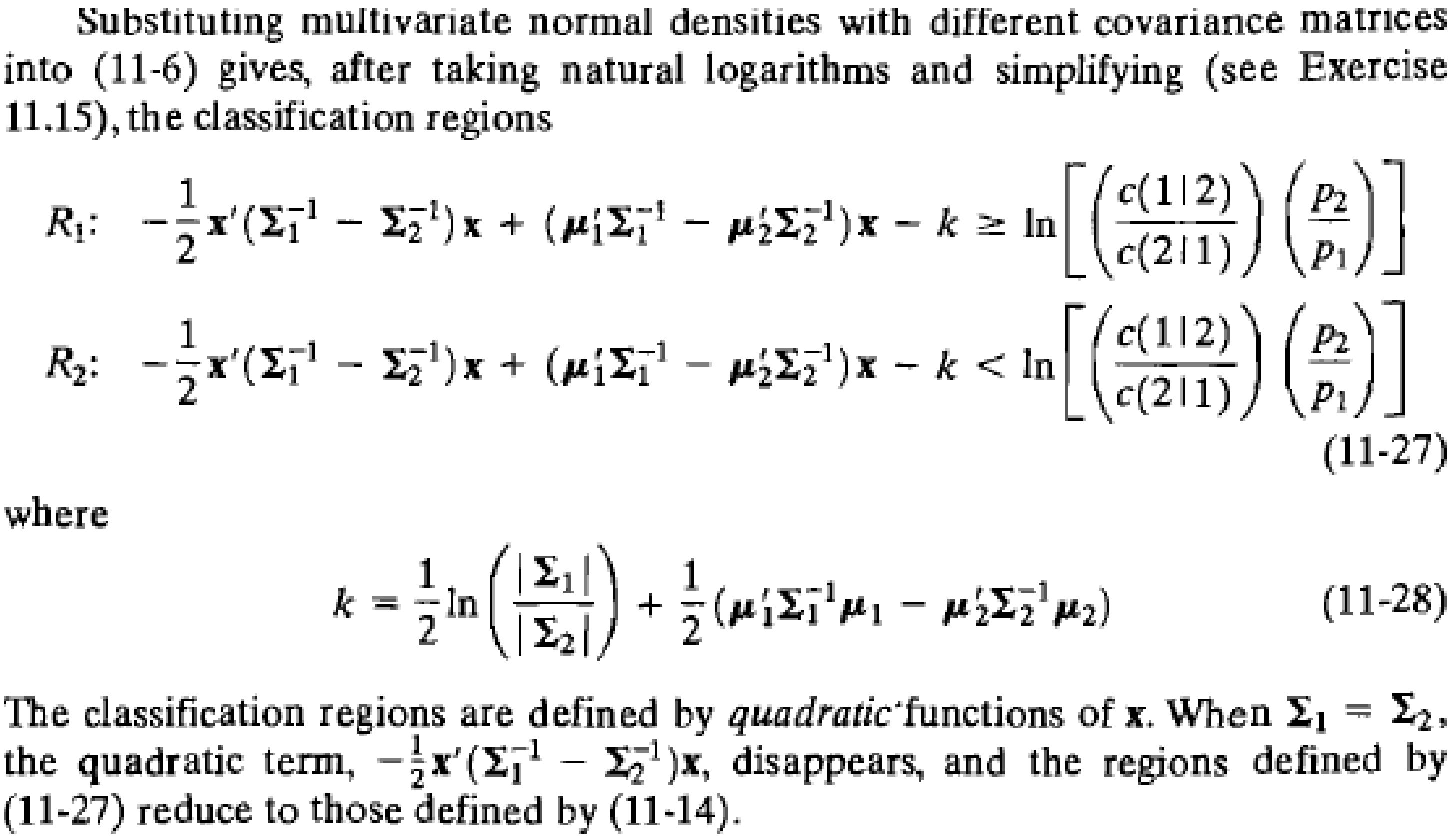
Suppose the populations $\pi_1$ and $\pi_2$ are multivariate normal *with a common covariance matrix* $\boldsymbol{\Sigma}$. Then, as in Section 6.3, a test of $H_0$: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1$: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ is accomplished by referring

$$\left(\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p}\right)\left(\frac{n_1 n_2}{n_1 + n_2}\right)D^2$$

to an $F$-distribution with $v_1 = p$ and $v_2 = n_1 + n_2 - p - 1$ d.f. If $H_0$ is rejected, we can conclude that the separation between the two populations $\pi_1$ and $\pi_2$ is significant.

*Comment.* Significant separation does not necessarily imply good classification. As we shall see in Section 11.4, the efficacy of a classification procedure can be evaluated independently of any test of separation. By contrast, if the separation is not significant, the search for a useful classification rule will probably prove fruitless.

# Classification Regions Defined by Quadratic Functions X When Covariance Matrices Not Equal



Substituting multivariate normal densities with different covariance matrices into (11-6) gives, after taking natural logarithms and simplifying (see Exercise 11.15), the classification regions

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

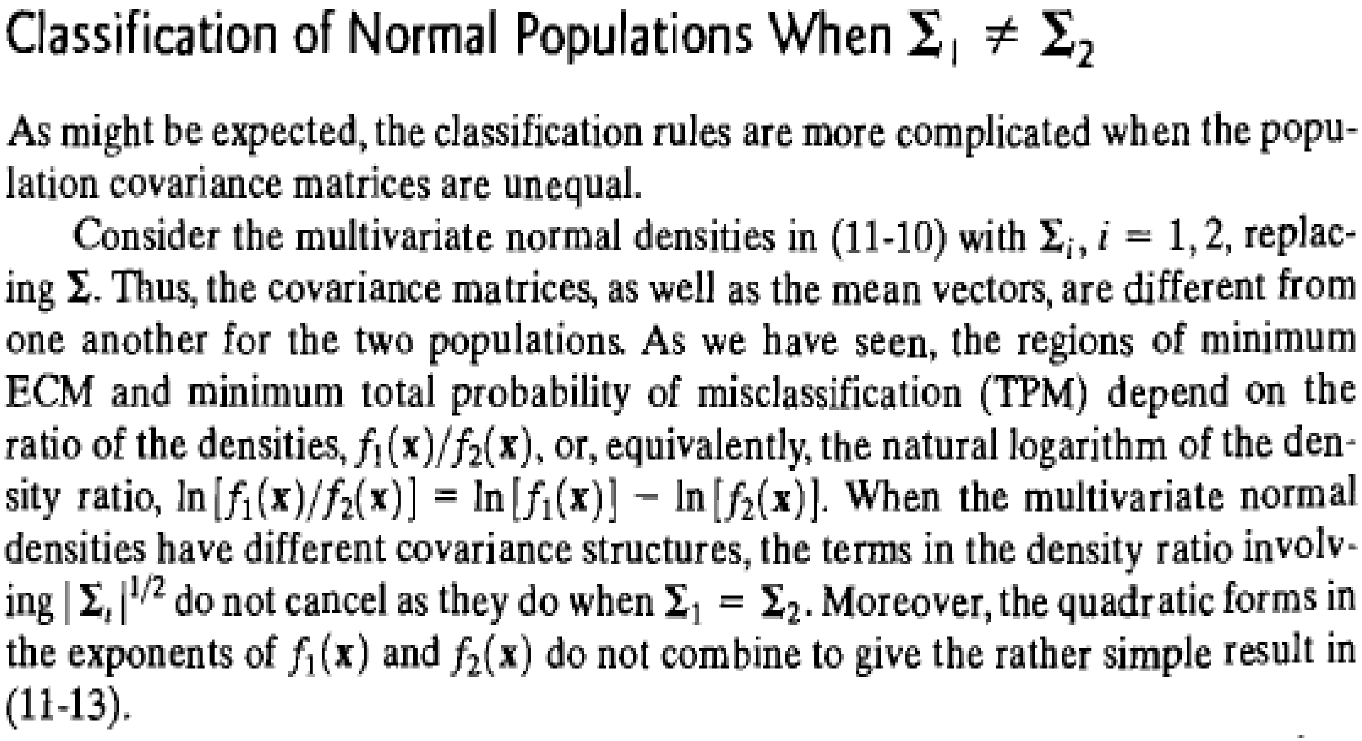
$$R_2: \quad -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{11-27}$$

where

$$k = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2) \tag{11-28}$$

The classification regions are defined by *quadratic* functions of $\mathbf{x}$. When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, the quadratic term, $-\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x}$, disappears, and the regions defined by (11-27) reduce to those defined by (11-14).

# Classification Normal Populations When Covariance Matrices Are not Equal

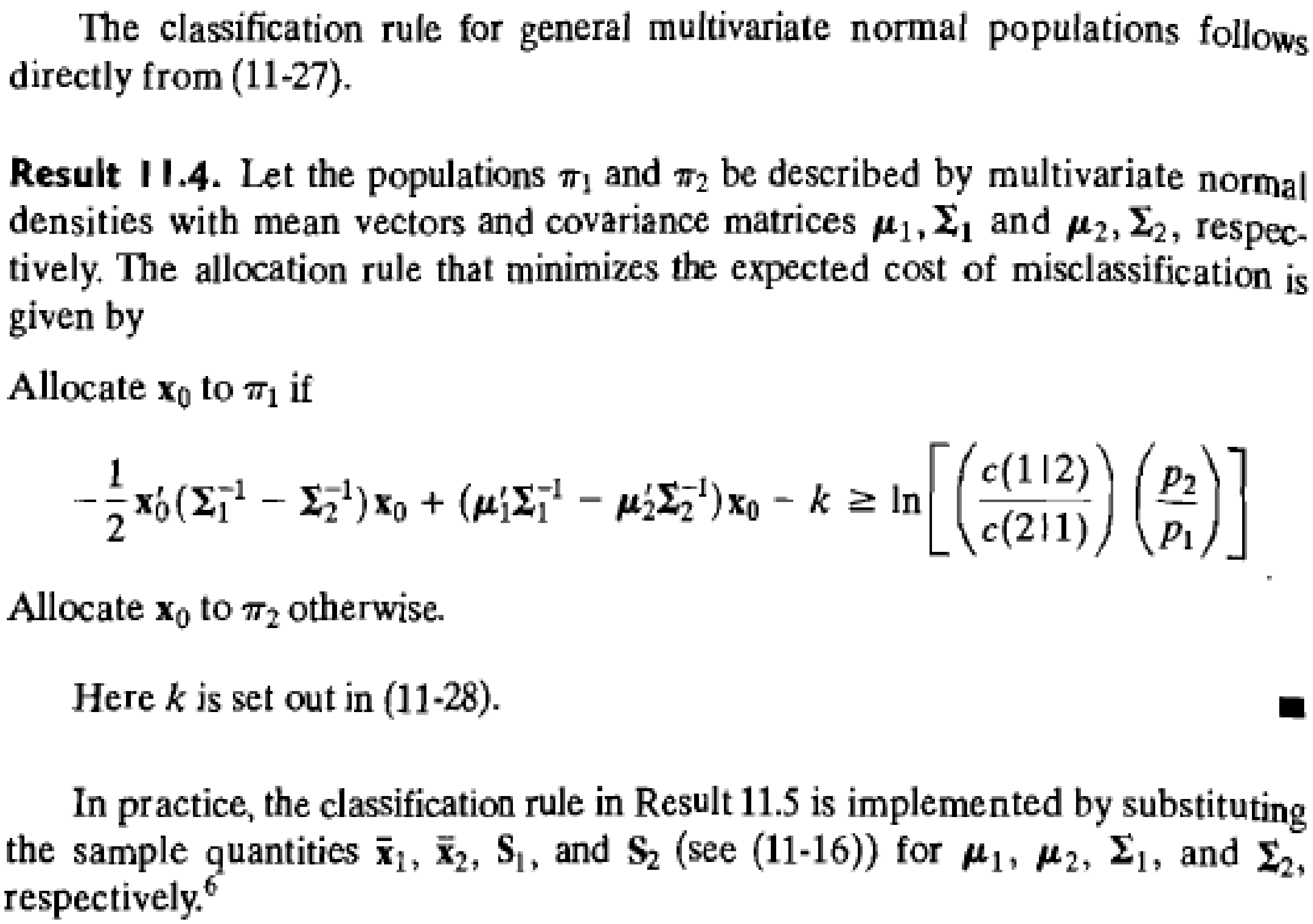


## Classification of Normal Populations When $\Sigma_1 \neq \Sigma_2$

As might be expected, the classification rules are more complicated when the population covariance matrices are unequal.

Consider the multivariate normal densities in (11-10) with $\mathbf{\Sigma}_i$, $i = 1, 2$, replacing $\mathbf{\Sigma}$. Thus, the covariance matrices, as well as the mean vectors, are different from one another for the two populations. As we have seen, the regions of minimum ECM and minimum total probability of misclassification (TPM) depend on the ratio of the densities, $f_1(\mathbf{x})/f_2(\mathbf{x})$, or, equivalently, the natural logarithm of the density ratio, $\ln[f_1(\mathbf{x})/f_2(\mathbf{x})] = \ln[f_1(\mathbf{x})] - \ln[f_2(\mathbf{x})]$. When the multivariate normal densities have different covariance structures, the terms in the density ratio involving $|\mathbf{\Sigma}_i|^{1/2}$ do not cancel as they do when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Moreover, the quadratic forms in the exponents of $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ do not combine to give the rather simple result in (11-13).

# Allocation Rule Minimizes Expected Cost of Misclassification



The classification rule for general multivariate normal populations follows directly from (11-27).

**Result 11.4.** Let the populations $\pi_1$ and $\pi_2$ be described by multivariate normal densities with mean vectors and covariance matrices $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2$, respectively. The allocation rule that minimizes the expected cost of misclassification is given by

Allocate $\mathbf{x}_0$ to $\pi_1$ if

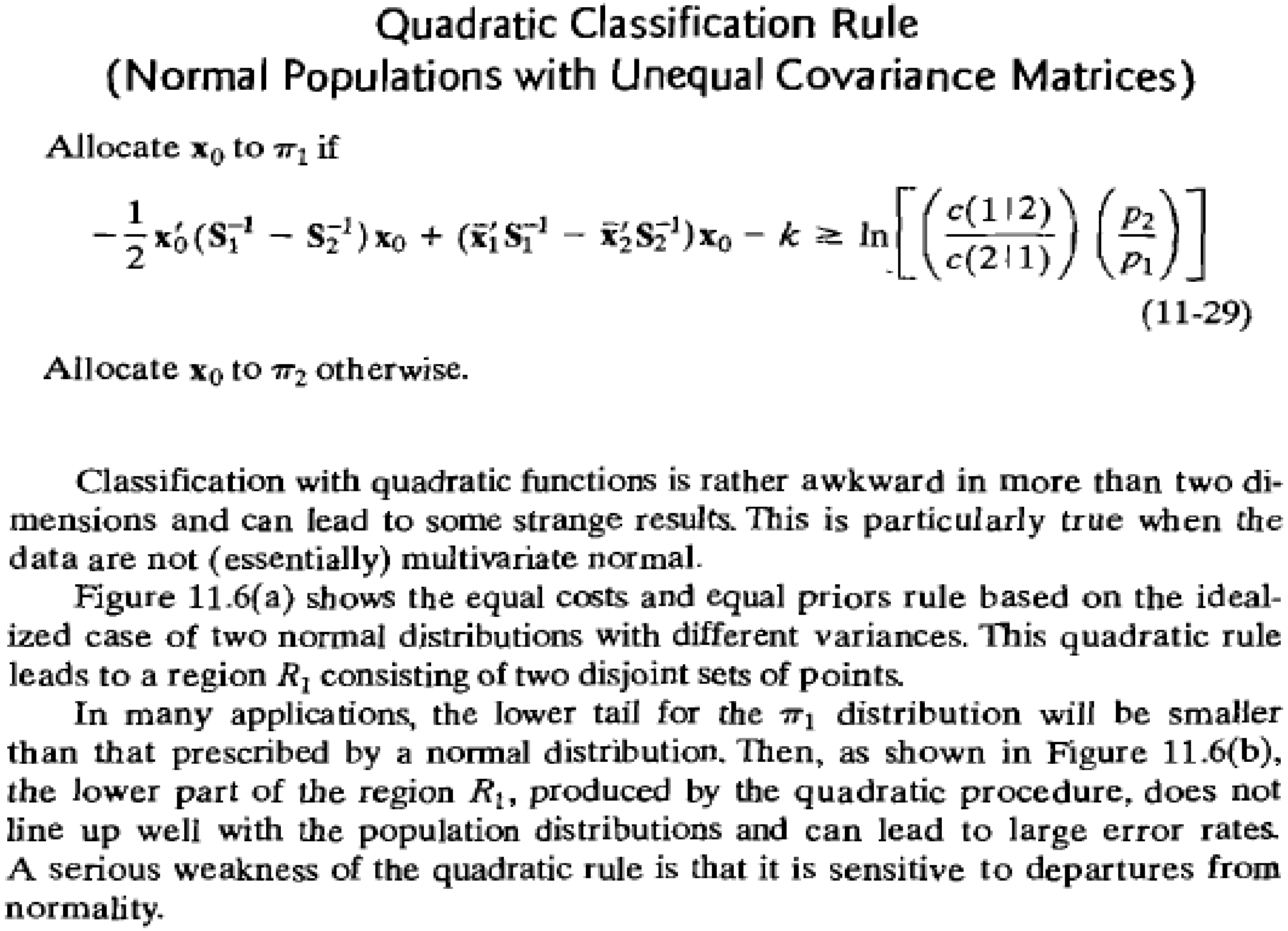
$$-\frac{1}{2}\mathbf{x}_0'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x}_0 + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Allocate $\mathbf{x}_0$ to $\pi_2$ otherwise.

Here $k$ is set out in (11-28). ■

In practice, the classification rule in Result 11.5 is implemented by substituting the sample quantities $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, $\mathbf{S}_1$, and $\mathbf{S}_2$ (see (11-16)) for $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\Sigma}_2$, respectively.[6]

# Quadratic Classification Rule: Normal Populations with Unequal Covariance Matrices



## Quadratic Classification Rule (Normal Populations with Unequal Covariance Matrices)

Allocate $\mathbf{x}_0$ to $\pi_1$ if

$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}_0 + (\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{11-29}$$

Allocate $\mathbf{x}_0$ to $\pi_2$ otherwise.

Classification with quadratic functions is rather awkward in more than two dimensions and can lead to some strange results. This is particularly true when the data are not (essentially) multivariate normal.

Figure 11.6(a) shows the equal costs and equal priors rule based on the idealized case of two normal distributions with different variances. This quadratic rule leads to a region $R_1$ consisting of two disjoint sets of points.

In many applications, the lower tail for the $\pi_1$ distribution will be smaller than that prescribed by a normal distribution. Then, as shown in Figure 11.6(b), the lower part of the region $R_1$, produced by the quadratic procedure, does not line up well with the population distributions and can lead to large error rates. A serious weakness of the quadratic rule is that it is sensitive to departures from normality.

# Quadratic Classification Rule: Normal Populations with Unequal Covariance

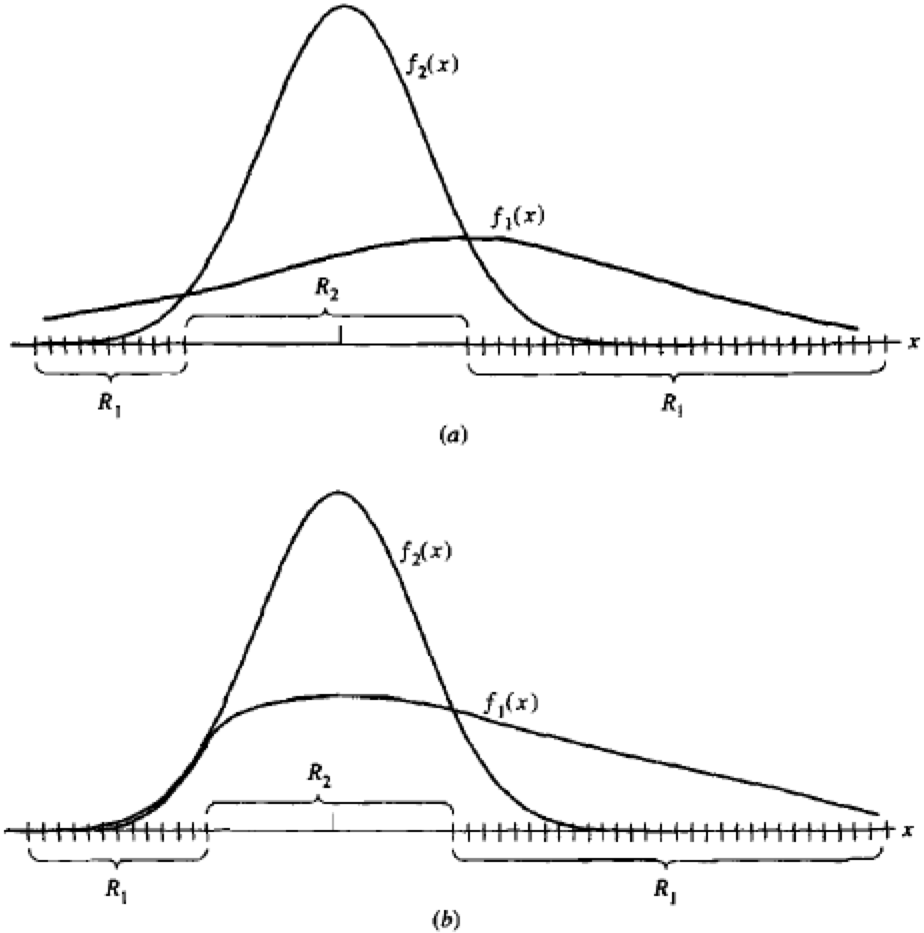
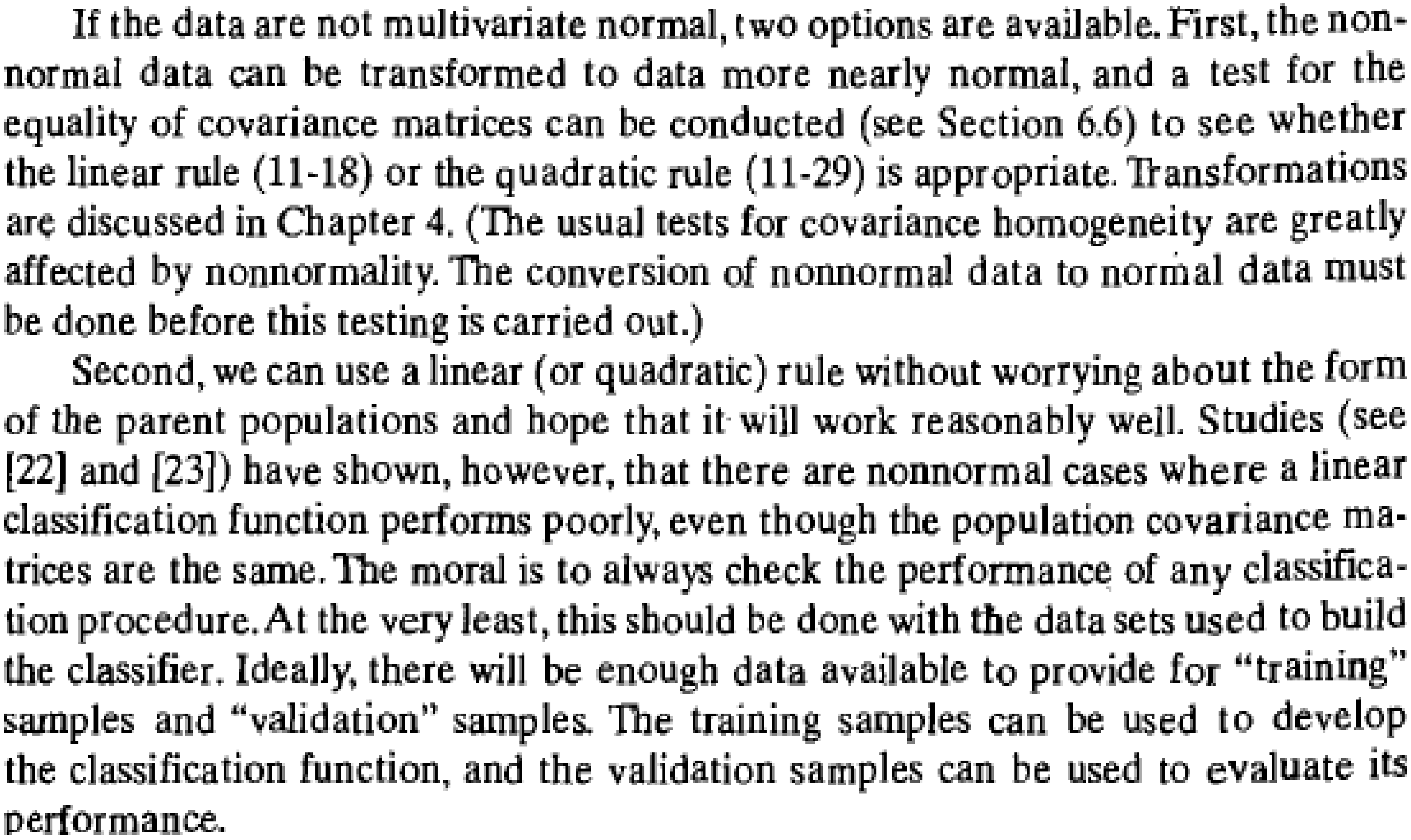


**Figure 11.6** Quadratic rules for (a) two normal distribution with unequal variances and (b) two distributions, one of which is nonnormal—rule not appropriate.

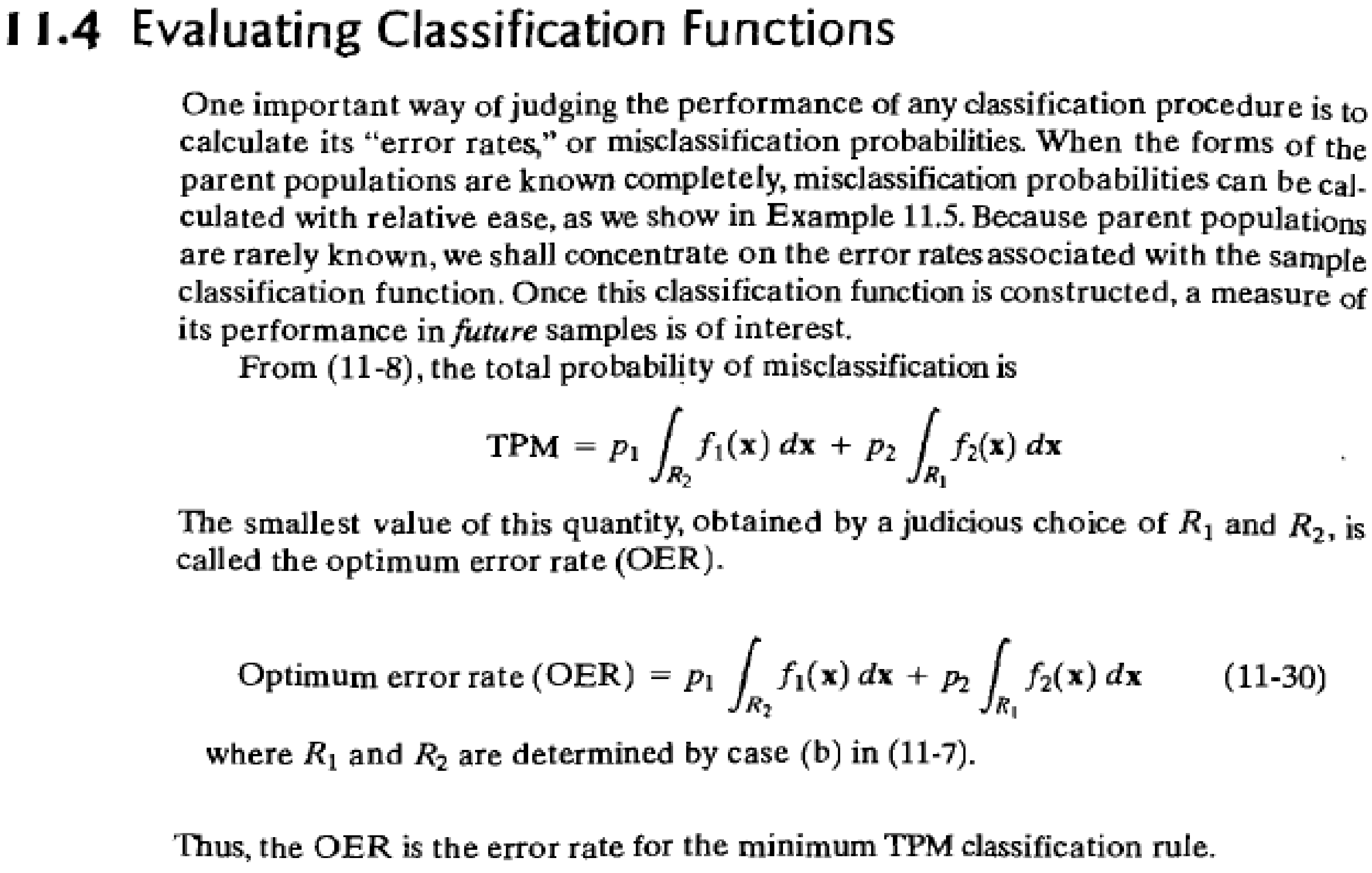
# If Data Not Multivariate Normal: Either Transform or Linear or Quadratic Rule



If the data are not multivariate normal, two options are available. First, the nonnormal data can be transformed to data more nearly normal, and a test for the equality of covariance matrices can be conducted (see Section 6.6) to see whether the linear rule (11-18) or the quadratic rule (11-29) is appropriate. Transformations are discussed in Chapter 4. (The usual tests for covariance homogeneity are greatly affected by nonnormality. The conversion of nonnormal data to normal data must be done before this testing is carried out.)

Second, we can use a linear (or quadratic) rule without worrying about the form of the parent populations and hope that it will work reasonably well. Studies (see [22] and [23]) have shown, however, that there are nonnormal cases where a linear classification function performs poorly, even though the population covariance matrices are the same. The moral is to always check the performance of any classification procedure. At the very least, this should be done with the data sets used to build the classifier. Ideally, there will be enough data available to provide for "training" samples and "validation" samples. The training samples can be used to develop the classification function, and the validation samples can be used to evaluate its performance.

# Evaluating Classification Functions



## 11.4 Evaluating Classification Functions

One important way of judging the performance of any classification procedure is to calculate its "error rates," or misclassification probabilities. When the forms of the parent populations are known completely, misclassification probabilities can be calculated with relative ease, as we show in Example 11.5. Because parent populations are rarely known, we shall concentrate on the error rates associated with the sample classification function. Once this classification function is constructed, a measure of its performance in *future* samples is of interest.

From (11-8), the total probability of misclassification is

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

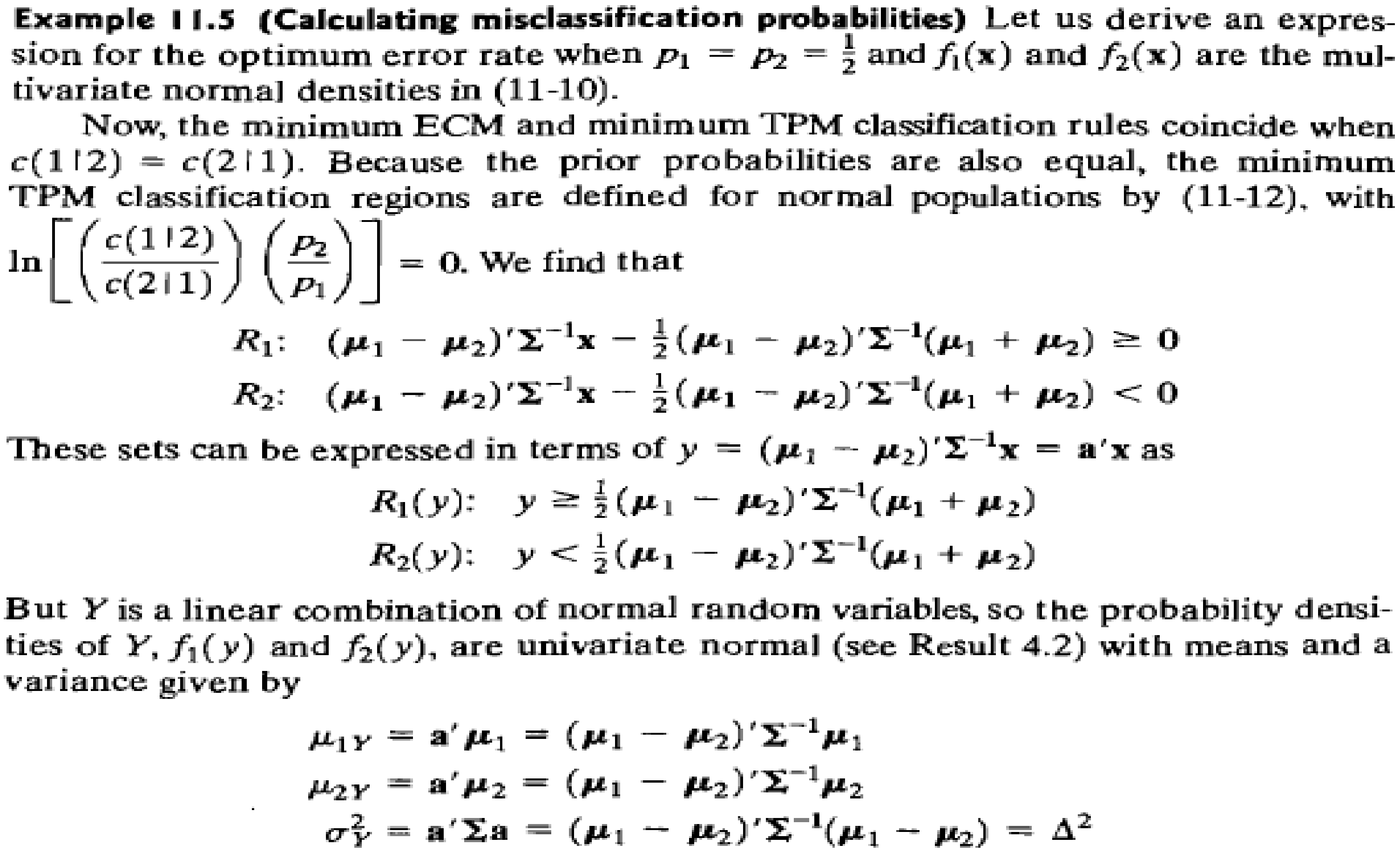
The smallest value of this quantity, obtained by a judicious choice of $R_1$ and $R_2$, is called the optimum error rate (OER).

$$\text{Optimum error rate (OER)} = p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} \qquad (11\text{-}30)$$

where $R_1$ and $R_2$ are determined by case (b) in (11-7).

Thus, the OER is the error rate for the minimum TPM classification rule.

# Calculating Misclassification Probabilities



**Example 11.5 (Calculating misclassification probabilities)** Let us derive an expression for the optimum error rate when $p_1 = p_2 = \frac{1}{2}$ and $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are the multivariate normal densities in (11-10).

Now, the minimum ECM and minimum TPM classification rules coincide when $c(1|2) = c(2|1)$. Because the prior probabilities are also equal, the minimum TPM classification regions are defined for normal populations by (11-12), with $\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] = 0$. We find that

$$R_1: \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \ge 0$$

$$R_2: \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < 0$$

These sets can be expressed in terms of $y = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} = \mathbf{a}'\mathbf{x}$ as

$$R_1(y): \quad y \ge \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

$$R_2(y): \quad y < \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

But $Y$ is a linear combination of normal random variables, so the probability densities of $Y$, $f_1(y)$ and $f_2(y)$, are univariate normal (see Result 4.2) with means and a variance given by

$$\begin{aligned} \mu_{1Y} &= \mathbf{a}'\boldsymbol{\mu}_1 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 \\ \mu_{2Y} &= \mathbf{a}'\boldsymbol{\mu}_2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 \\ \sigma_Y^2 &= \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \Delta^2 \end{aligned}$$

# Misclassification Probabilities Illustration Based on Y

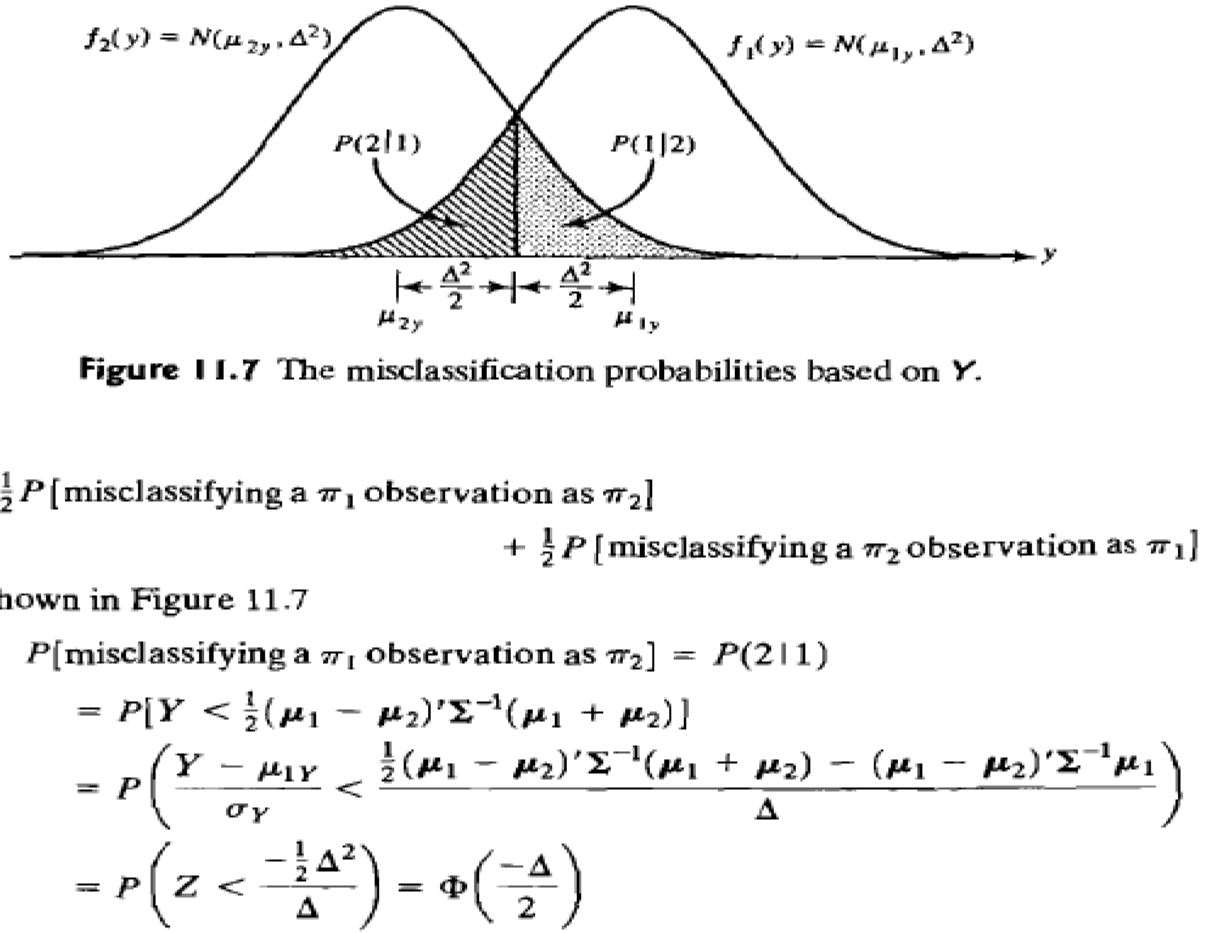

$f_2(y) = N(\mu_{2y}, \Delta^2)$
$f_1(y) = N(\mu_{1y}, \Delta^2)$
$P(2|1)$
$P(1|2)$
$y$
$\frac{\Delta^2}{2}$
$\frac{\Delta^2}{2}$
$\mu_{2y}$
$\mu_{1y}$

**Figure 11.7** The misclassification probabilities based on $Y$.

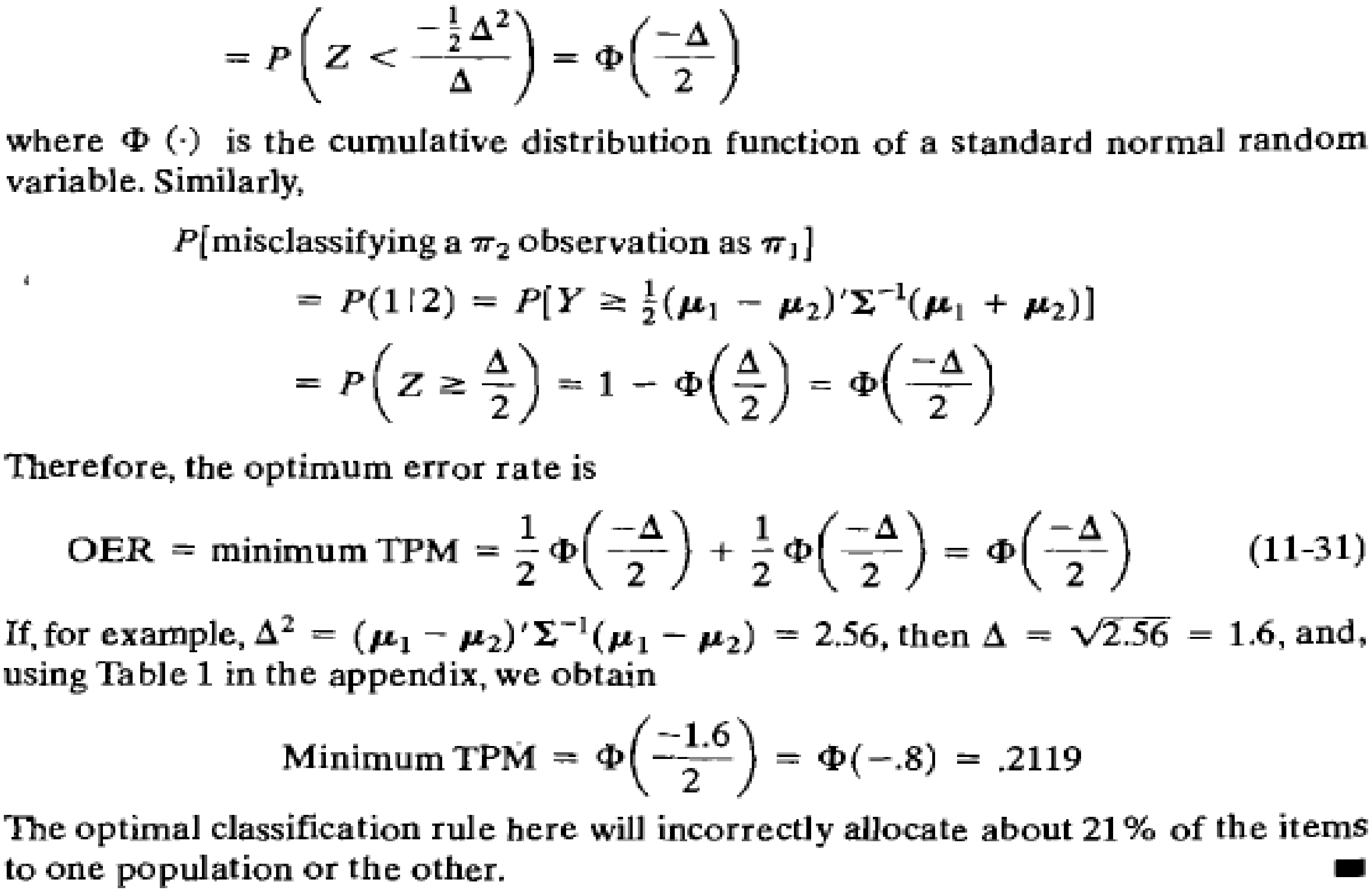
Now,

$$\text{TPM} = \tfrac{1}{2}P[\text{misclassifying a } \pi_1 \text{ observation as } \pi_2] + \tfrac{1}{2}P[\text{misclassifying a } \pi_2 \text{ observation as } \pi_1]$$

But, as shown in Figure 11.7

$$\begin{aligned} P[\text{misclassifying a } \pi_1 \text{ observation as } \pi_2] &= P(2|1) \\ &= P[Y < \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)] \\ &= P\left(\frac{Y - \mu_{1Y}}{\sigma_Y} < \frac{\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1}{\Delta}\right) \\ &= P\left(Z < \frac{-\frac{1}{2}\Delta^2}{\Delta}\right) = \Phi\left(\frac{-\Delta}{2}\right) \end{aligned}$$

# Optimum Error Rate When Population Density Function Known



$$= P\left(Z < \frac{-\frac{1}{2}\Delta^2}{\Delta}\right) = \Phi\left(\frac{-\Delta}{2}\right)$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Similarly,

$$P[\text{misclassifying a } \pi_2 \text{ observation as } \pi_1]$$

$$= P(1|2) = P[Y \geq \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)]$$

$$= P\left(Z \geq \frac{\Delta}{2}\right) = 1 - \Phi\left(\frac{\Delta}{2}\right) = \Phi\left(\frac{-\Delta}{2}\right)$$

Therefore, the optimum error rate is

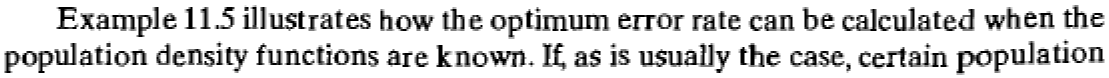
$$\text{OER} = \text{minimum TPM} = \frac{1}{2}\Phi\left(\frac{-\Delta}{2}\right) + \frac{1}{2}\Phi\left(\frac{-\Delta}{2}\right) = \Phi\left(\frac{-\Delta}{2}\right) \qquad (11\text{-}31)$$

If, for example, $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 2.56$, then $\Delta = \sqrt{2.56} = 1.6$, and, using Table 1 in the appendix, we obtain

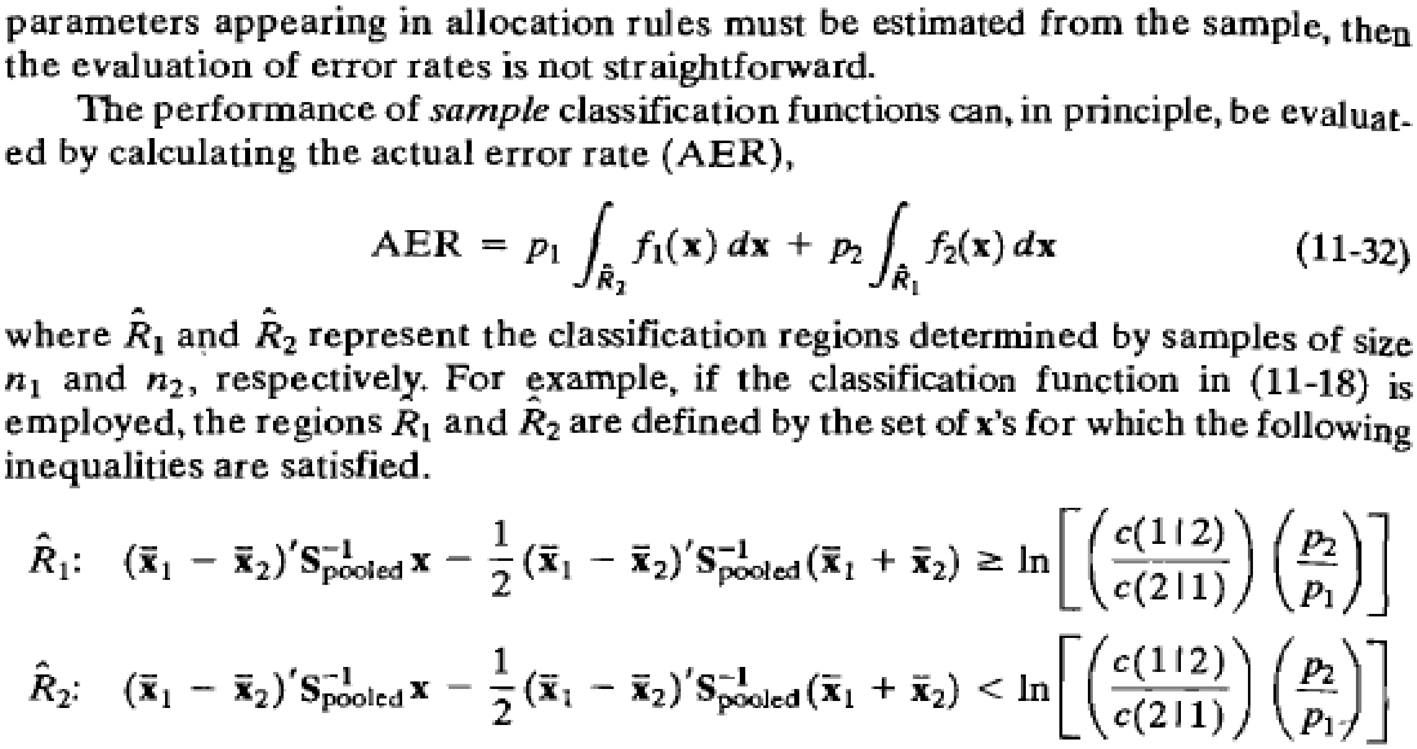
$$\text{Minimum TPM} = \Phi\left(\frac{-1.6}{2}\right) = \Phi(-.8) = .2119$$

The optimal classification rule here will incorrectly allocate about 21% of the items to one population or the other. ■

# Actual Error Rate



Example 11.5 illustrates how the optimum error rate can be calculated when the population density functions are known. If, as is usually the case, certain population



parameters appearing in allocation rules must be estimated from the sample, then the evaluation of error rates is not straightforward.

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

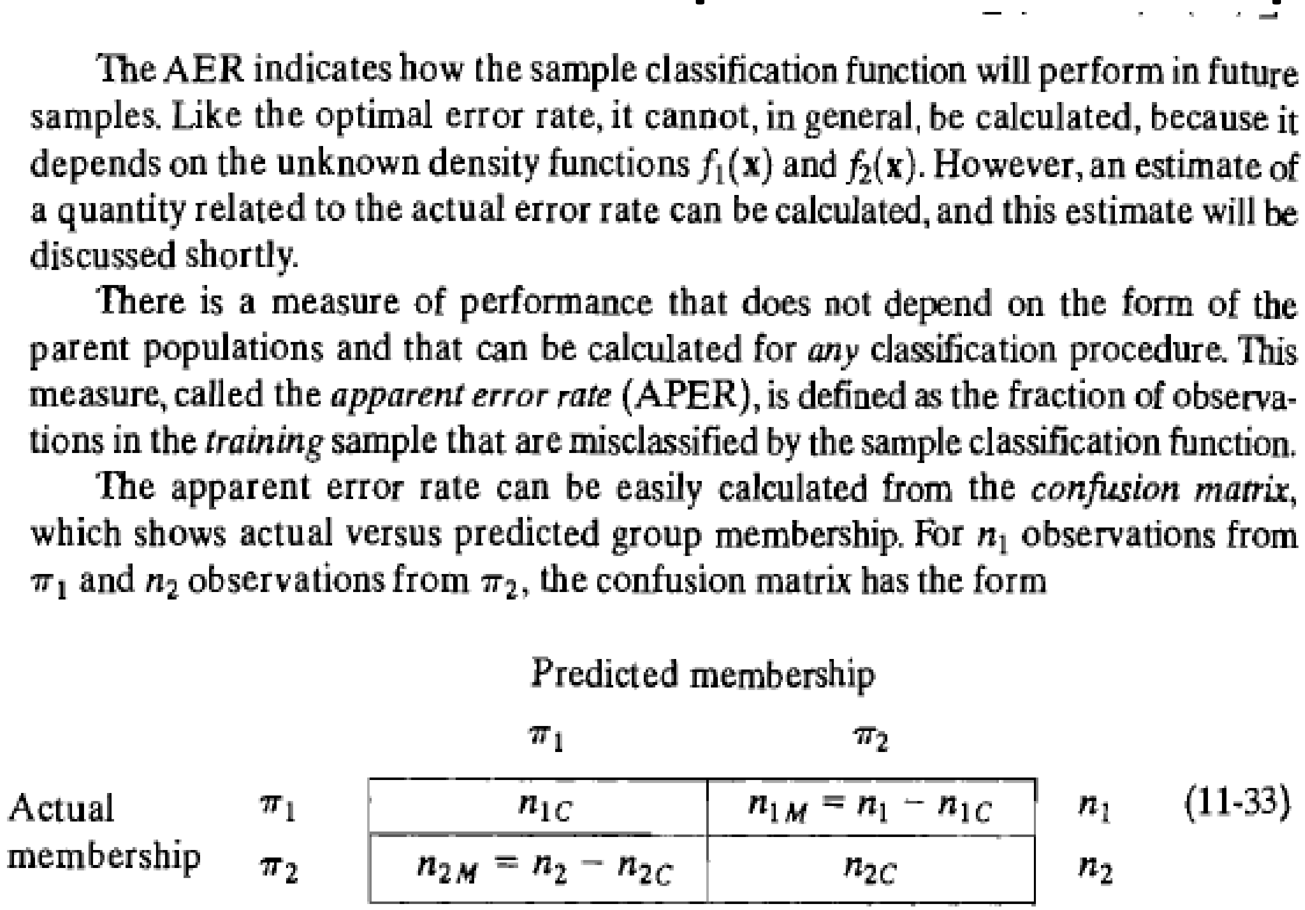
$$\text{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x})\, d\mathbf{x} \qquad (11\text{-}32)$$

where $\hat{R}_1$ and $\hat{R}_2$ represent the classification regions determined by samples of size $n_1$ and $n_2$, respectively. For example, if the classification function in (11-18) is employed, the regions $\hat{R}_1$ and $\hat{R}_2$ are defined by the set of $\mathbf{x}$'s for which the following inequalities are satisfied.

$$\hat{R}_1: \quad (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

$$\hat{R}_2: \quad (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

# Confusion Matrix: Actual Versus Predicted Group Membership



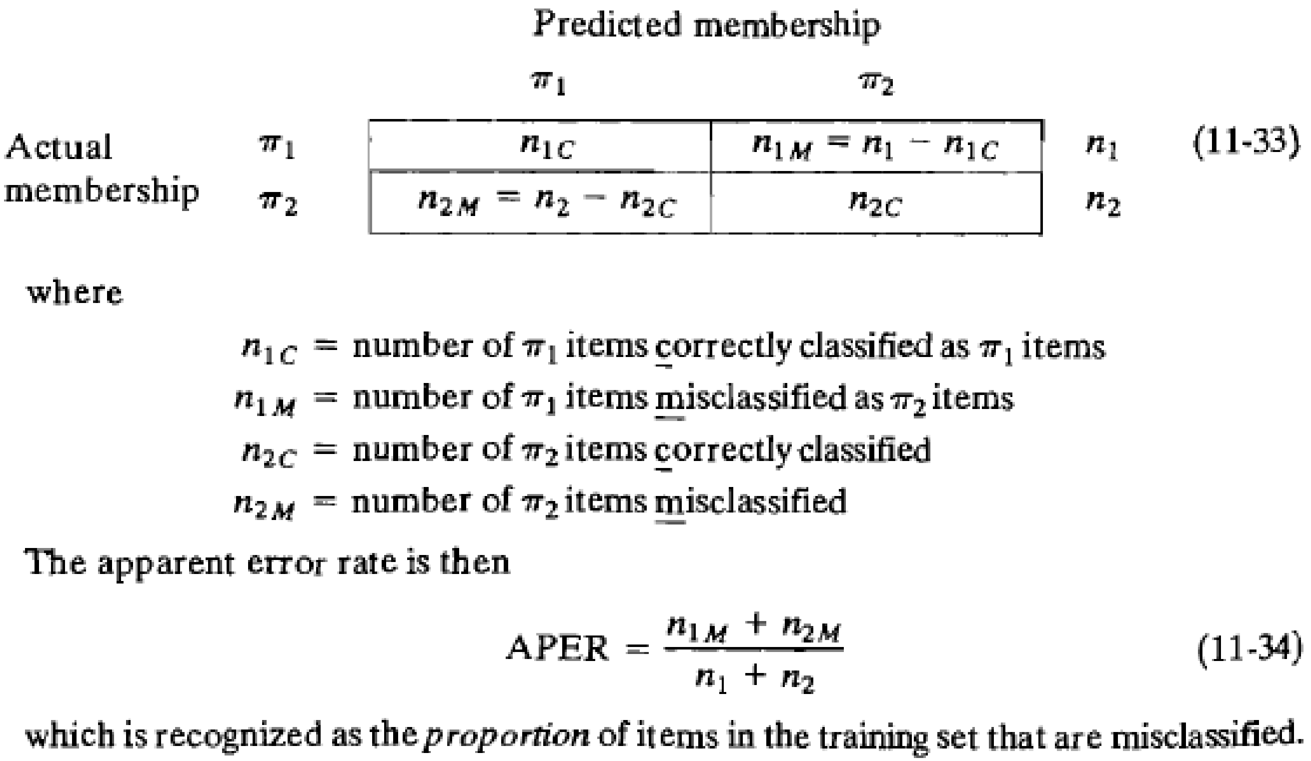
The AER indicates how the sample classification function will perform in future samples. Like the optimal error rate, it cannot, in general, be calculated, because it depends on the unknown density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$. However, an estimate of a quantity related to the actual error rate can be calculated, and this estimate will be discussed shortly.

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 $\pi_1$ and $n_2$ observations from $\pi_2$, the confusion matrix has the form

| | | Predicted membership | | | |
|---|---|---|---|---|---|
| | | $\pi_1$ | $\pi_2$ | | |
| Actual membership | $\pi_1$ | $n_{1C}$ | $n_{1M} = n_1 - n_{1C}$ | $n_1$ | (11-33) |
| | $\pi_2$ | $n_{2M} = n_2 - n_{2C}$ | $n_{2C}$ | $n_2$ | |

# Apparent Error Rate As Proportion of Items Misclassified in Training Data Set



| | | | | | |
|---|---|---|---|---|---|
| | | Predicted membership | | | |
| | | $\pi_1$ | $\pi_2$ | | |
| Actual membership | $\pi_1$ | $n_{1C}$ | $n_{1M} = n_1 - n_{1C}$ | $n_1$ | (11-33) |
| | $\pi_2$ | $n_{2M} = n_2 - n_{2C}$ | $n_{2C}$ | $n_2$ | |

where

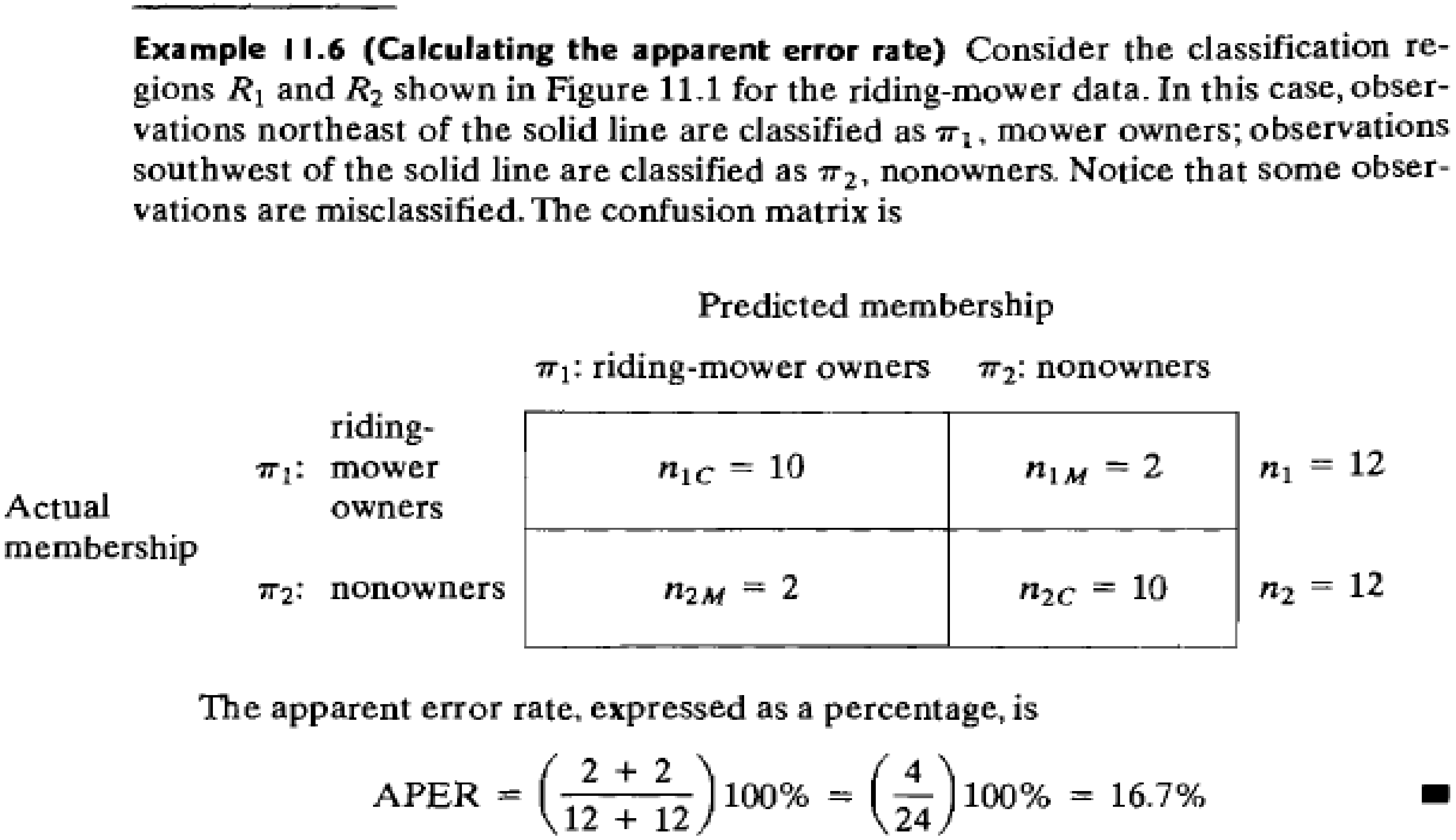
$n_{1C}$ = number of $\pi_1$ items correctly classified as $\pi_1$ items
$n_{1M}$ = number of $\pi_1$ items misclassified as $\pi_2$ items
$n_{2C}$ = number of $\pi_2$ items correctly classified
$n_{2M}$ = number of $\pi_2$ items misclassified

The apparent error rate is then

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \qquad (11\text{-}34)$$

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

# Calculating the Apparent Error Rate



**Example 11.6 (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 $\pi_1$, mower owners; observations southwest of the solid line are classified as $\pi_2$, nonowners. Notice that some observations are misclassified. The confusion matrix is

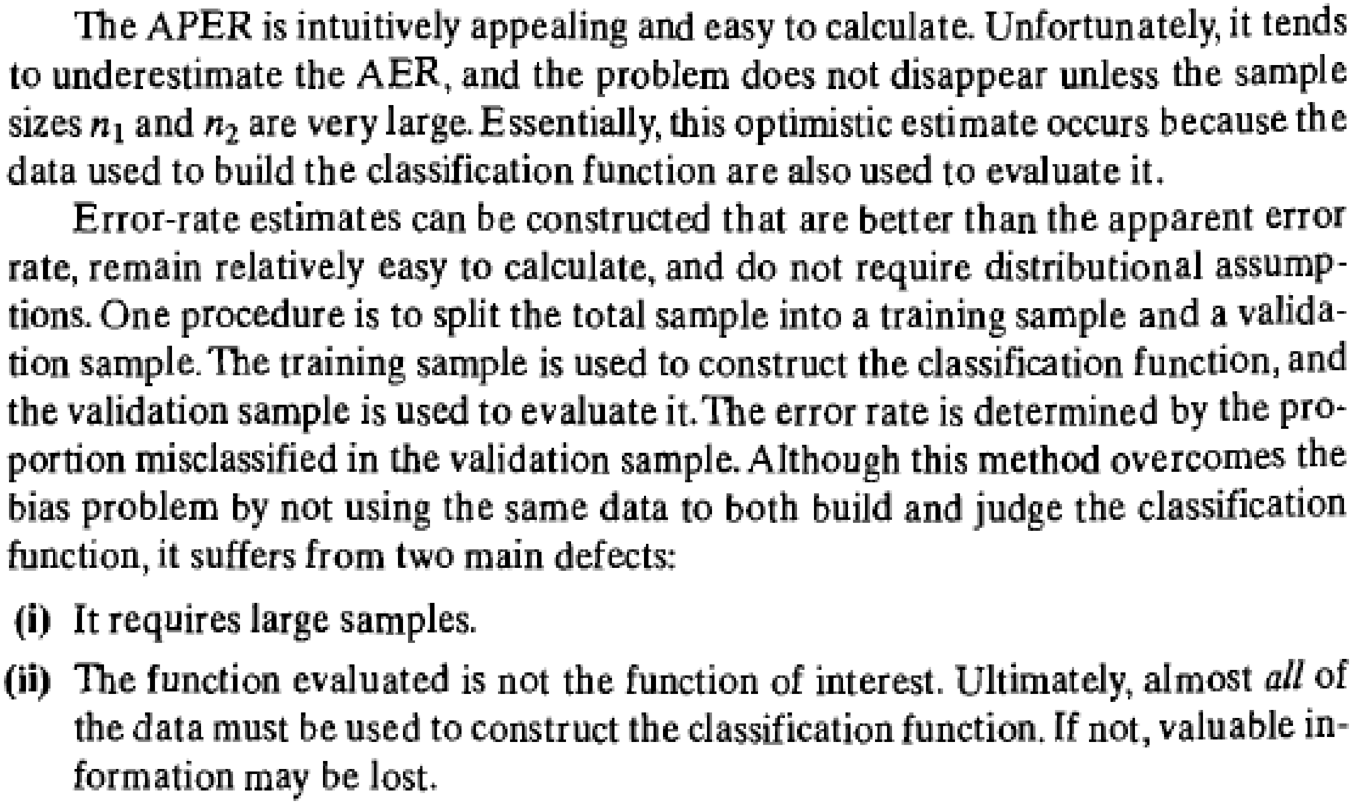
| | | Predicted membership | | |
|---|---|---|---|---|
| | | $\pi_1$: riding-mower owners | $\pi_2$: nonowners | |
| Actual membership | $\pi_1$: riding-mower owners | $n_{1C} = 10$ | $n_{1M} = 2$ | $n_1 = 12$ |
| | $\pi_2$: nonowners | $n_{2M} = 2$ | $n_{2C} = 10$ | $n_2 = 12$ |

The apparent error rate, expressed as a percentage, is

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

■

# Error Rate Determined by Proportion Misclassified in Validation Sample



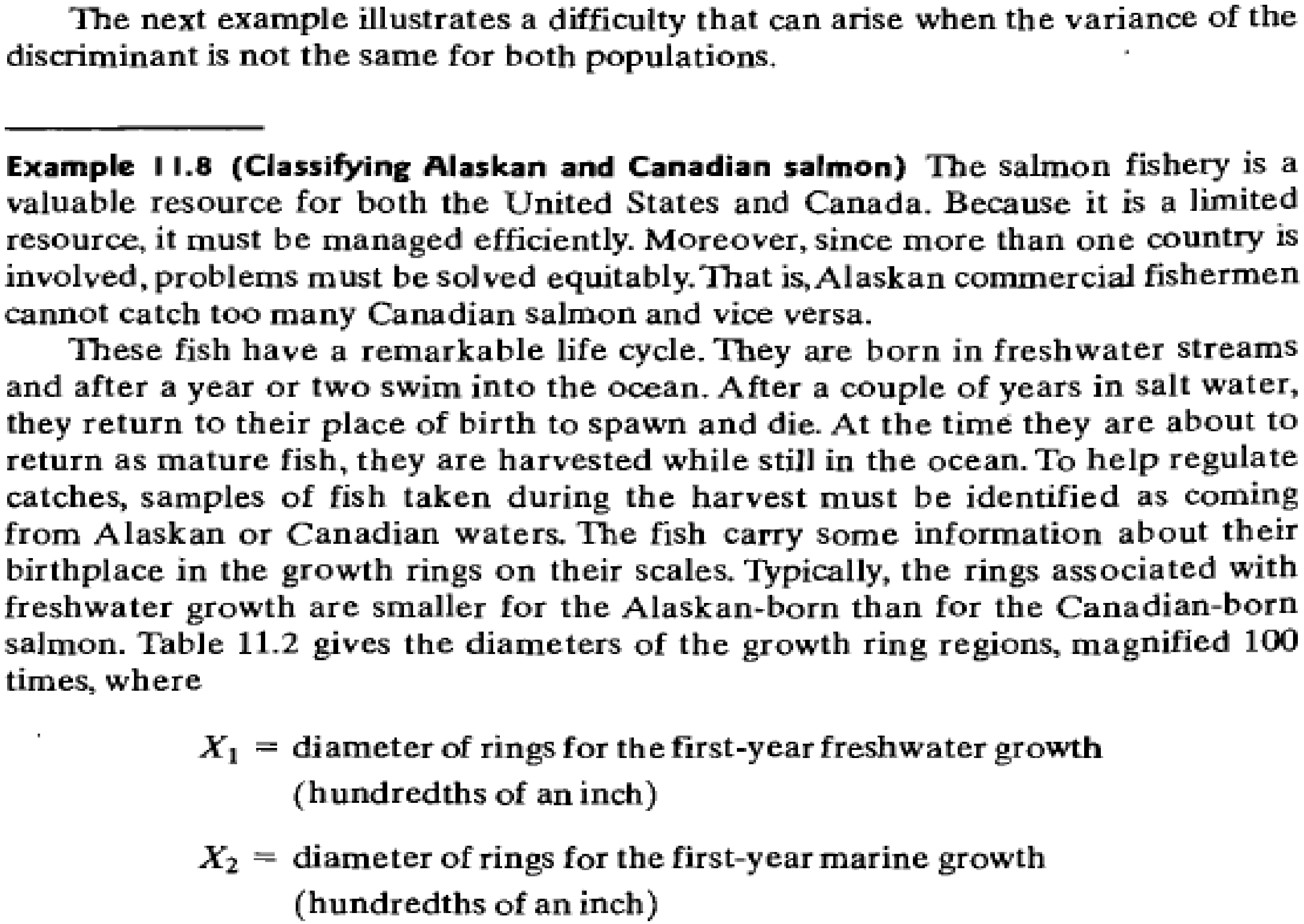
The APER is intuitively appealing and easy to calculate. Unfortunately, it tends to underestimate the AER, and the problem does not disappear unless the sample sizes $n_1$ and $n_2$ are very large. Essentially, this optimistic estimate occurs because the data used to build the classification function are also used to evaluate it.

Error-rate estimates can be constructed that are better than the apparent error rate, remain relatively easy to calculate, and do not require distributional assumptions. One procedure is to split the total sample into a training sample and a validation sample. The training sample is used to construct the classification function, and the validation sample is used to evaluate it. The error rate is determined by the proportion misclassified in the validation sample. Although this method overcomes the bias problem by not using the same data to both build and judge the classification function, it suffers from two main defects:

**(i)** It requires large samples.

**(ii)** The function evaluated is not the function of interest. Ultimately, almost *all* of the data must be used to construct the classification function. If not, valuable information may be lost.

# Example: Classifying Alaskan and Canadian Salmon



The next example illustrates a difficulty that can arise when the variance of the discriminant is not the same for both populations.

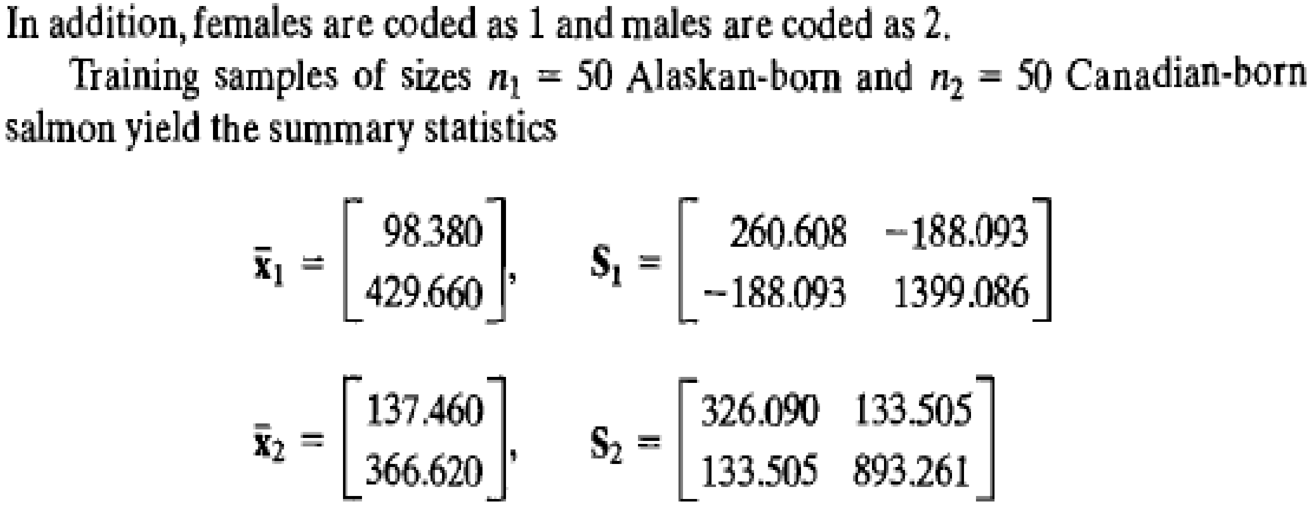
**Example 11.8 (Classifying Alaskan and Canadian salmon)** The salmon fishery is a valuable resource for both the United States and Canada. Because it is a limited resource, it must be managed efficiently. Moreover, since more than one country is involved, problems must be solved equitably. That is, Alaskan commercial fishermen cannot catch too many Canadian salmon and vice versa.

These fish have a remarkable life cycle. They are born in freshwater streams and after a year or two swim into the ocean. After a couple of years in salt water, they return to their place of birth to spawn and die. At the time they are about to return as mature fish, they are harvested while still in the ocean. To help regulate catches, samples of fish taken during the harvest must be identified as coming from Alaskan or Canadian waters. The fish carry some information about their birthplace in the growth rings on their scales. Typically, the rings associated with freshwater growth are smaller for the Alaskan-born than for the Canadian-born salmon. Table 11.2 gives the diameters of the growth ring regions, magnified 100 times, where

$X_1$ = diameter of rings for the first-year freshwater growth (hundredths of an inch)

$X_2$ = diameter of rings for the first-year marine growth (hundredths of an inch)

# Example: Classifying Alaskan and Canadian Salmon, Mean Vector and Covariance Matrices



In addition, females are coded as 1 and males are coded as 2.

Training samples of sizes $n_1 = 50$ Alaskan-born and $n_2 = 50$ Canadian-born salmon yield the summary statistics

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 98.380 \\ 429.660 \end{bmatrix}, \qquad \mathbf{S}_1 = \begin{bmatrix} 260.608 & -188.093 \\ -188.093 & 1399.086 \end{bmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 137.460 \\ 366.620 \end{bmatrix}, \qquad \mathbf{S}_2 = \begin{bmatrix} 326.090 & 133.505 \\ 133.505 & 893.261 \end{bmatrix}$$

# Example: Classifying Alaskan and Canadian Salmon (Data)

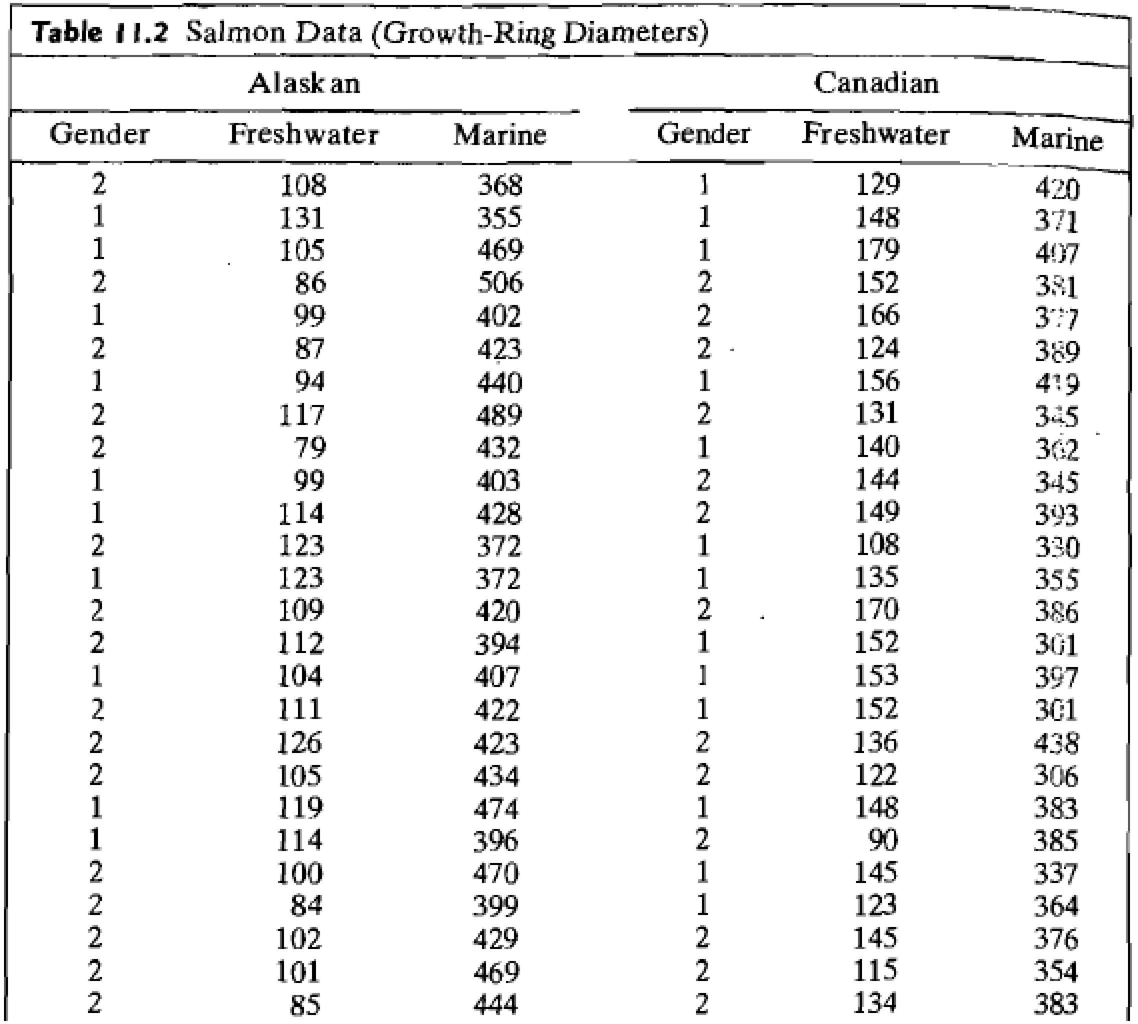
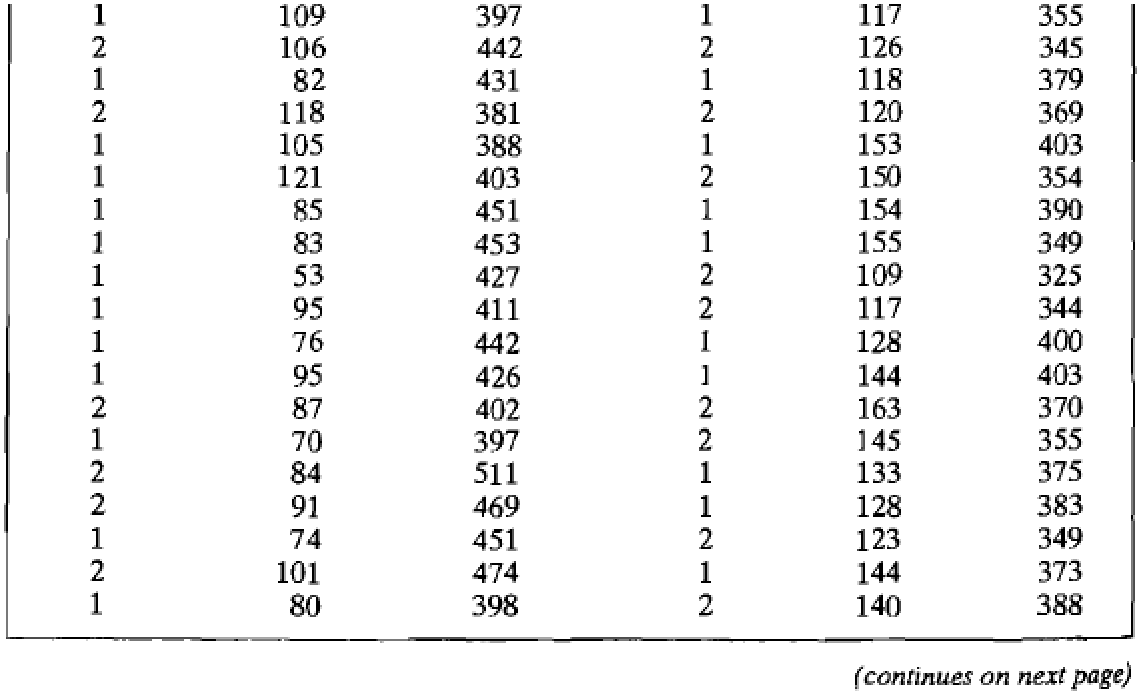


**Table 11.2** Salmon Data (Growth-Ring Diameters)

| Alaskan | | | Canadian | | |
|---|---|---|---|---|---|
| Gender | Freshwater | Marine | Gender | Freshwater | Marine |
| 2 | 108 | 368 | 1 | 129 | 420 |
| 1 | 131 | 355 | 1 | 148 | 371 |
| 1 | 105 | 469 | 1 | 179 | 407 |
| 2 | 86 | 506 | 2 | 152 | 381 |
| 1 | 99 | 402 | 2 | 166 | 377 |
| 2 | 87 | 423 | 2 | 124 | 389 |
| 1 | 94 | 440 | 1 | 156 | 419 |
| 2 | 117 | 489 | 2 | 131 | 345 |
| 2 | 79 | 432 | 1 | 140 | 362 |
| 1 | 99 | 403 | 2 | 144 | 345 |
| 1 | 114 | 428 | 2 | 149 | 393 |
| 2 | 123 | 372 | 1 | 108 | 330 |
| 1 | 123 | 372 | 1 | 135 | 355 |
| 2 | 109 | 420 | 2 | 170 | 386 |
| 2 | 112 | 394 | 1 | 152 | 301 |
| 1 | 104 | 407 | 1 | 153 | 397 |
| 2 | 111 | 422 | 1 | 152 | 301 |
| 2 | 126 | 423 | 2 | 136 | 438 |
| 2 | 105 | 434 | 2 | 122 | 306 |
| 1 | 119 | 474 | 1 | 148 | 383 |
| 1 | 114 | 396 | 2 | 90 | 385 |
| 2 | 100 | 470 | 1 | 145 | 337 |
| 2 | 84 | 399 | 1 | 123 | 364 |
| 2 | 102 | 429 | 2 | 145 | 376 |
| 2 | 101 | 469 | 2 | 115 | 354 |
| 2 | 85 | 444 | 2 | 134 | 383 |

# Example: Classifying Alaskan and Canadian Salmon (Data)



| | | | | | |
|---|---|---|---|---|---|
| 1 | 109 | 397 | 1 | 117 | 355 |
| 2 | 106 | 442 | 2 | 126 | 345 |
| 1 | 82 | 431 | 1 | 118 | 379 |
| 2 | 118 | 381 | 2 | 120 | 369 |
| 1 | 105 | 388 | 1 | 153 | 403 |
| 1 | 121 | 403 | 2 | 150 | 354 |
| 1 | 85 | 451 | 1 | 154 | 390 |
| 1 | 83 | 453 | 1 | 155 | 349 |
| 1 | 53 | 427 | 2 | 109 | 325 |
| 1 | 95 | 411 | 2 | 117 | 344 |
| 1 | 76 | 442 | 1 | 128 | 400 |
| 1 | 95 | 426 | 1 | 144 | 403 |
| 2 | 87 | 402 | 2 | 163 | 370 |
| 1 | 70 | 397 | 2 | 145 | 355 |
| 2 | 84 | 511 | 1 | 133 | 375 |
| 2 | 91 | 469 | 1 | 128 | 383 |
| 1 | 74 | 451 | 2 | 123 | 349 |
| 2 | 101 | 474 | 1 | 144 | 373 |
| 1 | 80 | 398 | 2 | 140 | 388 |

*(continues on next page)*

# Example: Classifying Alaskan and Canadian Salmon (Data)

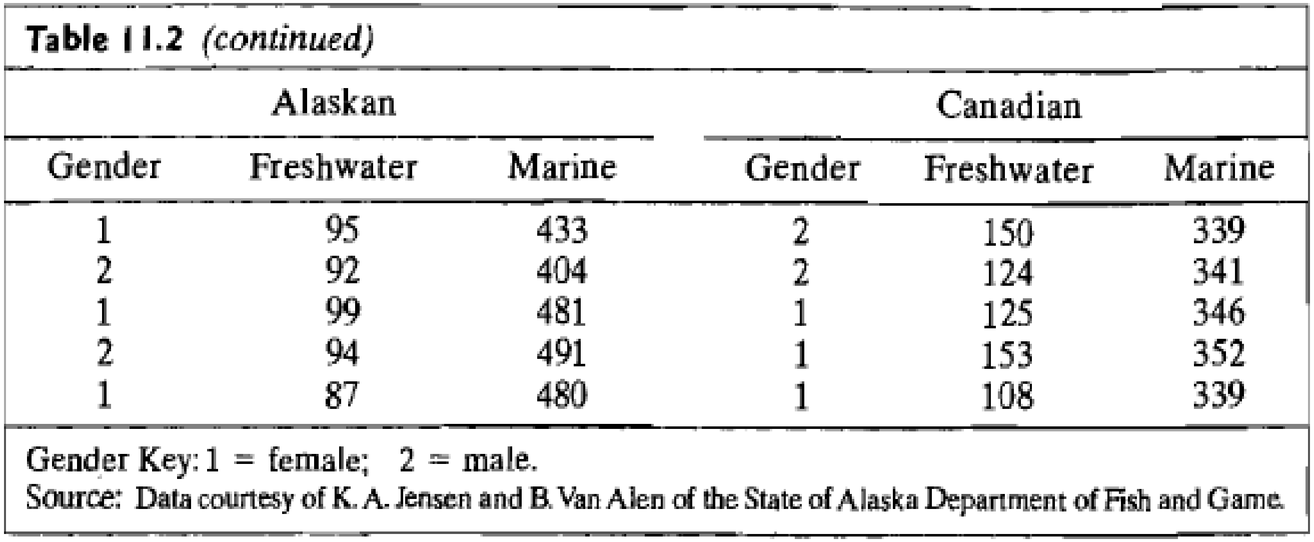
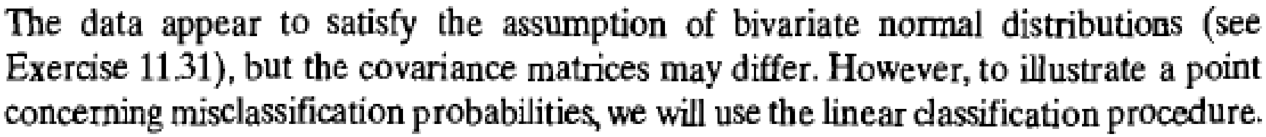


**Table 11.2** *(continued)*

| Alaskan |  |  | Canadian |  |  |
|---|---|---|---|---|---|
| Gender | Freshwater | Marine | Gender | Freshwater | Marine |
| 1 | 95 | 433 | 2 | 150 | 339 |
| 2 | 92 | 404 | 2 | 124 | 341 |
| 1 | 99 | 481 | 1 | 125 | 346 |
| 2 | 94 | 491 | 1 | 153 | 352 |
| 1 | 87 | 480 | 1 | 108 | 339 |

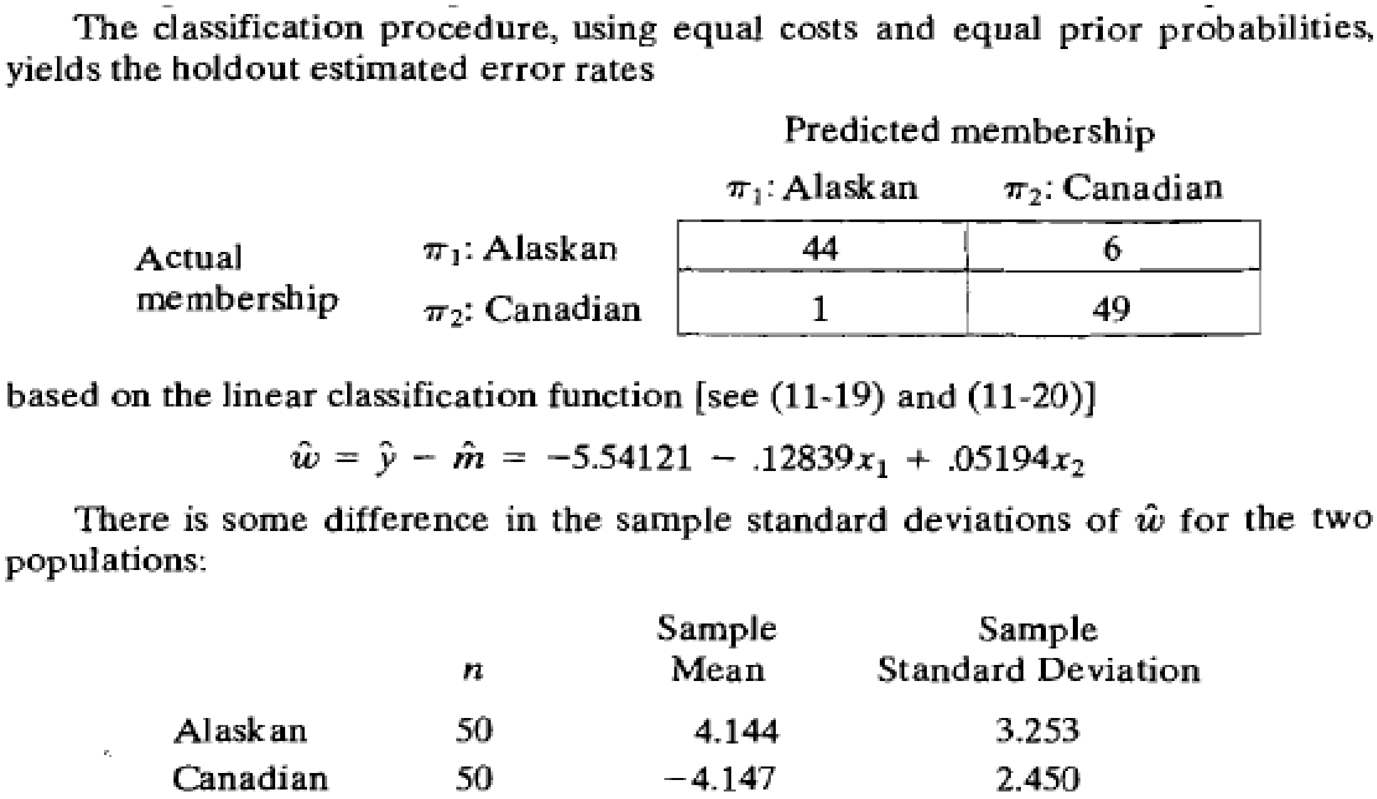
Gender Key: 1 = female; 2 = male.

Source: Data courtesy of K. A. Jensen and B. Van Alen of the State of Alaska Department of Fish and Game.



The data appear to satisfy the assumption of bivariate normal distributions (see Exercise 11.31), but the covariance matrices may differ. However, to illustrate a point concerning misclassification probabilities, we will use the linear classification procedure.

# Error Rates Estimated by Costs and Prior Probabilities



The classification procedure, using equal costs and equal prior probabilities, yields the holdout estimated error rates

| | | Predicted membership | |
|---|---|---|---|
| | | $\pi_1$: Alaskan | $\pi_2$: Canadian |
| Actual membership | $\pi_1$: Alaskan | 44 | 6 |
| | $\pi_2$: Canadian | 1 | 49 |

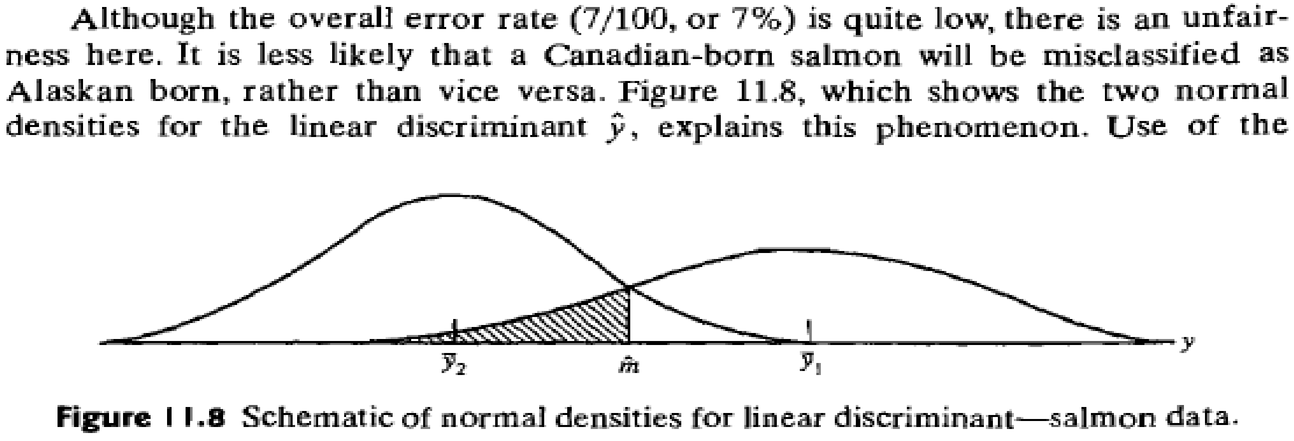
based on the linear classification function [see (11-19) and (11-20)]

$$\hat{w} = \hat{y} - \hat{m} = -5.54121 - .12839x_1 + .05194x_2$$

There is some difference in the sample standard deviations of $\hat{w}$ for the two populations:

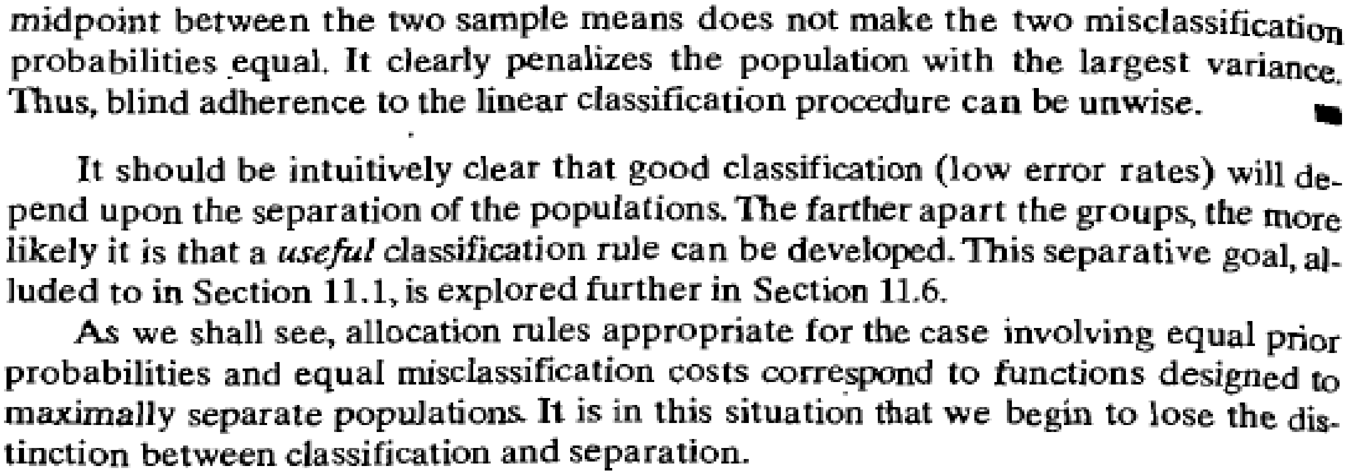
| | $n$ | Sample Mean | Sample Standard Deviation |
|---|---|---|---|
| Alaskan | 50 | 4.144 | 3.253 |
| Canadian | 50 | −4.147 | 2.450 |

# Normal Densities for Linear Discriminant Salmon Data



Although the overall error rate (7/100, or 7%) is quite low, there is an unfairness here. It is less likely that a Canadian-born salmon will be misclassified as Alaskan born, rather than vice versa. Figure 11.8, which shows the two normal densities for the linear discriminant $\hat{y}$, explains this phenomenon. Use of the midpoint between the two sample means does not make the two misclassification probabilities equal. It clearly penalizes the population with the largest variance. Thus, blind adherence to the linear classification procedure can be unwise. ■

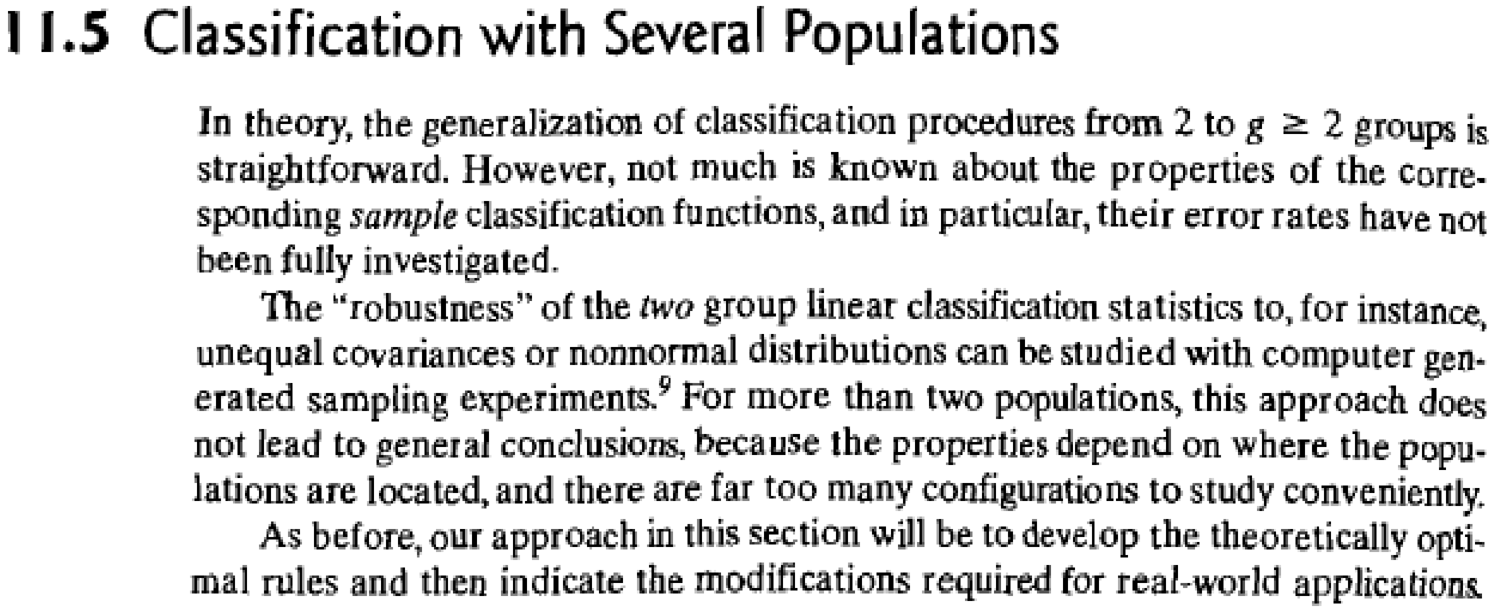
**Figure 11.8** Schematic of normal densities for linear discriminant—salmon data.



It should be intuitively clear that good classification (low error rates) will depend upon the separation of the populations. The farther apart the groups, the more likely it is that a *useful* classification rule can be developed. This separative goal, alluded to in Section 11.1, is explored further in Section 11.6.

As we shall see, allocation rules appropriate for the case involving equal prior probabilities and equal misclassification costs correspond to functions designed to maximally separate populations. It is in this situation that we begin to lose the distinction between classification and separation.

# Classification with Several Populations



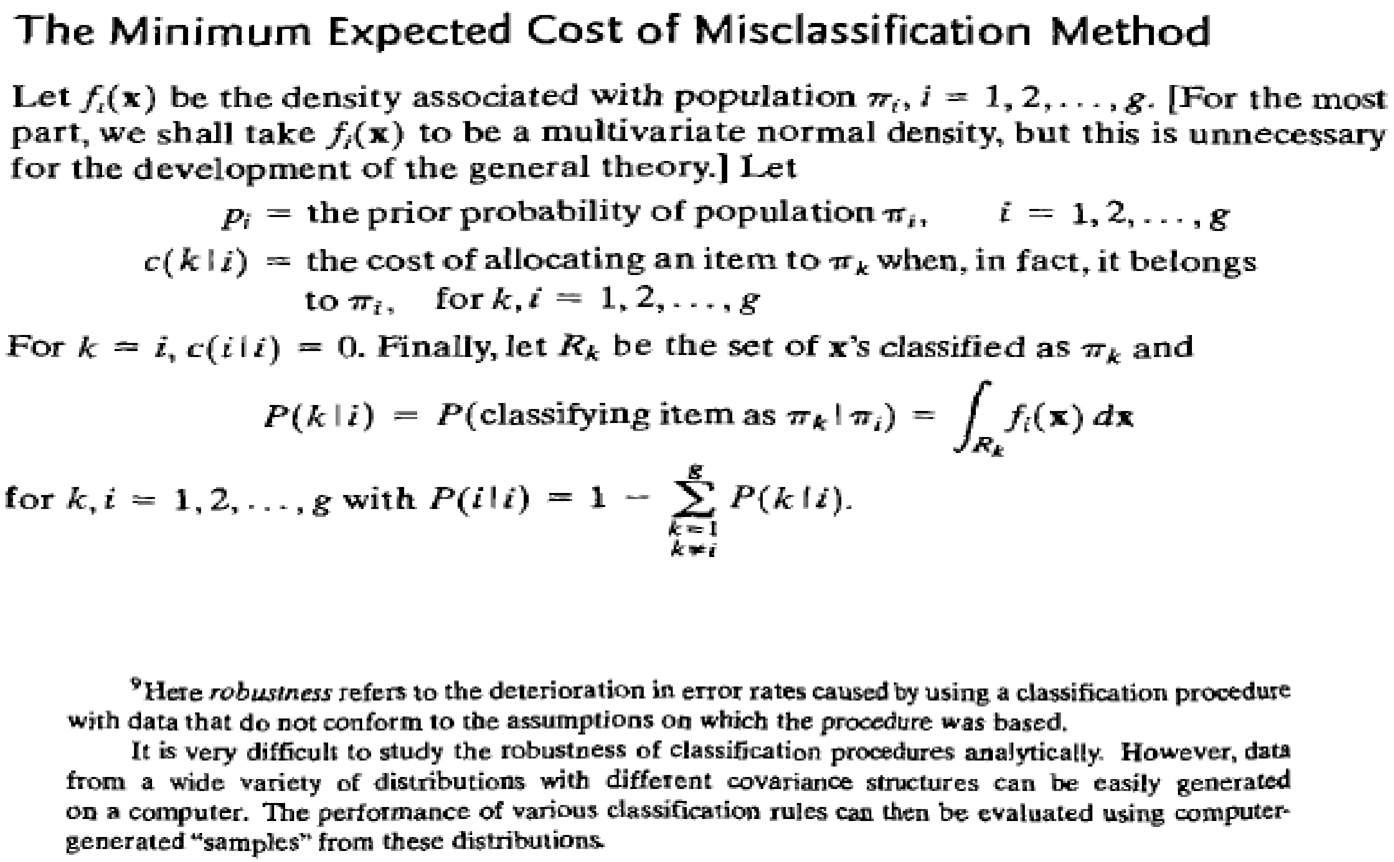
## 11.5 Classification with Several Populations

In theory, the generalization of classification procedures from 2 to $g \geq 2$ groups is straightforward. However, not much is known about the properties of the corresponding *sample* classification functions, and in particular, their error rates have not been fully investigated.

The "robustness" of the *two* group linear classification statistics to, for instance, unequal covariances or nonnormal distributions can be studied with computer generated sampling experiments.[9] For more than two populations, this approach does not lead to general conclusions, because the properties depend on where the populations are located, and there are far too many configurations to study conveniently.

As before, our approach in this section will be to develop the theoretically optimal rules and then indicate the modifications required for real-world applications.

# Minimum Expected Cost of Misclassification Method



## The Minimum Expected Cost of Misclassification Method

Let $f_i(\mathbf{x})$ be the density associated with population $\pi_i$, $i = 1, 2, \ldots, g$. [For the most part, we shall take $f_i(\mathbf{x})$ to be a multivariate normal density, but this is unnecessary for the development of the general theory.] Let

$$p_i = \text{the prior probability of population } \pi_i, \qquad i = 1, 2, \ldots, g$$

$$c(k \mid i) = \text{the cost of allocating an item to } \pi_k \text{ when, in fact, it belongs to } \pi_i, \quad \text{for } k, i = 1, 2, \ldots, g$$

For $k = i$, $c(i \mid i) = 0$. Finally, let $R_k$ be the set of $\mathbf{x}$'s classified as $\pi_k$ and

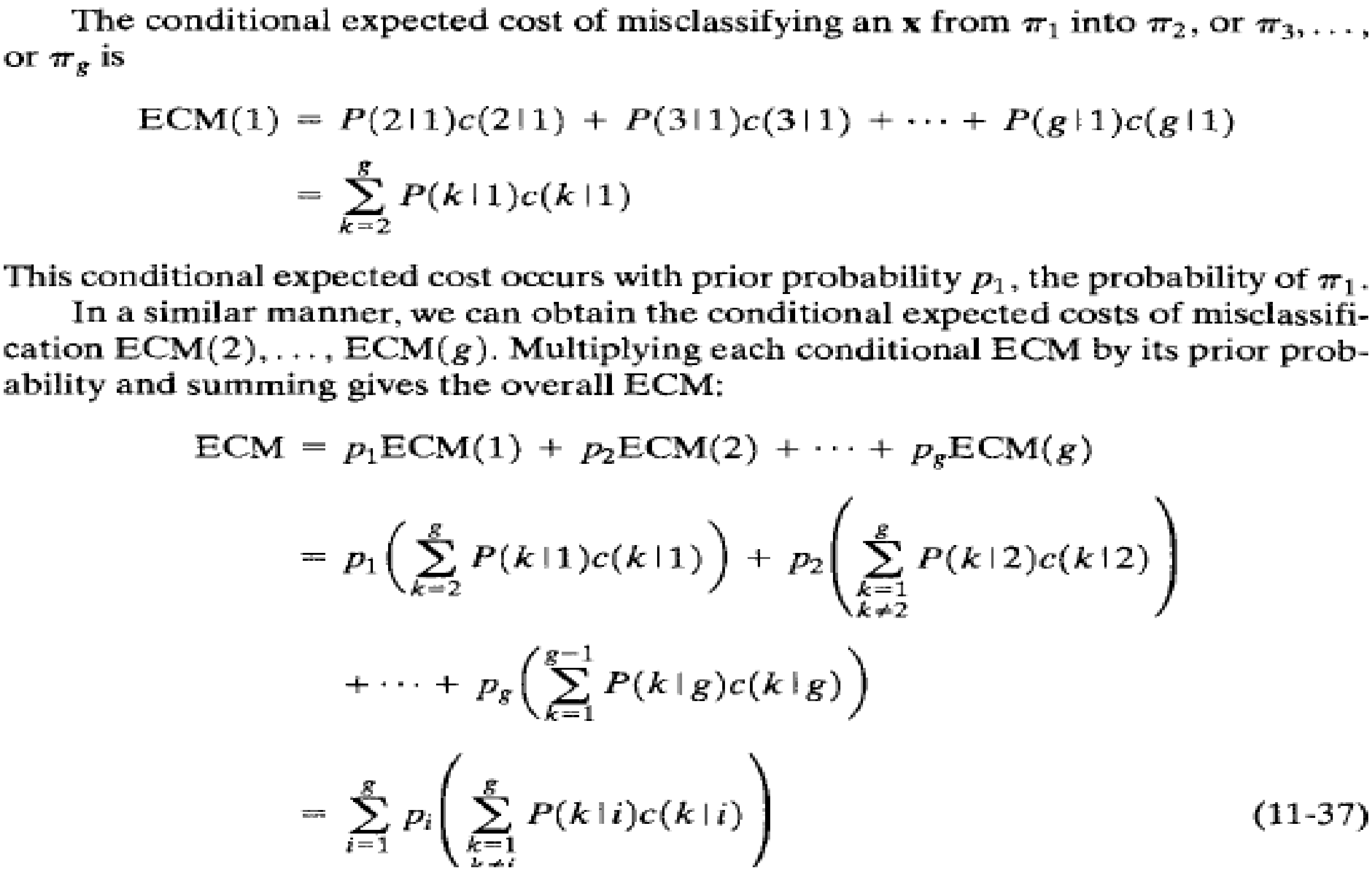
$$P(k \mid i) = P(\text{classifying item as } \pi_k \mid \pi_i) = \int_{R_k} f_i(\mathbf{x})\, d\mathbf{x}$$

for $k, i = 1, 2, \ldots, g$ with $P(i \mid i) = 1 - \sum_{\substack{k=1 \\ k \neq i}}^{g} P(k \mid i)$.

[9] Here *robustness* refers to the deterioration in error rates caused by using a classification procedure with data that do not conform to the assumptions on which the procedure was based.

It is very difficult to study the robustness of classification procedures analytically. However, data from a wide variety of distributions with different covariance structures can be easily generated on a computer. The performance of various classification rules can then be evaluated using computer-generated "samples" from these distributions.

# Expected Costs of Misclassification (ECM) Multiplied by Prior Probabilities and Summing Yields Over ECM



The conditional expected cost of misclassifying an $\mathbf{x}$ from $\pi_1$ into $\pi_2$, or $\pi_3, \ldots,$ or $\pi_g$ is

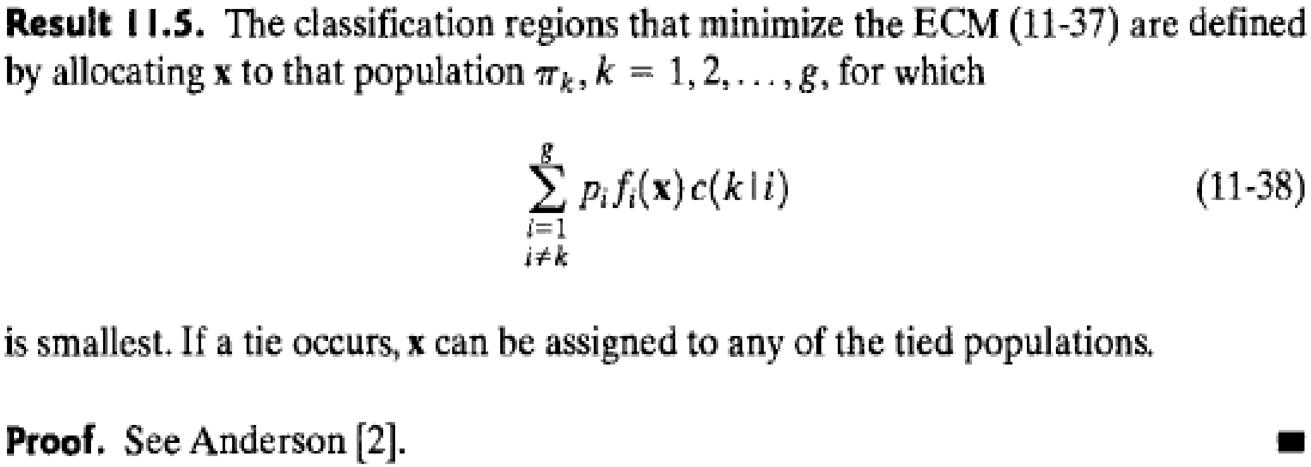
$$\begin{aligned}\text{ECM}(1) &= P(2|1)c(2|1) + P(3|1)c(3|1) + \cdots + P(g|1)c(g|1)\\ &= \sum_{k=2}^{g} P(k|1)c(k|1)\end{aligned}$$

This conditional expected cost occurs with prior probability $p_1$, the probability of $\pi_1$.

In a similar manner, we can obtain the conditional expected costs of misclassification $\text{ECM}(2), \ldots, \text{ECM}(g)$. Multiplying each conditional ECM by its prior probability and summing gives the overall ECM:

$$\begin{aligned}\text{ECM} &= p_1\text{ECM}(1) + p_2\text{ECM}(2) + \cdots + p_g\text{ECM}(g)\\ &= p_1\left(\sum_{k=2}^{g} P(k|1)c(k|1)\right) + p_2\left(\sum_{\substack{k=1\\k\neq 2}}^{g} P(k|2)c(k|2)\right)\\ &\quad + \cdots + p_g\left(\sum_{k=1}^{g-1} P(k|g)c(k|g)\right)\\ &= \sum_{i=1}^{g} p_i\left(\sum_{\substack{k=1\\k\neq i}}^{g} P(k|i)c(k|i)\right) \qquad (11\text{-}37)\end{aligned}$$

# Goal: Minimizing the ECM (Less Overlapping Region of the Two Groups)



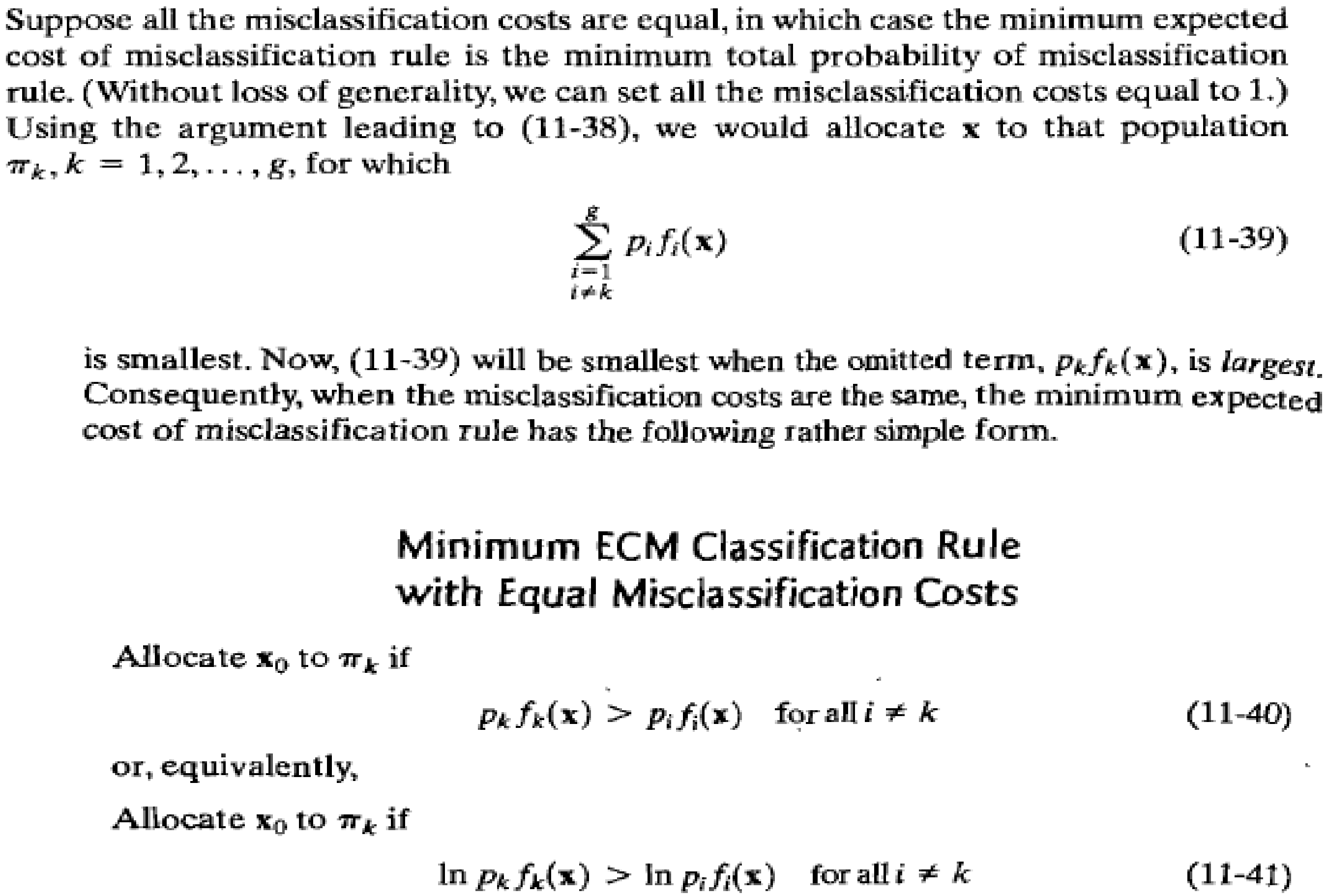
**Result 11.5.** The classification regions that minimize the ECM (11-37) are defined by allocating $\mathbf{x}$ to that population $\pi_k$, $k = 1, 2, \ldots, g$, for which

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(\mathbf{x}) c(k \mid i) \tag{11-38}$$

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

**Proof.** See Anderson [2]. ■

# Minimum ECM Classification Rule with Equal Misclassification Costs



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. (Without loss of generality, we can set all the misclassification costs equal to 1.) Using the argument leading to (11-38), we would allocate $\mathbf{x}$ to that population $\pi_k, k = 1, 2, \ldots, g$, for which

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(\mathbf{x}) \tag{11-39}$$

is smallest. Now, (11-39) will be smallest when the omitted term, $p_k f_k(\mathbf{x})$, is *largest*. Consequently, when the misclassification costs are the same, the minimum expected cost of misclassification rule has the following rather simple form.

### Minimum ECM Classification Rule with Equal Misclassification Costs

Allocate $\mathbf{x}_0$ to $\pi_k$ if

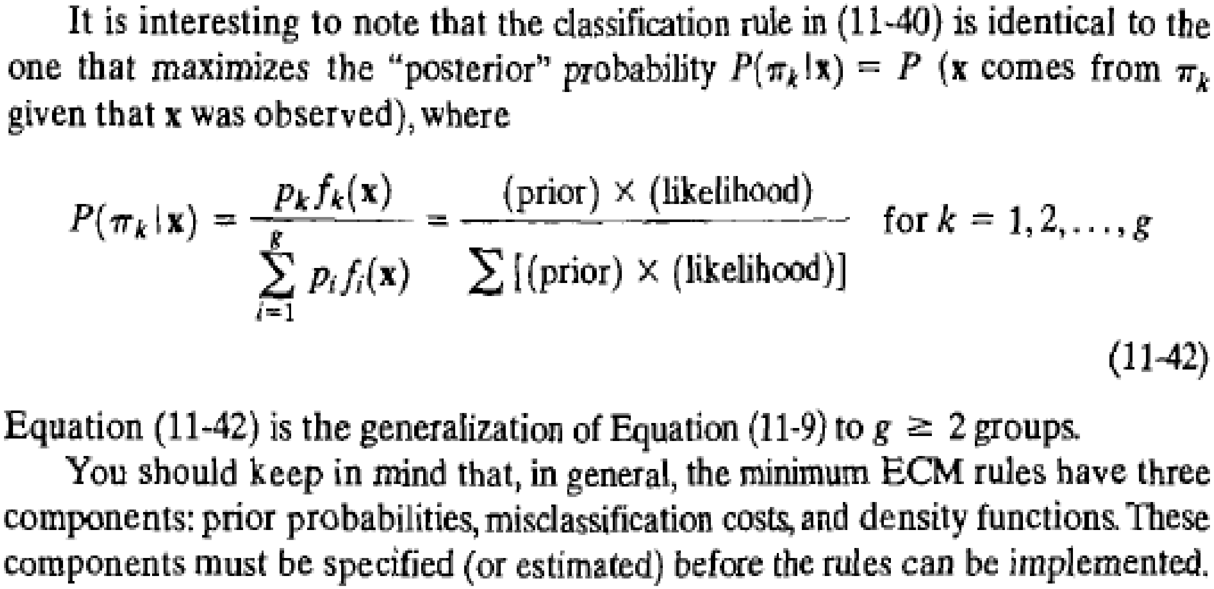
$$p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}) \quad \text{for all } i \neq k \tag{11-40}$$

or, equivalently,

Allocate $\mathbf{x}_0$ to $\pi_k$ if

$$\ln p_k f_k(\mathbf{x}) > \ln p_i f_i(\mathbf{x}) \quad \text{for all } i \neq k \tag{11-41}$$

# Three ECM Rules: Prior Probabilities, Misclassification Costs and Density Functions



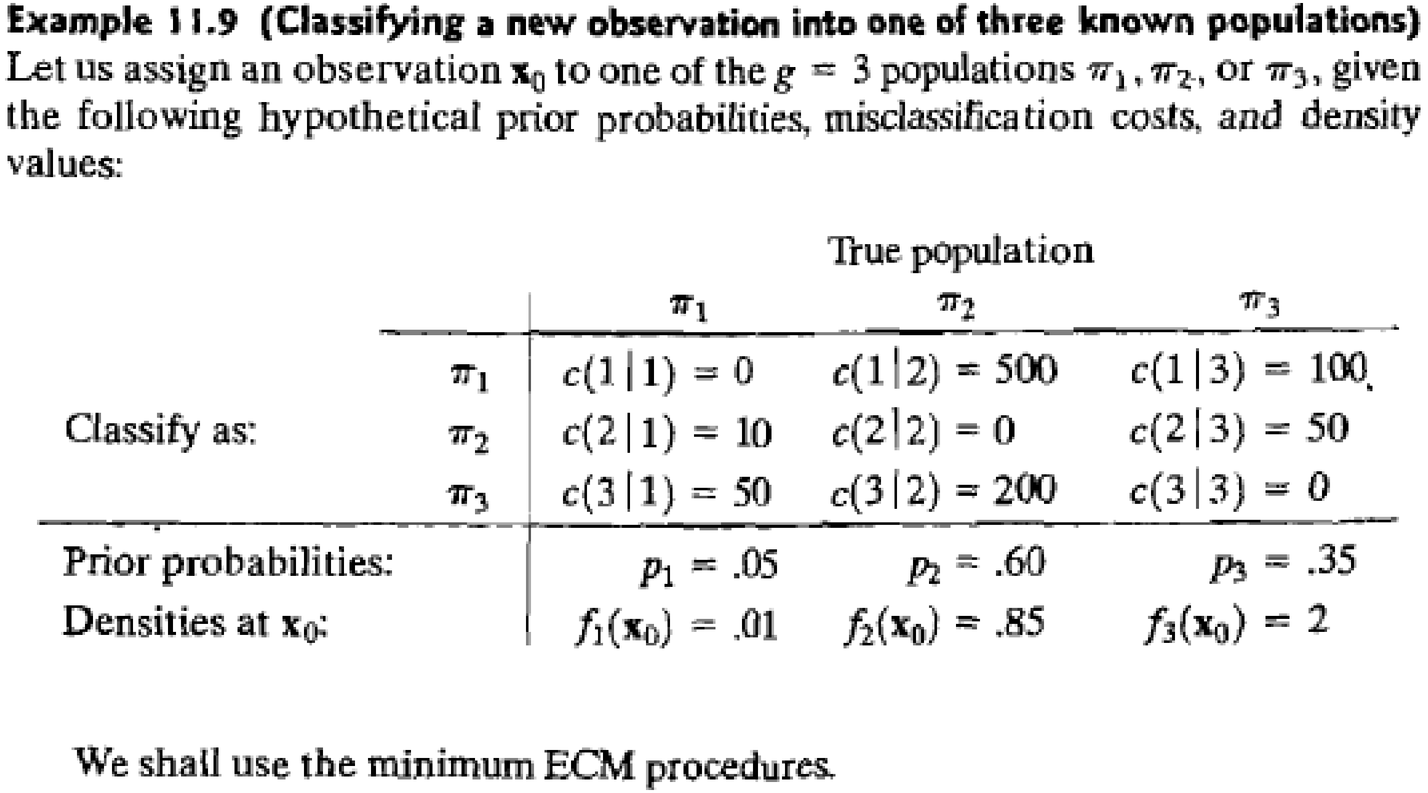
It is interesting to note that the classification rule in (11-40) is identical to the one that maximizes the "posterior" probability $P(\pi_k|\mathbf{x}) = P$ ($\mathbf{x}$ comes from $\pi_k$ given that $\mathbf{x}$ was observed), where

$$P(\pi_k|\mathbf{x}) = \frac{p_k f_k(\mathbf{x})}{\sum_{i=1}^{g} p_i f_i(\mathbf{x})} = \frac{(\text{prior}) \times (\text{likelihood})}{\sum[(\text{prior}) \times (\text{likelihood})]} \quad \text{for } k = 1, 2, \ldots, g \tag{11-42}$$

Equation (11-42) is the generalization of Equation (11-9) to $g \geq 2$ groups.

You should keep in mind that, in general, the minimum ECM rules have three components: prior probabilities, misclassification costs, and density functions. These components must be specified (or estimated) before the rules can be implemented.

# Classifying New Observation into One of Three Known Populations

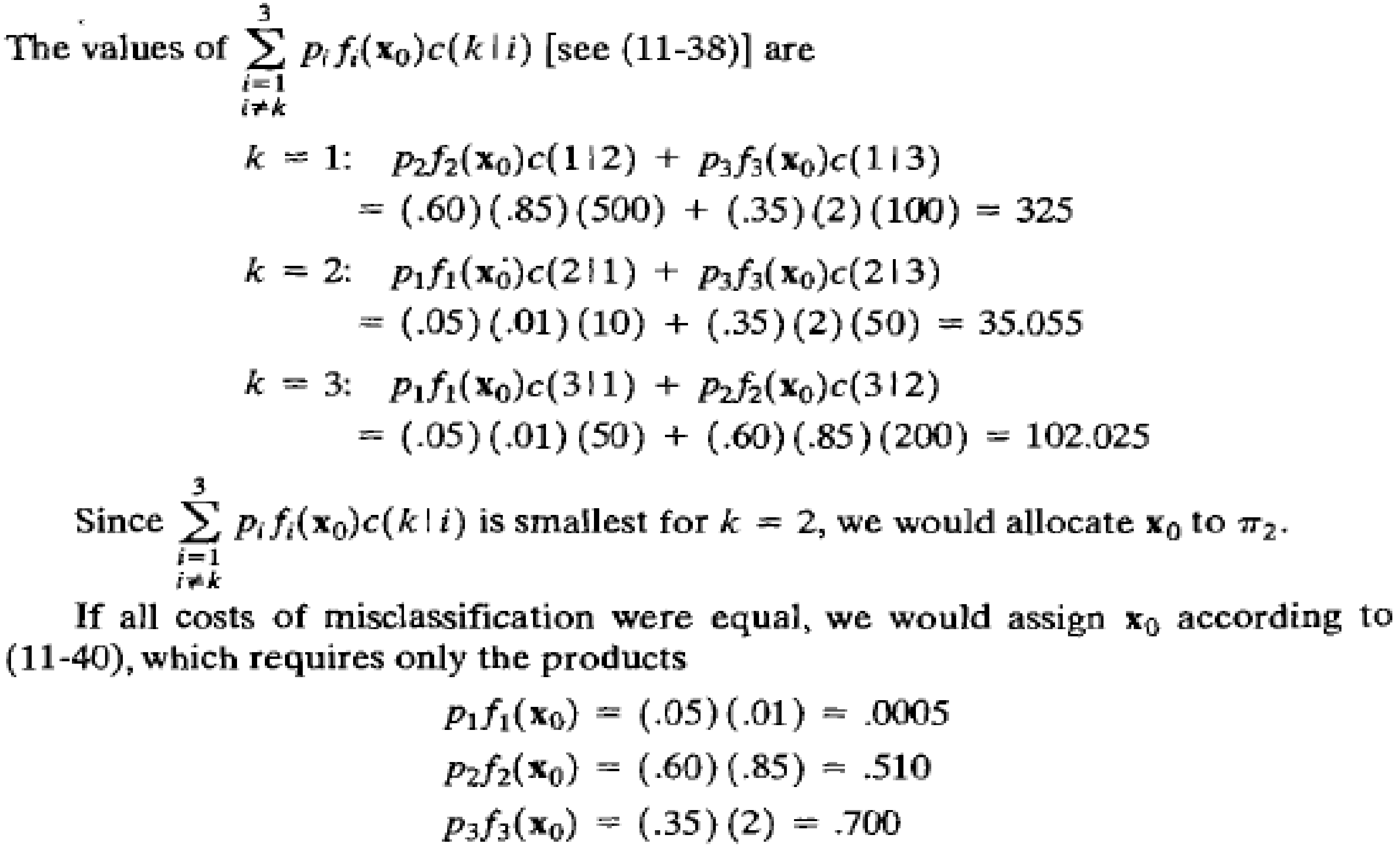


**Example 11.9 (Classifying a new observation into one of three known populations)** Let us assign an observation $\mathbf{x}_0$ to one of the $g = 3$ populations $\pi_1, \pi_2$, or $\pi_3$, given the following hypothetical prior probabilities, misclassification costs, and density values:

| | | True population | | |
|---|---|---|---|---|
| | | $\pi_1$ | $\pi_2$ | $\pi_3$ |
| | $\pi_1$ | $c(1\|1) = 0$ | $c(1\|2) = 500$ | $c(1\|3) = 100$ |
| Classify as: | $\pi_2$ | $c(2\|1) = 10$ | $c(2\|2) = 0$ | $c(2\|3) = 50$ |
| | $\pi_3$ | $c(3\|1) = 50$ | $c(3\|2) = 200$ | $c(3\|3) = 0$ |
| Prior probabilities: | | $p_1 = .05$ | $p_2 = .60$ | $p_3 = .35$ |
| Densities at $\mathbf{x}_0$: | | $f_1(\mathbf{x}_0) = .01$ | $f_2(\mathbf{x}_0) = .85$ | $f_3(\mathbf{x}_0) = 2$ |

We shall use the minimum ECM procedures.

# Classifying New Observation into One of Three Known Populations



The values of $\sum_{\substack{i=1\\i\neq k}}^{3} p_i f_i(\mathbf{x}_0)c(k\,|\,i)$ [see (11-38)] are

$$k = 1: \quad p_2 f_2(\mathbf{x}_0)c(1\,|\,2) + p_3 f_3(\mathbf{x}_0)c(1\,|\,3)$$
$$= (.60)(.85)(500) + (.35)(2)(100) = 325$$

$$k = 2: \quad p_1 f_1(\mathbf{x}_0)c(2\,|\,1) + p_3 f_3(\mathbf{x}_0)c(2\,|\,3)$$
$$= (.05)(.01)(10) + (.35)(2)(50) = 35.055$$

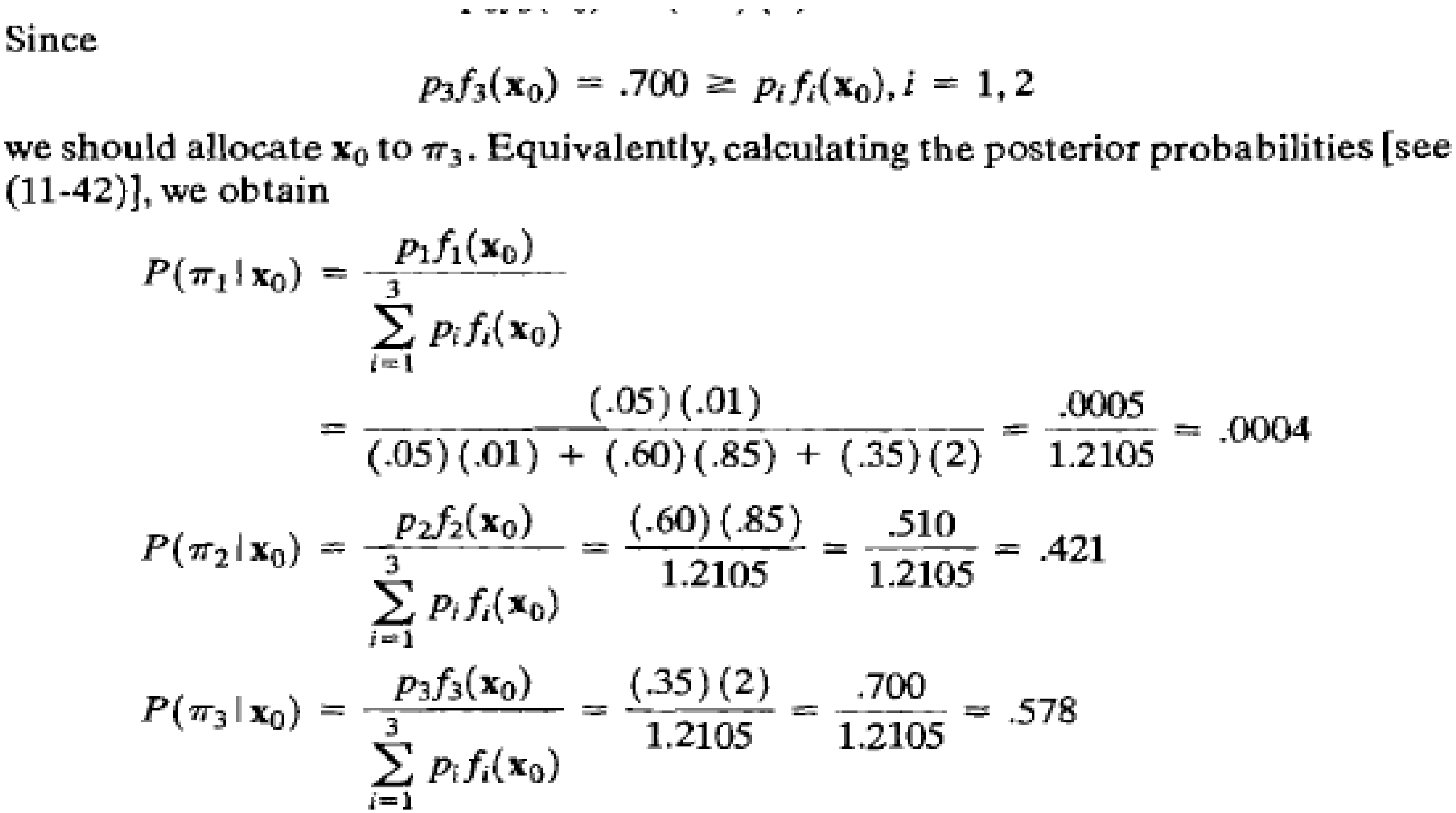
$$k = 3: \quad p_1 f_1(\mathbf{x}_0)c(3\,|\,1) + p_2 f_2(\mathbf{x}_0)c(3\,|\,2)$$
$$= (.05)(.01)(50) + (.60)(.85)(200) = 102.025$$

Since $\sum_{\substack{i=1\\i\neq k}}^{3} p_i f_i(\mathbf{x}_0)c(k\,|\,i)$ is smallest for $k = 2$, we would allocate $\mathbf{x}_0$ to $\pi_2$.

If all costs of misclassification were equal, we would assign $\mathbf{x}_0$ according to (11-40), which requires only the products

$$p_1 f_1(\mathbf{x}_0) = (.05)(.01) = .0005$$
$$p_2 f_2(\mathbf{x}_0) = (.60)(.85) = .510$$
$$p_3 f_3(\mathbf{x}_0) = (.35)(2) = .700$$

# Classifying New Observation into One of Three Known Populations



Since

$$p_3 f_3(\mathbf{x}_0) = .700 \geq p_i f_i(\mathbf{x}_0), i = 1, 2$$

we should allocate $\mathbf{x}_0$ to $\pi_3$. Equivalently, calculating the posterior probabilities [see (11-42)], we obtain

$$P(\pi_1 | \mathbf{x}_0) = \frac{p_1 f_1(\mathbf{x}_0)}{\sum_{i=1}^{3} p_i f_i(\mathbf{x}_0)}$$

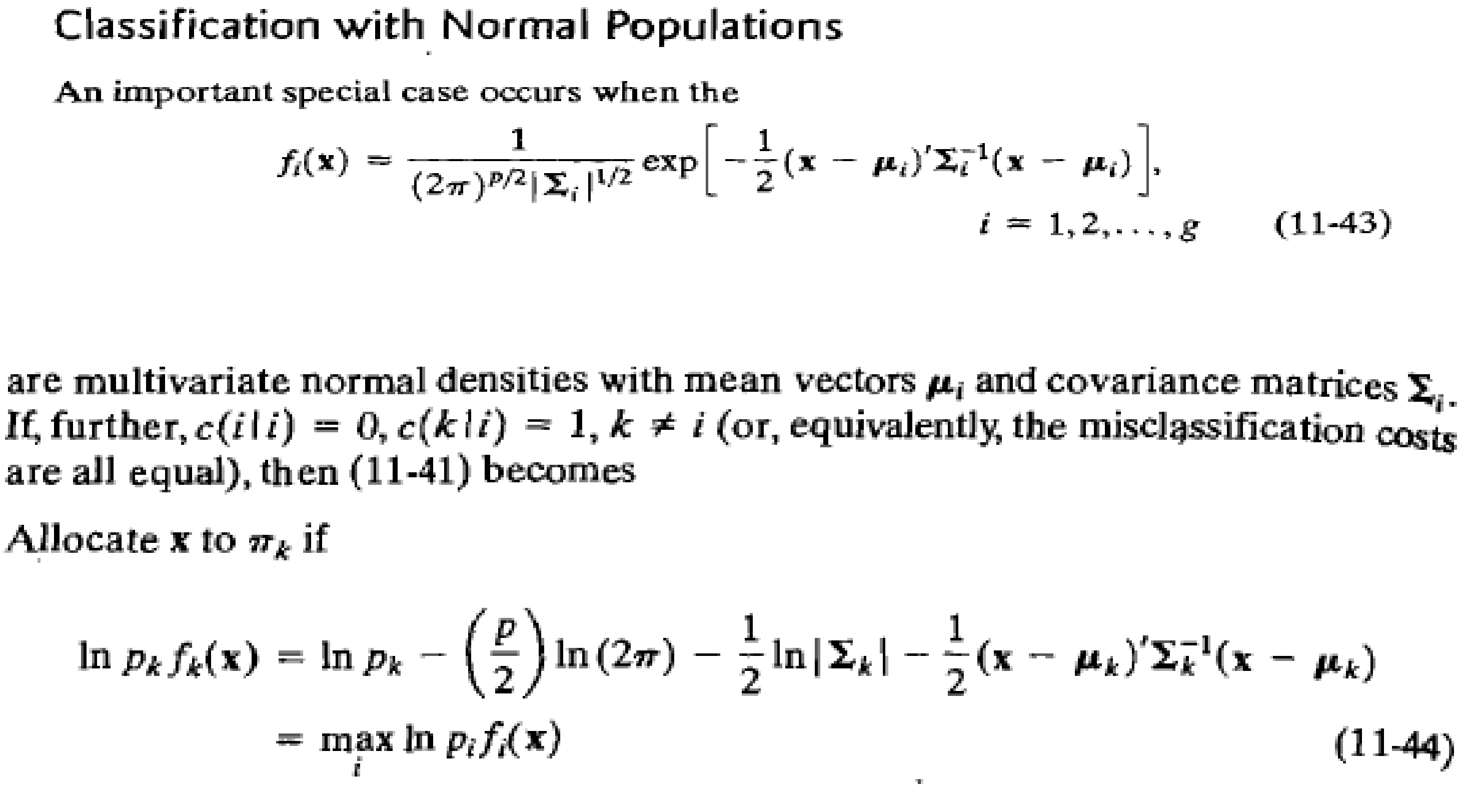
$$= \frac{(.05)(.01)}{(.05)(.01) + (.60)(.85) + (.35)(2)} = \frac{.0005}{1.2105} = .0004$$

$$P(\pi_2 | \mathbf{x}_0) = \frac{p_2 f_2(\mathbf{x}_0)}{\sum_{i=1}^{3} p_i f_i(\mathbf{x}_0)} = \frac{(.60)(.85)}{1.2105} = \frac{.510}{1.2105} = .421$$

$$P(\pi_3 | \mathbf{x}_0) = \frac{p_3 f_3(\mathbf{x}_0)}{\sum_{i=1}^{3} p_i f_i(\mathbf{x}_0)} = \frac{(.35)(2)}{1.2105} = \frac{.700}{1.2105} = .578$$

We see that $\mathbf{x}_0$ is allocated to $\pi_3$, the population with the largest posterior probability. ■

# Classification with Normal Populations



## Classification with Normal Populations

An important special case occurs when the

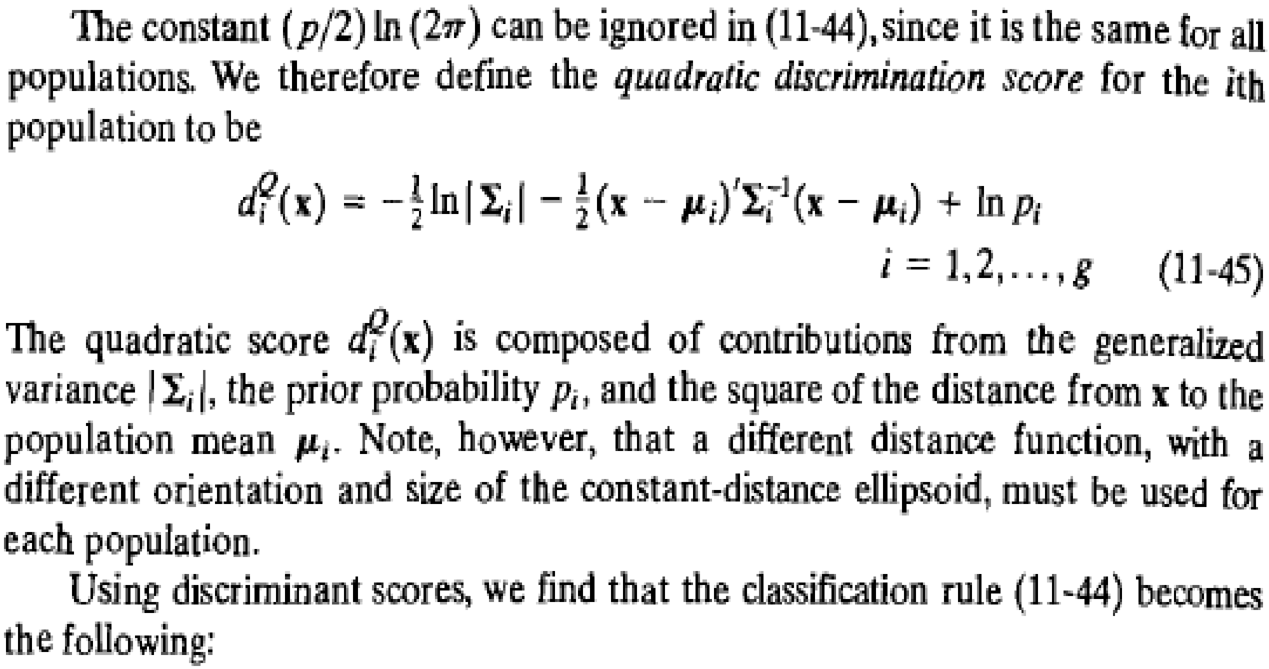
$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right], \qquad i = 1, 2, \ldots, g \tag{11-43}$$

are multivariate normal densities with mean vectors $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$. If, further, $c(i|i) = 0$, $c(k|i) = 1$, $k \neq i$ (or, equivalently, the misclassification costs are all equal), then (11-41) becomes

Allocate $\mathbf{x}$ to $\pi_k$ if

$$\begin{aligned} \ln p_k f_k(\mathbf{x}) &= \ln p_k - \left(\frac{p}{2}\right)\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) \\ &= \max_i \ln p_i f_i(\mathbf{x}) \end{aligned} \tag{11-44}$$

# Quadratic Score Composed of Contributions from ∑, Prior Probability and Square of Distance From x to the μ



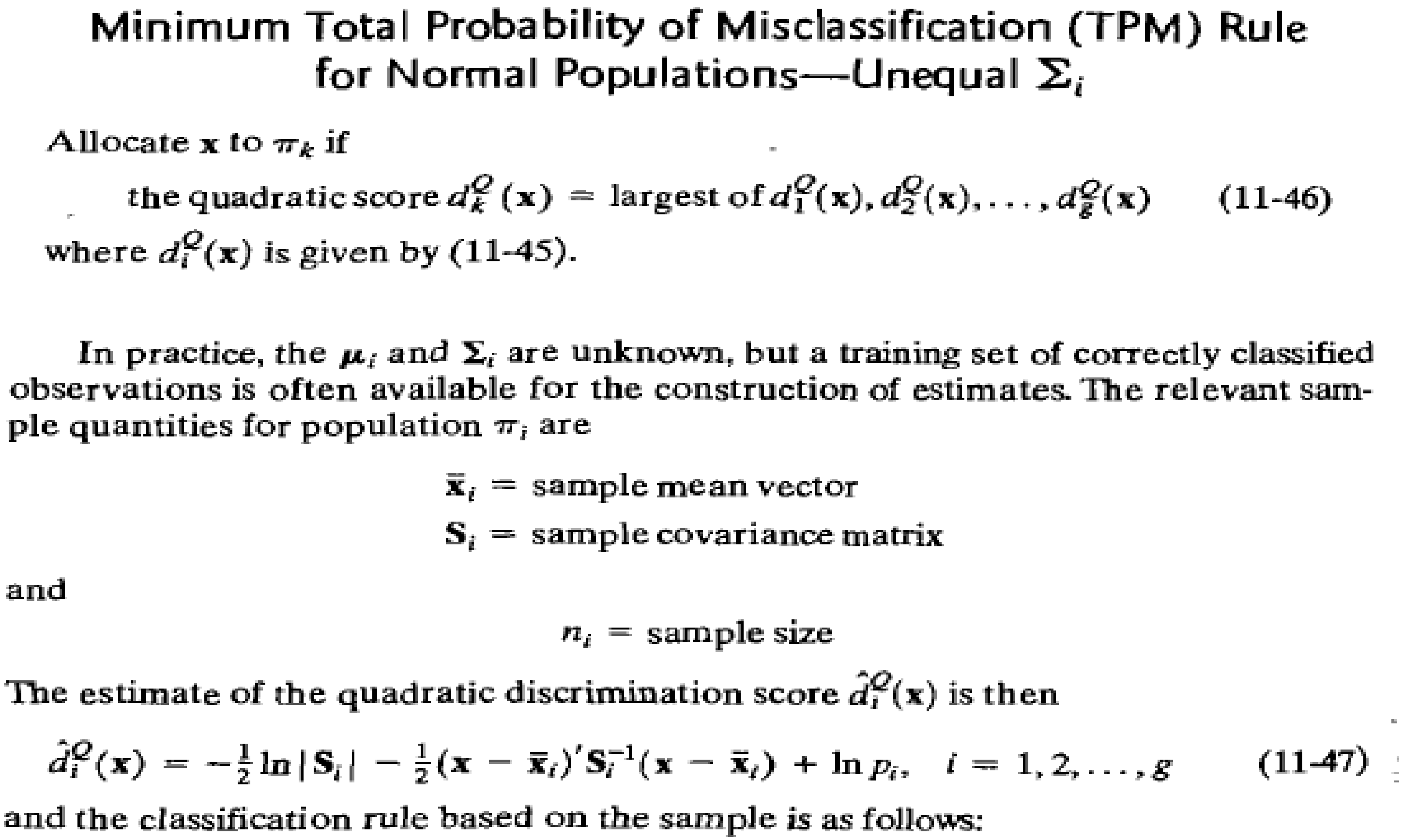
The constant $(p/2)\ln(2\pi)$ can be ignored in (11-44), since it is the same for all populations. We therefore define the *quadratic discrimination score* for the $i$th population to be

$$d_i^Q(\mathbf{x}) = -\tfrac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln p_i \qquad i = 1,2,\ldots,g \qquad (11\text{-}45)$$

The quadratic score $d_i^Q(\mathbf{x})$ is composed of contributions from the generalized variance $|\boldsymbol{\Sigma}_i|$, the prior probability $p_i$, and the square of the distance from $\mathbf{x}$ to the population mean $\boldsymbol{\mu}_i$. Note, however, that a different distance function, with a different orientation and size of the constant-distance ellipsoid, must be used for each population.

Using discriminant scores, we find that the classification rule (11-44) becomes the following:

# Minimum Total Probability of Misclassification Rule for Normal Populations: Unequal Variances



## Minimum Total Probability of Misclassification (TPM) Rule for Normal Populations—Unequal $\Sigma_i$

Allocate $\mathbf{x}$ to $\pi_k$ if

$$\text{the quadratic score } d_k^Q(\mathbf{x}) = \text{largest of } d_1^Q(\mathbf{x}), d_2^Q(\mathbf{x}), \ldots, d_g^Q(\mathbf{x}) \qquad (11\text{-}46)$$

where $d_i^Q(\mathbf{x})$ is given by (11-45).

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 $\pi_i$ are

$$\bar{\mathbf{x}}_i = \text{sample mean vector}$$
$$\mathbf{S}_i = \text{sample covariance matrix}$$

and

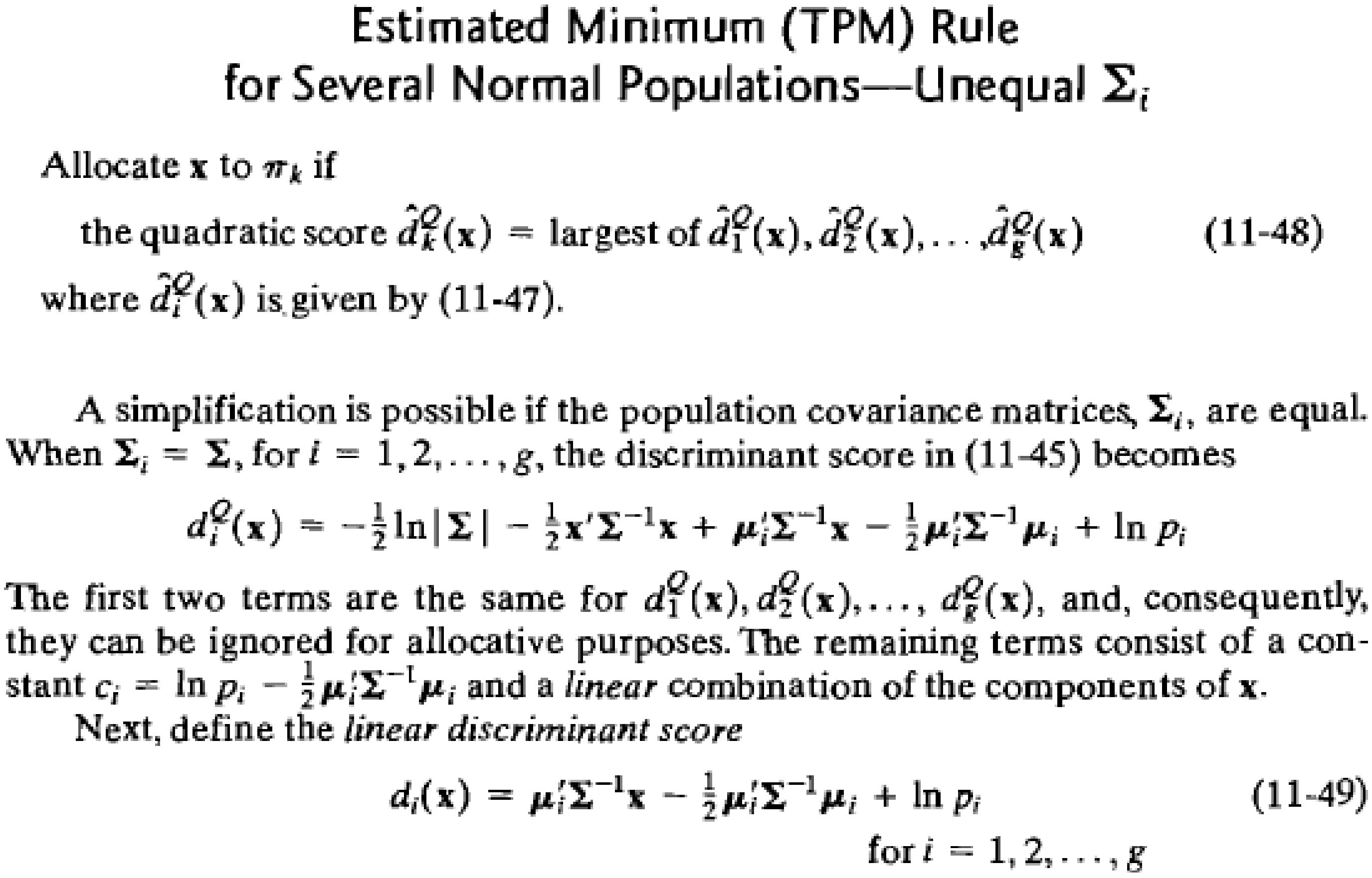
$$n_i = \text{sample size}$$

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

$$\hat{d}_i^Q(\mathbf{x}) = -\tfrac{1}{2}\ln|\mathbf{S}_i| - \tfrac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i) + \ln p_i, \quad i = 1, 2, \ldots, g \qquad (11\text{-}47)$$

and the classification rule based on the sample is as follows:

# Minimum Total Probability of Misclassification Rule for Normal Populations: Unequal Variances



### Estimated Minimum (TPM) Rule for Several Normal Populations—Unequal $\Sigma_i$

Allocate $\mathbf{x}$ to $\pi_k$ if

$$\text{the quadratic score } \hat{d}_k^Q(\mathbf{x}) = \text{largest of } \hat{d}_1^Q(\mathbf{x}), \hat{d}_2^Q(\mathbf{x}), \ldots, \hat{d}_g^Q(\mathbf{x}) \qquad (11\text{-}48)$$

where $\hat{d}_i^Q(\mathbf{x})$ is given by (11-47).

A simplification is possible if the population covariance matrices, $\boldsymbol{\Sigma}_i$, are equal. When $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$, for $i = 1, 2, \ldots, g$, the discriminant score in (11-45) becomes

$$d_i^Q(\mathbf{x}) = -\tfrac{1}{2}\ln|\boldsymbol{\Sigma}| - \tfrac{1}{2}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i$$

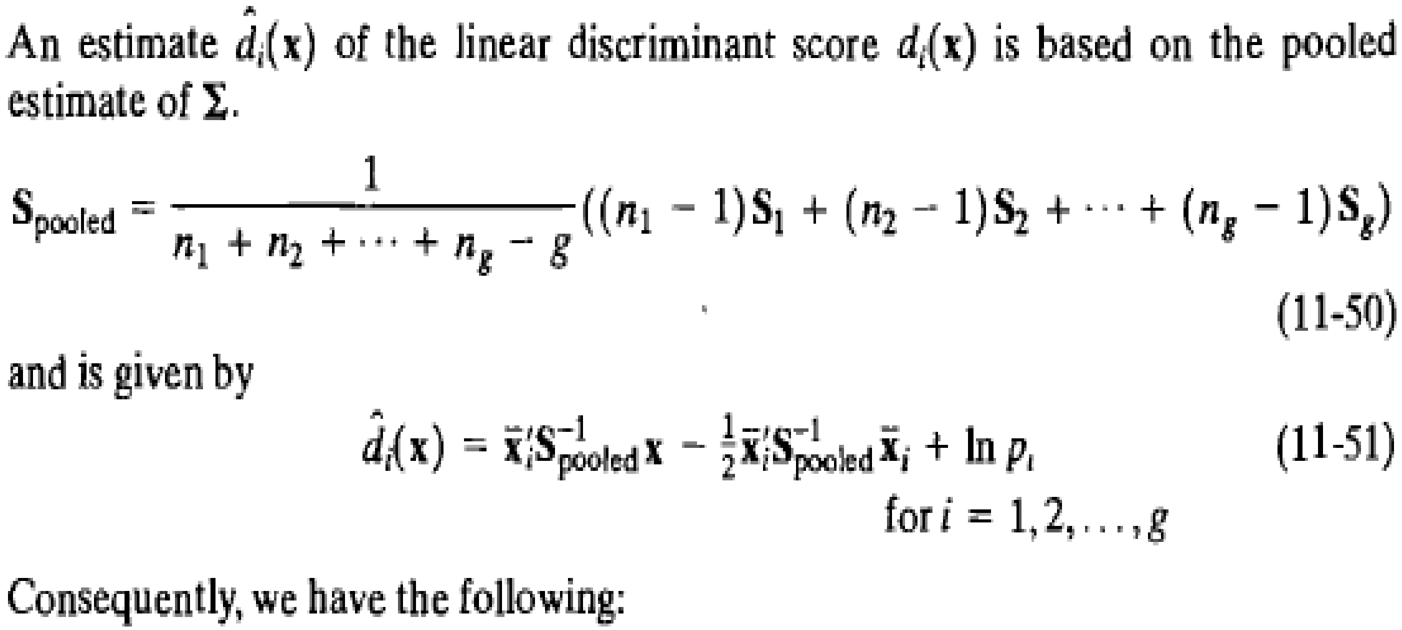
The first two terms are the same for $d_1^Q(\mathbf{x}), d_2^Q(\mathbf{x}), \ldots, d_g^Q(\mathbf{x})$, and, consequently, they can be ignored for allocative purposes. The remaining terms consist of a constant $c_i = \ln p_i - \frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$ and a *linear* combination of the components of $\mathbf{x}$.

Next, define the *linear discriminant score*

$$d_i(\mathbf{x}) = \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \qquad (11\text{-}49)$$

$$\text{for } i = 1, 2, \ldots, g$$

# Estimate of Linear Discriminant Score Based on Pooled Estimate of ∑



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

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

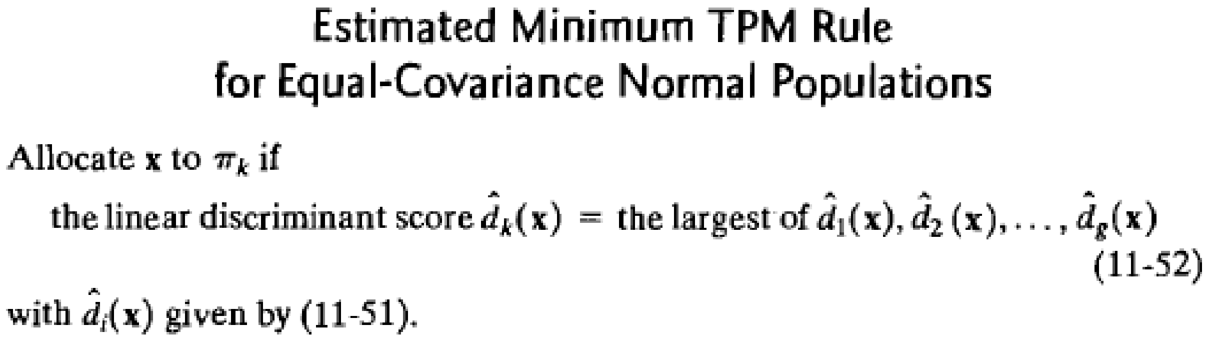
and is given by

$$\hat{d}_i(\mathbf{x}) = \bar{\mathbf{x}}_i'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \frac{1}{2}\bar{\mathbf{x}}_i'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_i + \ln p_i \tag{11-51}$$

for $i = 1, 2, \ldots, g$

Consequently, we have the following:

# Linear Discriminant Score Equals Calculation Squared Distances

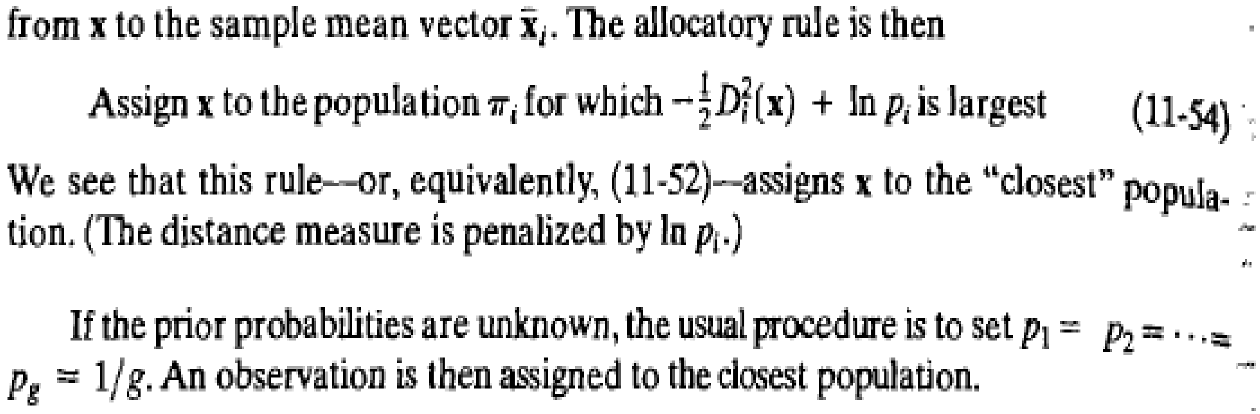


### Estimated Minimum TPM Rule for Equal-Covariance Normal Populations

Allocate $\mathbf{x}$ to $\pi_k$ 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).



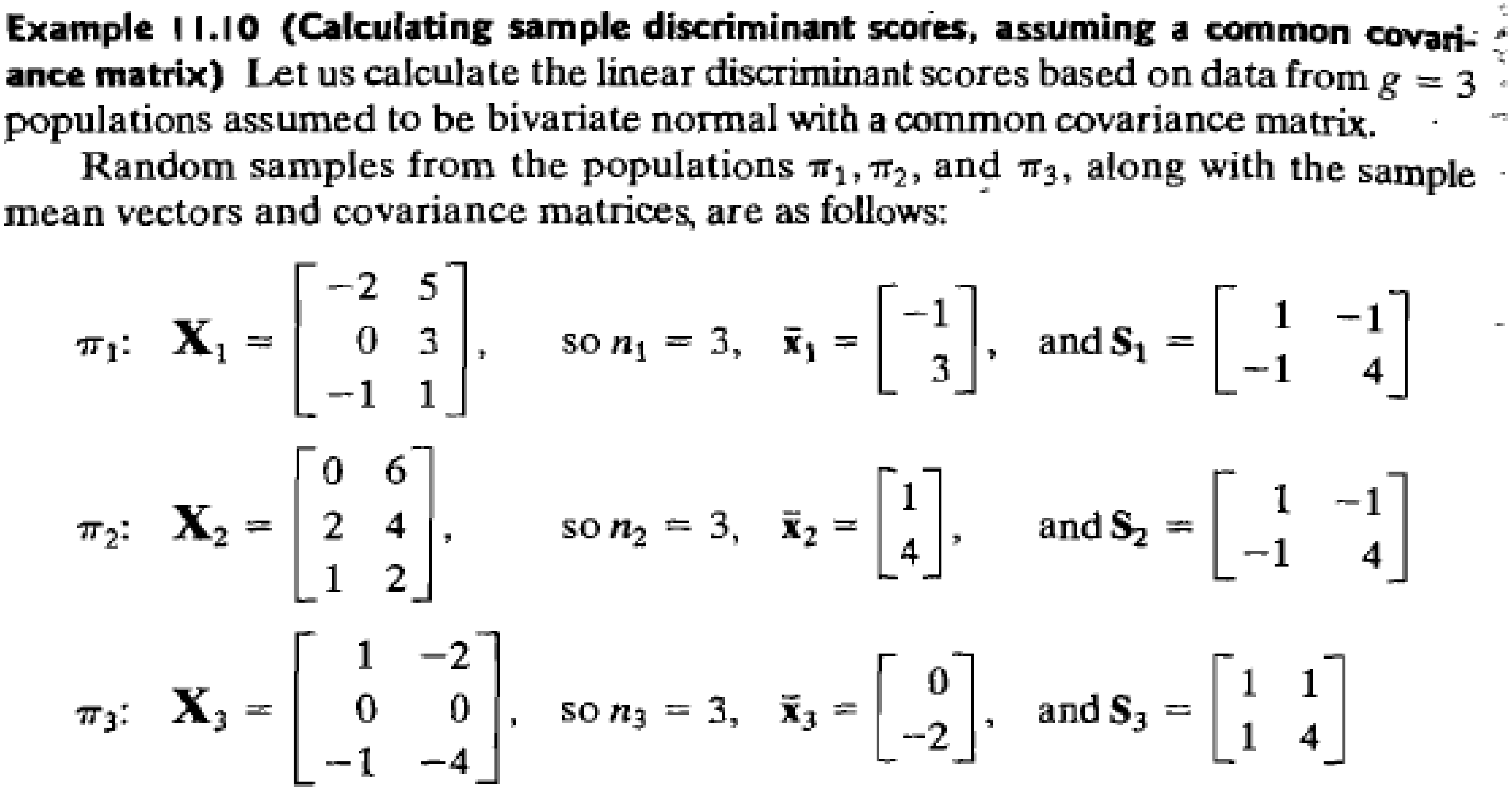
from $\mathbf{x}$ to the sample mean vector $\bar{\mathbf{x}}_i$. The allocatory rule is then

$$\text{Assign } \mathbf{x} \text{ to the population } \pi_i \text{ for which } -\tfrac{1}{2}D_i^2(\mathbf{x}) + \ln p_i \text{ is largest} \qquad (11\text{-}54)$$

We see that this rule—or, equivalently, (11-52)—assigns $\mathbf{x}$ to the "closest" population. (The distance measure is penalized by $\ln p_i$.)

If the prior probabilities are unknown, the usual procedure is to set $p_1 = p_2 = \cdots = p_g = 1/g$. An observation is then assigned to the closest population.

# Calculating Sample Discriminant Scores, Assume Common Covariance Matrix



**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.

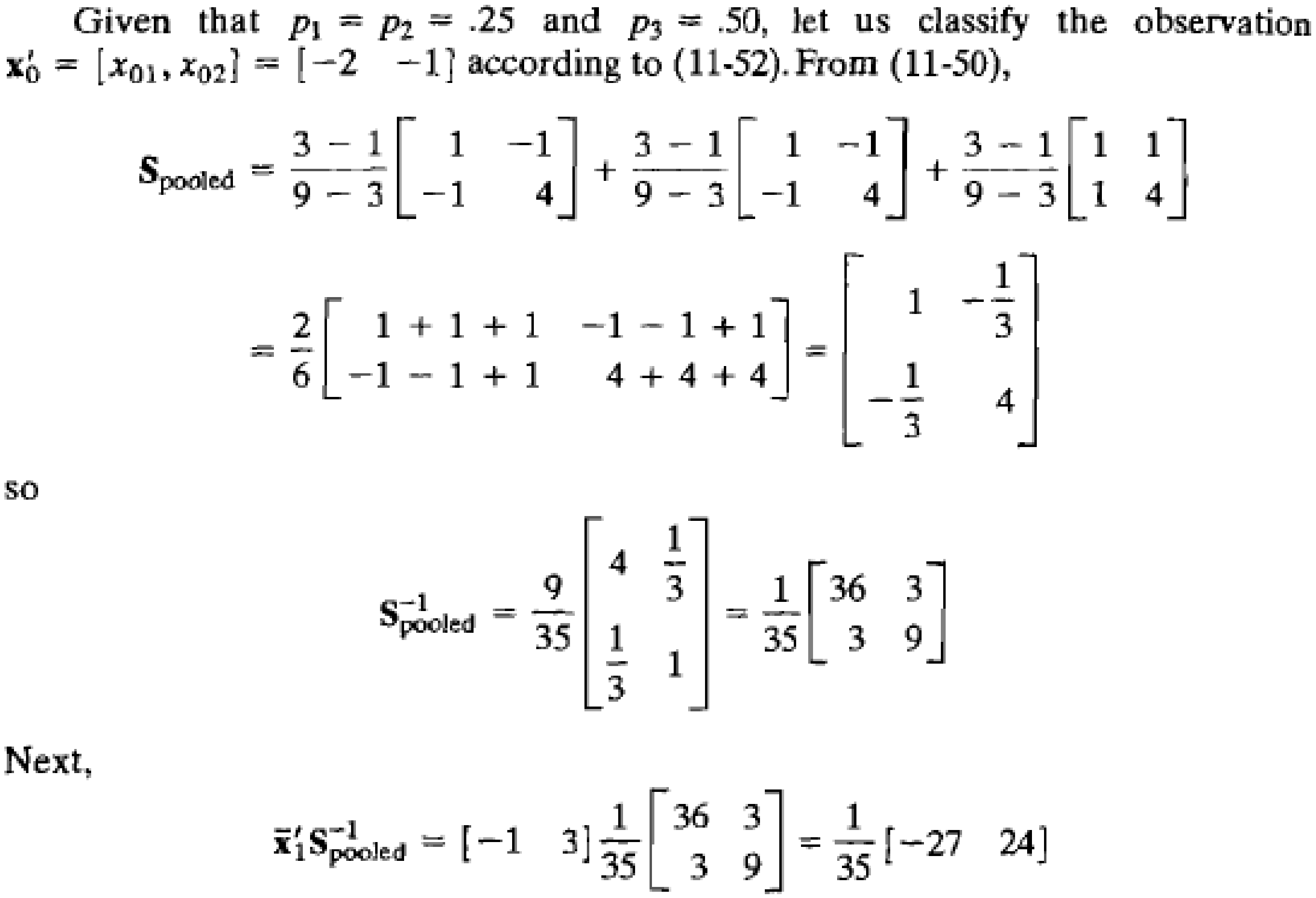
Random samples from the populations $\pi_1, \pi_2$, and $\pi_3$, along with the sample mean vectors and covariance matrices, are as follows:

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

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

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

# Calculating Sample Discriminant Scores, Assume Common Covariance Matrix



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),

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

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

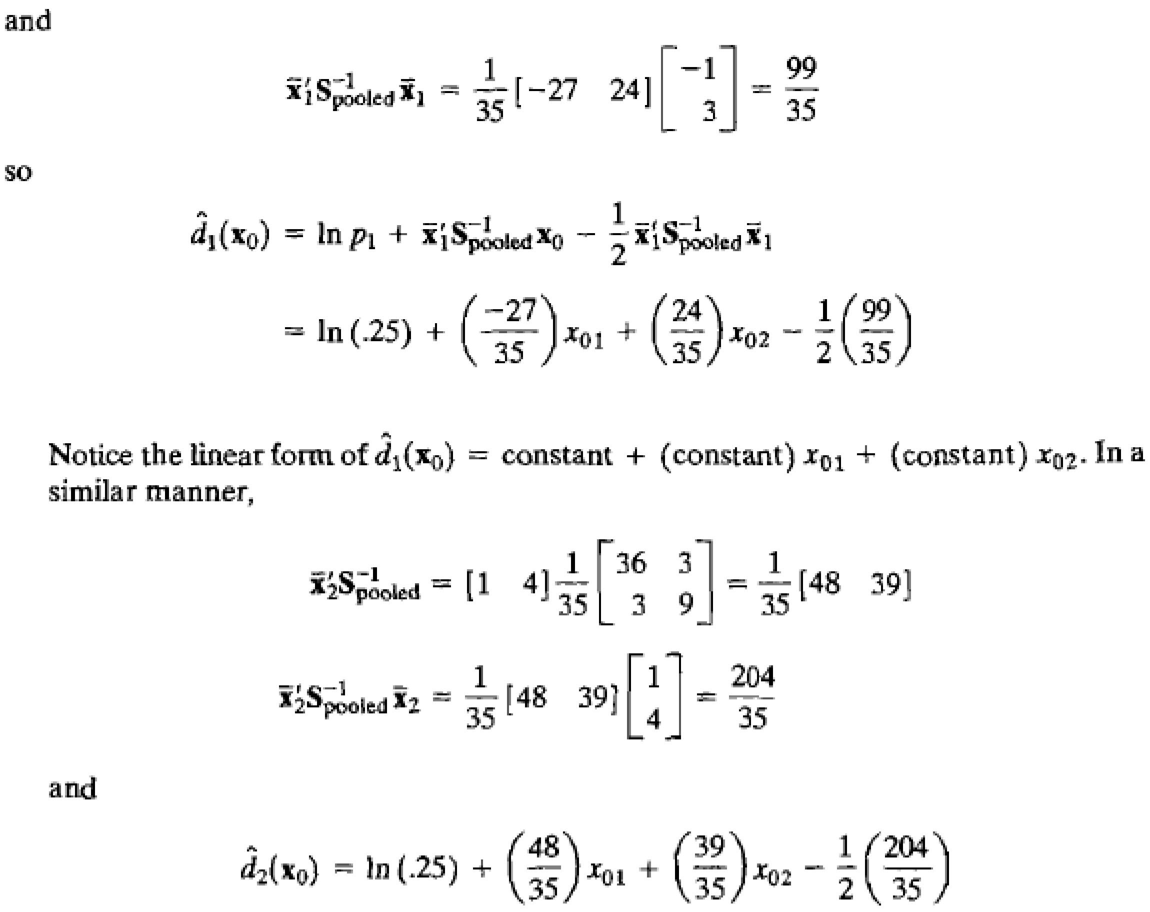
so

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

Next,

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

# Calculating Sample Discriminant Scores, Assume Common Covariance Matrix



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

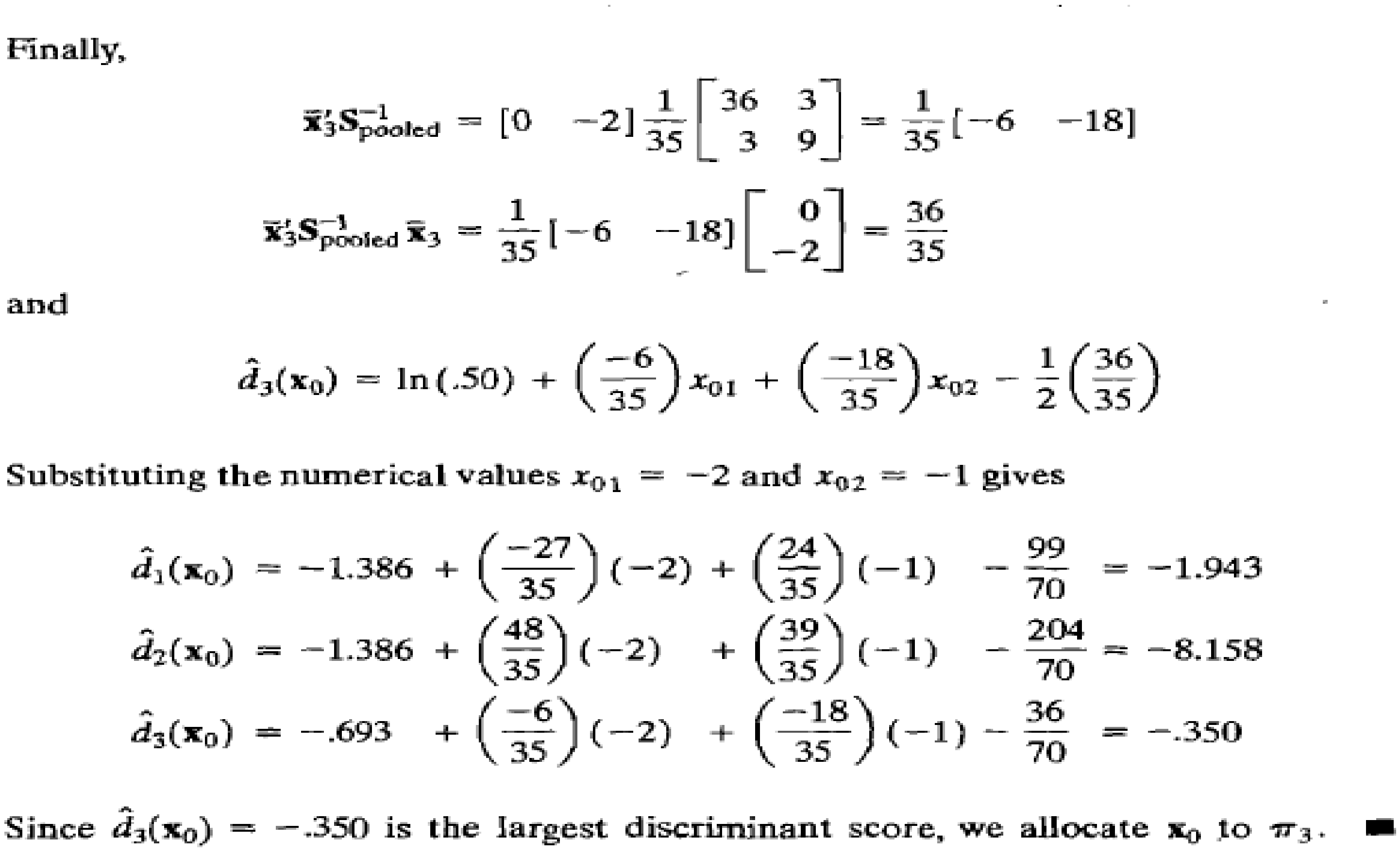
$$\bar{\mathbf{x}}_2' \mathbf{S}_{\text{pooled}}^{-1} = [1 \quad 4]\frac{1}{35}\begin{bmatrix} 36 & 3 \\ 3 & 9 \end{bmatrix} = \frac{1}{35}[48 \quad 39]$$

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

and

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

# Calculating Sample Discriminant Scores, Assume Common Covariance Matrix



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

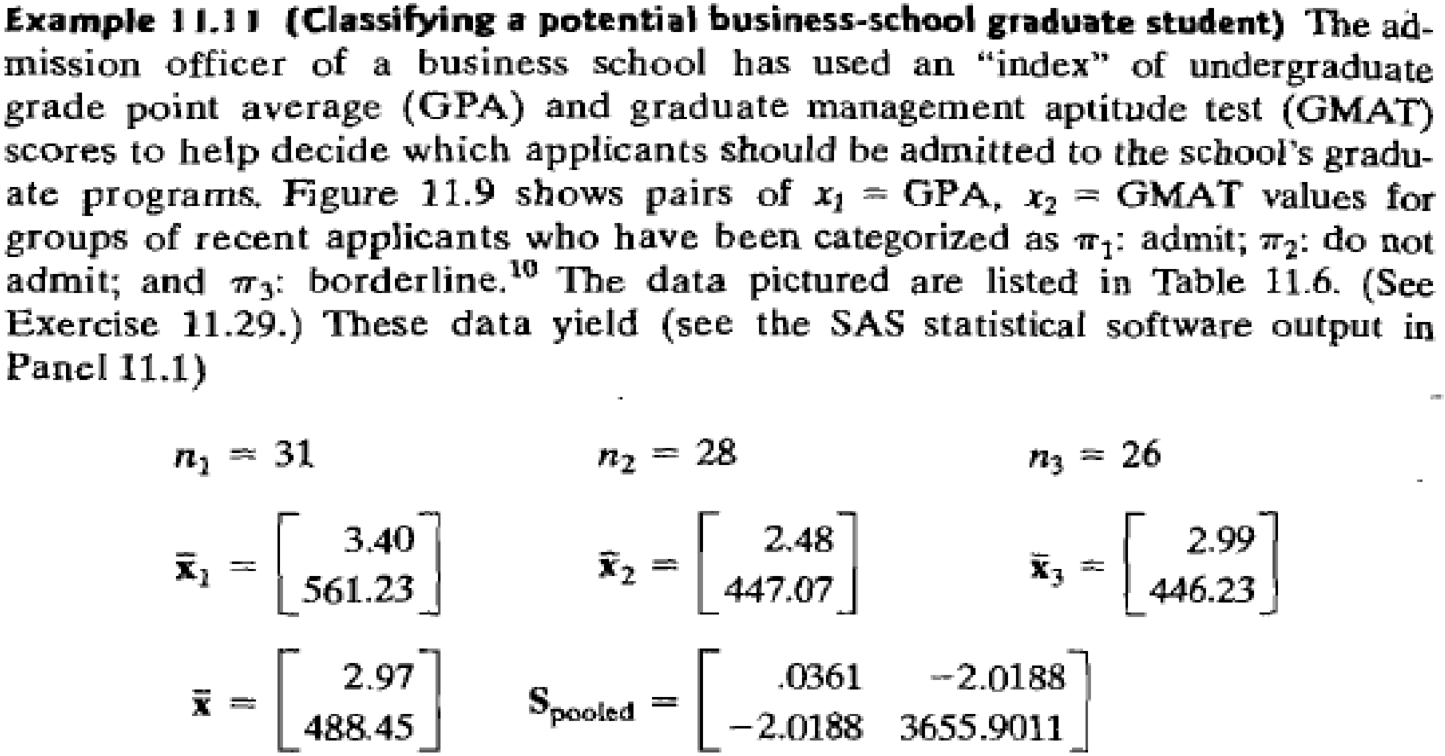
$$\hat{d}_1(\mathbf{x}_0) = -1.386 + \left(\frac{-27}{35}\right)(-2) + \left(\frac{24}{35}\right)(-1) - \frac{99}{70} = -1.943$$

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

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

Since $\hat{d}_3(\mathbf{x}_0) = -.350$ is the largest discriminant score, we allocate $\mathbf{x}_0$ to $\pi_3$. ■

# Example: Classifying a Potential Business-School Graduate Student



**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 $\pi_1$: admit; $\pi_2$: do not admit; and $\pi_3$: 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)

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

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

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

# Example: Classifying a Potential Business-School Graduate Student

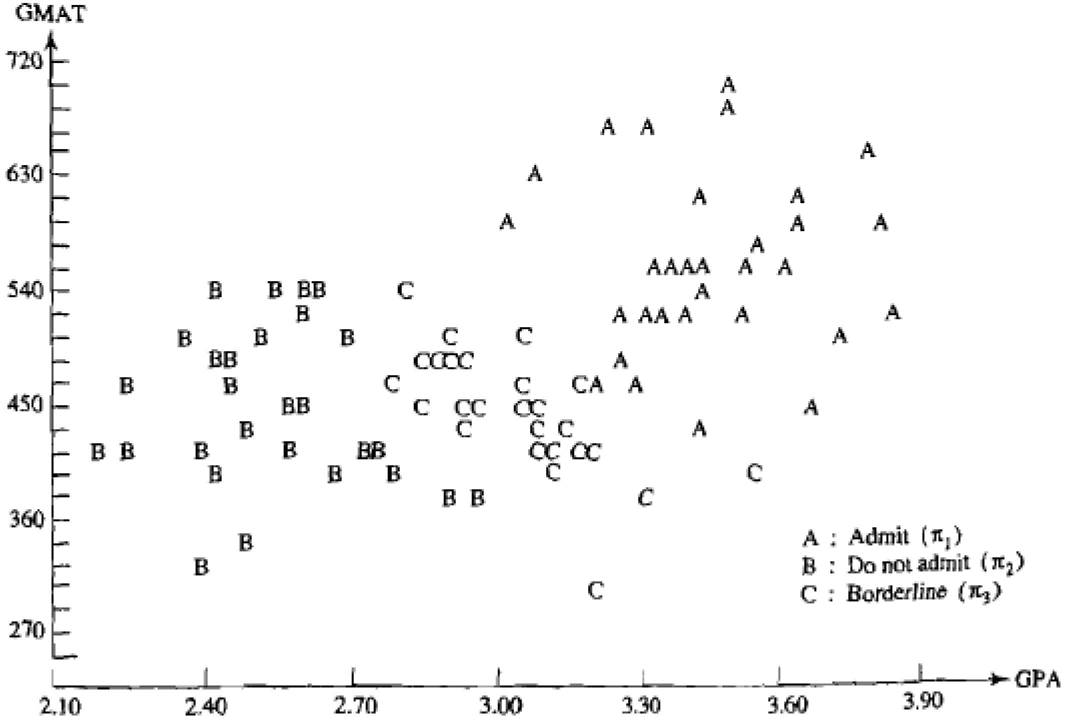
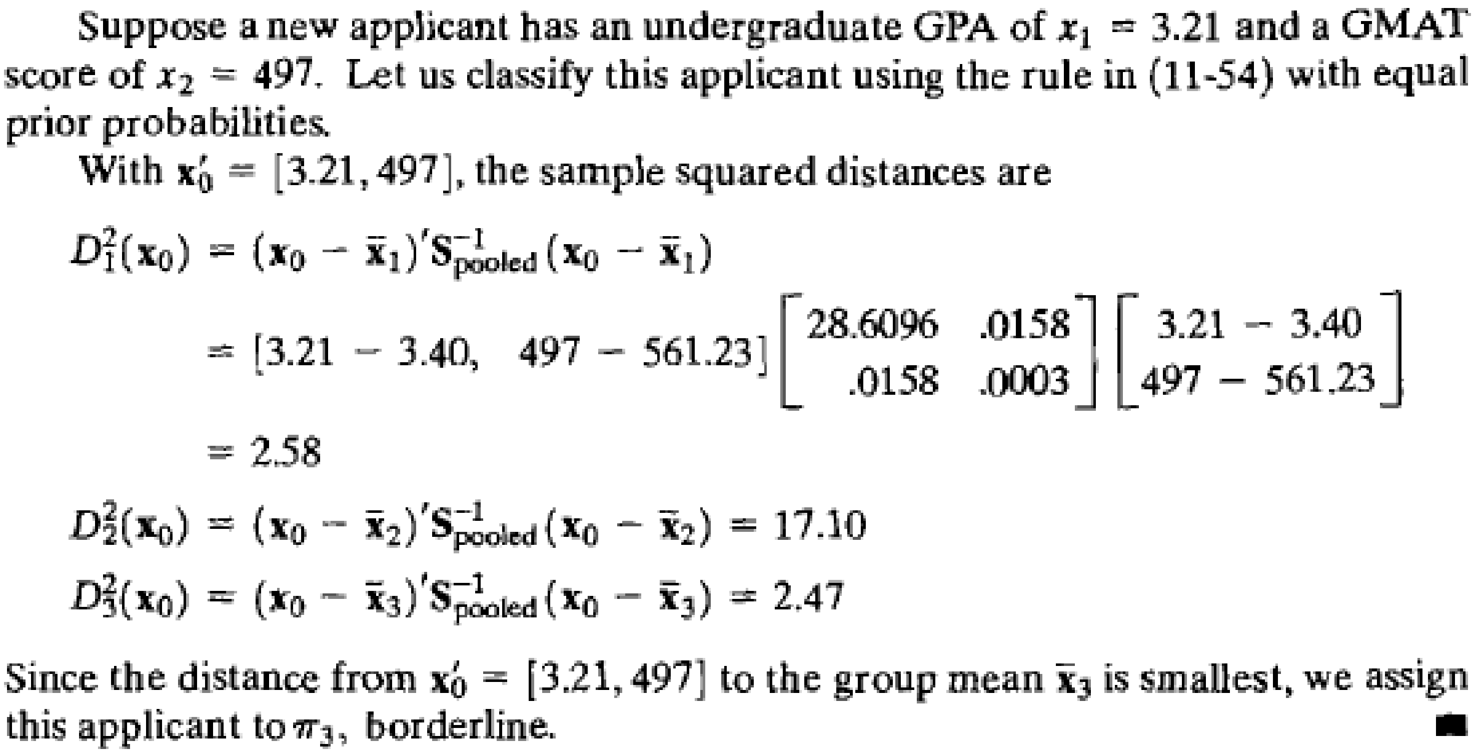


**Figure 11.9** Scatter plot of ($x_1$ = GPA, $x_2$ = GMAT) for applicants to a graduate school of business who have been classified as admit, do not admit, or borderline.

[10] In this case, the populations are artificial in the sense that they have been created by the admissions officer. On the other hand, experience has shown that applicants with high *GPA* and high GMAT scores generally do well in a graduate program; those with low readings on these variables generally experience difficulty.

# Example: Classifying a Potential Business-School Graduate Student



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

$$D_1^2(\mathbf{x}_0) = (\mathbf{x}_0 - \bar{\mathbf{x}}_1)'\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1)$$

$$= [3.21 - 3.40, \quad 497 - 561.23]\begin{bmatrix} 28.6096 & .0158 \\ .0158 & .0003 \end{bmatrix}\begin{bmatrix} 3.21 - 3.40 \\ 497 - 561.23 \end{bmatrix}$$

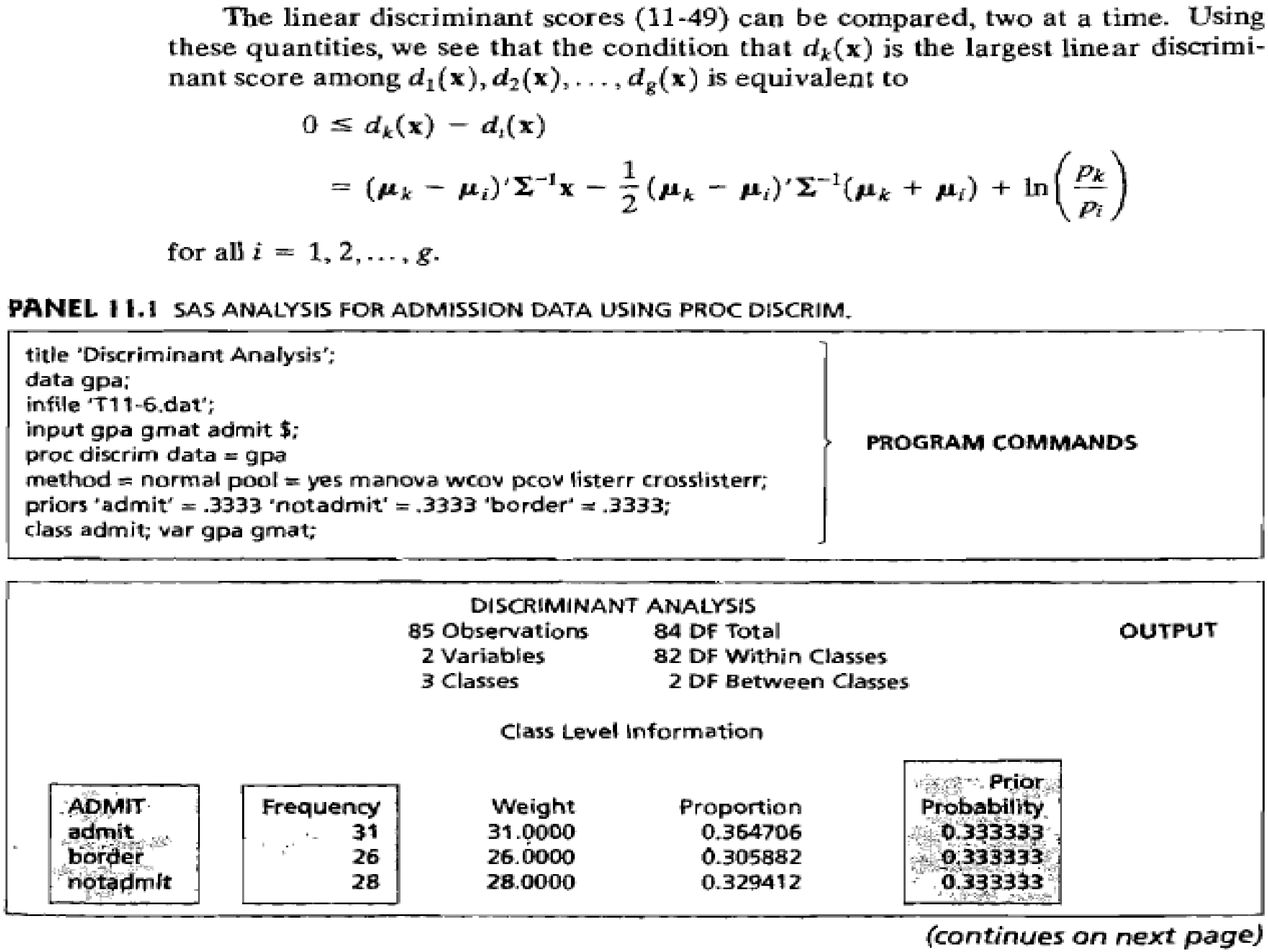
$$= 2.58$$

$$D_2^2(\mathbf{x}_0) = (\mathbf{x}_0 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_2) = 17.10$$

$$D_3^2(\mathbf{x}_0) = (\mathbf{x}_0 - \bar{\mathbf{x}}_3)'\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_3) = 2.47$$

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 $\pi_3$, borderline. ■

# Comparison of Discriminant scores



The linear discriminant scores (11-49) can be compared, two at a time. Using these quantities, we see that the condition that $d_k(\mathbf{x})$ is the largest linear discriminant score among $d_1(\mathbf{x}), d_2(\mathbf{x}), \ldots, d_g(\mathbf{x})$ is equivalent to

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

for all $i = 1, 2, \ldots, g$.

**PANEL 11.1** SAS ANALYSIS FOR ADMISSION DATA USING PROC DISCRIM.

```
title 'Discriminant Analysis';
data gpa;
infile 'T11-6.dat';
input gpa gmat admit $;
proc discrim data = gpa
method = normal pool = yes manova wcov pcov listerr crosslisterr;
priors 'admit' = .3333 'notadmit' = .3333 'border' = .3333;
class admit; var gpa gmat;
```

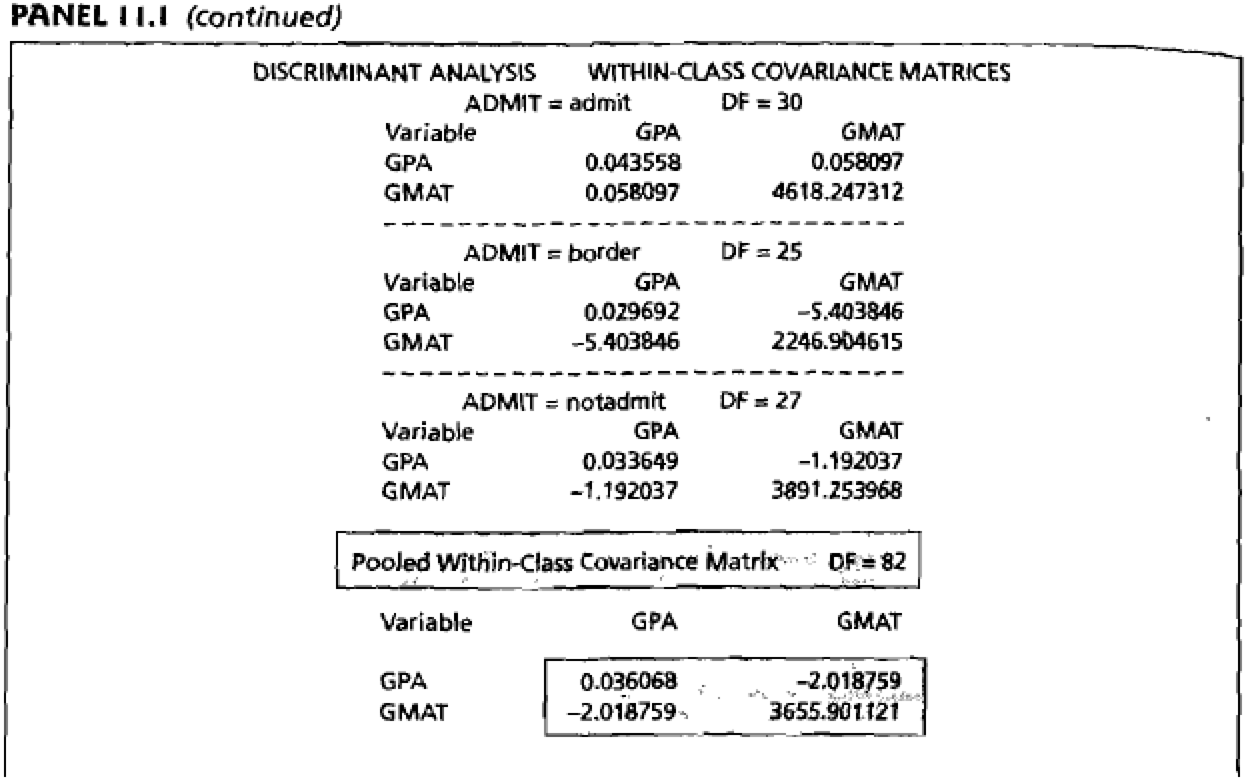
PROGRAM COMMANDS

```
                    DISCRIMINANT ANALYSIS
            85 Observations       84 DF Total
             2 Variables          82 DF Within Classes
             3 Classes             2 DF Between Classes

                   Class Level Information

                                                          Prior
ADMIT        Frequency        Weight       Proportion   Probability
admit               31       31.0000         0.364706      0.333333
border              26       26.0000         0.305882      0.333333
notadmit            28       28.0000         0.329412      0.333333
```

OUTPUT

(continues on next page)

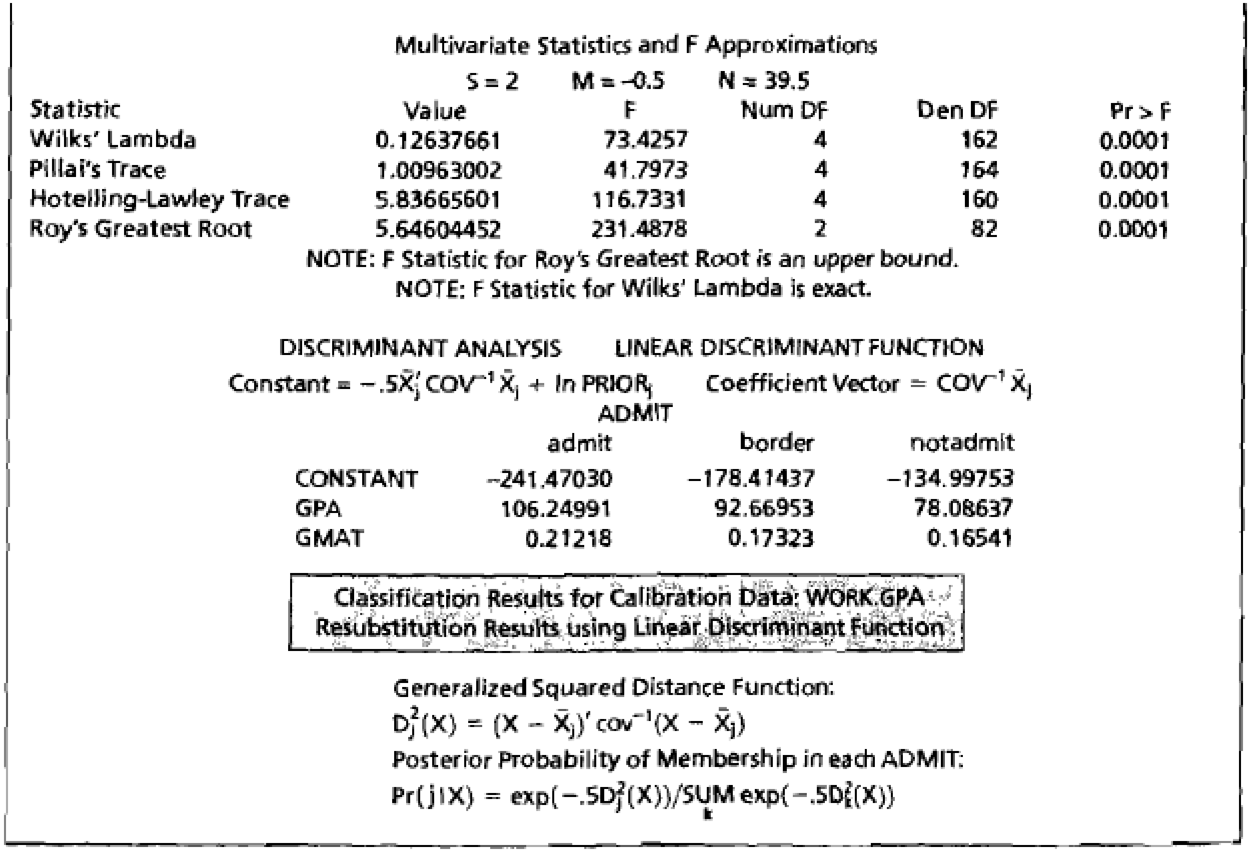
# Discriminant Analysis – SAS Output for Business School Student Admission Data



**PANEL 11.1** *(continued)*

```
          DISCRIMINANT ANALYSIS      WITHIN-CLASS COVARIANCE MATRICES
                        ADMIT = admit           DF = 30
                  Variable               GPA              GMAT
                  GPA               0.043558          0.058097
                  GMAT              0.058097       4618.247312
                  ---------------------------------------------
                        ADMIT = border          DF = 25
                  Variable               GPA              GMAT
                  GPA               0.029692         -5.403846
                  GMAT             -5.403846       2246.904615
                  ---------------------------------------------
                        ADMIT = notadmit        DF = 27
                  Variable               GPA              GMAT
                  GPA               0.033649         -1.192037
                  GMAT             -1.192037       3891.253968

               Pooled Within-Class Covariance Matrix    DF = 82

                  Variable               GPA              GMAT

                  GPA               0.036068         -2.018759
                  GMAT             -2.018759       3655.901121
```

# Discriminant Analysis – SAS Output for Business School Student Admission Data



```
                    Multivariate Statistics and F Approximations
                         S = 2      M = -0.5      N = 39.5
Statistic                   Value          F       Num DF      Den DF        Pr > F
Wilks' Lambda          0.12637661    73.4257          4         162        0.0001
Pillai's Trace         1.00963002    41.7973          4         164        0.0001
Hotelling-Lawley Trace 5.83665601   116.7331          4         160        0.0001
Roy's Greatest Root    5.64604452   231.4878          2          82        0.0001
               NOTE: F Statistic for Roy's Greatest Root is an upper bound.
                    NOTE: F Statistic for Wilks' Lambda is exact.
```

DISCRIMINANT ANALYSIS    LINEAR DISCRIMINANT FUNCTION

Constant = $-.5\bar{X}_j' \, COV^{-1} \bar{X}_j + \ln PRIOR_j$    Coefficient Vector = $COV^{-1} \bar{X}_j$

```
                                ADMIT
                      admit          border        notadmit
CONSTANT         -241.47030      -178.41437      -134.99753
GPA               106.24991        92.66953        78.08637
GMAT                0.21218         0.17323         0.16541
```

Classification Results for Calibration Data: WORK.GPA
Resubstitution Results using Linear Discriminant Function

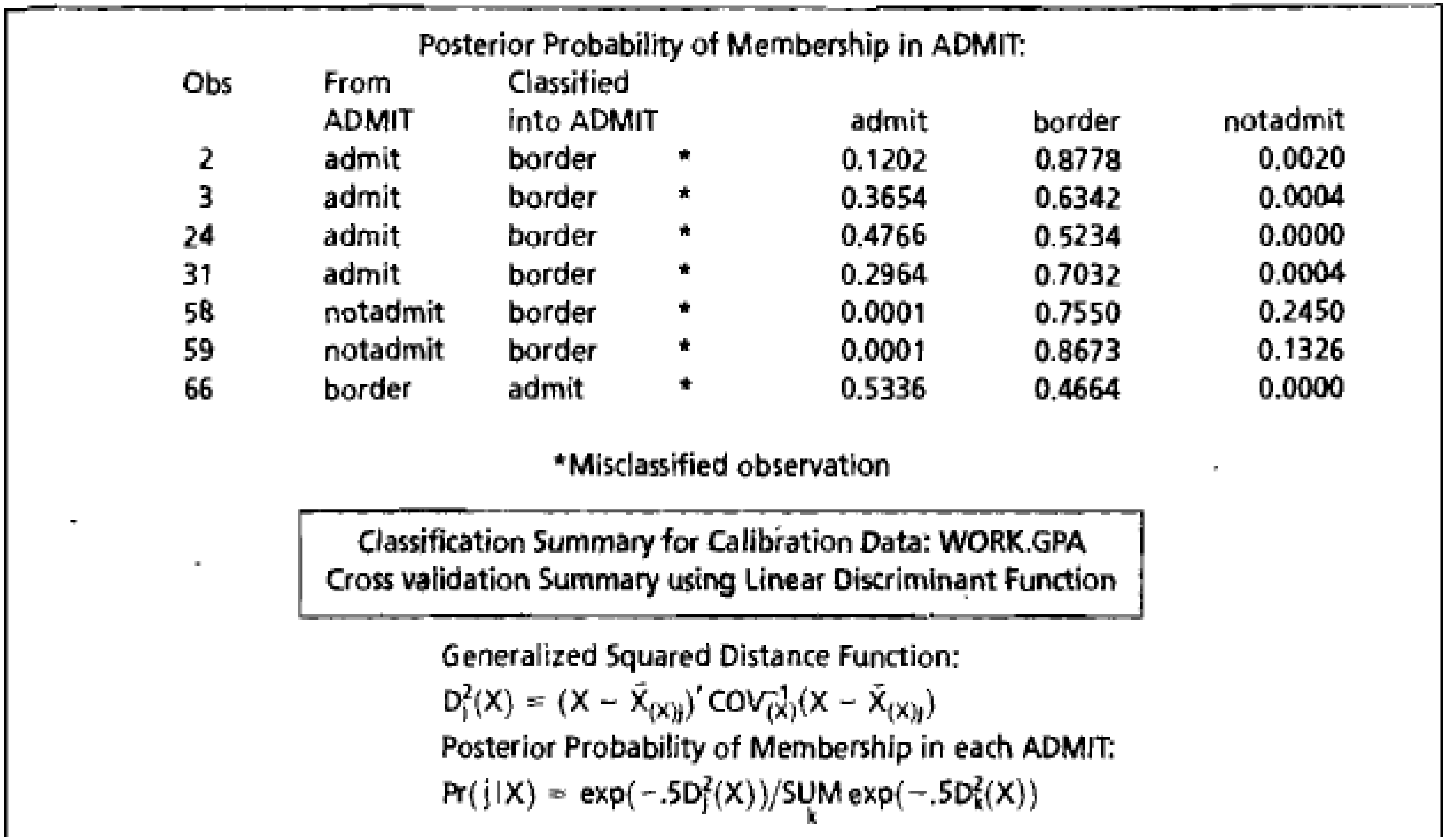
Generalized Squared Distance Function:

$D_j^2(X) = (X - \bar{X}_j)' \, cov^{-1}(X - \bar{X}_j)$

Posterior Probability of Membership in each ADMIT:

$Pr(j|X) = \exp(-.5D_j^2(X)) / \underset{k}{SUM} \exp(-.5D_k^2(X))$

# Discriminant Analysis – SAS Output for Business School Student Admission Data



```
                Posterior Probability of Membership in ADMIT:
Obs     From        Classified
        ADMIT       into ADMIT              admit      border      notadmit
  2     admit       border        *        0.1202      0.8778        0.0020
  3     admit       border        *        0.3654      0.6342        0.0004
 24     admit       border        *        0.4766      0.5234        0.0000
 31     admit       border        *        0.2964      0.7032        0.0004
 58     notadmit    border        *        0.0001      0.7550        0.2450
 59     notadmit    border        *        0.0001      0.8673        0.1326
 66     border      admit         *        0.5336      0.4664        0.0000

                      *Misclassified observation
```

Classification Summary for Calibration Data: WORK.GPA
Cross validation Summary using Linear Discriminant Function

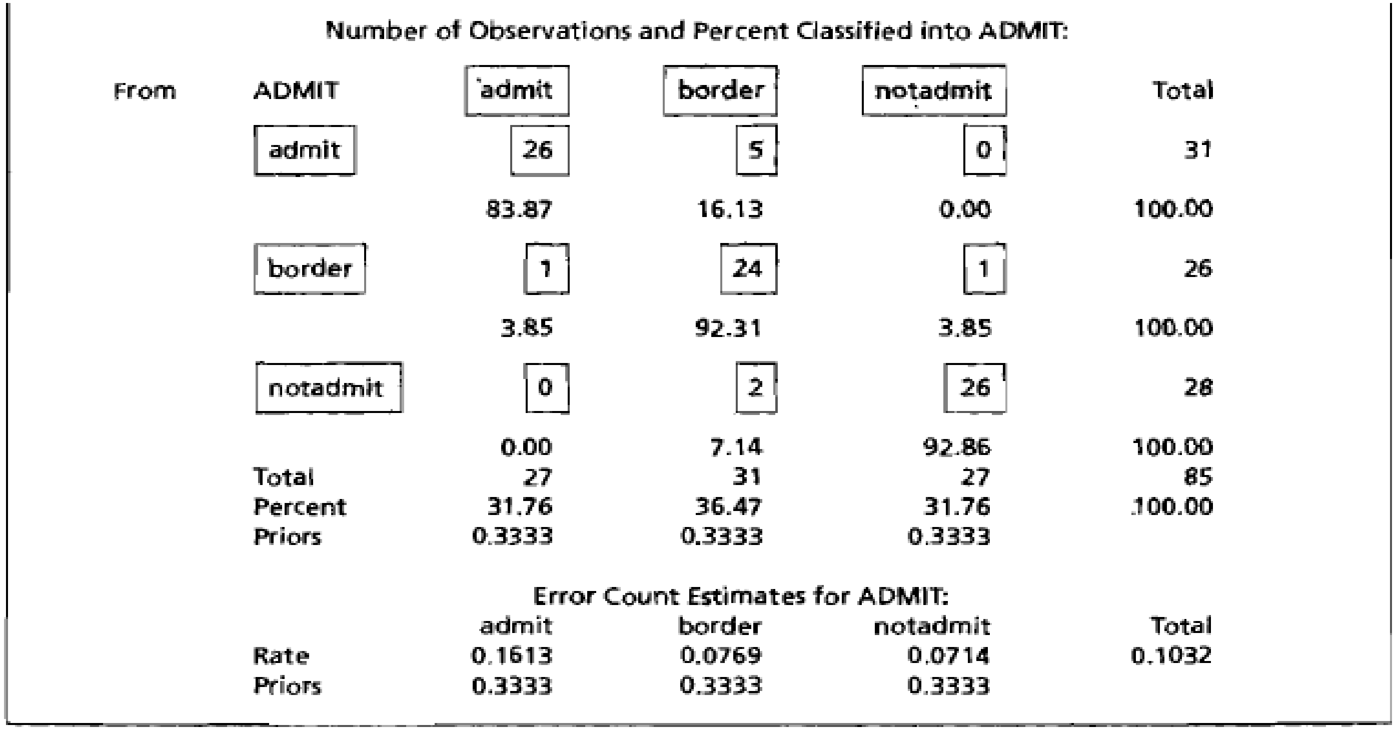
Generalized Squared Distance Function:

$$D_j^2(X) = (X - \bar{X}_{(X)j})' \, COV_{(X)}^{-1} (X - \bar{X}_{(X)j})$$

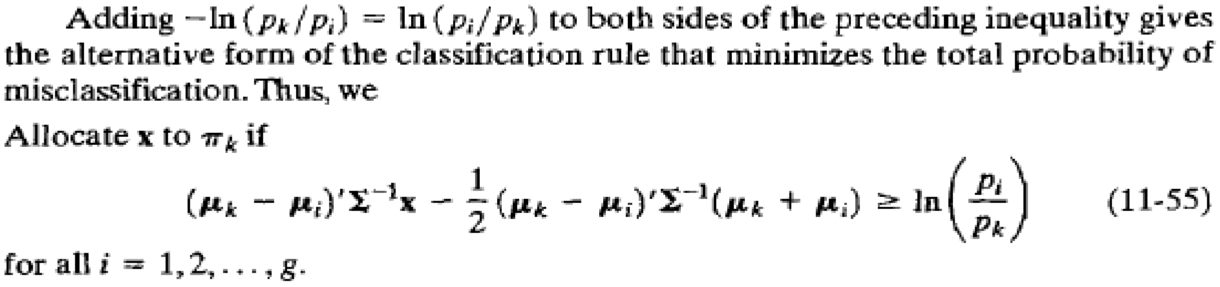
Posterior Probability of Membership in each ADMIT:

$$Pr(j|X) = \exp(-.5D_j^2(X)) / \underset{k}{SUM} \exp(-.5D_k^2(X))$$

# Discriminant Analysis – SAS Output for Business School Student Admission Data



```
                Number of Observations and Percent Classified into ADMIT:

From      ADMIT           admit         border        notadmit          Total

          admit              26              5               0             31

                          83.87          16.13            0.00         100.00

          border              1             24               1             26

                           3.85          92.31            3.85         100.00

          notadmit            0              2              26             28

                           0.00           7.14           92.86         100.00
          Total              27             31              27             85
          Percent         31.76          36.47           31.76         100.00
          Priors         0.3333         0.3333          0.3333

                         Error Count Estimates for ADMIT:
                          admit         border        notadmit          Total
          Rate           0.1613         0.0769          0.0714         0.1032
          Priors         0.3333         0.3333          0.3333
```



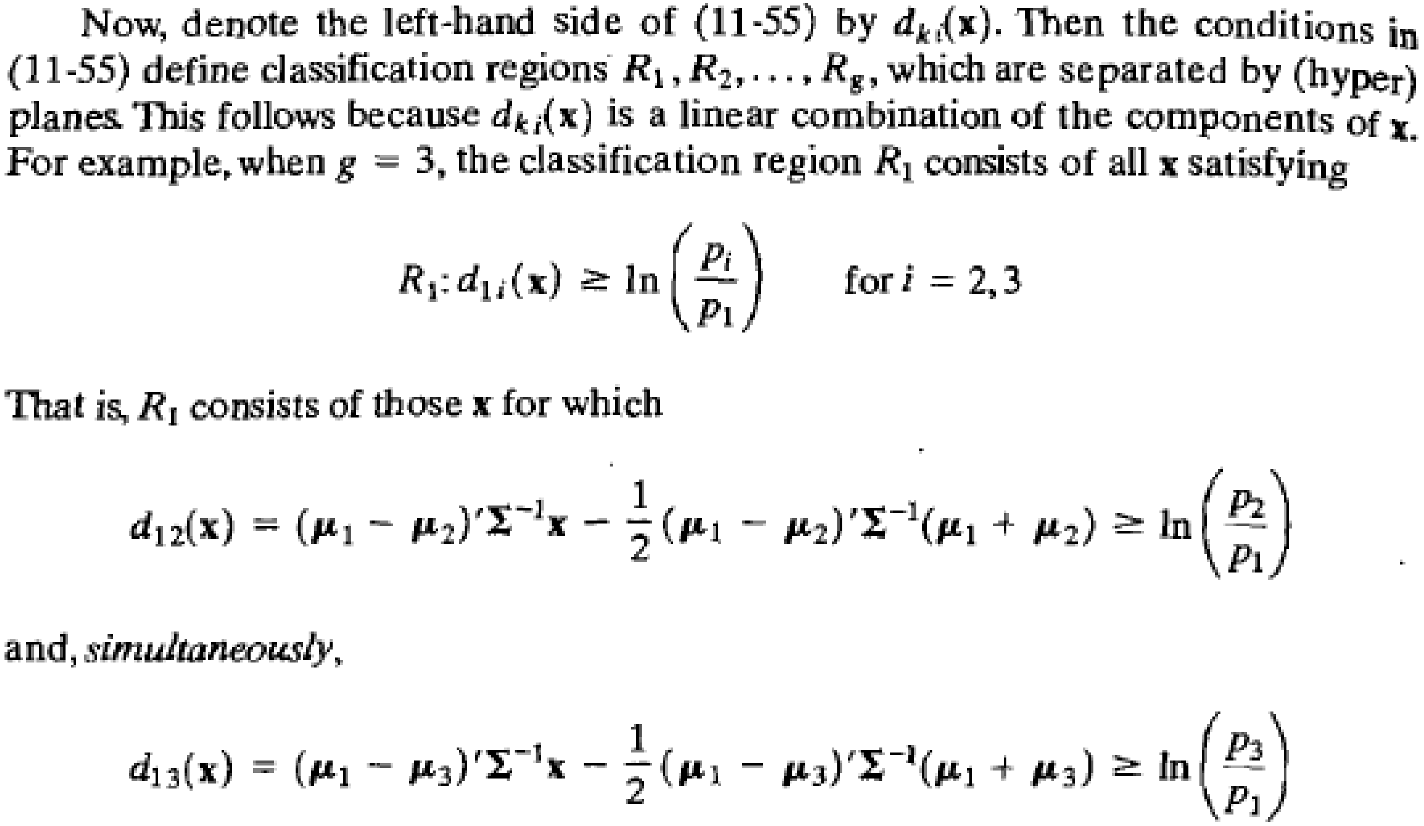
Adding $-\ln(p_k/p_i) = \ln(p_i/p_k)$ to both sides of the preceding inequality gives the alternative form of the classification rule that minimizes the total probability of misclassification. Thus, we

Allocate $\mathbf{x}$ to $\pi_k$ if

$$(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_i) \geq \ln\left(\frac{p_i}{p_k}\right) \qquad (11\text{-}55)$$

for all $i = 1, 2, \ldots, g$.

# Classification Regions R$_1$, R$_2$, etc. Separated by Hyper Plans



Now, denote the left-hand side of (11-55) by $d_{ki}(\mathbf{x})$. Then the conditions in (11-55) define classification regions $R_1, R_2, \ldots, R_g$, which are separated by (hyper) planes. This follows because $d_{ki}(\mathbf{x})$ is a linear combination of the components of $\mathbf{x}$. For example, when $g = 3$, the classification region $R_1$ consists of all $\mathbf{x}$ satisfying

$$R_1: d_{1i}(\mathbf{x}) \geq \ln\left(\frac{p_i}{p_1}\right) \qquad \text{for } i = 2, 3$$

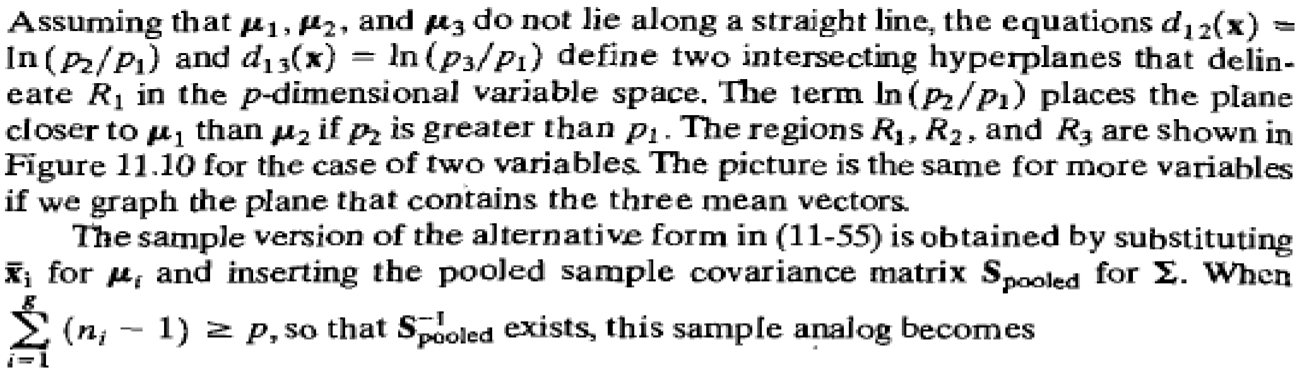
That is, $R_1$ consists of those $\mathbf{x}$ for which

$$d_{12}(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left(\frac{p_2}{p_1}\right)$$

and, *simultaneously*,

$$d_{13}(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_3) \geq \ln\left(\frac{p_3}{p_1}\right)$$

# Classification Regions for the Linear Minimum TPM Rule



Assuming that $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and $\boldsymbol{\mu}_3$ do not lie along a straight line, the equations $d_{12}(\mathbf{x}) = \ln(p_2/p_1)$ and $d_{13}(\mathbf{x}) = \ln(p_3/p_1)$ define two intersecting hyperplanes that delineate $R_1$ in the $p$-dimensional variable space. The term $\ln(p_2/p_1)$ places the plane closer to $\boldsymbol{\mu}_1$ than $\boldsymbol{\mu}_2$ if $p_2$ is greater than $p_1$. The regions $R_1, R_2$, and $R_3$ are shown in Figure 11.10 for the case of two variables. The picture is the same for more variables if we graph the plane that contains the three mean vectors.

The sample version of the alternative form in (11-55) is obtained by substituting $\bar{\mathbf{x}}_i$ for $\boldsymbol{\mu}_i$ and inserting the pooled sample covariance matrix $\mathbf{S}_{\text{pooled}}$ for $\boldsymbol{\Sigma}$. When $\sum_{i=1}^{g}(n_i - 1) \geq p$, so that $\mathbf{S}_{\text{pooled}}^{-1}$ exists, this sample analog becomes

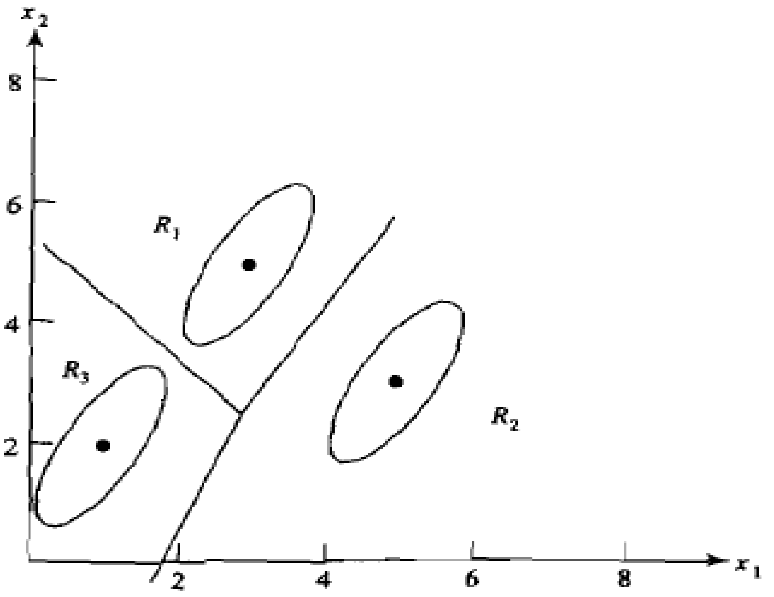
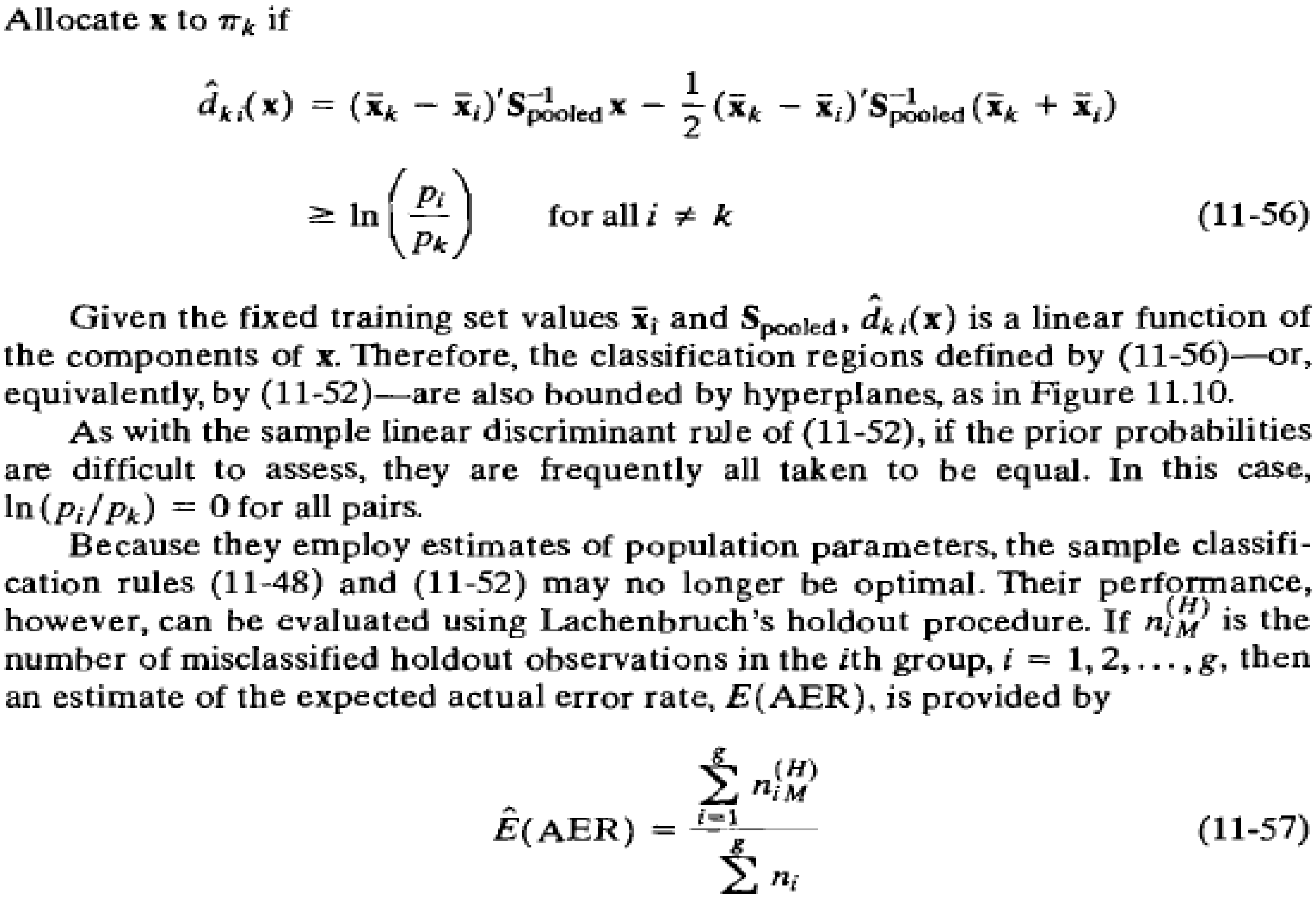


**Figure 11.10** The classification regions $R_1, R_2$, and $R_3$ for the linear minimum TPM rule $(p_1 = \frac{1}{4}, p_2 = \frac{1}{2}, p_3 = \frac{1}{4})$.

# Example: Effective Classification with Fewer Variables



Allocate $\mathbf{x}$ to $\pi_k$ if

$$\hat{d}_{ki}(\mathbf{x}) = (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_i)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_i)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_k + \bar{\mathbf{x}}_i)$$

$$\geq \ln\left(\frac{p_i}{p_k}\right) \qquad \text{for all } i \neq k \tag{11-56}$$

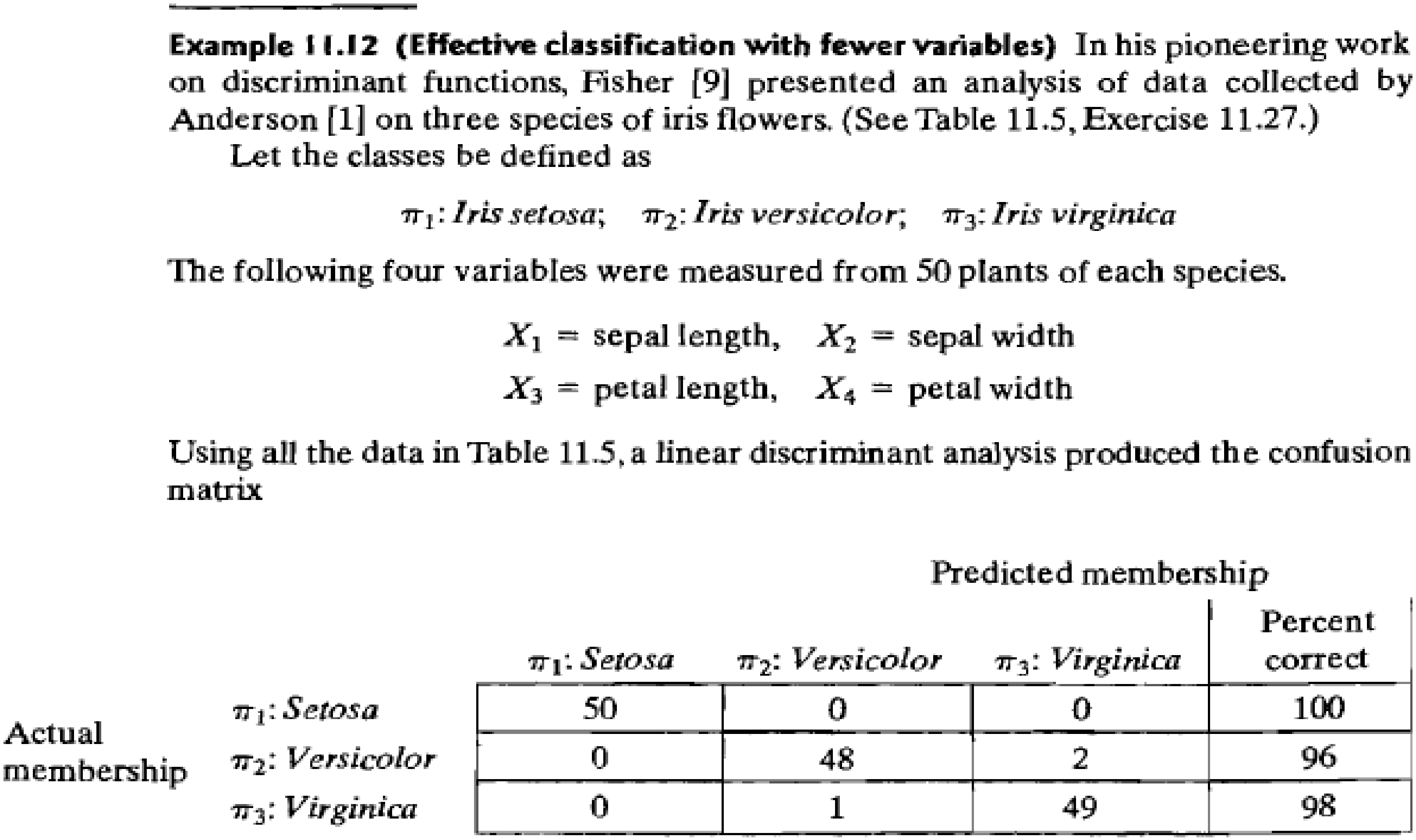
Given the fixed training set values $\bar{\mathbf{x}}_i$ and $\mathbf{S}_{\text{pooled}}$, $\hat{d}_{ki}(\mathbf{x})$ is a linear function of the components of $\mathbf{x}$. Therefore, the classification regions defined by (11-56)—or, equivalently, by (11-52)—are also bounded by hyperplanes, as in Figure 11.10.

As with the sample linear discriminant rule of (11-52), if the prior probabilities are difficult to assess, they are frequently all taken to be equal. In this case, $\ln(p_i/p_k) = 0$ for all pairs.

Because they employ estimates of population parameters, the sample classification rules (11-48) and (11-52) may no longer be optimal. Their performance, however, can be evaluated using Lachenbruch's holdout procedure. If $n_{iM}^{(H)}$ is the number of misclassified holdout observations in the $i$th group, $i = 1, 2, \ldots, g$, then an estimate of the expected actual error rate, $E(\text{AER})$, is provided by

$$\hat{E}(\text{AER}) = \frac{\sum_{i=1}^{g} n_{iM}^{(H)}}{\sum^{g} n_i} \tag{11-57}$$

# Example: Effective Classification with Fewer Variables



**Example 11.12 (Effective classification with fewer variables)** In his pioneering work on discriminant functions, Fisher [9] presented an analysis of data collected by Anderson [1] on three species of iris flowers. (See Table 11.5, Exercise 11.27.)

Let the classes be defined as

$$\pi_1: \textit{Iris setosa}; \quad \pi_2: \textit{Iris versicolor}; \quad \pi_3: \textit{Iris virginica}$$

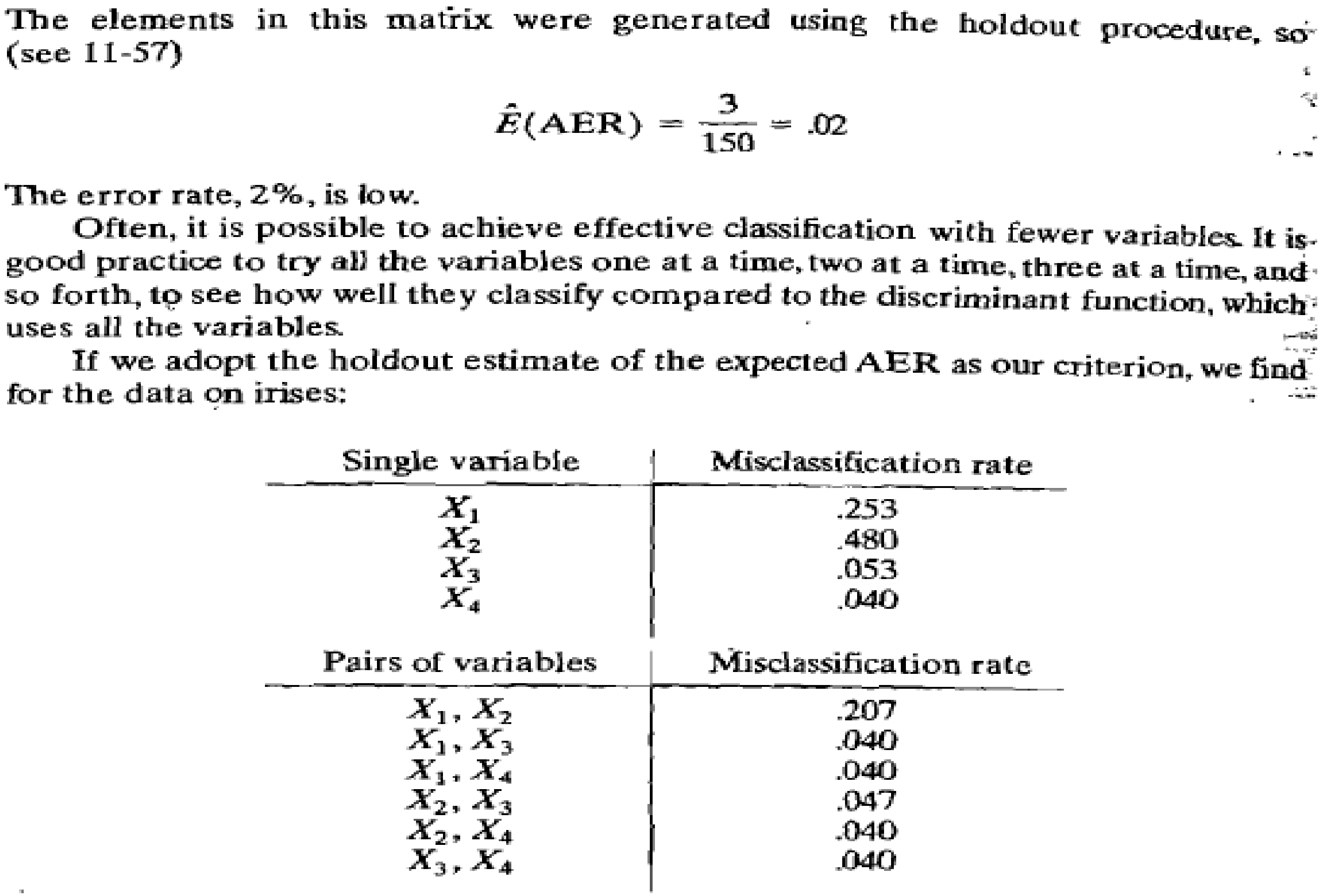
The following four variables were measured from 50 plants of each species.

$$X_1 = \text{sepal length}, \quad X_2 = \text{sepal width}$$
$$X_3 = \text{petal length}, \quad X_4 = \text{petal width}$$

Using all the data in Table 11.5, a linear discriminant analysis produced the confusion matrix

| | | Predicted membership | | | |
|---|---|---|---|---|---|
| | | $\pi_1$: *Setosa* | $\pi_2$: *Versicolor* | $\pi_3$: *Virginica* | Percent correct |
| Actual membership | $\pi_1$: *Setosa* | 50 | 0 | 0 | 100 |
| | $\pi_2$: *Versicolor* | 0 | 48 | 2 | 96 |
| | $\pi_3$: *Virginica* | 0 | 1 | 49 | 98 |

# Example: Effective Classification with Fewer Variables



The elements in this matrix were generated using the holdout procedure, so (see 11-57)

$$\hat{E}(\text{AER}) = \frac{3}{150} = .02$$

The error rate, 2%, is low.

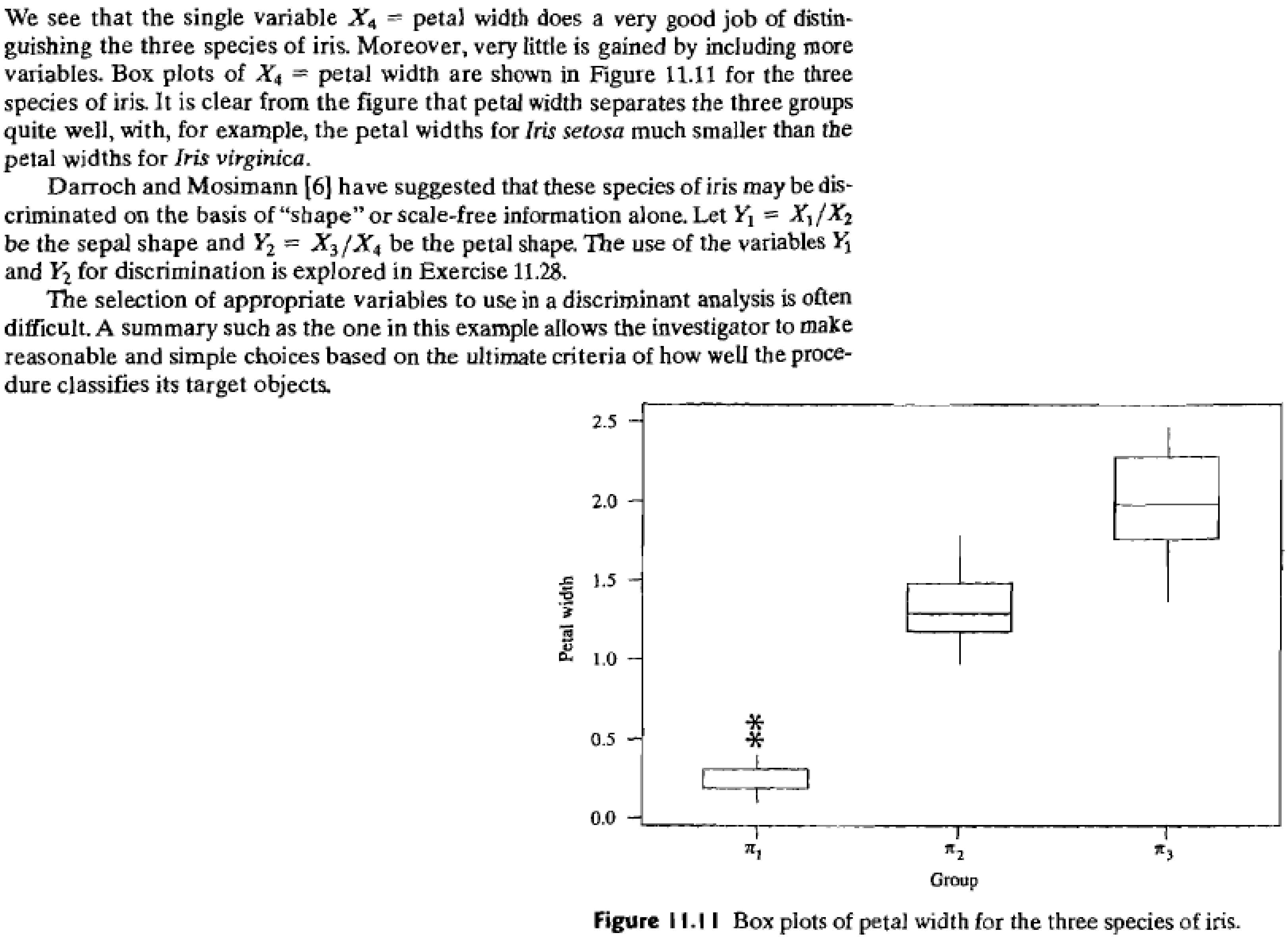
Often, it is possible to achieve effective classification with fewer variables. It is good practice to try all the variables one at a time, two at a time, three at a time, and so forth, to see how well they classify compared to the discriminant function, which uses all the variables.

If we adopt the holdout estimate of the expected AER as our criterion, we find for the data on irises:

| Single variable | Misclassification rate |
|---|---|
| $X_1$ | .253 |
| $X_2$ | .480 |
| $X_3$ | .053 |
| $X_4$ | .040 |

| Pairs of variables | Misclassification rate |
|---|---|
| $X_1, X_2$ | .207 |
| $X_1, X_3$ | .040 |
| $X_1, X_4$ | .040 |
| $X_2, X_3$ | .047 |
| $X_2, X_4$ | .040 |
| $X_3, X_4$ | .040 |

# Boxplots Petal Width 3 Species Irises



We see that the single variable $X_4$ = petal width does a very good job of distinguishing the three species of iris. Moreover, very little is gained by including more variables. Box plots of $X_4$ = petal width are shown in Figure 11.11 for the three species of iris. It is clear from the figure that petal width separates the three groups quite well, with, for example, the petal widths for *Iris setosa* much smaller than the petal widths for *Iris virginica*.

Darroch and Mosimann [6] have suggested that these species of iris may be discriminated on the basis of "shape" or scale-free information alone. Let $Y_1 = X_1/X_2$ be the sepal shape and $Y_2 = X_3/X_4$ be the petal shape. The use of the variables $Y_1$ and $Y_2$ for discrimination is explored in Exercise 11.28.

The selection of appropriate variables to use in a discriminant analysis is often difficult. A summary such as the one in this example allows the investigator to make reasonable and simple choices based on the ultimate criteria of how well the procedure classifies its target objects.

**Figure 11.11** Box plots of petal width for the three species of iris.