References

1. Andrews, D.F., and A. M. Herzberg. *Data*. New York: Springer-Verlag, 1985.
2. Bartlett, M.S. "Further Aspects of the Theory of Multiple Regression." *Proceedings of the Cambridge Philosophical Society*, **34** (1938), 33–40.
3. Bartlett, M. S. "A Note on Tests of Significance in Multivariate Analysis." *Proceedings of the Cambridge Philosophical Society*, **35** (1939), 180–185.
4. Dunham, R.B. "Reaction to Job Characteristics: Moderating Effects of the Organization." *Academy of Management Journal*, **20**, no. 1 (1977), 42–65.
5. Hotelling, H. "The Most Predictable Criterion." *Journal of Educational Psychology*, **26** (1935), 139–142.
6. Hotelling, H. "Relations between Two Sets of Variables." *Biometrika*, **28** (1936), 321–377.
7. Johnson, R. A., and T. Wehrly. "Measures and Models for Angular Correlation and Angular-Linear Correlation." *Journal of the Royal Statistical Society (B)*, **39** (1977), 222–229.
8. Kshirsagar, A. M. *Multivariate Analysis*. New York: Marcel Dekker, Inc., 1972.
9. Lawley, D. N. "Tests of Significance in Canonical Analysis." *Biometrika*, **46** (1959), 59–66.
10. Parker, R. N., and M. D. Smith. "Deterrence, Poverty, and Type of Homicide." *American Journal of Sociology*, **85** (1979), 614–624.
11. Rencher, A. C. "Interpretation of Canonical Discriminant Functions, Canonical Variates and Principal Components." *The American Statistician*, **46** (1992), 217–225.
12. Waugh, F. W. "Regression between Sets of Variates." *Econometrica*, **10** (1942), 290–310.

Chapter 11

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.

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 π_1 and π_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 π_1 and those from the second class as the population of $\mathbf{x}$ values for π_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.

Populations π_1 and π_2	Measured variables $\mathbf{X}$
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.

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 terminology of *observation* and *population*, we want to identify an observation of

[1] If the values of $\mathbf{X}$ were not very different for objects in π_1 and π_2, there would be no problem; that is, the classes would be indistinguishable, and new objects could be assigned to either class indiscriminately.

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 π_1, purchasers, or population π_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 π_1, and if it falls in R_2, we allocate it to population π_2. Thus, one set of observed values favors π_1, while the other set of values favors π_2.

You may wonder at this point how it is we *know* that some observations belong to a particular population, but we are unsure about others. (This, of course, is what makes classification a problem!) Several conditions can give rise to this apparent anomaly (see [20]):

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 π_2 object as belonging to π_1 or a π_1 object as belonging to π_2.

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

Table 11.1

π_1: Riding-mower owners		π_2: Nonowners	
x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)	x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)
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

These data are plotted in Figure 11.1. We see that riding-mower owners tend to have larger incomes and bigger lots than nonowners, although income seems to be a better "discriminator" than lot size. On the other hand, there is some overlap between the two groups. If, for example, we were to allocate those values of (x_1, x_2) that fall into region R_1 (as determined by the solid line in the figure) to π_1, mower owners, and those (x_1, x_2) values which fall into R_2 to π_2, nonowners, we would make some mistakes. Some riding-mower owners would be incorrectly classified as nonowners and, conversely, some nonowners as owners. The idea is to create a rule (regions R_1 and R_2) that minimizes the chances of making these mistakes. (See Exercise 11.2.) ■

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

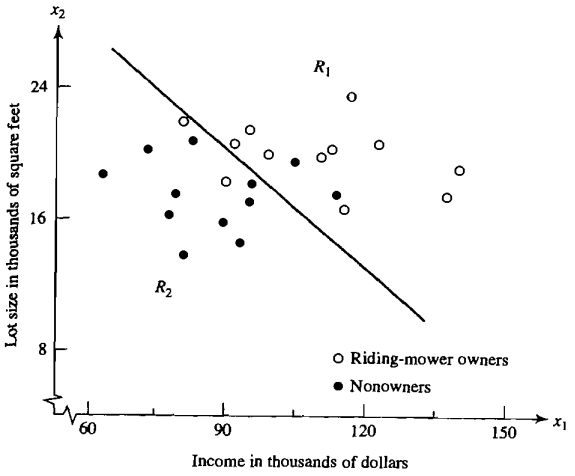

Figure 11.1 Income and lot size for riding-mower owners and nonowners.

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

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

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

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

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

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

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

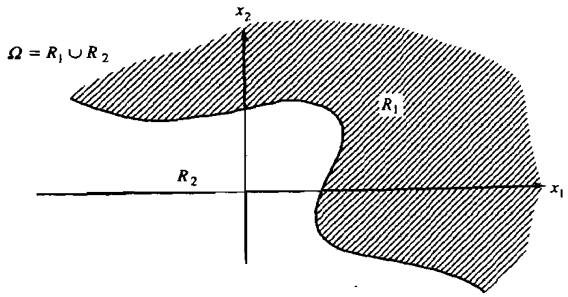

Figure 11.2 Classification regions for two populations.

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

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

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

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

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

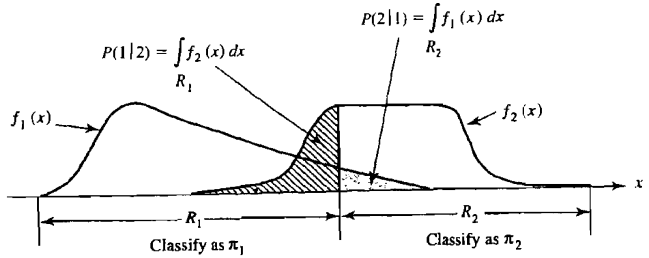

Figure 11.3 Misclassification probabilities for hypothetical classification regions when $p = 1$.

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

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

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

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

(11-4)

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

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

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

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

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:

$$\begin{aligned} R_1: \quad & \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right)\left(\frac{p_2}{p_1}\right) \\ & \begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} \geq \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix} \\ R_2: \quad & \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1 \mid 2)}{c(2 \mid 1)}\right)\left(\frac{p_2}{p_1}\right) \\ & \begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} < \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix} \end{aligned} \tag{11-6}$$

Proof. See Exercise 11.3. ■

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

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

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

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

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

(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)

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

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 π_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 π_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.

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 π_1 and π_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 π_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

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

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

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 π_1 or π_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 π_1. ■

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):

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

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).

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 π_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 π_1 and π_2 after observing $\mathbf{x}_0$ (hence the name *posterior* probabilities) is frequently useful for purposes of identifying the less clear-cut assignments.

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 π_1 and π_2 are given by

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

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

$$R_1: \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right] \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$
$$R_2: \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right] < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{11-11}$$

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 π_1 and π_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 π_1 if

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

Allocate $\mathbf{x}_0$ to π_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),

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

and, consequently,

$$R_1: (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$
$$R_2: (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{11-14}$$

The minimum ECM classification rule follows. ■

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 π_1 and n_2 measurements of this quantity from π_2, with $n_1 + n_2 - 2 \geq p$. Then the respective data matrices are

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

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

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

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

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

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

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

The Estimated Minimum ECM Rule for Two Normal Populations

Allocate $\mathbf{x}_0$ to π_1 if

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

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

If, in (11-18),

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

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

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

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

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

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

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

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

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

Example 11.3 (Classification with two normal populations—common $\boldsymbol{\Sigma}$ 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.

The investigators (see [4]) provide the information

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

[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 $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$, 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.)

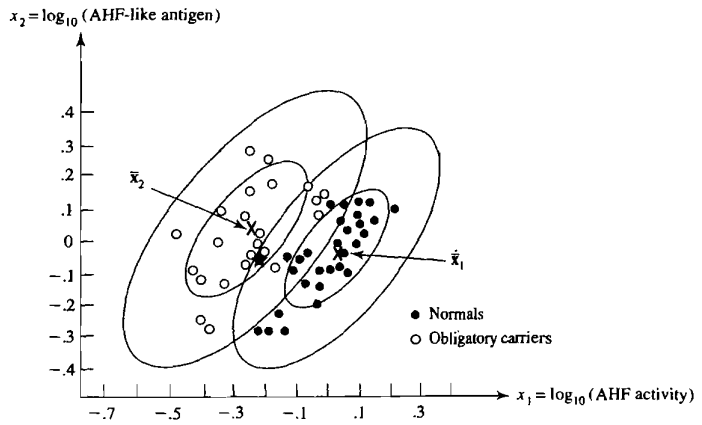

Figure 11.4 Scatter plots of [$\log_{10}$(AHF activity), $\log_{10}$(AHF-like antigen)] for the normal group and obligatory hemophilia A carriers.

and

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

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

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

Moreover,

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

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

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

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

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

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

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

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

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

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

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

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 π_2, an obligatory carrier. ■

Scaling

The coefficient vector $\hat{\mathbf{a}} = \mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ is unique only up to a multiplicative constant, so, for $c \neq 0$, any vector $c\hat{\mathbf{a}}$ will also serve as discriminant coefficients.

The vector $\hat{\mathbf{a}}$ is frequently "scaled" or "normalized" to ease the interpretation of its elements. Two of the most commonly employed normalizations are

1. Set

$$\hat{\mathbf{a}}^* = \frac{\hat{\mathbf{a}}}{\sqrt{\hat{\mathbf{a}}'\hat{\mathbf{a}}}} \tag{11-21}$$

so that $\hat{\mathbf{a}}^*$ has unit length.

2. Set

$$\hat{\mathbf{a}}^* = \frac{\hat{\mathbf{a}}}{\hat{a}_1} \tag{11-22}$$

so that the first element of the new coefficient vector $\hat{\mathbf{a}}^*$ is 1.

In both cases, $\hat{\mathbf{a}}^*$ is of the form $c\hat{\mathbf{a}}$. For normalization (1), $c = (\hat{\mathbf{a}}'\hat{\mathbf{a}})^{-1/2}$ and for (2), $c = \hat{a}_1^{-1}$.

The magnitudes of $\hat{a}_1^*, \hat{a}_2^*, \ldots, \hat{a}_p^*$ in (11-21) all lie in the interval $[-1, 1]$. In (11-22), $\hat{a}_1^* = 1$ and $\hat{a}_2^*, \ldots, \hat{a}_p^*$ are expressed as multiples of $\hat{a}_1^*$. Constraining the $\hat{a}_i^*$ to the interval $[-1, 1]$ usually facilitates a visual comparison of the coefficients. Similarly, expressing the coefficients as multiples of $\hat{a}_1^*$ allows one to readily assess the relative importance (vis-à-vis X_1) of variables $X_2, \ldots, X_p$ as discriminators.

Normalizing the $\hat{a}_i$'s is recommended only if the X variables have been standardized. If this is not the case, a great deal of care must be exercised in interpreting the results.

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 π_1 and π_2 were separated as much as possible. Fisher suggested taking linear combinations of $\mathbf{x}$ to create y's because they are simple enough functions of the $\mathbf{x}$ to be handled easily. Fisher's approach does not assume that the populations are normal. It does, however, implicitly assume that the population covariance matrices are equal, because a pooled estimate of the common covariance matrix is used.

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

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

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 $\bar{y}_2$.

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

$$\frac{\left(\begin{array}{c}\text{squared distance} \\ \text{between sample means of } y\end{array}\right)}{(\text{sample variance of } y)} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{\mathbf{a}}'\bar{\mathbf{x}}_1 - \hat{\mathbf{a}}'\bar{\mathbf{x}}_2)^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}} = \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}} \tag{11-23}$$

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 $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

Proof. The maximum of the ratio in (11-23) is given by applying (2-50) directly. Thus, setting $\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, we have

$$\max_{\hat{\mathbf{a}}} \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}} = \mathbf{d}'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = D^2$$

where D^2 is the sample squared distance between the two means. ■

Note that s_y^2 in (11-33) may be calculated as

$$s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2} \tag{11-24}$$

with $y_{1j} = \hat{\mathbf{a}}'\mathbf{x}_{1j}$ and $y_{2j} = \hat{\mathbf{a}}'\mathbf{x}_{2j}$.

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.

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

Allocate $\mathbf{x}_0$ to π_1 if

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

or

$$\hat{y}_0 - \hat{m} \geq 0 \tag{11-25}$$

Allocate $\mathbf{x}_0$ to π_2 if

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

or

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

[5] We must have $(n_1 + n_2 - 2) \geq p$; otherwise $\mathbf{S}_{\text{pooled}}$ is singular, and the usual inverse, $\mathbf{S}_{\text{pooled}}^{-1}$, does not exist.

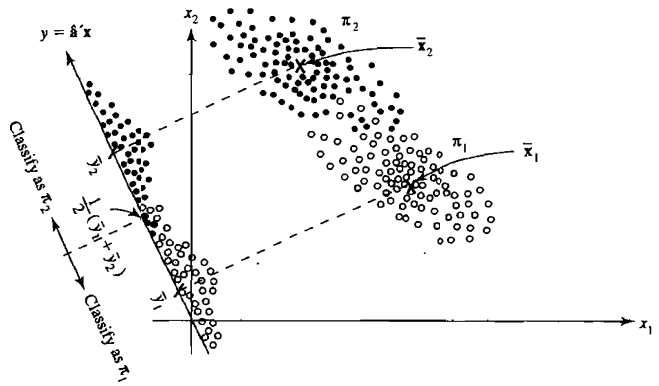

Figure 11.5 A pictorial representation of Fisher's procedure for two populations with $p = 2$.

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 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.

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 π_1 and π_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 π_1 and π_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 of Normal Populations When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

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 $\boldsymbol{\Sigma}_i$, $i = 1, 2$, replacing $\boldsymbol{\Sigma}$. 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 $|\boldsymbol{\Sigma}_i|^{1/2}$ do not cancel as they do when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. 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).

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

$$\begin{aligned}R_1:\ & -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \\ R_2:\ & -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]\end{aligned} \qquad (11\text{-}27)$$

where

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

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).

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

Result 11.4. Let the populations π_1 and π_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 π_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 π_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)

Allocate $\mathbf{x}_0$ to π_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 π_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 π_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.

[6] The inequalities $n_1 > p$ and $n_2 > p$ must both hold for $\mathbf{S}_1^{-1}$ and $\mathbf{S}_2^{-1}$ to exist. These quantities are used in place of $\boldsymbol{\Sigma}_1^{-1}$ and $\boldsymbol{\Sigma}_2^{-1}$, respectively, in the sample analog (11-29).

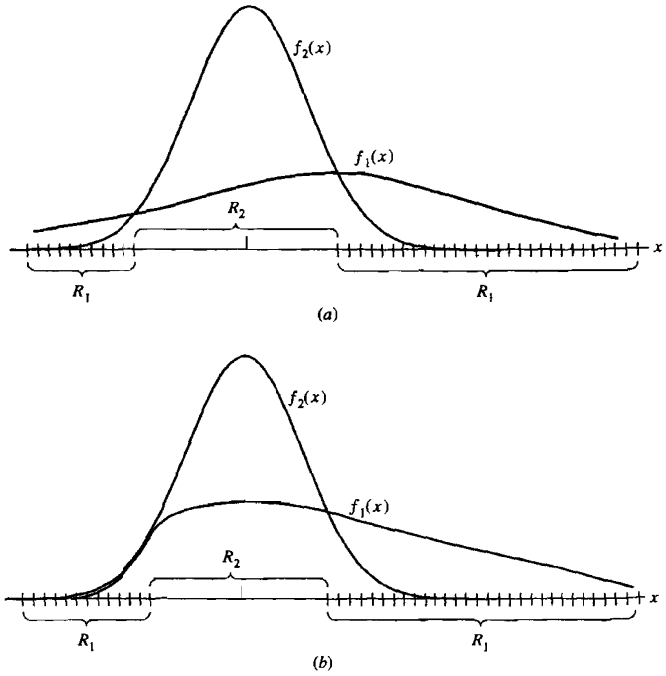

Figure 11.6 Quadratic rules for (a) two normal distribution with unequal variances and (b) two distributions, one of which is nonnormal—rule not appropriate.

If 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.

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.

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

$$R_1\text{: } (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq 0$$
$$R_2\text{: } (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < 0$$

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

$$R_1(y)\text{: } y \geq \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$
$$R_2(y)\text{: } y < \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

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

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

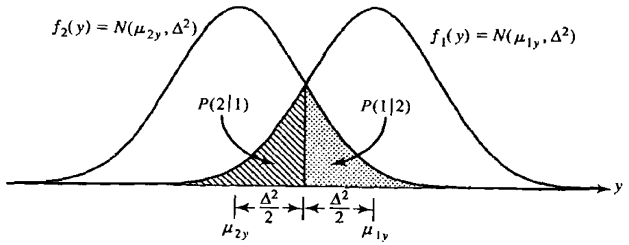

Figure 11.7 The misclassification probabilities based on Y.

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

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

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. ■

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),

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

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.

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

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

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 π_1 and n_2 observations from π_2, the confusion matrix has the form

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

(11-33)

where

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

The apparent error rate is then

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

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

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 π_1, mower owners; observations southwest of the solid line are classified as π_2, nonowners. Notice that some observations are misclassified. The confusion matrix is

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

The apparent error rate, expressed as a percentage, is

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

■

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.

A second approach that seems to work well is called Lachenbruch's "holdout" procedure[7] (see also Lachenbruch and Mickey [24]):

1. Start with the π_1 group of observations. Omit one observation from this group, and develop a classification function based on the remaining $n_1 - 1$, n_2 observations.
2. Classify the "holdout" observation, using the function constructed in Step 1.

[7]Lachenbruch's holdout procedure is sometimes referred to as *jackknifing* or *cross-validation*.

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

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

$$\hat{P}(2|1) = \frac{n_{1M}^{(H)}}{n_1}$$
$$\hat{P}(1|2) = \frac{n_{2M}^{(H)}}{n_2} \tag{11-35}$$

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

$$\hat{E}(\text{AER}) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} \tag{11-36}$$

Lachenbruch's holdout method is computationally feasible when used in conjunction with the linear classification statistics in (11-18) or (11-19). It is offered as an option in some readily available discriminant analysis computer programs.

Example 11.7 Calculating an estimate of the error rate using the holdout procedure) We shall illustrate Lachenbruch's holdout procedure and the calculation of error rate estimates for the equal costs and equal priors version of (11-18). Consider the following data matrices and descriptive statistics. (We shall assume that the $n_1 = n_2 = 3$ bivariate observations were selected randomly from two populations π_1 and π_2 with a common covariance matrix.)

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

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

The pooled covariance matrix is

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

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

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

and consequently,

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

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

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

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

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

with inverse[8]

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

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

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

$$\text{Squared distance from } \bar{\mathbf{x}}_{1H} = (\mathbf{x}_H - \bar{\mathbf{x}}_{1H})'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \bar{\mathbf{x}}_{1H})$$
$$= [2 - 3.5 \quad 12 - 9]\frac{1}{8}\begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix}\begin{bmatrix} 2 & -3.5 \\ 12 & -9 \end{bmatrix} = 4.5$$

$$\text{Squared distance from } \bar{\mathbf{x}}_2 = (\mathbf{x}_H - \bar{\mathbf{x}}_2)'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \bar{\mathbf{x}}_2)$$
$$= [2 - 4 \quad 12 - 7]\frac{1}{8}\begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix}\begin{bmatrix} 2 & -4 \\ 12 & -7 \end{bmatrix} = 10.3$$

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

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

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

[8] A matrix identity due to Bartlett [3] allows for the quick calculation of $\mathbf{S}_{H,\text{pooled}}^{-1}$ directly from $\mathbf{S}_{\text{pooled}}^{-1}$. Thus one does not have to recompute the inverse after withholding each observation. (See Exercise 11.20.)