Classification and Discriminant Analysis

Goldisse Fazeli

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfilment of the Requirements
for the Degree of Master of Science at
Concordia University
Montreal, Quebec, Canada

March 2000

0-612-47800-9

ABSTRACT

Classification and Discriminant Analysis

Goldisse Fazeli

This study provides a comprehensive review of the literature pertaining to the problem of classification. General concepts and principles of the classification problem are explored. These results are presented especially for populations under a normal distribution. Three major techniques of classification and discriminant analysis are presented: linear discriminant analysis, quadratic discriminant procedures and logistic regression. Logistic regression is reviewed in its general framework and as a classification tool. A few articles on the comparison of the efficiency of discriminant analysis and logistic regression are summarized. The discriminant approach is proven to be more efficient in the case of populations with a multivariate normal distribution. Under nonormality, logistic regression with maximum likelihood estimators outperforms discriminant analysis.

ACKNOWLEDGEMENTS

I would like to thank Dr.Y. P. Chaubey for sharing his statistical knowledge with me as a teacher and as my thesis supervisor. I thank him for suggesting the topic and providing the assistance essential to the completion of this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Introduction

The problem of classification

We are confronted with the problem of classification when we want to assign a unit into one of several categories (or populations) on the basis of measurements made on it. The goal of classification is to derive a rule that can be used to optimally assign a new observation to the labelled sets of observations (or populations).

Some examples of classification are as follows.

Example A Prospective students applying for admission into college: the problem is to classify a student applying for admission as *successful* or *unsuccessful* (fail to graduate) on the basis of his/her entrance examination scores, high-school grade-point average and number of high school activities.

Example B Classifying applicants for a bank loan as *low risk* or *high risk* on the basis of their income, age, number of credit cards, other existing loans and, family size.

Example C In anthropological studies, the problem of identifying a jawbone excavated from a burial ground as having belonged to *a male* or *a female*, based on measurements like circumference and volume.

In all the above examples, the problem is to assign a unit to one of a finite number of groups to which it may belong on the basis of a set of observed characteristics.

Discriminant analysis

Discriminant analysis is a general statistical tool in multivariate analysis, which separates distinct sets of objects (or observations) based on multivariate data. The first clear statement of the discrimination problem was given by Sir R.A. Fisher to classify skeletal remains. Fisher (1936) introduced the discriminant function for distinguishing between two multivariate normal observations with a common covariance matrix (see Chapter II).

The purpose of the thesis

There are three major techniques of discrimination

(1) *Linear* discriminant analysis: the classical approach by R.A. Fisher.

(2) *Quadratic* discriminant / classification procedure.

(3) *Logistic* regression / classification procedure.

The purpose of this thesis is to provide a comprehensive review of the literature pertaining to these methods. We also demonstrate the use of these methods through some examples.

In Chapter I we review the theory of classification in general terms. The standards of good classification are presented for two populations and then they are carried out for several populations. Finally, general classification procedures for populations involving the normal distribution are presented.

Chapter II is a review of the linear discriminant analysis, the method proposed by

Fisher (1936). Classification by Fisher's method is presented for two known multivariate normal populations and then they are extended to several populations. The distribution of the criterions of classification as well as the probabilities of misclassification are presented. Finally, some new diagnostic measures in linear discriminant analysis proposed by Fung (1995) are reviewed.

In Chapter III, the quadratic discriminant analysis, for classification and discrimination among two or more multivariate normal populations with unequal covariance matrices, is presented. The case of non-normal populations is also discussed.

In Chapter IV, the general framework of logistic regression is reviewed, followed by the relation between discriminant analysis and logistic regression.

Chapter V presents a summary of some results on the comparison of the efficiency of discriminant analysis and logistic regression.

Examples are provided to illustrate the results.

Chapter I General concepts and principles of the classification problem and discriminant analysis

1) Introduction

In this chapter, we introduce the concepts of classification and discriminant analysis (Johnson & Wichern, 1988). We investigate the standards of good classification through the optimal classification rule, some special cases of minimum expected costs, the total probability of misclassification and we evaluate classification functions for two populations. For more than two populations, we present the development of the optimal rules of classification. Finally, we elaborate the classification for normal populations.

An example of the separation-classification situation would be data collected on the sepal width, sepal length, petal width and petal length of three species of iris [see: Fisher (1936)]. The first goal would be to find "discriminant scores" such that the three classes of iris species are as separated as possible. And secondly, given a new iris, to classify it into one of the three classes.

Prior to the separation procedure, the probability distributions of the observations are checked. If the probability distributions are not known, we start by plotting the data for the pairs of observations in order to investigate their form. If the form of each distribution is known, then the parameters of the distributions are estimated from a sample of that population, called the *training sample*.

2) Classification for two populations: standards of good classification

We start by presenting classification for two populations π_1 and π_2, and later we shall treat the more general case.

Let $\mathbf{X}'=[X_1, X_2, \ldots, X_m]$ denote the vector of measurements of an observation. To classify $\mathbf{X}'$ into π_1 or π_2, we use the classification regions R_1 and R_2 obtained by the training sample. The training sample is the set of randomly selected objects known to come from each of the populations. We examine each object for their set of values $X_1, X_2, \ldots, X_m$ such that the set of all possible outcomes is divided into two regions R_1 and R_2.

Let Ω be the sample space divided into R_1 and R_2 such that $\Omega = R_1 \cup R_2$ and $R_1 \cap R_2 = \varnothing$. Hence R_1 and R_2 are mutually exclusive and exhaustive. If $\mathbf{X}'$ falls into R_1, we allocate it to population π_1, and if it falls in R_2, we allocate it to population π_2.

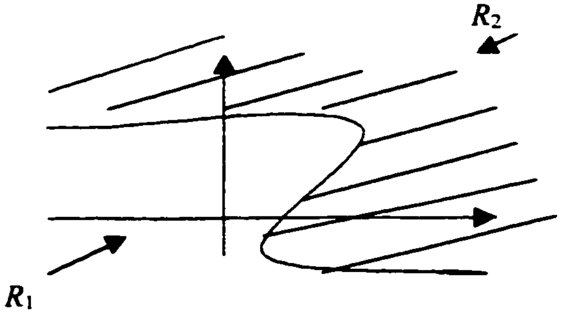

Figure 1.1 Classification regions R_1 and R_2 for two populations π_1 and π_2

Two kinds of errors in classification can be made when the sets of measured characteristics are not clearly distinct. One is to classify a π_2 object as belonging to π_1 and the other is to classify a π_1 object as belonging to π_2. A good classification procedure is one that minimises the probability of misclassification.

- **Optimal classification rules**

In the literature, two important features of an "optimal " classification rule are

1) the prior probabilities of occurrence,
2) the cost of misclassification.

1) If one population is relatively much larger than the other then it has a greater likelihood of occurrence. For example, there are more financially sound firms than bankrupt firms. Then the prior probability of a bankrupt firm is very small. A randomly selected firm should be classified as non-bankrupt unless the data overwhelmingly favors bankruptcy.

Let $f_1(\mathbf{X})$ and $f_2(\mathbf{X})$ be the probability density functions associated with $\mathbf{X}'$ for the populations π_1 and π_2, respectively. The prior probabilities for populations π_1 and π_2 are p_1 and p_2, respectively where $p_1 + p_2 = 1$.

The probabilities of correctly or incorrectly classifying observations are

Pr (observation is correctly classified as π_1)

$$= \Pr(\mathbf{X} \in R_1 \mid \pi_1)\Pr(\pi_1) = \Pr(1 \mid 1)\,p_1 = p_1 \int_{R_1} f_1(\mathbf{X})\,d\mathbf{X} \qquad (1.1)$$

Pr (observation is misclassified as π_1)

$$= \Pr(\mathbf{X} \in R_1 \mid \pi_2)\Pr(\pi_2) = \Pr(1 \mid 2)\,p_2 = p_2 \int_{R_1} f_2(\mathbf{X})\,d\mathbf{X} \qquad (1.2)$$

Pr (observation is correctly classified as π_2)

$$= \Pr(\mathbf{X} \in R_2 \mid \pi_2)\Pr(\pi_2) = \Pr(2 \mid 2)\, p_2 = p_2 \int_{R_2} f_2(\mathbf{X})\, d\mathbf{X} \qquad (1.3)$$

Pr (observation is misclassified as π_2)

$$= \Pr(\mathbf{X} \in R_2 \mid \pi_1)\Pr(\pi_1) = \Pr(2 \mid 1)\, p_1 = p_1 \int_{R_2} f_1(\mathbf{X})\, d\mathbf{X} \qquad (1.4)$$

where Pr (k | i) is the conditional probability of allocating an item to π_k when, in fact, it belongs to π_i.

As mentioned earlier, an optimal classification procedure is one that minimizes the probabilities of misclassification (1.2) & (1.4).

2) The cost of misclassification can be defined as a *cost matrix*.

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

These costs may be measured in any kind of unit. The costs are (1) zero for correct classification, (2) $c(1 \mid 2) > 0$ when an observation from π_2 is misclassified as π_1, and (3) $c(2 \mid 1)$ when an observation from π_1 is misclassified as π_2.

For any classification rule, the *expected cost of misclassification* (ECM) is the sum of the product of the misclassification costs and their probabilities of occurrence, i.e.

$$\text{ECM} = c(2 \mid 1)\Pr(2 \mid 1)\, p_1 + c(1 \mid 2)\Pr(1 \mid 2)\, p_2 \qquad (1.5)$$

An optimal classification rule should result in an ECM as small as possible. That is, we want to divide the sample space Ω into regions R_1 and R_2 such that the ECM is as small as possible. The regions R_1 and R_2 that minimize the ECM are defined by the values **X** for which the following inequalities hold.

$$\begin{aligned} R_1&: \ f_1(\mathbf{X}) / f_2(\mathbf{X}) \geq [c(1 \mid 2) / c(2 \mid 1)] \, [p_2 / p_1] \\ R_2&: \ f_1(\mathbf{X}) / f_2(\mathbf{X}) < [c(1 \mid 2) / c(2 \mid 1)] \, [p_2 / p_1] \end{aligned} \qquad (1.6)$$

Proof. From equation (1.5) we can write

$$\text{ECM} = c(2 \mid 1) \, p_1 \int_{R_2} f_1(\mathbf{X}) \, d\mathbf{X} + c(1 \mid 2) \, p_2 \int_{R_1} f_2(\mathbf{X}) \, d\mathbf{X}$$

Noting that $\Omega = R_1 \cup R_2$ so that

$$\int_{R_1} f_1(\mathbf{X}) \, d\mathbf{X} + \int_{R_2} f_1(\mathbf{X}) \, d\mathbf{X} = 1$$

we can write

$$\begin{aligned} \text{ECM} &= c(2 \mid 1) \, p_1 \left[1 - \int_{R_1} f_1(\mathbf{X}) \, d\mathbf{X} \right] + c(1 \mid 2) \, p_2 \int_{R_1} f_2(\mathbf{X}) \, d\mathbf{X} \\ &= \int_{R_1} \left[c(1 \mid 2) \, p_2 \, f_2(\mathbf{X}) - c(2 \mid 1) \, p_1 f_1(\mathbf{X}) \right] d\mathbf{X} + c(2 \mid 1) \, p_1 . \end{aligned}$$

where note that p_1. p_2, c(2 | 1), and c(1 | 2) are nonnegative. The density functions f_1(**X**) and f_2(**X**) are nonnegative for all **X** and are the only quantities in ECM that depend on **X**. Thus ECM is minimized if R_1 includes those values of **X** for which the integrand $[c(1 \mid 2) \, p_2 f_2(\mathbf{X}) - c(2 \mid 1) \, p_1 f_1(\mathbf{X})] \leq 0$ and excludes those **X** for which this quantity is positive. That is, R_1 must be the set of points **X** such that

$$c(1 \mid 2) \, p_2 f_2(\mathbf{X}) \leq c(2 \mid 1) \, p_1 f_1(\mathbf{X})$$

or

$$f_1(\mathbf{X}) / f_2(\mathbf{X}) \geq [c(1 \mid 2) / c(2 \mid 1)] \, [p_2 / p_1].$$

Since R_2 is the complement of R_1 in Ω, R_2 must be the set of points **X** for which

$$f_1(\mathbf{X}) / f_2(\mathbf{X}) < [c(1 \mid 2) / c(2 \mid 1)] [p_2 / p_1].$$

In the literature, a procedure that minimizes (1.5) for given p_1 and p_2 is called a *Bayes procedure* (Anderson, 1984).

Note. If $p_1 f_1(\mathbf{X})\, c(2 \mid 1) = p_2\, f_2(\mathbf{X})\, c(1 \mid 2)$, then **X** could be classified either as from π_1 or π_2.

If $\Pr \{f_1(\mathbf{X}) / f_2(\mathbf{X}) = [c(1 \mid 2) / c(2 \mid 1)] [p_2 / p_1] \mid \pi_i \} = 0$ for $i = 1,2$ (1.7)

then the procedure in (1.6) is unique except for the sets of probability zero (Anderson, 1984).

From (1.6) it is clear that the implementation of the minimum ECM rule requires (1) the density function ratio evaluated at a new observation, (2) the cost ratio, and (3) the prior probability ratio. In this chapter, we shall discuss special cases where each one of these components is unknown. The presence of ratios in (1.6) is significant because often it is much easier to specify the ratios than their component parts. For example, the cost to a credit company of classifying an applicant as a good client when, in fact he or she has no credit profile and classifying an applicant as a bad client when, in fact he or she has an excellent credit profile, is difficult to specify. However, a realistic number for the cost ratio of such misclassification can be obtained. Not admitting a client with a good credit profile may be four times more costly, over a determined period of time, than admitting a client with no credit profile. Thus, the cost ratio is four.

3) Special cases of minimum expected cost regions

- **The prior probabilities are unknown**

If the prior probabilities are unknown, they are often taken to be equal, i.e. $(p_1/p_2) = 1$. The regions R_1 and R_2 that minimize the ECM are defined by

$$R_1\text{: } f_1(\mathbf{X})/f_2(\mathbf{X}) \geq [c(1\mid 2)/c(2\mid 1)] \qquad (1.8)$$
$$R_2\text{: } f_1(\mathbf{X})/f_2(\mathbf{X}) < [c(1\mid 2)/c(2\mid 1)]$$

- **The misclassification cost ratio is indeterminate**

If the misclassification cost ratio is indeterminate, it is usually taken to be unity, $[c(1\mid 2)/c(2\mid 1)] = 1$. In this case, the optimal classification regions R_1 and R_2 are chosen to minimize the *total probability of misclassification* (TPM).

$$\text{TPM} = p_1 \int_{R_2} f_1(\mathbf{X})\, d\mathbf{X} + p_2 \int_{R_1} f_2(\mathbf{X})\, d\mathbf{X} \qquad (1.9)$$

$$R_1\text{: } f_1(\mathbf{X})/f_2(\mathbf{X}) \geq [p_2/p_1] \qquad (1.10)$$
$$R_2\text{: } f_1(\mathbf{X})/f_2(\mathbf{X}) < [p_2/p_1]$$

- **Equal prior probabilities and equal misclassification cost ratios**

When both the prior probability and misclassification cost ratios are unity or one ratio is the reciprocal of the other, i.e. $(p_1/p_2) = [c(1\mid 2)/c(2\mid 1)] = 1$ or $(p_1/p_2) = 1/[c(1\mid 2)/c(2\mid 1)]$, the optimal classification regions R_1 and R_2 are given by

$$R_1\text{: } f_1(\mathbf{X})/f_2(\mathbf{X}) \geq 1 \qquad (1.11)$$
$$R_2\text{: } f_1(\mathbf{X})/f_2(\mathbf{X}) < 1$$

- **Conditional or posterior probability**

Another way of minimizing the probability of misclassification is to allocate a new observation $\mathbf{X}_o$ to the population that has the higher conditional or posterior probability. Given a new observation $\mathbf{X}_o$, the conditional probability of coming from population π_1 is

$$\begin{aligned}\Pr(\pi_1 \mid \mathbf{X}_o) &= \Pr(\pi_1 \text{ and observe } \mathbf{X}_o) \;/\; \Pr(\text{observe } \mathbf{X}_o) \qquad (1.12)\\ &= \left[\Pr(\text{observe } \mathbf{X}_o \mid \pi_1)\Pr(\pi_1)\right] \;/\; \left[\Pr(\text{observe } \mathbf{X}_o \mid \pi_1)\Pr(\pi_1)\right.\\ &\quad \left. + \Pr(\text{observe } \mathbf{X}_o \mid \pi_2)\Pr(\pi_2)\right]\\ &= p_1 f_1(\mathbf{X}_o) \;/\; \left[\, p_1 f_1(\mathbf{X}_o) + p_2 f_2(\mathbf{X}_o)\right]\end{aligned}$$

$$\Pr(\pi_2 \mid \mathbf{X}_o) = 1 - \Pr(\pi_1 \mid \mathbf{X}_o) = p_2 f_2(\mathbf{X}_o) \;/\; \left[\, p_1 f_1(\mathbf{X}_o) + p_2 f_2(\mathbf{X}_o)\right]$$

If $\Pr(\pi_1 \mid \mathbf{X}_o) \geq \Pr(\pi_2 \mid \mathbf{X}_o)$, we classify $\mathbf{X}_o$ as π_1. Otherwise, we classify $\mathbf{X}_o$ as π_2. The optimal classification regions R_1 and R_2 are equivalent to (1.10).

4) Evaluating classification functions

In this section, we present an important feature, which plays an essential role in the performance of a classification procedure. That is the error rate or misclassification probabilities. If the forms of the parent populations are completely known then the smallest value of the *total probability of misclassification* (TPM) (1.9), obtained by a sensible choice of R_1 and R_2. is called the *optimum error rate* (OER).

$$\text{OER} = p_1 \int_{R_2} f_1(\mathbf{X})\, d\mathbf{X} + p_2 \int_{R_1} f_2(\mathbf{X})\, d\mathbf{X} \qquad (1.13)$$

where R_1 and R_2 are obtained by (1.10).

We can define the OER as the error rate for the minimum TPM classification rule. In the event that the parameters of the parent populations are not known, as mentioned earlier, they are estimated from the training sample. The performance of sample classification functions is evaluated by the actual error rate (AER).

$$\text{AER} = p_1 \int_{R_2} f_1(\mathbf{X})\, d\mathbf{X} + p_2 \int_{R_1} f_2(\mathbf{X})\, d\mathbf{X} \qquad (1.14)$$

where R_1 and R_2 are the classification regions by samples of size n_1 and n_2, respectively.

There are also error rate estimates that do not depend on the form of the parent populations and that can be calculated for any classification procedure. One of them is called the *apparent error rate* (APER). The APER is defined as the function of observations in the training sample that are misclassified by the sample classification function. This measure is calculated from the *confusion matrix*, which shows the actual versus predicted group membership.

		Predicted membership		
		π_1	π_2	
Actual membership	π_1	n_{1c}	$n_{1m} = n_1 - n_{1c}$	n_1
	π_2	$n_{2m} = n_2 - n_{2c}$	n_{2c}	n_2

n_1 = number of observations from π_1.

n_2 = number of observations from π_2.

n_{1c} = number of π_1 items correctly classified as π_1.

n_{2c} = number of π_2 items correctly classified as π_2.

n_{1m} = number of π_1 items misclassified.

n_{1m} = number of π_2 items misclassified.

$$\mathrm{APER} = (n_{1m} + n_{2m}) / (n_1 + n_2) \qquad (1.15)$$

As presented by (1.15), APER is the proportion of items in the training set that are misclassified.

Note. Unless the sample sizes n_1 and n_2 are very large. the APER underestimates the AER (Johnson & Wichern, 1988).

There exists other methods to estimate error rate, which are better than APER, are easy to calculate, and do not require distributional assumptions. One method is to split the total sample into a *training* sample and a *validation* sample, which are used to construct and evaluate the classification function, respectively. The disadvantages of this procedure are (1) it requires large samples, and (2) the function evaluated is not the function of interest. In order not to lose any valuable information, almost all of the data must be used to construct the classification function.

A method called *Lachenbruch's holdout procedure* (Lachenbruch & Mickey, 1968), which seems to work well is

1. Start with the π_1 group of observations. Omit one observation from this group and develop a classification function based on the remaining n_1-1, n_2 observations.
2. Classify the "holdout" observation using the function constructed in step 1.
3. Repeat steps 1 and 2 until all of the π_1 observations are classified. Let $n_{1m}^{(H)}$ be the number of holdout (H) observations misclassified in this group.
4. Repeat steps 1 through 3 for the π_2 observations. Let $n_{2m}^{(H)}$ be the number of holdout observations misclassified in this group.

Estimates Pr (2 | 1) and Pr (1 | 2) of the conditional misclassification probabilities in (1.2) and (1.4) are then given by

$$\Pr(2 \mid 1) = n_{1m}^{(H)} / n_1$$
$$\Pr(1 \mid 2) = n_{2m}^{(H)} / n_2 \qquad (1.16)$$

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(AER).

$$E(AER) = (n_{1m}^{(H)} + n_{2m}^{(H)}) / (n_1 + n_2) \qquad (1.17)$$

Lachenbruch's holdout procedure is computationally feasible when used in conjunction with linear classification statistics (see Chapter II).

As a conclusion, we note that a good classification rule depends on the separation of the population. Hence, it is important to effectively separate the groups as much as possible in order to develop good classification rules.

5) <u>Classification with several populations</u>

In this section, we present the development of the optimal rules to classify more than two populations. Let $f_i(\mathbf{X})$ be the density associated with population π_i, i =1,2,...,g.

Let p_i = the prior probability of population π_i, i =1,2,...,g.

c(k | i) = the cost of allocating an item to π_k when, in fact, it belongs to π_i,

for i, k = 1,2,...,g.

For k = i, c(i | i) = 0. Finally, let R_k be the set of **X**'s classified as π_k and

$$\Pr(k \mid i) = \Pr(\text{classify observation as } \pi_k \mid \pi_i) = \int_{R_k} f_i(\mathbf{X})\, d\mathbf{X}$$

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

The conditional expected cost of misclassifying an **X** from π_k to π_i, i, k = 1,2,...,g and k ≠ i is

$$\mathrm{ECM}(k) = \sum_{\substack{i=1 \\ k \neq i}}^{g} \Pr(i \mid k)\, c(i \mid k).$$

The overall ECM is given by

$$\begin{aligned} \mathrm{ECM} &= p_1\, \mathrm{ECM}(1) + \ldots + p_g\, \mathrm{ECM}(g) \\ &= \sum_{i=1}^{g} p_i \left(\sum_{\substack{i=1 \\ k \neq i}} \Pr(i \mid k)\, c(i \mid k) \right). \end{aligned} \qquad (1.18)$$

In order to develop an optimal classification rule, we must choose mutually exclusive and exhaustive classification regions $R_1, R_2, \ldots, R_g$ such that the overall ECM be a minimum. A judicious choice is to choose the classification regions by allocating **X** to that population π_k, k = 1,2,...,g for which

$$\sum_{\substack{i=1 \\ k \neq i}}^{g} p_i\, f_i(\mathbf{X})\, c(i \mid k) \qquad (1.19)$$

is smallest. We note that if a tie occurs, then **X** can be assigned to any of the tied populations (Anderson, 1984).

We look at the case where all the misclassification costs are equal. Without loss of generality, we set them equal to one. Following the same logic as for (1.19), we would allocate **X** to that population π_k, k = 1,2,...,g for which

$$\sum_{\substack{i=1 \\ k \neq i}}^{g} p_i\, f_i(\mathbf{X}) \quad \text{is smallest}$$

$$\text{or} \qquad p_k\, f_k(\mathbf{X}) \qquad \text{is largest.}$$

In that case, the minimum expected cost of misclassification rule has the following form:

Allocate $\mathbf{X}$ to π_k if $\quad p_k\, f_k(\mathbf{X}) > p_i\, f_i(\mathbf{X}) \quad$ for all $i \neq k$

or allocate $\mathbf{X}$ to π_k if $\quad \ln p_k\, f_k(\mathbf{X}) > \ln p_i\, f_i(\mathbf{X}) \quad$ for all $i \neq k$ (1.20)

We note that the components of the minimum ECM rules (prior probabilities, misclassification costs, and density functions) must be specified (or estimated) before the rules can be implemented.

Another approach to determine a minimum ECM rule with equal misclassification costs is to maximize the *posterior probability*.

$$\Pr(\pi_k \mid \mathbf{X}) = \Pr(\mathbf{X} \text{ comes from } \pi_k \text{ given that } \mathbf{X} \text{ was observed}).$$

where $\Pr(\pi_k \mid \mathbf{X}) = (p_k\, f_k(\mathbf{X}) \,/\, \sum_{n=1}^{g} p_n\, f_n(\mathbf{X}))$ (1.21)

$$= \big((\textit{prior}) \times (\textit{likelihood})\big) \,/\, \big(\Sigma\, [(\textit{prior}) \times (\textit{likelihood})]\big)$$

for $k = 1,2,\ldots,g$.

Equation (1.21) is the generalization of equation (1.12).

6) Classification with normal populations

In this section, we present an important special case. That is when $f_i(\mathbf{X})$ is a multivariate normal density with mean vectors μ_i and covariance matrices Σ_i.

$f_i(\mathbf{X}) = \left[1 / (2\pi)^{m/2} \mid \Sigma_i \mid ^{1/2}\right] \exp\left[(-1/2)\,(\mathbf{X}- \mu_i)' \,\Sigma_i^{-1}\, (\mathbf{X}- \mu_i)\right]$, for i = 1,2, ... ,g (1.22)

If the misclassification costs are all equal (or $c(k \mid i) = 1$ for $k \neq i$) then the rule in (1.20) becomes:

Allocate **X** to π_k if

$$\ln p_k f_k(\mathbf{X}) = \ln p_k - (\,m/2)\ln(2\pi) - (1/2)\ln\mid \Sigma_k \mid - (1/2)\,(\mathbf{X}- \mu_k)'\,\Sigma_k^{-1}\,(\mathbf{X}- \mu_k)$$

$$= \max_i \ln p_i f_i(\mathbf{X}) \qquad (1.23)$$

Eliminating the constant $(m/2)\ln(2\pi)$ since it is the same for all the populations, we define (1.23) as the *quadratic discrimination score* for the *i*th population to be

$$D_i^Q(\mathbf{X}) = (-1/2)\ln\mid \Sigma_i \mid - (1/2)\,(\mathbf{X}- \mu_i)'\,\Sigma_i^{-1}\,(\mathbf{X}- \mu_i) + \ln p_i \qquad i = 1,2,\ldots, g \qquad (1.24)$$

Hence, we obtain the following *minimum total probability of misclassification* rule for normal populations:

Allocate **X** to π_k if

the quadratic score $D_k^Q(\mathbf{X})$ = largest of $D_1^Q(\mathbf{X}), D_2^Q(\mathbf{X}), \ldots, D_g^Q(\mathbf{X})$ (1.25)

The estimates of μ_i and Σ_i, when they are unknown, are obtained through a training set of correctly classified observations.

The relevant sample quantities for populations π_i are

$\bar{\mathbf{X}}_i$ = sample mean vector

$\mathbf{S}_i$ = sample covariance matrix

$\mathbf{n}_i$ = sample size

The estimate of D_i^Q is then

$$d_i^Q(\mathbf{X}) = (-1/2)\ln|\mathbf{S}_i| - (1/2)(\mathbf{X}-\bar{\mathbf{X}}_i)'\,\mathbf{S}_i^{-1}\,(\mathbf{X}-\bar{\mathbf{X}}_i) + \ln p_i$$

Hence, the *estimated* minimum TPM rule for several normal populations is

Allocate $\mathbf{X}$ to π_k if

the quadratic score $d_k^Q(\mathbf{X}) = \max(d_1^Q(\mathbf{X}), d_2^Q(\mathbf{X}), \ldots, d_g^Q(\mathbf{X}))$ (1.26)

If the population covariance matrices, Σ_i, are equal, i.e. $\Sigma_i = \Sigma$ for $i = 1,2,\ldots,g$. the discriminant score in (1.24) becomes

$$D_i^Q(\mathbf{X}) = (-1/2)\ln|\Sigma| - (1/2)(\mathbf{X}'\Sigma^{-1}\mathbf{X}) + \mu_i'\Sigma^{-1}\mathbf{X} - (1/2)(\mu_i'\Sigma^{-1}\mu_i) + \ln p_i \quad (1.27)$$

We can ignore $((-1/2)\ln|\Sigma| - (1/2)(\mathbf{X}'\Sigma^{-1}\mathbf{X}))$ since it is the same for $D_1^Q(\mathbf{X})$, $D_2^Q(\mathbf{X}),\ldots, D_g^Q(\mathbf{X})$. Consequently, we get the *linear discriminant score* as

$$D_i^Q(\mathbf{X}) = \mu_i'\Sigma^{-1}\mathbf{X} - (1/2)(\mu_i'\Sigma^{-1}\mu_i) + \ln p_i \quad (1.28)$$

The minimum TPM rule for *equal covariance* normal populations is:

Allocate $\mathbf{X}$ to π_k if

the linear discriminant score $D_k(\mathbf{X})$ = largest of $D_1(\mathbf{X}), D_2(\mathbf{X}), \ldots, D_g(\mathbf{X})$ (1.29)

The estimate $d_i(\mathbf{X})$, of the linear discriminant score $D_i(\mathbf{X})$ is based on the pooled estimate of Σ.

$$\mathbf{S}_{\text{pooled}} = ((n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2 + \ldots + (n_g-1)\mathbf{S}_g) \,/\, (n_1 + n_2 + \ldots + n_g) \quad (1.30)$$

and is given by

$$d_i(\mathbf{X}) = (\bar{\mathbf{X}}_i'\,\mathbf{S}^{-1}_{\text{pooled}}\,\mathbf{X}) - ((1/2)(\bar{\mathbf{X}}_i'\,\mathbf{S}^{-1}_{\text{pooled}}\bar{\mathbf{X}}_i)) + \ln p_i \quad (1.31)$$

Consequently, the *estimated* minimum TPM rule for *equal covariance* normal populations is:

Allocate **X** to π_k if

the linear discriminant score $d_k(\mathbf{X}) = \max(d_1(\mathbf{X}), d_2(\mathbf{X}), \ldots, d_g(\mathbf{X}))$ (1.32)

Another approach for the equal covariance case is obtained from (1.24) by ignoring the constant term, $(-1/2)\ln|\Sigma|$, where the allocatory rule is given by:

Allocate **X** to the population π_i for which

$$(-1/2)\,(\mathbf{X}-\overline{\mathbf{X}}_i)'\,\mathbf{S}^{-1}_{\text{pooled}}\,(\mathbf{X}-\overline{\mathbf{X}}_i)+\ln p_i \quad \text{is largest.} \qquad (1.33)$$

We can interpret $(\mathbf{X}-\overline{\mathbf{X}}_i)'\,\mathbf{S}^{-1}_{\text{pooled}}\,(\mathbf{X}-\overline{\mathbf{X}}_i)$ as the squared distance from **X** to the sample mean vector $\mathbf{X}_i$. We note that both rules, (1.32) and (1.33), assign **X** to the closest population.

Remark. If the prior probabilities are unknown, they are set to be $p_1 = p_2 = \ldots = p_g = 1/g$.

Chapter II Linear discriminant analysis: *Fisher's method*

1) Classification for two populations by *Fisher's method*

In this section we present the method proposed by Fisher (1936) which consists of transforming the multivariate observations $\mathbf{X}'=[X_1, X_2, \ldots, X_m]$ (the vector of measurements of m relevant variables of an observation) to univariate observations $\mathbf{Y}$ such that the $\mathbf{Y}$'s derived from populations π_1 and π_2 are separated as much as possible.

Fisher's idea was to take linear combinations of $\mathbf{X}$ in order to create the univariate observation $\mathbf{Y}$ to create a single index for classifying observations. Let μ_{1y} and μ_{2y} be the means of the $\mathbf{Y}$'s obtained from $\mathbf{X}$'s belonging to π_1 and π_2, respectively. Let the mean and covariance matrix of $\mathbf{X}$ be denoted by

$\mu_1 = E\ (\mathbf{X} | \pi_1)$ = expected value of a multivariate observation from π_1

$\mu_2 = E\ (\mathbf{X} | \pi_2)$ = expected value of a multivariate observation from π_2

$\Sigma_i = E\ (\mathbf{X} - \mu_i)\ (\mathbf{X} - \mu_i)',\ i = 1,2.$

We consider the case $\Sigma_1 = \Sigma_2 = \Sigma$ and the linear combination

$$\underset{(1\times 1)}{\mathbf{Y}} = \underset{(1\times m)}{L'}\ \underset{(m\times 1)}{\mathbf{X}} \tag{2.1}$$

with

$$\begin{aligned} \mu_{1Y} &= E\ (\mathbf{Y} | \pi_1) = E\ (L'\mathbf{X} | \pi_1) = L'\mu_1 \\ \mu_{2Y} &= E\ (\mathbf{Y} | \pi_2) = E\ (L'\mathbf{X} | \pi_2) = L'\mu_2 \end{aligned} \tag{2.2}$$

$$\sigma_Y^2 = \mathrm{Var}\ (L'\mathbf{X}) = L'\ \mathrm{Cov}\ (\mathbf{X})\ L = L'\ \Sigma\ L \tag{2.3}$$

Fisher's idea was to choose the linear combinations that maximized the (squared) distance between μ_{1Y} and μ_{2Y} relative to the variability of the $\mathbf{Y}$'s, σ_Y^2:

$$[(\mu_{1Y} - \mu_{2Y})^2 / \sigma_Y^2] = [(L'\mu_1 - L'\mu_2)^2 / (L'\Sigma L)]$$

$$= [L'(\mu_1 - \mu_2)(\mu_1 - \mu_2)' L] / (L'\Sigma L)$$

$$= (L'\delta)^2 / (L'\Sigma L) \qquad (2.4)$$

where $\delta = \mu_1 - \mu_2$.

The coefficients $L' = [L_1, L_2, \ldots, L_m]$ which maximize the ratio (2.4), are called the *Fisher's linear combination coefficients.* We maximize the numerator in (2.4) with respect to L and we hold the denominator constant (Anderson, 1984). If λ is a *Lagrange multiplier*, we seek the maximum of

$$L'(\mu_1 - \mu_2)(\mu_1 - \mu_2)' L - \lambda (L'\Sigma L - 1) \qquad (2.5)$$

Taking the derivatives of (2.5) with respect to the components of L and equating them to zero, we get

$$2 [(\mu_1 - \mu_2)(\mu_1 - \mu_2)'] L = 2 \lambda \Sigma L \qquad (2.6)$$

Since $(\mu_1 - \mu_2)' L$ is a scalar, say v, we can write (2.6) as

$$\mu_1 - \mu_2 = (\lambda / v) \Sigma L \qquad (2.7)$$

$$\Rightarrow \quad \Sigma^{-1} (\mu_1 - \mu_2) = (\lambda / v) L$$

$$\Rightarrow \quad L = (v/\lambda) \Sigma^{-1} (\mu_1 - \mu_2) \qquad (2.8)$$

The ratio (2.4) is maximized by choice of L in (2.8), for any $(v/\lambda) \neq 0$.

Choosing $(v/\lambda) = 1$ produces the linear combination

$$\mathbf{Y} = L'\mathbf{X} = (\mu_1 - \mu_2)' \Sigma^{-1} \mathbf{X} \qquad (2.9)$$

which is known as ***Fisher's linear discriminant function.***

Note. The maximum of the ratio in (2.4) is given by

$$\max_L \left[(L'\delta)^2 / (L'\Sigma L)\right] = \delta'\Sigma^{-1}\delta.$$

Fisher's discriminant function does not depend on the form of the parent populations π_1 and π_2. However, there are non-normal cases where Fisher's discriminant function performs poorly.

2) Classification into one of two known multivariate normal populations

In this section we use the optimal classification rule for two populations outlined in Chapter I, in the case of two multivariate normal populations with equal covariance matrices, Σ (Anderson, 1984). The vector of means of the *i*th population is $\mu_i' = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im})$, $i = 1,2$. This approach was first given by Wald (1944).

The *i*th density is

$$f_i(\mathbf{X}) = \left[1 / (2\pi)^{m/2} |\Sigma|^{1/2}\right] \exp\left[(-1/2)(\mathbf{X} - \mu_i)' \Sigma^{-1} (\mathbf{X} - \mu_i)\right] \tag{2.10}$$

The ratio of the densities is

$$(f_1(\mathbf{X}) / f_2(\mathbf{X})) = (\exp\left[(-1/2)(\mathbf{X} - \mu_1)' \Sigma^{-1} (\mathbf{X} - \mu_1)\right] / \exp\left[(-1/2)(\mathbf{X} - \mu_2)' \Sigma^{-1} (\mathbf{X} - \mu_2)\right])$$

$$= \exp\left\{(-1/2)\left[(\mathbf{X} - \mu_1)' \Sigma^{-1} (\mathbf{X} - \mu_1) - (\mathbf{X} - \mu_2)' \Sigma^{-1} (\mathbf{X} - \mu_2)\right]\right\} \tag{2.11}$$

The regions of classification R_1 and R_2 are given by

$$\begin{aligned} R_1&: (f_1(\mathbf{X}) / f_2(\mathbf{X})) \geq k \\ R_2&: (f_1(\mathbf{X}) / f_2(\mathbf{X})) < k \end{aligned} \tag{2.12}$$

for k suitably chosen.

In order to simplify the ratio in (2.11), we rewrite it in term of its logarithm function:

$$(-1/2)\left[(\mathbf{X}-\mu_1)'\Sigma^{-1}(\mathbf{X}-\mu_1)-(\mathbf{X}-\mu_2)'\Sigma^{-1}(\mathbf{X}-\mu_2)\right]=\ln\left(f_1(\mathbf{X})/f_2(\mathbf{X})\right) \qquad (2.13)$$

Rearranging the terms we obtain

$$\ln\left(f_1(\mathbf{X})/f_2(\mathbf{X})\right)=\mathbf{X}'\,\Sigma^{-1}(\mu_1-\mu_2)-(1/2)(\mu_1+\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2) \qquad (2.14)$$

The first term is the *Fisher's linear discriminant function.* The second term is the midpoint M between the two-univariate population means.

$$\begin{aligned} M &= (1/2)\,(\mu_{1Y}+\mu_{2Y}) = (1/2)\,(\mathrm{L}'\mu_1+\mathrm{L}'\mu_2) \\ &= (1/2)\,(\mu_1+\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2) \end{aligned} \qquad (2.15)$$

Hence the best regions of classification that minimize the expected cost of misclassification, are

$$\begin{aligned} &R_1\text{: } \mathbf{X}'\,\Sigma^{-1}(\mu_1-\mu_2)-(1/2)(\mu_1+\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2)\ \geq\ \ln k \\ &R_2\text{: } \mathbf{X}'\,\Sigma^{-1}(\mu_1-\mu_2)-(1/2)(\mu_1+\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2)\ <\ \ln k \end{aligned} \qquad (2.16)$$

If prior probabilities p_1 and p_2 are known, then k is given by

$$k=\left[c(1\,|\,2)/c(2\,|\,1)\right]\left[p_2/p_1\right] \qquad (2.17)$$

In the case of two populations being equally likely and the costs being equal, $k = 1$ and $\ln k = 0$. Then the regions of classification are

$$\begin{aligned} &R_1\text{: } \mathbf{X}'\,\Sigma^{-1}(\mu_1-\mu_2)\geq\ (1/2)(\mu_1+\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2) \\ &R_2\text{: } \mathbf{X}'\,\Sigma^{-1}(\mu_1-\mu_2)<\ (1/2)(\mu_1+\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2) \end{aligned} \qquad (2.18)$$

If we do not have the prior probabilities, we may select $\ln k = c$, say, on the basis of making the expected losses due to misclassification equal.

3) The distribution of the criterion of classification

Let $\mathbf{X}$ be a random observation. We are interested to find the distribution of

$$U = \mathbf{X}' \Sigma^{-1} (\mu_1 - \mu_2) - (1/2) (\mu_1 + \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \qquad (2.19)$$

For $\mathbf{X}$ being distributed as $N(\mu_1, \Sigma)$, U is normally distributed with mean

$$\begin{aligned} \mathrm{E}(U) &= \mu_1' \Sigma^{-1} (\mu_1 - \mu_2) - (1/2) (\mu_1 + \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \\ &= (1/2) (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \\ &= (1/2) \Delta^2 \qquad (2.20) \end{aligned}$$

and variance

$$\begin{aligned} \mathrm{Var}(U) &= \mathrm{E}\,[(\mu_1 - \mu_2)' \Sigma^{-1} (\mathrm{X} - \mu_1) (\mathrm{X} - \mu_1)' \Sigma^{-1} (\mu_1 - \mu_2)] \\ &= (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \\ &= \Delta^2 \qquad (2.21) \end{aligned}$$

where Δ^2 is the *Mahalanobis squared distance* between $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$.

As a conclusion, if $\mathbf{X}$ is distributed according to $N(\mu_1, \Sigma)$ then U is distributed according to $N((1/2)\Delta^2, \Delta^2)$. Similarly, if $\mathbf{X}$ is distributed according to $N(\mu_2, \Sigma)$ then U is distributed according to $N((-1/2)\Delta^2, \Delta^2)$.

4) The probabilities of misclassifications

The probability of misclassifying an observation from π_1 as π_2 is

$$\Pr(2|1) = \int_{\infty}^{c} (1/(2\pi)^{1/2} \Delta) \exp[-(\mathbf{Z} - (-1/2)\Delta^2)^2 / (2\Delta^2)]\, d\mathbf{Z}$$

$$= \int_{\infty}^{(c-(1/2)\,\Delta^2)/\Delta} (1/(2\pi)^{1/2}) \exp[(-1/2)\,Y^2]\, dY \qquad (2.22)$$

The probability of misclassifying an observation from π_2 as π_1 is

$$\Pr(2 \mid 1) = \int_{c}^{\infty} (1/(2\pi)^{1/2}\,\Delta) \exp[-(Z + (-1/2)\,\Delta^2)^2 / (2\,\Delta^2)]\, dZ$$

$$= \int_{(c-(1/2)\,\Delta^2)/\Delta}^{\infty} (1/(2\pi)^{1/2}) \exp[(-1/2)\,Y^2]\, dY \qquad (2.23)$$

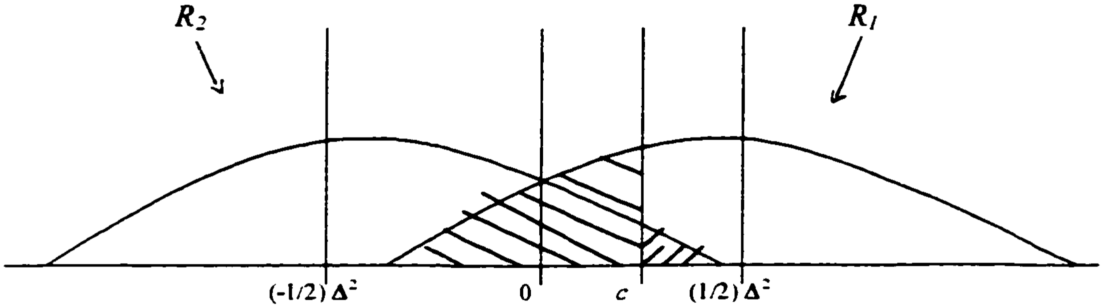

Figure 2.1 Pr (1 | 2) and Pr (2 | 1) are indicated by the shaded portion in the tails.

For the *minimax* solution (a solution where the maximum expected loss is a minimum) we choose c so that

$$c(1 \mid 2) \int_{(c-(1/2)\,\Delta^2)/\Delta}^{\infty} (1/(2\pi)^{1/2}) \exp[(-1/2)\,Y^2]\, dY$$

$$= c(2 \mid 1) \int_{\infty}^{(c-(1/2)\,\Delta^2)/\Delta} (1/(2\pi)^{1/2}) \exp[(-1/2)\,Y^2]\, dY \qquad (2.24)$$

where $c(i \mid j)$ are the costs of misclassification, $i, j = 1,2$.

As a conclusion, the *minimax* regions of classification for the two multivariate normal populations are given by (2.16) where $c = \ln k$ is chosen by the condition (2.24). If the costs of misclassification are equal then $c = 0$ and the probability of misclassification is

$$\int_{\Delta/2}^{\infty} (1/(2\pi)^{1/2}) \exp[(-1/2)\,\mathbf{Y}^2]\, d\mathbf{Y} \tag{2.25}$$

If the costs of misclassification are unequal, c could be determined to sufficient accuracy by a *trial-and-error* method with the normal tables.

5) Classification into one of two multivariate normal populations when the parameters are estimated

In most cases the population quantities $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ are not known. Hence they are inferred from samples, one from each populations, π_1 and π_2.

Suppose we have the data matrices

$$\underset{(m \times n1)}{\mathbf{X}_1} = [\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}] \quad \text{from } N(\boldsymbol{\mu}_1, \Sigma)$$

$$\text{and} \quad \underset{(m \times n2)}{\mathbf{X}_2} = [\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}] \quad \text{from } N(\boldsymbol{\mu}_2, \Sigma) \tag{2.26}$$

which represent the *training* sample.

On the basis of this information we want to classify the observation $\mathbf{X}$ as coming from π_1 or π_2. The sample mean vectors and covariance matrices are

$$\underset{(m \times 1)}{\overline{\mathbf{X}}_1} = (1/n_1) \sum_{j=1}^{n1} \mathbf{X}_{1j} \; ; \qquad \underset{(m \times m)}{\mathbf{S}_1} = (1/(n_1 - 1)) \sum_{j=1}^{n1} (\mathbf{X}_{1j} - \overline{\mathbf{X}}_1)(\mathbf{X}_{1j} - \overline{\mathbf{X}}_1)'$$

$$\underset{(m \times 1)}{\overline{\mathbf{X}}_2} = (1/n_2) \sum_{j=1}^{n2} \mathbf{X}_{2j} \; ; \qquad \underset{(m \times m)}{\mathbf{S}_2} = (1/(n_2 - 1)) \sum_{j=1}^{n2} (\mathbf{X}_{2j} - \overline{\mathbf{X}}_2)(\mathbf{X}_{2j} - \overline{\mathbf{X}}_2)' \tag{2.27}$$

and

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

$$= \left[(n_1 - 1)\, \mathbf{S}_1 + (n_2 - 1)\, \mathbf{S}_2\right] / (n_1 + n_2 - 2) \qquad (2.28)$$

Remark. $\mathbf{S}_{pooled}$ is an unbiased estimate of Σ and it represents a weighted average of $\mathbf{S}_1$ and $\mathbf{S}_2$.

The estimate of L is given by

$$\hat{L} = \mathbf{S}^{-1}{}_{pooled}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) \qquad (2.29)$$

We substitute these estimates for the parameters in (2.14) and we obtain

$$w(\mathbf{X}) = \mathbf{X}'\, \mathbf{S}^{-1}{}_{pooled}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) - (1/2)\, (\overline{\mathbf{X}}_1 + \overline{\mathbf{X}}_2)'\, \mathbf{S}^{-1}{}_{pooled}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)$$

$$= (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)'\, \mathbf{S}^{-1}{}_{pooled}\left[\mathbf{X} - (1/2)(\overline{\mathbf{X}}_1 + \overline{\mathbf{X}}_2)\right] \qquad (2.30)$$

$w(\mathbf{X})$ is often called *Anderson's classification function (statistic)* and it is used as a criterion of classification in the same way that (2.14) is.

Another case is when we have a sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ from either π_1 or π_2, and we wish to classify the sample as a whole. Then, an unbiased estimate of the covariance matrix Σ is defined by

$$\mathbf{S}_{pooled} = \left[1 / (n_1 + n_2 + n - 3)\right] \left[\sum_{j=1}^{n1} (\mathbf{X}_{1j} - \overline{\mathbf{X}}_1)\, (\mathbf{X}_{1j} - \overline{\mathbf{X}}_1)' + \sum_{j=1}^{n2} (\mathbf{X}_{2j} - \overline{\mathbf{X}}_2)\, (\mathbf{X}_{2j} - \overline{\mathbf{X}}_2)'\right.$$

$$\left. + \sum_{j=1}^{n} (\mathbf{X}_j - \overline{\mathbf{X}})\, (\mathbf{X}_j - \overline{\mathbf{X}})'\right] \qquad (2.31)$$

where

$$\overline{\mathbf{X}} = (1/n) \sum_{j=1}^{n} \mathbf{X}_j \qquad (2.32)$$

Then the criterion of classification is

$$[\mathbf{X} - (1/2)(\overline{\mathbf{X}}_1 + \overline{\mathbf{X}}_2)]' \, \mathbf{S}^{-1}_{pooled}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) \qquad (2.33)$$

Note. The larger n is, the smaller are the probabilities of misclassification.

Similarly to (2.5) and (2.6), the linear combination

$$\mathbf{Y} = \hat{L}' \, \mathbf{X} = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)' \, \mathbf{S}^{-1}_{pooled}\mathbf{X}$$

which is the ***Fisher's sample linear discriminant function***, maximizes the ratio

$$[(\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2)^2 / \mathbf{S}^2{}_{\mathbf{Y}}] = [(L' \, \overline{\mathbf{X}}_1 - L' \, \overline{\mathbf{X}}_2)^2 / (\hat{L}' \, \mathbf{S}_{pooled} \, \hat{L})] = [(\hat{L}'d) / (\hat{L}' \, \mathbf{S}_{pooled} \, \hat{L})] \qquad (2.34)$$

where $d = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)$.

Note. We must have $(n_1 + n_2 - 2) > m$, otherwise $\mathbf{S}_{pooled}$ is singular and the usual inverse, $\mathbf{S}^{-1}_{pooled}$ does not exist.

The maximum value of the sample ratio (2.34) is given by

$$\max_L \, [(\hat{L}'d)^2 / (\hat{L}' \, \mathbf{S}_{pooled} \, \hat{L})] = d' \, \mathbf{S}^{-1}_{pooled} \, d = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)' \, \mathbf{S}^{-1}_{pooled}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) \qquad (2.35)$$

which is the sample squared distance.

The midpoint, m, between the two univariate sample means, $\mathbf{Y}_1 = \hat{L}' \, \overline{\mathbf{X}}_1$ and $\mathbf{Y}_2 = \hat{L}' \, \overline{\mathbf{X}}_2$ is given by:

$$m = (1/2)\,(\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2) = (1/2)\,(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)' \, \mathbf{S}^{-1}_{pooled}(\overline{\mathbf{X}}_1 + \overline{\mathbf{X}}_2) \qquad (2.36)$$

Hence the regions of classification that minimizes the expected cost of misclassification, are given by:

$$R_1\text{:} \quad (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)' \, \mathbf{S}^{-1}_{pooled}\mathbf{X} - (1/2)\,(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)' \, \mathbf{S}^{-1}_{pooled}(\overline{\mathbf{X}}_1 + \overline{\mathbf{X}}_2) \geq \ln k \qquad (2.37)$$

R_2: $(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)' \mathbf{S}^{-1}_{pooled} \mathbf{X} - (1/2) (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)' \mathbf{S}^{-1}_{pooled} (\overline{\mathbf{X}}_1 + \overline{\mathbf{X}}_2) < \ln k$

where $k = [c(1 \mid 2) / c(2 \mid 1)] \; [p_2 / p_1]$.

6) **The distribution of the criterion**

Let $w(\mathbf{X}) = \mathbf{X}' \mathbf{S}^{-1}_{pooled} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) - (1/2) (\overline{\mathbf{X}}_1 + \overline{\mathbf{X}}_2)' \mathbf{S}^{-1}_{pooled} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)$. The distribution of $w(\mathbf{X})$ is said to be extremely complicated. It depends on the sample sizes and the unknown Δ^2.

Anderson (1984) gives the following result: if $n_1 = n_2$, the distribution of w for $\mathbf{X}$ from π_1 is the same as that of $-w$ of $\mathbf{X}$ from π_2. Thus, if $w \geq 0$ is the region of classification as π_1, then the probability of misclassifying $\mathbf{X}$ when it is from π_1 is equal to the probability of that when it is from π_2.

- **The asymptotic distribution of the criterion**

Wald (1944) was the first one to conclude that the limiting distribution of w as $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$ is the same as the distribution of U given in equation (2.19). For sufficiently large samples from π_1 and π_2, we can use the criterion as if we knew the population exactly and we make only a small error. This result is presented in the following theorem.

Theorem1

Let w be given by (2.30) with $\overline{\mathbf{X}}_1$ the mean of a sample of size n_1 from $N(\mu_1, \Sigma)$, $\overline{\mathbf{X}}_2$ the mean of a sample of size n_2 from $N(\mu_2, \Sigma)$, and $\mathbf{S}$ the estimate of Σ based on the pooled sample. The limiting distribution of w as $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$ is $N((1/2)\,\Delta^2, \Delta^2)$ if $\mathbf{X}$ is distributed according to $N(\mu_1, \Sigma)$ and is $N((-1/2)\,\Delta^2, \Delta^2)$ if **X** is distributed according to $N(\mu_2, \Sigma)$.

7) ***Fisher's method* for discriminating among several populations**

In this section we outline a several population extension of *Fisher's discriminant method.* The purpose of this is to obtain a reasonable representation of the population that involves only a few linear combinations of the observations, such as $\mathrm{L}'_1\mathbf{X}$, $\mathrm{L}'_2\mathbf{X}$ and $\mathrm{L}'_3\mathbf{X}$. The primary purpose of this method is to separate populations. It can also be used to classify observations.

In this case, we have g populations, which are not necessary multivariate normal. We assume that the population covariance matrices are equal and of full rank, i.e.

$\Sigma_1 = \Sigma_2 = \ldots = \Sigma_g = \Sigma$ and $\mathrm{Rank}(\Sigma) = m$.

Remark. If Σ is not of full rank then we let $\mathbf{P} = [\mathbf{e}_1, \ldots, \mathbf{e}_q]$ be the eigenvectors of Σ corresponding to nonzero eigenvalues $[\lambda_1, \ldots, \lambda_q]$. Then we replace **X** by $\mathbf{P}'\,\mathbf{X}$. which has a full rank covariance matrix $\mathbf{P}'\Sigma\mathbf{P}$.

Let $\bar{\mu} = (1/g)\,\sum_{i=1}^{g} \mu_i$ be the mean vector of the combined groups.

$$\mathbf{B}_0 = \sum_{i=1}^{g} (\mu_i - \bar{\mu})\,(\mu_i - \bar{\mu})' \tag{2.38}$$

be the between groups sum of crossproducts.

We consider the same linear combination as in (2.1) with expected value

$$\mathrm{E}\,(\mathbf{Y}) = \mathrm{L}'\,\mathrm{E}\,(\mathbf{X} \mid \pi_i) = \mathrm{L}'\,\mu_i = \mu_{iY} \quad \text{for population } i, \tag{2.39}$$

and variance

$$\mathrm{Var}\,(\mathbf{Y}) = L'\,\mathrm{Cov}(\mathbf{X})\,L = L'\,\Sigma\,L \quad \text{for all populations.} \qquad (2.40)$$

The overall mean is defined by

$$\bar{\mu}_Y = (1/g)\,\sum_{i=1}^{g} \mu_{iY} = (1/g)\,\sum_{i=1}^{g} L'\mu_i = L'\left((1/g)\,\sum_{i=1}^{g} \mu_i\right) = L'\,\bar{\mu} \qquad (2.41)$$

Fisher's idea was to find the linear combinations that maximized the sum of squared distances from populations to the overall mean of $\mathbf{Y}$ relative to the variance of $\mathbf{Y}$, i.e. to minimize

$$\sum_{i=1}^{g} (\mu_{iY} - \bar{\mu}_Y)^2 / (L'\,\Sigma\,L) = \sum_{i=1}^{g} (L'\mu_i - L'\bar{\mu})^2 \,/\, (L'\,\Sigma\,L)$$

$$= L'\left(\sum_{i=1}^{g} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'\right) L \,/\, (L'\,\Sigma\,L)$$

$$= (L'\,\mathbf{B}_0\,L) \,/\, (L'\,\Sigma\,L) \qquad (2.42)$$

Fisher showed that we could select L such that the ratio in (2.42) is maximized in the following result. It is convenient to scale L so that $L'\,\Sigma\,L = 1$, without loss of generality.

Lemma1 Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s > 0$ denote the $s \leq \min\,(g\text{-}1, m)$ nonzero eigenvalues of $\Sigma^{-1}\mathbf{B}_0$ and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_s$ the corresponding eigenvectors (scaled so that $\mathbf{e}'\,\Sigma\,\mathbf{e} = 1$). Then the vector of coefficients L that maximizes the ratio $(L'\,\mathbf{B}_0\,L) \,/\, (L'\,\Sigma\,L)$ is given by $L_1 = \mathbf{e}_1$. The linear combination $L'_1\mathbf{X}$ is called the *first discriminant.* The value $L_2 = \mathbf{e}_2$ maximizes the ratio subject to $\mathrm{Cov}(L'_1\mathbf{X}, L'_2\,\mathbf{X}) = 0$. The linear combination $L'_2\,\mathbf{X}$ is called the *second discriminant.* Continuing, $L_k = \mathbf{e}_k$ maximizes the ratio subject to $\mathrm{Cov}(L'_k\mathbf{X}, L'_i\mathbf{X}) = 0$, $i < k$, and, $L'_k\mathbf{X}$ is called the *kth discriminant.* Also $\mathrm{Var}\,(L'_i\mathbf{X}) = 1$,

$i = 1,\ldots,s$.

In most applications, Σ and $\boldsymbol{\mu}_i$ are not known. Hence, their estimates are obtained through the *training sample* of size n_i from populations π_i, $i = 1,2,\ldots,g$. The data set from populations π_i is denoted by the mx n_i matrix, $\mathbf{X}_i$.

Let $\overline{\mathbf{X}}_i = (1/n_i) \sum_{i=1}^{n_i} \mathbf{X}_i$ be the sample mean vector of population π_i. (2.43)

$S_i = (1/n_i -1) \sum_{i=1}^{n_i} (\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)'$ the covariance matrix of population π_i. (2.44)

$\mathbf{X} = (\sum_{i=1}^{g} n_i \overline{\mathbf{X}}_i) / (\sum_{i=1}^{g} n_i)$

$= (\sum_{i=1}^{g} \sum_{i=1}^{n_i} \mathbf{X}_{ij}) / (\sum_{i=1}^{g} \mathrm{n}_i)$ the overall sample average vector. (2.45)

The *sample between groups* matrix is defined by

$$\mathbf{b}_0 = \sum_{i=1}^{g} (\overline{\mathbf{X}}_i - \overline{\mathbf{X}})(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})' \qquad (2.46)$$

An estimate of Σ is given by

$\mathbf{S}_{pooled} = (1/(n_1 + n_2 + \ldots + n_g - g)) \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)'$

$= (1/(n_1 + n_2 + \ldots + n_g - g))\, w_0$

w_0 is the *sample within groups* matrix.

Since $w_0 = (1/(n_1 + n_2 + \ldots + n_g - g))\, \mathbf{S}_{pooled}$, then the same $\hat{L}$ that maximizes $(\hat{L}'\, \mathbf{b}_0\, \hat{L}) / (\hat{L}'\, \mathbf{S}_{pooled}\, \hat{L})$ also maximizes $(\hat{L}'\, \mathbf{b}_0\, \hat{L}) / (\hat{L}'\, w_0\, \hat{L})$. As a result, the optimizing $\hat{L}$ is given by eigenvectors $\mathbf{e}_i$ of $w_0^{-1}\mathbf{b}_0$, because if $w_0^{-1}\mathbf{b}_0\, \mathbf{e} = \bar{\lambda}\, \mathbf{e}$ then

$\mathbf{S}_{pooled}\,\mathbf{b}_0\,\mathbf{e} = \bar{\lambda}\,(n_1 + n_2 + \ldots + n_g - g)\,\mathbf{e}$. The *fisher's sample discriminants* are outlined in the following result.

Lemma 2 Let $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \ldots \geq \bar{\lambda}_s > 0$ denote the s $\leq$ min $(g\text{-}1, m)$ nonzero eigenvalues of $w_0^{-1}\mathbf{b}_0$ and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_s$ be the corresponding eigenvectors (scaled so that $\mathbf{e}'\,\mathbf{S}_{pooled}\,\mathbf{e} = 1$). Then the vector of coefficients $\hat{L}$ that maximizes the ratio $(\hat{L}'\,\mathbf{b}_0\,\hat{L})\,/\,(\hat{L}'\,w_0\,\hat{L})$

$$= [\hat{L}'\,(\textstyle\sum_{i=1}^{g}(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})\,(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})')\,\hat{L}]\,/\,[\hat{L}'\,(\textstyle\sum_{i=1}^{g}\sum_{j=1}^{n}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)\,(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)')\,\hat{L}] \quad (2.48)$$

is given by $\hat{L}_1 = \mathbf{e}_1$. The linear combination $\hat{L}'_1\mathbf{X}$ is called the *sample first discriminant.* The choice $\hat{L}_2 = \mathbf{e}_2$ produces the *sample second discriminant,* $\hat{L}'_2\mathbf{X}$. Continuing, $\hat{L}'_k\mathbf{X} = \mathbf{e}_k\mathbf{X}$ is the *sample kth discriminant,* $k \leq s$. Unlike the population result, the discriminants will not have zero covariance for each random sample $\mathbf{X}_i$. Rather, the condition

$$\hat{L}_i'\,\mathbf{S}_{pooled}\,\hat{L}_k = 1 \quad \text{if } i = k \leq s$$
$$= 0 \quad \text{otherwise}$$

will be satisfied.

8) Classification by *Fisher's discriminants*

Fisher's discriminant also provides the basis for a classification rule.

Let $\mathbf{Y}' = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_s]$ where $\mathbf{Y}_k = \hat{L}'_k\,\mathbf{X} = kth$ discriminant (2.49)

where $k \leq s$ and $s = \min(g\text{-}1, m)$.

$\mathbf{Y}$ has mean vector $\boldsymbol{\mu}_{iY}' = [\boldsymbol{\mu}_{iY_1}, \boldsymbol{\mu}_{iY_2}, \ldots, \boldsymbol{\mu}_{iY_s}] = [\hat{L}'_1\boldsymbol{\mu}_i, \hat{L}'_2\boldsymbol{\mu}_i, \ldots, \hat{L}'_s\boldsymbol{\mu}_i]$ under population π_i and covariance matrix $\mathbf{I}$ (Identity matrix), for all populations (see Lemma1).

Since the components of $\mathbf{Y}$ have unit variances and zero covariances, the appropriate measure of squared distance from $\mathbf{Y} = \mathbf{Y}$ to $\mu_{i\mathbf{Y}}$ is

$$(\mathbf{Y} - \mu_{i\mathbf{Y}})' \, (\mathbf{Y} - \mu_{i\mathbf{Y}}) = \sum_{j=1}^{s} (y_j - \mu_{i\mathbf{Y}_j})^2$$

A reasonable classification rule is one that assigns $\mathbf{Y}$ to population π_k if the squared distance from $\mathbf{Y}$ to $\mu_{k\mathbf{Y}}$ is smaller than the squared distance from $\mathbf{Y}$ to $\mu_{i\mathbf{Y}}$ for $i \neq k$.

If only r of the discriminants are used for allocation, the rule is:

Allocate $\mathbf{X}$ to π_k if

$$\sum_{j=1}^{r} (y_j - \mu_{k\mathbf{Y}_j})^2 = \sum_{j=1}^{r} [\hat{L}'_j(\mathbf{X} - \mu_k)]^2$$

$$\leq \sum_{j=1}^{r} [\hat{L}'_j(\mathbf{X} - \mu_i)]^2 \qquad \text{for all } i \neq k \tag{2.50}$$

Remark. The restriction on the number of discriminants is explained by the number of nonzero eigenvalues of $\Sigma^{-1}\mathbf{B}_0$ or $\Sigma^{-1/2}\, \mathbf{B}_0\, \Sigma^{-1/2}$ (see Lemma1).

We know that $\Sigma^{-1}\mathbf{B}_0$ is $m \times m$, hence $s \leq m$. Furthermore, the g vectors

$$\mu_1 - \bar{\mu},\ \mu_2 - \bar{\mu},\ \ldots,\ \mu_g - \bar{\mu} \tag{2.51}$$

satisfy $(\mu_1 - \bar{\mu}) + (\mu_2 - \bar{\mu}) + \ldots + (\mu_g - \bar{\mu}) = \sum_{i=1}^{g} \mu_i - g\,\bar{\mu} = g\,\bar{\mu} - g\,\bar{\mu} = 0$.

That is any of the differences $\mu_i - \bar{\mu}$, $i = 1, \ldots, g$, can be written as a linear combination of the other $(g - 1)$ differences. Linear combinations of the g vectors in (2.51) determines a hyperplane of dimension $q \leq g - 1$. Taking any vector $\mathbf{e}$ perpendicular to every $\mu_i - \bar{\mu}$, and hence the hyperplane, gives

$$\mathbf{B}_0\,\mathbf{e} = \sum_{i=1}^{g} (\mu_i - \bar{\mu})\, (\mu_i - \bar{\mu})'\, \mathbf{e} = \sum_{i=1}^{g} (\mu_i - \bar{\mu})\, 0 = \mathbf{0}$$

So $\Sigma^{-1}\mathbf{B}_0\,\mathbf{e} = 0\mathbf{e}$.

There are $(m - q)$ orthogonal eigenvectors corresponding to the zero eigenvalue. This implies that there are q or fewer ***nonzero*** eigenvalues. Since it is always true that $q \leq g - 1$, the number of nonzero eigenvalues s must satisfy $s \leq \min(m, g - 1)$. Thus there is no loss of discriminant information by plotting in two dimensions if the following conditions hold.

Number of variables	Number of populations	Maximum number of discriminants
any m	$g = 2$	1
any m	$g = 3$	2
$m = 2$	any g	2

Given the classification rule in (2.50) and the ***normal theory*** discriminant scores.

$$\mathrm{D}_i(\mathbf{X}) = \mu'_i\, \Sigma^{-1}\mathbf{X} - (1/2)\, \mu'_i\, \Sigma^{-1}\mu_i + \ln p_i, \tag{2.52}$$

or, equivalently,

$$\mathrm{D}_i(\mathbf{X}) - (1/2)\, \mathbf{X}'\, \Sigma^{-1}\mathbf{X} = -(1/2)\, (\mathbf{X} - \mu_i)'\, \Sigma^{-1}\, (\mathbf{X} - \mu_i) + \ln p_i,$$

Obtained by adding the same constant $-(1/2)\, \mathbf{X}'\, \Sigma^{-1}\mathbf{X}$ to each $\mathrm{D}_i(\mathbf{X})$, we present the following important lemma.

Lemma 3 Let $y_j = \mathrm{L}'_j\mathbf{X}$ where $\mathrm{L}_j = \Sigma^{-1}\mathbf{e}_j$ and $\mathbf{e}_j$ is an eigenvector of $\Sigma^{-1/2}\, \mathbf{B}_0\, \Sigma^{-1/2}$. Then

$$\sum_{j=1}^{m} (y_j - \mu_{iY_j})^2 = \sum_{j=1}^{m} [L'_j(\mathbf{X} - \mu_i)]^2 = (\mathbf{X} - \mu_i)'\, \Sigma^{-1}\, (\mathbf{X} - \mu_i)$$

$$= -\, \mathrm{D}_i(\mathbf{X}) + (1/2)\, \mathbf{X}'\, \Sigma^{-1}\, \mathbf{X} + \ln p_i$$

If $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s > 0 = \lambda \mathrm{s}_{+1} = \lambda\ \mathrm{s}_{+2} = \ldots = \lambda_m,$

$\sum_{j=s+1}^{m} (y_j - \mu_{iY_j})^2$ is constant for all populations $i = 1, 2, \ldots, g$ so only the first s y_j, or $\sum_{j=1}^{s} (y_j - \mu_{iY_j})^2$, $i = 1, 2, \ldots, g$ contribute to the classification.

Also, if the prior probabilities are such that $p_1 = p_2 = \ldots = p_g = 1/g$, the rule in (2.50) with $r = s$ is equivalent to the minimum TPM rule (1.29).

Fisher's classification procedure based on sample discriminants is:

Allocate **X** to π_k if

$$\sum_{j=1}^{r} (\mathbf{Y}_j - \bar{\mathbf{Y}}_{kj})^2 = \sum_{j=1}^{r} [\hat{L}'_j(\mathbf{X} - \bar{\mathbf{X}}_k)]^2$$

$$\leq \sum_{j=1}^{r} [\hat{L}'_j(\mathbf{X} - \bar{\mathbf{X}}_i)]^2 \qquad \text{for all } i \neq k \qquad (2.53)$$

where $\hat{L}_j$ is defined in (2.48) (see Lemma 2) and $r \leq s$.

When the prior probabilities are such that $p_1 = p_2 = \ldots = p_g = 1/g$ and $r = s$, the rule (2.53) is equivalent to the rule based on the largest linear discriminant score of (1.32). In addition, if $r < s$ discriminants are used for classification, there is a loss of squared distance, or score, of

$$\sum_{j=r+1}^{m} [\hat{L}'_j(\mathbf{X} - \mu_i)]^2 \quad \text{for each population } \pi_i$$

where $\sum_{j=r+1}^{s} [L'_j(\mathbf{X} - \mu_i)]^2$ is the part useful for classification.

9) Diagnostics in linear discriminant analysis

In this paper, Fung (1995) proposed some new diagnostic measures in linear discriminant analysis. For simplicity, a common prior probability and misclassification cost function is taken for both populations π_1 and π_2 .The *Fisher's linear discriminant* rule (2.9) is:

Allocate an observation **X** to π_1 if

$$(\mu_1 - \mu_2)' \Sigma^{-1}\mathbf{X} - (1/2)(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2) \geq 0$$

and to π_2 if otherwise.

We have $M = (1/2)(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2)$ and $L' = (\mu_1 - \mu_2)' \Sigma^{-1}$. Hence we have

$L'\mathbf{X} - L'(\mu_1 + \mu_2)(1/2) \geq 0.$

The discriminant coefficients L can be estimated in two ways:

a) by the usual sample estimates $\hat{L}' = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)' \mathbf{S}_{pooled}^{-1}$.

b) By using the regression model $\mathbf{Z} = \mathbf{Y}\gamma + \varepsilon$, where $\mathbf{Z}$ and ε are n x 1,

$\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n]^T$ is n x $(m+1)$ and $\hat{L}$ is a $(m+1)$ vector.

Let γ be the least squares estimator for γ. The residual is $r_i = \mathbf{Z}_i + \mathbf{Y}_i^T\bar{\gamma}$, and the leverage is $h_i = \mathbf{Y}_i^T(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}_i$. Many diagnostic measures in regression can be expressed in terms of them. One example is the statistic of Cook (1977)

$$C_i = [(\bar{\gamma} - \bar{\gamma}_{(i)})^T \mathbf{Y}^T \mathbf{Y} (\bar{\gamma} - \bar{\gamma}_{(i)})] / [(p + 1)\bar{\sigma}^2] \qquad (2.54)$$

where $\bar{\sigma}^2$ is the unbiased error variance estimate and $\bar{\gamma}_{(i)}$ is the least squares estimate for α using the sample without observation i in the regression model $\mathbf{Z} = \mathbf{Y}\gamma + \varepsilon$. (2.54a)

Under the linear discriminant analysis framework where the first column of **Y** contains unities and the remaining columns contain n observations $\mathbf{X}_{ij}, j = 1, \ldots, n_i, i = 1,2$, and **Z** has the first n_1 elements as an arbitrary constant b_1 and the other elements as b_2. Let γ^T be partitioned as (γ_1, γ_2^{T}) then the least squares estimates $\bar{\gamma}_2$ is known to be proportional to L (Anderson, 1984, sec. 6.5; Cox and Snell, 1989, sec. 4.4; Mclachlan, 1992, sec. 3.3.4).

Replacing the parameters by the sample estimates, we get

$$\hat{L}' \mathbf{X} - \hat{L}' (\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2)(1/2) \geq 0$$

and the allocation rule is identical to (2.15) where $\hat{L}' = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' \, \mathbf{S}_{pooled}^{-1}$. The quantity $\hat{L}' \mathbf{X} - \hat{L}' (\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2)(1/2)$ is the discriminant score, which is also the estimated *log-odds*, Log $[\Pr(\mathbf{X}_0 \in \pi_1) \,/\, \Pr(\mathbf{X}_0 \in \pi_2)]$, for observation **X**. Fung denotes it as $\beta^{T}\mathbf{Y}$, where

$$\bar{\beta}^{T} = (-\hat{L}' (\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2)(1/2), \hat{L}') \quad \text{and} \quad \mathbf{Y}' = (1, \mathbf{X}') \tag{2.55}$$

Fung is interested in the effect of the omission of observation *i* (for simplicity, we assume *i* is from π_1) on the parameter estimate $\bar{\beta}_{(i)}$. He studied this through the mean squared difference of the discriminant scores for the full sample and the sample without observation *i*, i.e.

$$\mathrm{E}\,(\bar{\beta}^{T}\mathbf{Y} - \bar{\beta}^{T}_{(i)}\mathbf{Y})^2 \tag{2.56}$$

The expectation is taken with respect to the estimated density of **X**, which is evaluated in two ways: *parametrically* and *nonparametrically*.

<u>Parametrically</u> **X** is distributed as $t\,N(\mu_1, \Sigma) + (1-t)\,N(\mu_2, \Sigma)$. Let $t = n_1 / n_2$ and plug-in estimates $\bar{\mathbf{X}}_1$, $\bar{\mathbf{X}}_2$, and **S** in (2.56). After some calculations, the expectation is given as

$$\mathrm{E2} = t\,\mathrm{B}_1^{2} + (1-t)\,\mathrm{B}_2^{2} + \mathrm{V} \tag{2.57}$$

where

$$\begin{aligned} B_1 &= (\hat{L} - \hat{L}_{(i)})^T (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)(1/2) - \hat{L}_{(i)}^T (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_{1(i)})(1/2) \\ B_2 &= -(\hat{L} - \hat{L}_{(i)})^T (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)(1/2) - \hat{L}_{(i)}^T (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_{1(i)})(1/2) \end{aligned} \qquad (2.58)$$

and

$$V = (\hat{L} - \hat{L}_{(i)})^T \mathbf{S} (\hat{L} - \hat{L}_{(i)})$$

are the bias and the variance.

Nonparametrically The empirical (*nonparametric*) distribution function for **X** is used to evaluate (2.56) as

$$F2 = \Sigma_j \left[(\bar{\beta} - \bar{\beta}_{(i)})^T \mathbf{Y}_j\right]^2 / n$$

or, equivalently

$$F2 = \left[(\bar{\beta} - \bar{\beta}_{(i)})^T \mathbf{Y}^T \mathbf{Y} (\bar{\beta} - \bar{\beta}_{(i)})\right] / n \qquad (2.59)$$

Fung makes the remark that F2 is in analogy to the well-known *Cook statistic* in (2.54).

After some calculations, F2 is also expressed as

$$F2 = t\, B_1^2 + (1 - t)\, B_2^2 + (n - 2)\, V / n \qquad (2.60)$$

Note. Since F2 and E2 are very close, especially for a large size *n*, the later discussion is mainly on F2.

F2 and E2 can be expressed in terms of the two fundamental statistics in discriminant analysis

$$d_i^2 = (\mathbf{X}_{1i} - \overline{\mathbf{X}}_1)^T \mathbf{S}^{-1} (\mathbf{X}_{1i} - \overline{\mathbf{X}}_1)$$

and

$$\overline{\Psi}_i = \bar{\alpha}^T (\mathbf{X}_{1i} - \overline{\mathbf{X}}_1) \qquad (2.61)$$

which are like the residual and leverage measure in regression, on which many influence measures depend. The following theorem is useful for getting the asymptotic distribution for the proposed measures.

Theorem 2 The statistics DIF = $d_i^2 - (\overline{\Psi}_i / D)^2$ and $\overline{\Psi}_i / D$, where

$$D = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)' S^{-1} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2),$$

are asymptotically independent and are distributed as χ^2_{p-1} and $N(0, 1)$.

By the use of this theorem, it could be shown that d_i^2 and $\overline{\Psi}_i / D$ are asymptotically χ^2_p and $N(0, 1)$ distributed respectively. Hence critical values and expected quantiles of the measures can be approximated using numerical integration.

(2.56) could also be evaluated *non-parametrically* based on the empirical distribution function estimated from the sample without observation *i*, giving a measure

$$F2I = [(\bar{\beta} - \bar{\beta}_{(i)})^T \mathbf{Y}_{(i)}{}^T \mathbf{Y}_{(i)} (\bar{\beta} - \bar{\beta}_{(i)})] / (n - 1) \qquad (2.62)$$

Which is analogous to the statistic of Welsch (1982) in regression diagnostics.

Similarly, if we evaluate (2.56) *parametrically*, treating the *leave-one-out* estimates $\overline{\mathbf{Y}}_{1(i)}$, $\overline{\mathbf{Y}}_2$, and $S_{(i)}$ as parameters, then we obtain the measure E2*I* with a form similar to that of E2 in (2.57).

F2I and E2I could also be expressed in terms of the basic statistics (2.61). These four measures are aimed at detecting influential observations that have an unusually high influence on the estimated *log-odds* or the discriminant score. They are asymptotically equivalent, under the null case, to having no influential observations. F2 and E2 could give different results from F2*I* and E2*I* as the *Cook* and *Welsch statistics* do in regression diagnostics.

In the study of the possibility of generalizing regression diagnostics to linear discriminant analysis, Fung discusses one basic distinction between regression and discriminant analysis. In both *linear* and *logistic regression*, **Z** is assumed to be random and **Y** is non-random, but linear discriminant analysis instead models **Y** (random) given **Z** (random). Therefore, regression diagnostic measures, such as the covariance ratio, being constructed under the foregoing randomness assumptions for regression, are inappropriate in the discriminant analysis situation.

Setting the **Z** and **Y** in model (2.54a), both the regression residual r_i and the leverage statistic h_i can be expressed in terms of d_i^2 and $\overline{\Psi}_i$ in discriminant analysis. But r_i is arbitrarily determined by the constants b_1 and b_2 in **Z**, whereas h_i is equivalent to a multivariate outlier test for a single population (Rousseeuw and Van Zomeren, 1990), without taking into account the special structure of discriminant analysis. Thus it would be hard to have a simple interpretation for r_i and h_i in the context of discriminant analysis.

Fung shows that although F2 is in analogy to the *Cook statistic* C_i, in regression, they are not in proportion over all possible indices i, $i = 1,\ldots,$n. The *Cook-like statistics* C_i' in logistic discriminant analysis, apart from the weights, have the same interpretation as F2. The vector $\mathbf{Y}(\gamma' - \gamma_{(i)}')$ in C_i' contains the differences of the *log-odds*, having the same meaning as $\mathbf{Y}\,(\overline{\beta} - \overline{\beta}_{(i)})$ in F2. But the meaning of $\mathbf{Y}(\overline{\gamma} - \overline{\gamma}_{(i)})$ in C_i, under the linear discriminant analysis, is ambiguous and different. The interpretations of C_i and other regression diagnostics are not simple when applied to linear discriminant analysis.

Fung comes to the conclusion that the discriminant coefficients can be determined using a regression model, whereas the well-known regression diagnostic measures cannot be used under a discriminant analysis framework. The proposed measures are useful for detecting single influential observations. By sequential application, they could be useful to identify multiple influential observations. They could be extended to detect multiple influential observations in blocks avoiding the masking problem. However, the computation requirements are increased. The approximation suggested by Critchley and Vitiello (1991) could be applied to reduce the load of computation. Moreover, the methods suggested by Rousseuw and Van Zomeren (1990) and Fung (1993) may be extended for detecting multiple outliers in discriminant analysis.

10) Examples

We have generated multivariate normal data for two populations, π_1 and π_2, with m=3, different means, μ_1 and μ_2 , and equal covariance matrices, Σ.

We consider nine different cases:

a) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,1)$, $\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

b) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,1)$, $\Sigma = \begin{bmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}$

c) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,1)$, $\Sigma = \begin{bmatrix} 1 & .9 & .9 \\ .9 & 1 & .9 \\ .9 & .9 & 1 \end{bmatrix}$

d) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,2)$, $\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

e) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,2)$, $\Sigma = \begin{bmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}$

f) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,2)$, $\Sigma = \begin{bmatrix} 1 & .9 & .9 \\ .9 & 1 & .9 \\ .9 & .9 & 1 \end{bmatrix}$

g) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,5)$, $\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

h) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,5)$, $\Sigma = \begin{bmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}$

i) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,5)$, $\Sigma = \begin{bmatrix} 1 & .9 & .9 \\ .9 & 1 & .9 \\ .9 & .9 & 1 \end{bmatrix}$

For each case, the training samples are of sizes 15, 30, 100. For each training sample, there are 50 validation samples of 1000 observations. The Mahalanobis distance, $\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$, and the "optimum error rate", i.e., the "minimum total probability of misclassification", $\Phi(-\Delta/2)$ (where $\Phi(.)$ is the cumulative distribution function of a standard normal random variable) are tabulated for each case (Tables I, II, and III). The lowest total probability of misclassifications (TPM) are obtained for cases f, g, h, and i. These cases have in common that $\Phi(-\Delta/2) < 3$. The size of the training samples does not have an important impact on the total probabilities of misclassification. Hence, we can conclude that the probabilities of misclassification between two multivariate normal populations with equal covariance matrices and unequal mean vectors are influenced by their Mahalanobis distance. As Δ^2 increases, $\Delta/2$ increases and $\Phi(-\Delta/2)$, i.e., the *optimum error rate*, decreases.

Table I. Probabilities of misclassification of two multivariate normal populations π_1 and π_2, (m=3) with equal covariance matrices Σ and mean vectors μ_1'=(0,0,0) and different values of μ_2.

$\Sigma=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$	μ_2'=(0,1,1)			μ_2'=(0,1,2)			μ_2'=(0,1,5)		
Training sample sizes	15	30	100	15	30	100	15	30	100
Δ^2	2			5			26		
$\Delta/2$	0.71			1.12			2.55		
$\Phi(-\Delta/2)$ (%)	23.89			13.14			0.54		
Total number of misclassified observations	13457	12679	12205	7706	7125	6734	437	306	285
TPM (%)	26.91	25.36	24.41	15.41	14.25	13.47	0.87	0.61	0.57

Table II. Probabilities of misclassification of two multivariate normal populations π_1 and π_2, (m=3) with equal covariance matrices Σ and mean vectors $\mu_1'=(0,0,0)$ and different values of μ_2.

$\Sigma=\begin{bmatrix}1 & .5 & .5\\ .5 & 1 & .5\\ .5 & .5 & 1\end{bmatrix}$	$\mu_2'=(0,1,1)$			$\mu_2'=(0,1,2)$			$\mu_2'=(0,1,5)$		
Training sample sizes	15	30	100	15	30	100	15	30	100
Δ^2	2			5.5			34		
$\Delta/2$	0.71			1.17			2.92		
$\Phi(-\Delta/2)$ (%)	23.89			12.1			0.18		
Total number of misclassified observations	12989	12658	12107	6839	6508	6067	149	132	95
TPM (%)	25.98	25.32	24.21	13.68	13.02	12.13	0.30	0.26	0.19

Table III. Probabilities of misclassification of two multivariate normal populations π_1 and π_2, (m=3) with equal covariance matrices Σ and mean vectors μ_1'=(0,0,0) and different values of μ_2.

$\Sigma=\begin{bmatrix} 1 & .9 & .9 \\ .9 & 1 & .9 \\ .9 & .9 & 1 \end{bmatrix}$	**μ_2'=(0,1,1)**			**μ_2'=(0,1,2)**			**μ_2' =(0,1,5)**		
Training sample sizes	15	30	100	15	30	100	15	30	100
Δ^2	7.14			21.07			144.29		
$\Delta/2$	1.34			2.30			6.01		
$\Phi(-\Delta/2)$ (%)	9.01			1.07			0		
Total number of misclassified observations	5620	5097	4781	848	661	594	0	0	0
TPM (%)	11.24	10.19	9.56	1.70	1.32	1.19	0	0	0

Chapter III Quadratic discriminant analysis

So far we have outlined procedures of classification and discriminant analysis for two or more multivariate normal populations with equal covariance matrices. In this chapter, we analyse the special case of multivariate normal populations where covariance matrices Σ_i are not equal.

1) Case of two multivariate normal populations

The *ith* density is

$$f_i(\mathbf{X}) = \left[1 / (2\pi)^{m/2} |\Sigma_i|^{1/2}\right] \exp\left[(-1/2)\,(\mathbf{X}-\mu_i)'\,\Sigma_i^{-1}\,(\mathbf{X}-\mu_i)\right],\quad i = 1,2. \tag{3.1}$$

The ratio of the densities is

$$\left(f_1(\mathbf{X}) / f_2(\mathbf{X})\right) = \tag{3.2}$$

$$\left(|\Sigma_2|^{1/2} \exp\left[(-1/2)\,(\mathbf{X}-\mu_1)'\,\Sigma_1^{-1}\,(\mathbf{X}-\mu_1)\right] / |\Sigma_1|^{1/2} \exp\left[(-1/2)\,(\mathbf{X}-\mu_2)'\,\Sigma_2^{-1}\,(\mathbf{X}-\mu_2)\right]\right)$$

The ratio in (3.2) is known as the *likelihood ratio*.

The natural logarithm of (3.2) is

$$\ln\left(f_1(\mathbf{X}) / f_2(\mathbf{X})\right) =$$

$$(-1/2)\ln\left(|\Sigma_1| / |\Sigma_2|\right) - (1/2)\left(\mu_1'\,\Sigma_1^{-1}\,\mu_1 - \mu_2'\,\Sigma_2^{-1}\,\mu_2\right) - (1/2)\,\mathbf{X}'\left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)\mathbf{X}$$

$$+ \left(\mu_1'\,\Sigma_1^{-1} - \mu_2'\,\Sigma_2^{-1}\right)\mathbf{X} \tag{3.3}$$

The result in (3.3) is a *quadratic* function of **X**. Substituting (3.3) in (1.8) gives the following classification regions (3.4)

R_1: $(-1/2)\,\mathbf{X}'\left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)\mathbf{X} + \left(\mu_1'\,\Sigma_1^{-1} - \mu_2'\,\Sigma_2^{-1}\right)\mathbf{X} - k \geq \ln\left[\left(c(1\,|\,2) / c(2\,|\,1)\right)\left(p_2 / p_1\right)\right]$

R_2: $(-1/2)\,\mathbf{X}'\left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)\mathbf{X} + \left(\mu_1'\,\Sigma_1^{-1} - \mu_2'\,\Sigma_2^{-1}\right)\mathbf{X} - k < \ln\left[\left(c(1\,|\,2) / c(2\,|\,1)\right)\left(p_2 / p_1\right)\right]$

where $k = (1/2) \ln (|\Sigma_1| / |\Sigma_2|) + (1/2) (\mu_1' \Sigma_1^{-1} \mu_1 - \mu_2' \Sigma_2^{-1} \mu_2)$ (3.5)

We note that the classification regions are defined by *quadratic* functions of **X**.

The classification rule that minimises the expected cost of misclassification is as follows:

Allocate **X** to π_1 if (3.6)

$$(-1/2)\, \mathbf{X}'(\Sigma_1^{-1} - \Sigma_2^{-1})\, \mathbf{X} + (\mu_1' \Sigma_1^{-1} - \mu_2' \Sigma_2^{-1})\, \mathbf{X} - k \geq \ln \left[(c(1\,|\,2) / c(2\,|\,1)) (p_2 / p_1) \right]$$

and allocate **X** to π_2 otherwise.

In most applications μ_i and Σ_i, $i = 1,2$ are unknown. Then they are estimated through a *training sample*. The sample quantities $\overline{\mathbf{X}}_1$, $\overline{\mathbf{X}}_2$, $\mathbf{S}_1$ and $\mathbf{S}_2$ are substituted in (3.6). As a result, the sample analogue of the *quadratic* classification rule is as follows:

Allocate **X** to π_1 if (3.7)

$$(-1/2)\, \mathbf{X}'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\, \mathbf{X} + (\overline{\mathbf{X}}_1' \mathbf{S}_1^{-1} - \overline{\mathbf{X}}_2' \mathbf{S}_2^{-1})\, \mathbf{X} - k \geq \ln \left[(c(1\,|\,2) / c(2\,|\,1)) (p_2 / p_1) \right]$$

And allocate **X** to π_2 otherwise.

Where $k = (1/2) \ln (|\mathbf{S}_1| / |\mathbf{S}_2|) + (1/2) (\overline{\mathbf{X}}_1' \mathbf{S}_1^{-1} \overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2' \mathbf{S}_2^{-1} \overline{\mathbf{X}}_2)$ (3.8)

Note. For $\mathbf{S}_1^{-1}$ and $\mathbf{S}_2^{-1}$ to exist, it's very important that the inequalities $n_1 > m$ and $n_2 > m$ hold. n_1 and n_2 are the sizes of the *training samples* from π_1 and π_2, respectively. Classification with *quadratic* functions is rather awkward in more than two dimensions and can lead to some strange results, especially if the data are not multivariate normal.

The probabilities of misclassification are difficult to compute. In that effect, Anderson (1984) suggests a linear transformation of **X** so that its covariance matrix is **I** and the matrix of the *quadratic* form is diagonal; then the result in (3.3) has the distribution of a linear combination of non-central χ^2-variables plus a constant. Another way of computing the probabilities of misclassification which is easy and appealing, would be the *apparent error rate* (APER) (see Chapter I). Unfortunately, it tends to underestimate the *actual error rate* (AER) unless the sample sizes n_1 and n_2 are very large.

2) Case of non-normal multivariate populations

In the presence of non-normal multivariate populations, two options are suggested. First, the non-normal data can be transformed to data more nearly normal. Then a test for the equality of covariance matrices can be conducted to see if the linear rule in (2.16) or the *quadratic* rule in (3.5) is appropriate. The second option is to use a linear (or *quadratic*) rule without considering the form of the parent populations. For example, Fisher's procedure did not depend on the form of the parent populations. It only requires that the populations have identical covariance structures. However, studies by Krzanowski (1977) and Lachenbruch (1975) have shown non-normal cases where Fisher's linear classification function performs poorly. Therefore, we always have to check the performance of any classification procedure.

3) Case of several multivariate normal populations

This case has been outlined earlier in the introductory chapter. We briefly recall the results.

The multivariate normal densities are

$$f_i(\mathbf{X}) = \left[1 / (2\pi)^{m/2} | \Sigma_i |^{1/2}\right] \exp\left[(-1/2)\ (\mathbf{X}- \boldsymbol{\mu}_i)'\ \Sigma_i^{-1}\ (\mathbf{X}- \boldsymbol{\mu}_i)\right],\ i = 1,2,\ldots g. \qquad (3.9)$$

If $c(i|\,i) = 0$ and $c(k|\,i) = 1$, $k = i$ then the *quadratic discriminant score* for the *ith* population is as follows:

$$D_i^Q(\mathbf{X}) = (-1/2)\ln|\,\Sigma_i| - (1/2)\,(\mathbf{X}-\mu_i)'\,\Sigma_i^{-1}\,(\mathbf{X}-\mu_i) + \ln p_i \qquad i = 1,2,\ldots,g \qquad (3.10)$$

where p_i is the prior probability.

The *minimum total probability of misclassification* rule for several normal populations is as follows:

Allocate $\mathbf{X}$ to π_k if

the quadratic score $D_k^Q(\mathbf{X})$ = largest of $D_1^Q(\mathbf{X}), D_2^Q(\mathbf{X}), \ldots, D_g^Q(\mathbf{X})$ (3.11)

In most applications, μ_i and Σ_i are unknown. Then the estimate of the *quadratic discriminant score*, obtained through the *training sample*, is

$$d_i^Q(\mathbf{X}) = (-1/2)\ln|\,\mathbf{S}_i| - (1/2)\,(\mathbf{X}-\overline{\mathbf{X}}_i)'\,\mathbf{S}_i^{-1}\,(\mathbf{X}-\overline{\mathbf{X}}_i) + \ln p_i$$

Hence, the *estimated* minimum TPM rule for several normal populations is as follows:

Allocate $\mathbf{X}$ to π_k if

the quadratic score $d_k^Q(\mathbf{X})$ = largest of $d_1^Q(\mathbf{X}), d_2^Q(\mathbf{X}), \ldots, d_g^Q(\mathbf{X})$ (3.12)

4) Examples

We have generated data for two multivariate normal populations, π_1 and π_2, with m=3, different means, μ_1 and μ_2, and unequal covariance matrices, Σ_1 and Σ_2.

We consider six different cases:

a) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,1)$, $\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}$

b) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,5)$, $\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}$

c) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,1)$, $\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} 1 & .9 & .9 \\ .9 & 1 & .9 \\ .9 & .9 & 1 \end{bmatrix}$

d) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,5)$, $\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} 1 & .9 & .9 \\ .9 & 1 & .9 \\ .9 & .9 & 1 \end{bmatrix}$

e) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,1)$, $\Sigma_1 = \begin{bmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} 1 & .9 & .9 \\ .9 & 1 & .9 \\ .9 & .9 & 1 \end{bmatrix}$

f) $\mu_1' = (0,0,0)$, $\mu_2' = (0,1,5)$, $\Sigma_1 = \begin{bmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} 1 & .9 & .9 \\ .9 & 1 & .9 \\ .9 & .9 & 1 \end{bmatrix}$

For each case, training samples of sizes 15, 30, and 100 are generated. For each training sample, 50 validation samples of 1000 observations are generated. The probabilities of misclassification are evaluated by the *apparent error rate* (APER). These probabilities are presented in tables IV, V, and VI. The probabilities of misclassification substantially decrease with a choice of $\mu_2'=(0,1,5)$ (cases b, d, and f). For these same cases, an increase in the size of the training samples dramatically decreases the probabilities of misclassification.

Table IV. Probabilities of misclassification (APER) for two multivariate normal populations (m=3) with mean vectors μ_1'=(0,0,0) and μ_2, and covariance matrices Σ_1 and Σ_2.

$\Sigma_1=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$, and $\Sigma_2=\begin{bmatrix}1 & .5 & .5\\ .5 & 1 & .5\\ .5 & .5 & 1\end{bmatrix}$	μ_2'=(0,1,1)			μ_2' =(0,1,5)		
Training sample size	15	30	100	15	30	100
Total number of misclassified observations	13881	12740	12131	519	330	258
APER (%)	27.76	25.48	24.26	1.04	0.66	0.52

Table V. Probabilities of misclassification (APER) for two multivariate normal populations (*m*=3) with mean vectors μ_1'=(0,0,0) and μ_2, and covariance matrices Σ_1 and Σ_2.

$\Sigma_1=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$, and $\Sigma_2=\begin{bmatrix}1&.9&.9\\.9&1&.9\\.9&.9&1\end{bmatrix}$	μ_2'=(0,1,1)			μ_2' =(0,1,5)		
Training sample size	15	30	100	15	30	100
Total number of misclassified observations	6908	6159	5545	222	88	46
APER (%)	13.82	12.32	11.09	0.44	0.18	0.09

Table VI. Probabilities of misclassification (APER) for two multivariate normal populations (*m*=3) with mean vectors μ_1'=(0,0,0) and μ_2, and covariance matrices Σ_1 and Σ_2.

$\Sigma_1 = \begin{bmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} 1 & .9 & .9 \\ .9 & 1 & .9 \\ 9 & .9 & 1 \end{bmatrix}$	**μ_2'=(0,1,1)**			**μ_2' =(0,1,5)**		
Training sample size	15	30	100	15	30	100
Total number of misclassified observations	8646	7783	7119	55	13	7
APER (%)	17.29	15.57	14.24	0.11	0.03	0.01

Chapter IV Logistic regression: an alternative method for the discriminant analysis

In this chapter first we consider the general framework for logistic regression. And then we consider the relation between discriminant analysis and logistic regression.

1) Logistic regression (Cox and Snell, 1989)

We suppose there are n individuals, usually assumed to be independent. On each individual we have a binary observation or response, $\mathbf{Y}_i = 1$ a "success" or $\mathbf{Y}_i = 0$ a "failure". Also for each individual there is a row vector, $\mathbf{X}_i$, of explanatory variables. The probability of a binary variable, $\mathbf{Y}_i$, on a vector $\mathbf{X}_i$, of explanatory variables is θ_i. Hence Pr $(\mathbf{Y}_i = 1; \mathbf{X}_i) = \theta_i$ and Pr $(\mathbf{Y}_i = 0; \mathbf{X}_i) = 1 - \theta_i$.

The problem is to develop good methods of analysis for assessing any dependence of θ_i on the explanatory variables, $\mathbf{X}_i$, representing, for example, groupings of the individuals or quantitative explanatory variables. The simplest empirical relation is to suppose that θ_i is linearly dependent to the explanatory variables, $\mathbf{X}_i$

$$\theta_i = \alpha + \mathbf{X}_i \beta = \alpha + \sum \mathbf{X}_{is} \beta_s \qquad (4.1)$$

where β is a column vector of unknown regression coefficients and α is an unknown intercept.

The most serious restriction on the usefulness of (4.1) arises from

$$0 \leq \theta_i \leq 1 \qquad (4.2)$$

We discuss the models in which the constraint (4.2) is automatically satisfied.

The notion of a distribution of a *latent response variable* is used to motivate some alternatives. Suppose that there is a *latent variable* $\mathbf{u}$, which has a continuous cumulative distribution function $F(\mathbf{u}; \mathbf{X})$, for a given vector $\mathbf{X}$ of explanatory variables. The binary response $\mathbf{Y} = 1$ is recorded if and only if $\mathbf{u} > 0$. That is

$$\theta = \Pr(\mathbf{Y} = 1; \mathbf{X}) = 1 - F(0; \mathbf{X}) \qquad (4.3)$$

Note that there is no loss of generality in taking the critical point to be zero since $\mathbf{u}$ is not directly observed and also we may take the standard deviation of $\mathbf{u}$ or some other measure of dispersion, if constant, to be unity.

In this formulation, the critical level of $\mathbf{u}$ is regarded as fixed and the distribution of $\mathbf{u}$ as changing with $\mathbf{X}$. The complementary formulation in which the distribution of $\mathbf{u}$ is fixed and the critical level varies with $\mathbf{X}$ is more natural in bioassay when *dose*, or *log dose*, is the explanatory variable. For this version, we take $\mathbf{v}$ as the dose that would just produce a response, also called the *tolerance*. If the *dose* is $\alpha + \mathbf{X}\,\beta$, then

$$\Pr(\mathbf{Y} = 1; \mathbf{X}) = \Pr(\mathbf{v} \leq \alpha + \mathbf{X}\,\beta) \qquad (4.4)$$

which relates the probability that $\mathbf{Y} = 1$ directly to the distribution function of $\mathbf{v}$.
It is recommended, however, to use the first formulation because $\mathbf{u}$ thereby is more directly related to the observed binary response.

There are few possibilities for the distribution of $\mathbf{u}$. One is where $\mathbf{u}$ has a logistic distribution with location $\alpha + \mathbf{X}\,\beta$ and unit scale. This has cumulative distribution function (*c.d.f*)

$$F(\mathbf{u}; \mathbf{X}) = \exp(\mathbf{u} - \alpha - \mathbf{X}\,\beta) \,/\, \{1 + \exp(\mathbf{u} - \alpha - \mathbf{X}\,\beta)\} \qquad (4.5)$$

so that $$F(0; \mathbf{X}) = 1 \,/\, \{1 + \exp(\alpha + \mathbf{X}\,\beta)\} \qquad (4.6)$$

from which it follows

$$\theta = \Pr(\mathbf{Y} = 1; \mathbf{X}) = \exp(\alpha + \mathbf{X}\,\beta) \,/\, \{1 + \exp(\alpha + \mathbf{X}\,\beta)\} \qquad (4.7)$$

$$1 - \theta = \Pr(\mathbf{Y} = 0; \mathbf{X}) = 1 / \{1 + \exp(\alpha + \mathbf{X}\beta)\}$$

The relationship is *linearized* by the transformation

$$\text{Log}\ \{\theta / (1 - \theta)\} = \alpha + \mathbf{X}\beta \tag{4.8}$$

For scalar **X** and $\beta > 0$, (4.7) is said to define via (4.4) a probability density function on differentiation with respect to **X**, namely

$$\beta \exp(\alpha + \mathbf{X}\beta) / (1 + \exp(\alpha + \mathbf{X}\beta)^2) \tag{4.9}$$

The logistic regression model is formulated mathematically by relating the probability of some event, $\mathbf{Y} = 1$ or 0, conditional on a vector, **X**, of explanatory variables, to the vector **X**, through the functional form of a logistic *c.d.f.* This model is given by (4.7) where (α, β) are unknown parameters that are estimated from the data. The *linearized* relation in (4.8) is called the *linear logistic model.*

$$\lambda_t = (\theta_i / (1 - \theta_i)) = \alpha + \mathbf{X}_i \beta = \alpha + \sum_{t=1}^{m} \mathbf{X}_{it} \beta_t \tag{4.10}$$

In this model, there are unknown parameters α and the mx1 column vector β. For general purposes, it is convenient to change the notation slightly by writing $\beta_0 = \alpha$ and $\mathbf{X}_{i0} = 1$, when (4.9) is equivalent to $\lambda = \mathbf{X}\beta$, where λ is an nx1 column, **X** is an nxd matrix ($d = m+1$), and β is a dx1 column of parameters, $\beta^T = (\beta_0, \ldots, \beta_m)$. We shall assume that $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are n distinct individuals, mutually independent.

To estimate α and β, we can maximize the *conditional likelihood function*

$$f_{\alpha,\beta}(\mathbf{Y}_1, \ldots, \mathbf{Y}_n) = \prod_{i=1}^{n} \theta_i^{\mathbf{Y}_i} (1 - \theta_i)^{1 - \mathbf{Y}_i}$$

$$= \prod_{i=1}^{n} \left[e^{\alpha + \mathbf{X}_i \beta} / \left(1 + e^{\alpha + \mathbf{X}_i \beta}\right)\right]^{\mathbf{Y}_i} \left[1 / \left(1 + e^{\alpha + \mathbf{X}_i \beta}\right)\right]^{(1 - \mathbf{Y}_i)}$$

$$= \prod_{i=1}^{n} \left[\left(e^{\alpha + \mathbf{X}_i \beta}\right)^{\mathbf{Y}_i} / \left(1 + e^{\alpha + \mathbf{X}_i \beta}\right)\right]$$

$$= \exp\left[\Sigma_{i=1}^{n} (\alpha + \mathbf{X}_i \beta)\, \mathbf{Y}_i\right] / \prod_{i=1}^{n} \left(1 + e^{\alpha + \mathbf{X}_i \beta}\right) \tag{4.11}$$

with respect to (α, β).

Note. In most situations it is preferable to work directly with the probabilities of success. The concept of a *latent distribution* has proved useful when the *latent variable* has an intrinsic physical significance and also when the idea is useful in suggesting models for more complex problems.

2) <u>Relation between discriminant analysis and logistic regression</u>

We shall consider the relation between two intimately related and yet conceptually quite different techniques, namely discriminant analysis and logistic regression.

In discriminant analysis (see Introduction, Chapter I) there are two distinct populations, defined by 1 or 0. Within each of these populations, there is a set of properties **X**. That is, there are two probability densities $f_0(\mathbf{X})$ and $f_1(\mathbf{X})$. The focus in discriminant analysis is on how those distributions differ most sharply. The problem could be formulated as follows. Given a new vector **X′** from an individual of unknown **Y**, we wish to find out, in some optimal way, the population from which the individual was drawn. The emphasis is strongly on the distributions of **X** within the two populations. On the other hand, logistic regression presupposes a stable statistical relation such that once a vector of explanatory variables, **X**, is given a probability that a binary response, **Y**, is equal to one, is determined. The distribution of **X** is not directly relevant to the definition.

At this point, two rather different situations are considered. In the first, the relative frequencies with which the two populations generate data are not defined, since they may change relatively under hypothetical repetition. Then we cannot consider the probability distribution of **Y**, either marginally or conditionally, on **X**. Thus logistic regression is not applicable. However, discriminant analysis is applicable and the statistic for assessing **X**′ is the *log-likelihood ratio*

$$\log f_1(\mathbf{X}') - \log f_0(\mathbf{X}') \qquad (4.12)$$

If the two densities come from the same exponential family with canonical statistic **X** and with two different parameter values, then (4.11) is a linear function of the components of **X**. The resulting function is called a *linear discriminant function*. The most important special case is when f_0 and f_1 are multivariate normal densities with the same covariance matrix Σ and means μ_0 and μ_1 (see Chapter II). Then (4.11) becomes

$$(-1/2)\ (\mu_1\, \Sigma^{-1}\, \mu_1' - \mu_0\, \Sigma^{-1}\, \mu_0') + \mathbf{X}\, \Sigma^{-1}\, (\mu_1 - \mu_0)'$$

which is the population *linear discriminant function.*

In the second situation, still within the framework of discriminant analysis, there are physically defined probabilities π_0 and π_1 such that **Y** is 0, 1 with $\pi_0 + \pi_1 = 1$. Then we can represent membership of a population for an arbitrary individual by a random variable, **Y**. The full properties of **Y** are represented by a vector of random variable (**Y**, **X**).

The functions $f_0(\mathbf{X})$ and $f_1(\mathbf{X})$ specify conditional densities of **X** given **Y** = 0, 1. For the new individual of known **X′** but unknown **Y**, we have by *Bayes Theorem* that

$$\Pr(\mathbf{Y} = 1 \mid \mathbf{X}') = (f_1(\mathbf{X}')\,\pi_1) \,/\, (f_0(\mathbf{X}')\,\pi_0 + f_1(\mathbf{X}')\,\pi_1) \qquad (4.13)$$

So that

$$\log\{\Pr(\mathbf{Y} = 1 \mid \mathbf{X} = \mathbf{X}') / \Pr(\mathbf{Y} = 0 \mid \mathbf{X} = \mathbf{X}')\} = \log(\pi_1 / \pi_0) + \log(f_1(\mathbf{X}') / f_0(\mathbf{X}')) \qquad (4.14)$$

defining a logistic regression in which the prior probabilities are isolated into a single term. Hence from a ***linear discriminant function***, in the sense mentioned above, results a ***linear logistic regression***. It is noted that this happens only when the conditional distributions of **X** are normal with the same covariance matrix.

3) Comparison of the efficiency of discriminant analysis and logistic regression

We shall compare the efficiency of discriminant analysis and logistic regression. The literature on this topic is numerous. We shall summarize a few articles.

As observed earlier, relating qualitative variables to other variables through a ***logistic cumulative density function*** (*c.d.f*) functional form is logistic regression. Classifying an observation into one of several populations is discriminant analysis. In most discriminant analysis applications, at least one variable is qualitative (ruling out multivariate normality). If the populations are normal with common covariance matrices, discriminant analysis estimators are preferred to logistic regression estimators for the discriminant analysis problem. However, under non-normality, the logistic regression model with ***maximum likelihood estimators*** is preferred.

We start the comparison by summarizing the discussion presented by Cox and Snell (1989). As discussed in Chapter II, one approach to linear discriminant analysis is to find the linear combination of the components of **X** which most strongly separates the two populations, by maximizing the square of the difference between the population means divided by its covariance matrix, assumed common but arbitrary for the components of **X**.

The result is the *Fisher's linear discriminant function*

$$\mathbf{Y} = \mathrm{L}'\mathbf{X}$$

$$\mathbf{Y} = (\mu_1 - \mu_2)'\,\Sigma^{-1}\,\mathbf{X}$$

L might be estimated by replacing μ_i by the sample mean of population *i* and Σ by the pooled sample covariance matrix **S**.

As a consequence of the geometry of the estimation problem this technique is considered equivalent to the formal linear regression of the binary variable **Y** on the vector **X**, treated as fixed. And also under normal theory assumptions, exact tests of regression coefficients are obtained by pretending that the *fixed* binary **Y** is normal and that the random multivariate normal **X** is fixed. With a slight loss of generality, the two conditional densities are taken to be of common parametric form with some common parameters. That is

$$f_0(\mathbf{X}) = g(\Psi, \lambda_0)\,,\;\; f_1(\mathbf{X}) = g(\Psi, \lambda_1) \qquad (4.15)$$

for some known function g and unknown parameters Ψ, λ_0, λ_1 and , π_0. The multivariate normal case with the common covariance matrix is clearly included. We assume there are n_0, n_1 individuals respectively from the two populations I and II. All individuals are assumed to be independent and generated by the full probability model described. Each individual has its full **X** vector of observations.

The method of *maximum likelihood* is applied to estimate the unknown parameters. The contribution of the *ith* individual to the likelihood can be written as

$$\Pr(\mathbf{Y} = \mathbf{Y}_i) f_Y(\mathbf{X}_i) \quad \text{for } \mathbf{Y}_i = 0,1 \tag{4.16}$$

$$\text{or} \quad \Pr(\mathbf{Y} = \mathbf{Y}_i \mid \mathbf{X} = \mathbf{X}_i)\ f_X(\mathbf{X}_i) \tag{4.17}$$

The maximization of either expression leads to the same result, in the case of a normal population to *Fisher's linear discriminant function.* From the maximization of (4.16) we see that $\pi_0 = n_0/(n_0 + n_1)$ and the remaining parameters are estimated from the two samples of sizes n_0 and n_1, from the densities f_0 and f_1, respectively.

For the second version, (4.17), if we maximize only the first factor we would be using techniques of logistic regression analysis. It follows that logistic regression is inefficient, under the full assumptions, in that the second factor of (4.17) does contain information about relevant parameters and this information has been totally discarded. Efron (1975) and Ruiz (1989) have investigated this loss, the former by using discriminant misclassification rates as a criterion and the latter estimating efficiency; up to one-third loss in efficiency can occur.

The most important special case considered is where f_0 and f_1 are multivariate normal densities with the same covariance matrix Σ and means μ_0 and μ_1, respectively. Then the *log-likelihood ratio*, for assessing $\mathbf{X}'$, given by

$$\log f_1(\mathbf{X}') - \log f_0(\mathbf{X}') \tag{4.18}$$

becomes $(-1/2)\ (\boldsymbol{\mu}_1\, \boldsymbol{\Sigma}^{-1}\, \boldsymbol{\mu}_1' - \boldsymbol{\mu}_0\, \boldsymbol{\Sigma}^{-1}\, \boldsymbol{\mu}_0') + \mathbf{X}\ \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ (4.19)

which is the population *linear discriminant function.*

Under the same assumptions, Cox and Snell (1989) proved that the estimate $\bar{\beta}_{(d)}$ of coefficients determined by a discriminant approach is directly proportional to $\bar{\beta}_{(\hat{L})}$ of obtained by substituting the ***maximum likelihood*** estimates $\bar{\boldsymbol{\Sigma}}$, $\bar{\boldsymbol{\mu}}_0$ and $\bar{\boldsymbol{\mu}}_1$ into the population discriminant (4.19). Thus from (4.19)

$$\bar{\beta}_{(\hat{L})} = \bar{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0)^{\mathrm{T}} \tag{4.20}$$

The estimate $\beta_{(d)}$ is given by solution of the least squares equations

$$\mathbf{S}\bar{\beta}_{(d)} = (n_0\, n_1/n)\ (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0)^{\mathrm{T}} \tag{4.21}$$

Where S is the matrix of total sums of squares and products and n_0, n_1 ($n = n_0 + n_1$) are the numbers responding to $\mathbf{Y} = 0, 1$. Since

$$\mathbf{S} = n\,\bar{\boldsymbol{\Sigma}}^{-1} + (n_0\, n_1/n)\ (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0)^{\mathrm{T}}\ (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0) \tag{4.22}$$

Thus, we have

$$\bar{\beta}_{(d)} = (n_0\, n_1/n)\ \{1 - (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0)\ \bar{\beta}_{(d)}\}\ \bar{\beta}_{(\hat{L})} = k\,\bar{\beta}_{(\hat{L})} \tag{4.23}$$

where k is equal to the difference between the total sum of squares $(n_0\, n_1/n)$ and the sum of squares due to regression. Hence,

$$\bar{\beta}_{(d)} = \bar{\beta}_{(\hat{L})}\, SS_{res}/n \cong \bar{\beta}_{(\hat{Lr})}\, SS_{res}/n \tag{4.24}$$

where $\bar{\beta}_{(\hat{Lr})}$ is the ***maximum likelihood logistic regression*** estimate and SS_{res} is the residual sum of squares obtained when **Y** is regressed on **X**.

In summary, Cox and Snell (1989) state that the key issue in choosing one of the two techniques, is the stability of the conditional probability of **Y** given **X** and of the distributions of **X** within the two sub-populations. If the two techniques are applicable, and the logistic regression is effectively linear, logistic regression assumes less, in that given the linear regression the forms of the distributions of **X** within sub-populations are irrelevant. On the other hand, if multivariate normality or some other specific distributional form can be taken, then the discriminant approach is more efficient.

4) Summary of few articles

In the article "Choosing between logistic regression and discriminant analysis", Press and Wilson (1978) present theoretical arguments for using logistic regression with maximum likelihood estimation compared to using linear discriminant analysis, in the classification problem and the problem of relating qualitative to explanatory variables. The presence of qualitative variables rules out multivariate normality. The authors concluded that under non-normality, the logistic regression model with maximum likelihood estimators outperforms discriminant analysis. The related arguments are supported by the results of several empirical comparisons of the MLE logistic regression and discriminant analysis estimators involving breast cancer, and population changes across states of the U.S.

Discriminant function estimators have been used in logistic regression, in both theory and applications (see: Truett, Cornfield, and Kannel,1967).
These estimators were compared empirically with maximum likelihood estimators for logistic regression problems, and they were found to be generally inferior, but not by substantial amounts (Halperin, Blackwelder, and Velter, 1971, and D'Agostino et al.1978).

The discriminant function estimators have been used as starting values in iterative maximum likelihood estimation and in exploratory data analysis, for logistic regression models. There are alternative estimators for the logistic regression problem, as well as for the non-normal discriminant problem such as the "reverse Taylor series approximations" and the "conditional estimators" (Nerlove and Press, 1973).

"Conditional estimators" are obtained by maximizing the conditional likelihood (conditional on the explanatory variables). "Reverse Taylor series approximations" arise from the logistic cumulative density function,

$$F(\mathbf{X}) = 1 \,/\, [\, 1 \,/\, (1 + e^{-(a+b\mathbf{X})})\,] \,, \quad b \neq 0 \,, \quad -\infty < x < \infty \,.$$

From the Taylor series expansion about $\mathbf{X} = (\mathbf{X} - \mathbf{X_0}) + \mathbf{X}$, we get

$$\begin{aligned} F(\mathbf{X}) &= (1 \,/\, [1 + e^{-(a+b\mathbf{X})}]\,) \\ &= F(\mathbf{X} - \mathbf{X_0} + \mathbf{X_0}) \\ &= F(\mathbf{X_0}) + (\mathbf{X} - \mathbf{X_0})\, F'(\mathbf{X_0}) \\ &= F(\mathbf{X_0}) - \mathbf{X_0}\, F'(\mathbf{X_0}) + \mathbf{X}\, F'(\mathbf{X_0}) \\ &= A + B\mathbf{X} \end{aligned}$$

where $A = \{1 \,/\, [1 + e^{-(a+b\mathbf{X})}]\,\} - B\,\mathbf{X}$

and $B = \{\, b\, e^{-(a+b\mathbf{X_0})} / [1 + e^{-(a+b\mathbf{X_0})}]^2 \}.$

Solving these equations for a and b, we get

$$b = B \,/\, [\,(A + B\mathbf{X})\,(1 - A - B\mathbf{X})\,]$$

and $$a = -\, b\mathbf{X} - \log\, (\,(1 \,/\, A + B\mathbf{X}) - 1)$$

as the reverse Taylor series approximation.

The reverse Taylor series estimators are appropriate regardless of the underlying distribution of explanatory variables. By contrast, the discriminant function estimators are appropriate only when the explanatory variables have a multivariate normal distribution, with equal covariance.

At this point, the authors address two general questions. The first is, why use a logistic formulation rather some other functional form? The second is, how should the parameters of the model be estimated? As Anderson (1972) pointed out the logistic formulation results not only from assuming that the explanatory variables have multivariate normal distribution with equal covariance matrices, but also from assuming that the explanatory variables are independent and dichotomous zero-or-one variables, or that some are multivariate normal and some dichotomous. By contrast, the linear discriminant approach is applicable only when the explanatory variables are multivariate normal with equal covariance matrices. Thus, one advantage of using the logistic model rather than the linear discriminant function, for discriminant analysis, is that the former is robust; i.e., many types of underlying assumptions lead to the same logistic formulation.

Another advantage of the logistic model would be its use as an alternative to contingency table analysis in biological and medical applications. This was pointed out by Gordon (1974); Cross-classified tables with large numbers of cells, and usually too few observations per cell, are replaced by a logistic or log-linear relationship among the variables. One possible hazard of the linear combination of variables in a multivariate logistic formulation is that some types of interaction may not be expressible in that form.

However, the logistic function can be appropriately used in many such applications.

Efron (1975) has shown that logistic regression estimators are between one-half to two-thirds as efficient as discriminant function estimators when the data are multivariate normal with equal covariance matrices. Halperin, Blackwelder, and Verter (1971) compared maximum likelihood estimation and linear discriminant estimation, for a logistic regression, and found that "the times required for compilation and execution of the programs were higher for the maximum likelihood method than for the discriminant method by factors ranging from 1.3 to 2".

The authors present the following arguments against the use of discriminant function estimators:

- If the explanatory variables are binary (they don't follow a multivariate normal distribution with equal covariance matrices) discriminant function estimators of the slope coefficients in the logistic regression will not be consistent. Even in large samples there is no guarantee that good prediction will be obtained by this method. The solution is to use a consistent method of estimation, such as MLE.

Halperin, Blackwelder, and verter (1971) have proven the inconsistency of discriminant function estimators in logistic regression, for various cases, under non-normality.

- Under non-normality of the explanatory variables, discriminant function estimation can give misleading results regarding significance of the logistic regression coefficients. For example, a slope coefficient which is really zero, is not necessarily estimated as zero by the discriminant function method.

- Halperin, Blackwelder, and Verter (1971) found that, under non-normal conditions, the "maximum likelihood method usually gives slightly better fits to the model, as evaluated from observed and expected numbers of cases per decile of risk." They also found that "there is a theoretical basis for the possibility that the discriminant function will give a very poor fit, even if the logistic regression model holds."

- The use of estimators based on sufficient statistics result in smaller mean squared error (Rao-Blackwell theorem, see Rao, 1965) compared to estimators not based on sufficient statistics. The MLEs are functions of the sufficient statistics, while the discriminant function estimators are not.

- An interesting and desirable property of the maximum likelihood estimation of the logistic regression is that the expected number of cases equals the observed number of cases; i.e., $\Sigma Y_i = \Sigma P(\mathbf{X}_{1i}, \ldots, \mathbf{X}_{ki})$ (Halperin, Blackwelder, and Verter, 1971). The discriminant approach does not satisfy this property.

- In a Bayesian analysis, McFadden (1976) concludes that the use of discriminant function estimators may tend to generate substantial bias in some applications.

The ideas and arguments presented by the authors are illustrated through two examples of classification problem. In each case, both logistic regression and linear discriminant analysis were carried out on empirical data. We illustrate one of the two studies. The data in this example comes from breast cancer patients initially treated at the British Columbia Cancer Institute between 1955 and 1963.

The variables are mixed: continuous, discrete, and binary. The binary grouping variable is defined to be 0 if metastatic carcinoma is not present in the lymph nodes, and 1 if it is present. The independent variables are

- Number of births, X_1.
- A history of hysterectomy (0-1), X_2.
- A history of benign breast disease during lactation (0-1), X_3.
- Presence of nipple changes as the first disease symptom (0-1), X_4.
- Duration of symptoms in months, X_5.

There were 173 patients of which 115 were used in the training set and 58 in the validation set. The patients' nodal status had been determined by a surgical procedure. There were no missing data. The estimated functions for logistic regression were given by

$$Y(X) = .058 - .233\, X_1 - 1.096\, X_2 + .713\, X_3 - .028\, X_4 + .995\, X_5 .$$

$$U(X) = .362 - .251\, X_1 - 1.245\, X_2 + 1.104\, X_3 - .036\, X_4 + 2.114\, X_5 .$$

The prior probabilities used were the approximate proportions of actual cases in the data: 0.66 of having no metastases and 0.34 of having metastases. The logistic regression classified correctly 71% of the patients into the training set and 62% of the patients into the validation set. The discriminant analysis correct classification rate was of 67% for the training set and 59% for the validation set. The correct classification rate of logistic regression is higher compared to discriminant analysis. Moreover, there was a difference in the types of cases misclassified by the two procedures. The discriminant function consistently misclassify more patients into the group having no metatases than the

logistic regression. Hence logistic regression with MLE outperforms classical linear discriminant analysis, in the presence of non-normality, but not by a large amount.
A similar result is obtained in the second empirical example.

Thus, Press and Wilson agreed with the conclusion of Halperin, Blackwelder, and Verter (1971) that "use of maximum likelihood method would be preferable, whenever practical, in situations where the normality assumptions are violated, especially when many of the independent variables are qualitative".

O'Neill (1980) showed that the efficiency of logistic regression in some non-normal cases is low. In his article, the asymptotic distribution of the *error rates* of an estimator of the optimal classification rule

$$R^m = D_0 \cup D_1$$

such that $\mathbf{Y} = 1$, *i.e.* the individual belongs to π_1, if $\mathbf{X} \in D_1$,

and $\mathbf{Y} = 0$, *i.e.* the individual belongs to π_0, if $\mathbf{X} \in D_0$.

with the optimal partition

$$D_1 = \{\mathbf{X} \in R^m: \pi_1 f_1(\mathbf{X}) / \pi_0 f_0(\mathbf{X}) > 1\}$$

$$D_0 = \{\mathbf{X} \in R^m: \pi_1 f_1(\mathbf{X}) / \pi_0 f_0(\mathbf{X}) \leq 1\}$$

where $f_i(\mathbf{X}) = f(\mathbf{X} \mid \mathbf{Y} = i)$, $i = 0, 1$, for arbitrary f_0 and f_1 is given.
Once the asymptotic distribution of the *logistic regression estimators* was obtained, this enabled the comparison of logistic regression and *maximum likelihood discrimination* for arbitrary distributions other than the normal distribution with constant covariance studied by Efron (1975). O'Neill also compared the efficiency of logistic regression and *maximum likelihood discrimination* in two cases: the *exponential distribution* with $m = 2$

and *quadratic normal discrimination.* He concluded that the inefficiency of *logistic regression discrimination* is more marked in both cases considered. The poor performance for situations in which good discrimination is possible casts doubt on the use of the logistic regression discrimination rule and suggests that the maximum likelihood estimation of optimal discriminant rule for the specific distributions at hand should be used whenever possible.

Efron (1975) computes the asymptotic relative efficiency of the *normal discriminant analysis*, i.e., linear discriminant analysis, and the logistic regression in his article "The efficiency of logistic regression compared to normal discriminant analysis". The author shows that logistic regression is between one-half to two-thirds as effective as normal discrimination for statistically interesting values of the parameters.

The framework of the article is as follows: there are two m-dimensional normal Populations, 1 and 0, differing in mean but not in covariance

$$\begin{aligned} &\mathbf{X} \sim N_m(\boldsymbol{\mu}_1, \Sigma) \text{ with prior probability } p_1, \\ &\mathbf{X} \sim N_m(\boldsymbol{\mu}_0, \Sigma) \text{ with prior probability } p_0, \end{aligned} \tag{4.25}$$

where $p_1 + p_0 = 1$.

The Anderson's classification function is $\lambda(\mathbf{X}) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}$, where

$$\begin{aligned} &\beta_0 \equiv \log(p_1/p_0) - (1/2)(\boldsymbol{\mu}_1' \Sigma^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0' \Sigma^{-1} \boldsymbol{\mu}_0) \\ &\boldsymbol{\beta}' \equiv (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)' \Sigma^{-1}. \end{aligned} \tag{4.26}$$

A random vector $\mathbf{X}$, which arises from one of the two populations, is assigned to population 1 if $\lambda(\mathbf{X}) > 0$ and to population 0 if $\lambda(\mathbf{X}) < 0$.

When the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, p_1, p_0 and Σ are unknown, they are substituted by their maximum likelihood estimates, through a training set $(Y_1, \mathbf{X}_1), (Y_2, \mathbf{X}_2), \ldots,$

$(Y_n, \mathbf{X}_n)$,

where Y_j indicates which population $\mathbf{X_j}$ comes from.

$$\begin{aligned} Y_j &= 1 \text{ with probability } p_1, \\ &= 0 \text{ with probability } p_0, \end{aligned} \qquad (4.27)$$

and $\mathbf{X_j} \mid Y_j \sim N_p(\mu_{Yj}, \Sigma)$. (4.28)

The maximum likelihood estimates of the parameters are

$$\hat{p}_1 = n_1 / n, \quad \hat{p}_0 = n_0 / n$$

$$\hat{\mu}_1 = \overline{\mathbf{X}}_1 \equiv \Sigma_{Y_j=1} \mathbf{X_j} / n_1, \quad \hat{\mu}_0 = \overline{\mathbf{X}}_0 \equiv \Sigma_{Y_j=0} \mathbf{X_j} / n_0, \qquad (4.29)$$

and $\hat{\Sigma} = [\Sigma_{Y_j=1} (\mathbf{X_j} - \overline{\mathbf{X}}_1)(\mathbf{X_j} - \overline{\mathbf{X}}_1)' + \Sigma_{Y_j=0} (\mathbf{X_j} - \overline{\mathbf{X}}_0)(\mathbf{X_j} - \overline{\mathbf{X}}_0)'] / n$

where $n_1 \equiv \Sigma_{j=1}^{n} Y_j$ and $n_0 \equiv n - n_1$.

The values $(\hat{\beta}, \hat{\beta}_0)$ gives a version of Anderson's estimated linear discriminant function $\hat{\lambda}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}' \mathbf{X}$, such that a new observation $\mathbf{X}$ is assigned to population 1 or 0 if $\hat{\lambda}(\mathbf{X})$ is greater than or less than zero.

If the functions $f_1(\mathbf{X})$ and $f_0(\mathbf{X})$ specify the conditional densities of $\mathbf{X}$ given y equal to one or zero, for a new observation of known $\mathbf{X_j}$ but unknown Y_j, we have by the Bayes' theorem that

$$\Pr\{Y_j = 1 \mid \mathbf{X_j}\} = p_1 f_1(\mathbf{X_j}) / (p_1 f_1(\mathbf{X_j}) + p_0 f_0(\mathbf{X_j})) \qquad (4.30)$$

So that $\log\{\Pr\{Y_j = 1 \mid \mathbf{X_j}\} / \Pr\{Y_j = 0 \mid \mathbf{X_j}\}\} = \log(p_1/p_0) + \log\{f_1(\mathbf{X_j}) / f_0(\mathbf{X_j})\}$

Denote $p_{1j} = \Pr\{Y_j = 1 \mid \mathbf{X_j}\}$ and $p_{0j} = \Pr\{Y_j = 0 \mid \mathbf{X_j}\}$, then $\lambda(\mathbf{X}) \equiv \lambda(\mathbf{X}_j) = \log (p_{1j}/ p_{0j})$. (4.31)

Hence, $\lambda(\mathbf{X})$ is shown to be the a *posteriori* log odds ratio for population 1 versus population 0 having observed $\mathbf{X}$.

Given the values $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, the Y_j are conditionally independent binary random variables,

$$p_{1j} = \Pr\{\ Y_j = 1 \mid \mathbf{X}_j\ \} = \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_j) \,/\, [1 + \exp(\beta_0 + \boldsymbol{\beta}'\,\mathbf{X}_j)\,]$$

$$p_{0j} = \Pr\{\ Y_j = 0 \mid \mathbf{X}_j\ \} = 1 \,/\, [1 + \exp(\beta_0 + \boldsymbol{\beta}'\,\mathbf{X}_j)\,]. \tag{4.32}$$

Estimates of $(\beta_0, \boldsymbol{\beta})$ are obtained by maximization of the conditional likelihood

$$f_{\beta 0,\beta}(\ Y_1, Y_2, \ldots, Y_n \mid \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$$

$$= \prod_{j=1}^{n} p_{1j}^{\ Yj}\ p_{0j}^{\ (1-Yj)}\ , \tag{4.33}$$

$$= \prod_{j=1}^{n} \exp[\ (\beta_0 + \boldsymbol{\beta}'\mathbf{X}_j)\, Y_j\ ]\ /\ [1 + \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_j)\ ]$$

with respect to $(\beta_0, \boldsymbol{\beta})$.

The values $(\bar{\beta}_0, \bar{\boldsymbol{\beta}})$ give $\bar{\lambda}(\mathbf{X}) = \bar{\beta}_0 + \bar{\boldsymbol{\beta}}'\mathbf{X}$ as an estimate of the linear discriminant function $\lambda(\mathbf{X})$. The discriminant procedure which chooses population 1 if $\bar{\lambda}(\mathbf{X}) > 0$ and population 0 if $\bar{\lambda}(\mathbf{X}) < 0$, will be referred to as the *logistic regression procedure*.

The normal discrimination procedure is based on the full maximum likelihood estimator for $\lambda(\mathbf{X})$ whereas the logistic regression procedure is based on the conditional likelihood estimator for $\lambda(\mathbf{X})$. Thus, the logistic regression must be less efficient than than the normal discrimination, at least asymptotically, as n goes to infinity.

Under a variety of situations and measures of efficiency, the central result for the asymptotic relative efficiencies is

$$\mathrm{ARE} = (2\pi)^{(-1/2)}\,(1 + \Delta^2 p_0 p_1)\exp(-\Delta^2/8)\int_{-\infty}^{\infty}\exp(-\mathbf{X}^2/2)\,/\,(p_1\exp(\Delta\mathbf{X}/2) + p_0\exp(-\Delta\mathbf{X}/2)) \qquad (4.34)$$

where Δ is the square root of the Mahalanobis distance between of population 1 and 0. The author gives values of ARE for reasonable values of Δ, with $p_0 = p_1 = ½$, which is the case most favorable to the logistic regression.

Δ	0	.5	1	1.5	2	2.5	3	3.5	4
ARE	1.0000	1.0000	.995	.968	.899	.786	.641	.486	.343

For Δ between 2.5 and 3.5, good discrimination becomes possible but at the same time, the ARE of logistic regression decreases sharply.

Although the logistic regression is less efficient and also more difficult to calculate, it is more robust than normal discrimination. The conditional likelihood (4.33) is valid under general exponential family assumptions on the density $f(\mathbf{X})$ of **X**,

$$\begin{aligned} f(\mathbf{X}) &= g(\theta_1, \eta)\, h(\mathbf{X}, \eta)\exp(\theta_1{}'\,\mathbf{X}) && \text{with probability } p_1, \\ f(\mathbf{X}) &= g(\theta_0, \eta)\, h(\mathbf{X}, \eta)\exp(\theta_0{}'\,\mathbf{X}) && \text{with probability } p_0, \end{aligned} \qquad (4.35)$$

where $p_1 + p_0 = 1$.

η is an arbitrary nuisance parameter, like **Σ** in (4.25). Equation (4.25) is a special case of equation (4.35).

Efron used the linear transformation $\tilde{\mathbf{X}} = \mathrm{a} + \mathbf{AX}$ to reduce (4.25) to the case

$\tilde{\mathbf{X}} \sim N_m((\Delta/2)\,\mathbf{e}_1, \mathrm{I})$ with probability p_1,

$\tilde{\mathbf{X}} \sim N_m(-(\Delta/2)\,\mathbf{e}_1, \mathrm{I})$ with probability p_0,

where $p_1 + p_0 = 1$.

And $\mathbf{e}'_1=(1,0,0,\ldots,0)$; I is the mxm identity matrix; and Δ is the square root of the Mahalanobis distance (1.11).

The boundary $B = \{\mathbf{X}: \lambda(\mathbf{X}) = 0\}$ between Fisher's optimum decision regions for the two populations to the new optimum boundary

$$\tilde{B} = \{\tilde{\mathbf{X}}: \tilde{\lambda}(\tilde{\mathbf{X}}) = 0\} = \{\tilde{\mathbf{X}}: \tilde{\mathbf{X}} = \mathrm{a} + \mathrm{A}\mathbf{X}, \mathbf{X} \in B\}. \qquad (4.37)$$

Both estimated boundaries $\hat{B} = \{\mathbf{X}: \hat{\lambda}(\mathbf{X}) = 0\}$ and $\bar{B} = \{\mathbf{X}: \bar{\lambda}(\mathbf{X}) = 0\}$ for logistic regression and normal discrimination, respectively, are transformed as in (4.37). In other words, for both procedures, the estimated discrimination procedure based on the transformed data is the transform of that based on the original data.

For a partition of the m-dimensional space E^m into the regions R_0 and R_1, such that we choose population 0 or 1 as $\mathbf{X}$ falls into R_0 or R_1, respectively, the error rate (or the probability of misclassification) under assumption (4.25) is

$$\begin{aligned}\text{Error Rate} \equiv\ & p_1 \Pr\{\mathbf{X} \in R_0 \mid \mathbf{X} \sim \eta_m(\mu_1, \Sigma)\} \\ & + p_0 \Pr\{\mathbf{X} \in R_1 \mid \mathbf{X} \sim \eta_m(\mu_0, \Sigma)\} \qquad (4.38)\end{aligned}$$

Error rate is a random variable since the partition is chosen randomly by the logistic regression and normal discrimination procedures. For either procedure, the error rate will have the same distribution under (4.25) and (4.36).

Henceforth, the simpler assumptions (4.36) (calling it the "standard situation", and $\tilde{\mathbf{X}}$ will be referred to as **X**) will be worked with.

Fisher's linear discriminant function (4.26), under the standard situation, becomes

$$\lambda(\mathbf{X}) = \lambda + \Delta\, \mathbf{X}_1 \qquad (4.39)$$

The optimal boundary $B(0,0) = \{\lambda(\mathbf{X}) = 0\}$ is the $(m\text{-}1)$ dimensional plane orthogonal to the $\mathbf{X}_1$ axis and intersecting it at the value $\tau \equiv -\lambda / \Delta$.

Let $B(d\tau, d\alpha)$ be another boundary, intersecting the $\mathbf{X}_1$ axis at $\tau + d\tau$, with normal vector at an angle $d\alpha$ from the $\mathbf{X}_1$ axis. The error rate (4.38) of the regions separated by $B(d\tau, d\alpha)$ will be denoted by $\mathrm{ER}(d\tau, d\alpha)$. $d\tau$ and $d\alpha$ denote small discrepancies from optimal, which will be the case in the large sample theory.

The error rate of the optimal boundary $B(0,0)$ is

$$\mathrm{ER}(0,0) = p_1\phi(-\mathrm{D}_1) + p_0\phi(-\mathrm{D}_0) \qquad (4.40)$$

where $\quad \mathrm{D}_1 = (\Delta/2) - \tau, \quad \mathrm{D}_0 = (\Delta/2) + \tau, \qquad (4.41)$

and $\quad \phi(\mathbf{Z}) = \int_{\infty}^{\mathbf{Z}} \varphi(\mathbf{t})\, d\mathbf{t}, \quad \varphi(\mathbf{t}) = (2\pi)^{-1/2} \exp(-\mathbf{t}^2/2).$

The distances from $\mu_1$and μ_0 to $B(d\tau, d\alpha)$ are defined as

$$d_1 = (\mathrm{D}_1 - d\tau) \cos(d\alpha),$$

$$d_0 = (\mathrm{D}_0 - d\tau) \cos(d\alpha). \qquad (4.42)$$

Then $\quad \mathrm{ER}(d\tau, d\alpha) = p_1\phi(-\mathrm{D}_1) + p_0\phi(-\mathrm{D}_0). \qquad (4.43)$

From the Taylor expansions,

$$\cos(d\alpha) = 1 - (d\alpha)^2/2 + \dots$$

and $\quad \phi(-\mathrm{D} + d\tau) = \phi(-\mathrm{D}) + \varphi(\mathrm{D}) + \mathrm{D}\,\varphi(\mathrm{D})\,(d\tau)^2/2 + \dots.$

We get the following lemma.

Lemma 4 Ignoring differential terms of third and higher orders,

$$\begin{aligned} ER(d\tau, d\alpha) &= p_1\phi(-D_1) + p_0\phi(-D_0) + (\Delta/2)\, p_1\, \varphi(D_1)\, [(d\tau)^2 + (d\alpha)^2] \\ &= ER(0,0) + (\Delta/2)\, p_1\, \varphi(D_1)\, [(d\tau)^2 + (d\alpha)^2]. \qquad (4.44) \end{aligned}$$

Suppose that the boundary $B(d\tau, d\alpha)$ is given by those $\mathbf{X}$ satisfying

$$(\lambda + d\beta_0) + (\Delta e_1 + d\beta)'\, \mathbf{X} = 0 \qquad (4.45)$$

where $d\beta_0$ and $d\beta = (d\beta_1, d\beta_2, \ldots, d\beta_m)'$, indicating small discrepancies from the optimal linear function (4.39).

The expansion of $d\tau$ and $d\tau^2$, ignoring higher-order terms, are

$$\begin{aligned} d\tau &= (1/\Delta)\, (-d\beta_0 + (\lambda/\Delta)\, d\beta_1), \\ (d\tau)^2 &= (1/\Delta^2)\, (\, (d\beta_0)^2 - (2\lambda/\Delta)\, d\beta_0\, d\beta_1 + (\lambda/\Delta)^2\, (d\beta_1)^2\,). \qquad (4.46) \end{aligned}$$

Similarly, expansion of

$$d\alpha = \arctan\, [(\, (d\beta_2)^2 + \ldots + (d\beta_m)^2\,)^{1/2} \,/\, (\Delta + d\beta_1)]$$

and so $$(d\alpha)^2 = (\, (d\beta_2)^2 + (d\beta_3)^2 + \ldots + (d\beta_m)^2\,) \,/\, \Delta^2. \qquad (4.47)$$

Suppose that under some method of estimation, the $(m+1)$ vector of errors $(d\beta_0, d\beta)$ has a Limiting normal distribution with mean vector 0 and covariance matrix Σ/n,

$$\mathbf{L}: \sqrt{n} \begin{bmatrix} d\beta_0 \\ \\ d\beta \end{bmatrix} \rightarrow \eta_{m+1}\, (0, \Sigma). \qquad (4.48)$$

Hence the differential term

$$(d\tau)^2 + (d\alpha)^2 = (1/\Delta^2)\, [\, (d\beta_0)^2 - (2\lambda/\Delta)\, d\beta_0\, d\beta_1 + (\lambda/\Delta)^2\, (d\beta_1)^2 + (d\beta_2)^2 + \ldots + (d\beta_m)^2\,]$$

$$(4.49)$$

will have a limiting distribution of $1/n$ times the normal quadratic form

$$(1/\Delta)^2[\mathbf{Z}_0^2 - (2\lambda/\Delta)\,\mathbf{Z}_0\mathbf{Z}_1 + (\lambda/\Delta)^2\,\mathbf{Z}_1^2 + \mathbf{Z}_2^2 + \ldots + \mathbf{Z}_m^2],$$

Where $\mathbf{Z} \sim \eta_{m+1}(0, \Sigma)$.

As moments converge correctly for the logistic regression and normal discriminant procedures lemma gives a simple expression for the expected error rate in terms of the elements σ_{ij} of Σ.

Theorem 3 Ignoring terms of order less than 1/n,

$$E\{ER(d\tau, d\alpha) - ER(0, 0)\} = (p_1\,\varphi(D_1) / 2\Delta n)\,[\sigma_{00} - (2\lambda/\Delta)\,\sigma_{11} + \sigma_{22} + \ldots + \sigma_{mm}\,]. \quad (4.50)$$

The quantity $E\{ER(d\tau, d\alpha) - ER(0, 0)\}$ is a measure of our expected regret, in terms of increased error rate, when using some estimated discrimination procedure. Next, Σ for the logistic regression procedure and the normal discriminant procedure is being evaluated.

Asymptotic error rates of the two procedures

For the normal discriminant procedure described after (4.29), we have

Lemma 5 In the standard situation, the normal discriminant procedures estimates $(\hat{\beta}_0, \hat{\beta}') = (\lambda_0, e_1') + (d\hat{\beta}_0, d\hat{\beta}')$ satisfying

$$\mathbf{L}: \sqrt{n}\begin{bmatrix} d\hat{\beta}_0 \\ d\hat{\beta} \end{bmatrix} \rightarrow \eta_{m-1}(0, \hat{\Sigma}), \quad (4.51)$$

where

$$\hat{\Sigma} = (1/p_0 p_1)\begin{bmatrix} 1 + (\Delta^2/4) & (-\Delta/2)(p_0 - p_1) & 0 & \dots & 0 & \dots & 0 \\ (-\Delta/2)(p_0 - p_1) & 1 + 2\Delta^2 p_0 p_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & 0 & 1 + \Delta^2 p_0 p_1 & & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \dots & 0 & & 0 & 0 & 1 + \Delta^2 p_0 p_1 \end{bmatrix}$$

(4.52)

For the logistic regression estimates defined at (4.29), we have

<u>Lemma 6</u> In the standard situation, the logistic regression procedure produces estimates $(\bar{\beta}_0, \bar{\beta}') = (\lambda_0, e_1') + (d\bar{\beta}_0, d\bar{\beta}')$ satisfying

$$\mathbf{L}: \sqrt{n}\begin{bmatrix} d\bar{\beta}_0 \\ d\bar{\beta} \end{bmatrix} \rightarrow \eta_{m+1}(0, \bar{\Sigma}), \tag{4.53}$$

where

$$\bar{\Sigma} = (1/p_0 p_1)\begin{bmatrix} A_2/(A_0A_2 - A_1^2) & -A_1/(A_0A_2 - A_1^2) & 0 \dots & 0 \\ -A_1/(A_0A_2 - A_1^2) & A_0/(A_0A_2 - A_1^2) & 0 \dots & 0 \\ 0 & 0 & 1/A_0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1/A_0 \end{bmatrix}$$

(4.54)

$A_i = A_i(p_1, \Delta)$ is defined as

$$A_i(p_1, \Delta) = \int (e^{-\Delta^2/8}\, \mathbf{X}^i \varphi(\mathbf{X}) / p_1 e^{\Delta X/2} + p_0 e\text{-}^{\Delta X/2}) d\mathbf{X}, \quad i = 0, 1, 2, \dots \tag{4.55}$$

Denote the errors for the logistic regression procedure and the normal discrimination procedure by $(d\bar{\tau}, d\bar{\alpha})$ and $(d\hat{\tau}, d\hat{\alpha})$, respectively.

Define the efficiency measure,

$$\mathrm{EFF}_m(\lambda, \Delta) = \lim E\{\mathrm{ER}(d\hat{\tau}, d\hat{\alpha}) - \mathrm{ER}(0, 0)\} \ / \ E\{\mathrm{ER}(d\bar{\tau}, d\bar{\alpha}) - \mathrm{ER}(0, 0)\} \qquad (4.56)$$

Theorem 3, lemmas 5 and 6 then give

$$\mathrm{EFF}_m = (Q_1 + (m-1)\,Q_2) \ / \ (Q_3 + (m-1)\,Q_4), \qquad 4.57)$$

Where

$$Q_1 = (1, \lambda/\Delta) \begin{bmatrix} 1 + \Delta^2/4 & (p_0 - p_1)(\Delta/2) \\ (p_0 - p_1)(\Delta/2) & 1 + 2\,p_0 p_1\,\Delta^2 \end{bmatrix} \begin{bmatrix} 1 \\ \lambda/\Delta \end{bmatrix}$$

$$Q_2 = 1 + p_0 p_1\,\Delta^2 \qquad (4.58)$$

$$Q_3 = (1, \lambda/\Delta) \ \ (1 \,/\, (A_0 A_2 - A_1^2)) \begin{bmatrix} A_2 & A_1 \\ A_1 & A_0 \end{bmatrix}$$

$$Q_4 = 1 \,/\, A_0$$

The following theorem gives a simple expression for $\mathrm{EFF}_m(\lambda, \Delta)$ as a weighted average of the relative efficiencies when $m = 1$ and $m \to \infty$.

<u>Theorem 4</u> The relative efficiency of logistic regression to normal discrimination is

$$\mathrm{EFF}_m(\lambda, \Delta) = [q(\lambda, \Delta)\,\mathrm{EFF}_1(\lambda, \Delta) + (m-1)\,\mathrm{EFF}_\infty(\lambda, \Delta)] \ / \ [q(\lambda, \Delta) + (m-1)] \qquad (4.59)$$

Where $\mathrm{EFF}_\infty(\lambda, \Delta) = Q_2 / Q_4$ is by (4.56) the asymptotic efficiency as $m \to \infty$.

$\mathrm{EFF}_1(\lambda, \Delta) = Q_1 / Q_3$ follows from (4.56) and lemma 1 for $m = 1$. (Note: $d\alpha$ can always be taken equal to zero when $m = 1$).

For the case $\lambda = 0$ (then $A_1 = 0$) we have the following result.

Corollary When $\lambda = 0$, i.e. when $p_0 = p_1 = \frac{1}{2}$,

$$\mathrm{EFF}_m(\lambda, \Delta) = \mathrm{EFF}_\infty(\lambda, \Delta) = A_0\,(1 + \Delta^2/4), \tag{4.60}$$

for all values of m.

Note that when $\lambda \neq 0$ (i.e. $p_0, p_1 \neq \frac{1}{2}$), $\mathrm{EFF}_1(\lambda, \Delta) > \mathrm{EFF}_\infty(\lambda, \Delta)$, and q is near unity (see Table 1). Under these conditions, $\mathrm{EFF}_m(\lambda, \Delta)$ in (4.59) shows that it will be nearer $\mathrm{EFF}_\infty(\lambda, \Delta)$ than $\mathrm{EFF}_1(\lambda, \Delta)$, for $m \geq 3$.

$q=1,\ m=1 \qquad \mathrm{EFF}_m(\lambda, \Delta) = [\mathrm{EFF}_1(\lambda, \Delta) + (1-1)\,\mathrm{EFF}_\infty(\lambda, \Delta)] \;/\; [1 + (1-1)]$

$$= \mathrm{EFF}_1(\lambda, \Delta)$$

$q=1,\ m=2 \qquad \mathrm{EFF}_m(\lambda, \Delta) = [\mathrm{EFF}_1(\lambda, \Delta) + (2-1)\,\mathrm{EFF}_\infty(\lambda, \Delta)] \;/\; [1 + (2-1)]$

$$= \tfrac{1}{2}\,[\mathrm{EFF}_1(\lambda, \Delta) + \mathrm{EFF}_\infty(\lambda, \Delta)]$$

$q=1,\ m=3 \qquad \mathrm{EFF}_m(\lambda, \Delta) = [\mathrm{EFF}_1(\lambda, \Delta) + (3-1)\,\mathrm{EFF}_\infty(\lambda, \Delta)] \;/\; [1 + (3-1)]$

$$= (1/3)\,\mathrm{EFF}_1(\lambda, \Delta) + (2/3)\,\mathrm{EFF}_\infty(\lambda, \Delta)$$

$q=1,\ m=4 \qquad \mathrm{EFF}_m(\lambda, \Delta) = [\mathrm{EFF}_1(\lambda, \Delta) + (4-1)\,\mathrm{EFF}_\infty(\lambda, \Delta)] \;/\; [1 + (4-1)]$

$$= (\tfrac{1}{4})\,\mathrm{EFF}_1(\lambda, \Delta) + (\tfrac{3}{4})\,\mathrm{EFF}_\infty(\lambda, \Delta)$$

Angle and intercept error

$\mathrm{EFF}_\infty(\lambda, \Delta)$ could also be interpreted as the asymptotic relative efficiency of logistic regression to normal discrimination for estimating the angle of the discriminant boundary,

$$\mathrm{EFF}_\infty(\lambda, \Delta) = \lim_{n\to\infty} \mathrm{Var}\,(d\hat{\alpha}) \;/\; \mathrm{Var}\,(d\bar{\alpha}) \tag{4.61}$$

Likewise, $EFF_I(\lambda, \Delta)$ is the asymptotic relative efficiency for estimating the intercept of the discriminant boundary,

$$EFF_I(\lambda, \Delta) = \lim_{n \to \infty} \mathrm{Var}\,(d\hat{\tau}) \;/\; \mathrm{Var}\,(d\bar{\tau})$$

These results follow from (4.46), (4.47), (4.52) and (4.54). A comparison of (4.47) and lemmas 5 and 6 shows that

$$\begin{aligned} &L\!: \; n\,(d\hat{\tau})^2 \rightarrow (1 \,/\, p_0 p_1 \Delta^2)\; (1 + p_0 p_1 \Delta^2)\; \chi^2_{m-1}, \\ &L\!: \; n\,(d\bar{\alpha})^2 \rightarrow (1 \,/\, p_0 p_1 \Delta^2)\; (1 \,/\, A_0)\; \chi^2_{m-1}. \end{aligned} \qquad (4.62)$$

In terms of the angular error, the asymptotic relative efficiency of logistic regression to normal discrimination is

$$\begin{aligned} ARE &= (1 + p_0 p_1 \Delta^2)\; A_0 \\ &= [(1 + p_0 p_1 \Delta^2) \,/\, (2\pi)^{1/2}]\; e^{-\Delta^2/8} \int_{-\infty}^{\infty} e^{-X^2/2} \,/\, (p_1 e^{\Delta X/2} + p_0 e^{-\Delta X/2})\, d\mathbf{X} \end{aligned}$$

Hence, a sample of size $\bar{n}$ using logistic regression produces asymptotically the same angular error distribution as a sample of size $\hat{n}$ = ARE x $\bar{n}$, using normal discrimination. For example, for $\lambda = 0$ (i.e. $p_0 = p_1 = ½$) and $\Delta = 2.5$, $\bar{n} = 1000$ is approximately equivalent to $\hat{n} = 0.786$ (see (1.12), Efron 1975).

The above statement is not valid for intercept error because the two matrices involved in the definition of Q_1 and Q_3, are not proportional. However, $\lambda = 0$, i.e. when $p_0 = p_1 = ½$, (4.46) and lemmas 5 and 6 show that

$$\begin{aligned} &L\!: \; n\,(d\hat{\tau})^2 \rightarrow (4 \,/\, \Delta^2)\; (1 + \Delta^2/4)\; \chi^2_1, \\ &L\!: \; n\,(d\bar{\alpha})^2 \rightarrow (4 \,/\, \Delta^2)\; (1 \,/\, A_0)\; \chi^2_1. \end{aligned} \qquad (4.63)$$

In this case, the ARE by (4.62) again gives asymptotically equivalent sample sizes.

When $\lambda = 0$, then $\tau = 0$ and so $D_1 = D_0 = \Delta/2$. Now combining (4.62) and (4.63) with lemma 4 we get

$$\begin{aligned} &L\!: \; n\,\{ER(d\hat{\tau}, d\hat{\alpha}) - ER(0, 0)\} \rightarrow (\,\varphi(\Delta/2) / \Delta)\,(1 + \Delta^2/4)\; \chi^2_m, \\ &L\!: \; n\,\{ER(d\bar{\tau}, d\bar{\alpha}) - ER(0, 0)\} \rightarrow (\,\varphi(\Delta/2) / \Delta)\,(1 / A_0)\; \chi^2_m. \end{aligned} \tag{4.64}$$

So for $p_0 = p_1 = ½$, error rate for samples of size $\hat{n} = ARE \times \bar{n}$ and n will have asymptotically equivalent distributions. This statement is not true for $p_0, p_1 \neq ½$, however it becomes true as the dimension m gets large. Then error rates for the two procedures will have the same asymptotic distribution if $\hat{n} = ARE \times \bar{n}$, when $m \rightarrow \infty$ and $\bar{n} / m \rightarrow \infty$.

This result follows from (4.49) and lemmas 5 and 6.

Distorted sampling proportions

In some situations, the probabilities p_0 and p_1 may be distorted due to the sampling scheme employed. Let $\tilde{p}_0$ and $\tilde{p}_1$ be the distorted values, then $\lambda = \log(p_1 / p_0)$,
$\tilde{\lambda} = \log(\tilde{p}_1 / \tilde{p}_0)$and for some known constant $\lambda = \tilde{\lambda} + C$. (4.65)
For example, due to some experimental constraints, the statistician might have to randomly exclude from his training set nine out of ten members of population 0. In this case, $C = \log 10$. Then the normal discrimination procedure assigns a new **X** to population 1 or 0 as $\hat{\lambda}$ (**X**) is greater or less than C. The logistic regression procedure is modified similarly.

The relative efficiency of logistic regression to normal discrimination (Theorem 4) remains true. Only, λ is replaced by $\tilde{\lambda}$ for the vector $[1, \lambda/\Delta]'$ and its transpose, which appear in the definition of $Q_1$and Q_3. With a choice of $C \neq 0$, the intercept is changed from ($-\lambda$ / C) to $-(\lambda + C)/\Delta$. The effect of this change is to reduce EFF$_1(\lambda, \Delta)$, as shown in the following tabulation.

Δ=2, p_1 = 0.5					Δ=3, p_1 = 0.5			
C	0	±1	±2	±3	O	±1	±2	±3
EFF$_1$	.899	.869	.836	.819	.641	.604	.550	.516

The angular efficiency. EFF$_\infty(\lambda, \Delta)$, remains unchanged for any choice of C since the corresponding discrimination boundary is parallel to that for $C = 0$.

References

Anderson, T.W. (1984), "An Introduction to Multivariate Statistical Methods," New York: John Wiley.

Anderson, J.A. (1972), "Separate Sample Logistic Discrimination," *Biometrika.*

Cook, R.D. (1977), "Detection of Influential Observation in Linear Regression," *Technometrics,* **19**, 15-18.

Cox, D.R., and Snell, E. J. (1989), "Analysis of Binary Data," (2nd Edition), London: Chapman and Hall.

Critchley, F., and Vitiello, C. (1991), "The Influence of Observations on Misclassification Probability Estimates in Linear Discriminant Analysis," *Biometrika,* **78**, 677-690.

Efron, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of American Statistical Association,* **70**, 892-898.

Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics,* **7**, 179-188.

Fung, W.K. (1992), "Some Diagnostic Measures in Discriminant Analysis," *Statistics and Probability Letters,* **13**, 279-285.
(1993), "Unmasking Outliers and Leverage Points: A Confirmation," *Journal of The American Statistical Association,* **88**, 515-519.
(1995), "Diagnostics in Linear Discriminant analysis," *Journal of The American Statistical Association,* **90**, 952-956.

Gordon,T. (1974), "Hazards in the Use of The Logistic Function with Special Reference to Data from Prospective Cardiovascular Studies," *Journal of Chronic Diseases,* **27**, 97-102.

Halperin, M., Blackwelder, W.C., and Verter, J.I. (1971), "Estimation of The Multivariate Logistic Risk Function: A Comparison of The Discriminant Function and Maximum Likelihood Approaches," *Journal of Chronic Diseases,* **24**, 125-128.

Johnson, R.A., and Wichern, D.W. (1988), "Applied Multivariate Statistical Analysis," (3rd Edition).

Krzanowski, W.J. (1977), "The Performance of Fisher's Linear Discriminant Function Under Non-Optimal Conditions," *Technometrics,* **19**, 191-200.

Lachenbruch, P.A. (1975), "Discriminant Analysis," New York: Hafner Press.

Lachenbruch, P.A. and Mickey, M.R. (1968), "Estimation of Error Rates in Discriminant Analysis," *Technometrics*, **10**, 1-11.

McFadden, D. (1976), "A Comment on Discriminant Analysis 'Versus' Logit Analysis," *Annals of Economic and Social Measurement*, **5**, 511-523.

McLachlan, G.J. (1992), "Discriminant Analysis and Statistical Pattern Recognition," New York: John Wiley.

Nerlove, M., and Press, S.J. (1973), "Univariate and Multivariate Log-Linear and Logistic Models," R-1306, Santa Monica, Calif.: The Rand Corporation.

O'Neill, T.J. (1980), "The General Distribution of The Error Rate of A Classification Procedure with Application to Logistic Regression Discrimination," *Journal of The American Statistical Association*, **75**, 154-160.

Press, S.J., and Wilson, S. (1978), "Choosing Between Logistic Regression and Discriminant Analysis," *Journal of The American Statistical Association*, **73**, 699-705.

Rao, C.R. (1965), "Linear Statistical Inference and Its Applications," New York: John Wiley & Sons.

Rousseeuw, P.J., and Van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," (with discussion), *Journal of The American Statistical Association*, **85**. 633-651.

Ruiz, S. (1989), "Relation Between Logistic Regression and Discriminant Analysis," Unpublished Chapter of *Imperial College Thesis.*

Truett, J., Cornfiels, J., and Kannel, W. (1967), "A Multivariate Analysis of The Risk of Coronary Heart Disease in Framingham," *Journal of Chronic Diseases*, **20**, 511-524.

Wald, A. (1944), "On A Statistical Problem Arising in The Classification of An Individual into One of Two Groups," *Annals of Mathematical Statistics*, **15**, 145-162.

Welsh, R.E. (1982), "Influence Functions and Regression Diagnostics," in Modern Data Analysis, eds. R.L. Launer and F.A. Siegel, New York: Academic Press.