Discriminant Analysis and Statistical Pattern Recognition

This Page Intentionally Left Blank

Discriminant Analysis and Statistical Pattern Recognition

This Page Intentionally Left Blank

Discriminant Analysis and Statistical Pattern Recognition

GEOFFRY J. McLACHLAN
The University of Queensland

A JOHN WILEY & SONS, INC., PUBLICATION

A NOTE TO THE READER
This book has been electronically reproduced from digital information stored at John Wiley & Sons, Inc. We are pleased that the use of this new technology will enable us to keep works of enduring scholarly value in print as long as there is a reasonable demand for them. The content of this book is identical to previous printings.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data is available.

ISBN 0-471-69115-1

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To
Beryl, Jonathan, and Robbie

This Page Intentionally Left Blank

Contents

This Page Intentionally Left Blank

Preface

Over the years a not inconsiderable body of literature has accumulated on discriminant analysis, with its usefulness demonstrated over many diverse fields, including the physical, biological and social sciences, engineering, and medicine. The purpose of this book is to provide a modern, comprehensive, and systematic account of discriminant analysis, with the focus on the more recent advances in the field. Discriminant analysis or (statistical) discrimination is used here to include problems associated with the statistical separation between distinct classes or groups and with the allocation of entities to groups (finite in number), where the existence of the groups is known *a priori* and where typically there are feature data on entities of known origin available from the underlying groups. It thus includes a wide range of problems in statistical pattern recognition, where a pattern is considered as a single entity and is represented by a finite dimensional vector of features of the pattern.

In recent times, there have been many new advances made in discriminant analysis. Most of them, for example those based on the powerful but computer-intensive bootstrap methodology, are now computationally feasible with the relatively easy access to high-speed computers. The new advances are reported against the background of the extensive literature already existing in the field. Both theoretical and practical issues are addressed in some depth, although the overall presentation is biased toward practical considerations.

Some of the new advances that are highlighted are regularized discriminant analysis and bootstrap-based assessment of the performance of a sample-based discriminant rule. In the exposition of regularized discriminant analysis, it is noted how some of the sample-based discriminant rules that have been proposed over the years may be viewed as regularized versions of the normal-based quadratic discriminant rule. Recently, there has been proposed a more sophisticated regularized version, known as regularized discriminant analysis. This approach, which is a sample-based compromise between normal-based linear and quadratic discriminant analyses, is considered in some detail, given the highly encouraging results that have been reported for its performance in such difficult circumstances, as when the group-sample sizes are small relative to the number of feature variables. On the role of the bootstrap in estimation

problems in discriminant analysis, particular attention is given to its usefulness in providing improved nonparametric estimates of the error rates of sample-based discriminant rules in their applications to unclassified entities.

With the computer revolution, data are increasingly being collected in the form of images, as in remote sensing. As part of the heavy emphasis on recent advances in the literature, an account is provided of extensions of discriminant analysis motivated by problems in statistical image analysis.

The book is a monograph, not a textbook. It should appeal to both applied and theoretical statisticians, as well as to investigators working in the many diverse areas in which relevant use can be made of discriminant techniques. It is assumed that the reader has a fair mathematical or statistical background.

The book can be used as a source of reference on work of either a practical or theoretical nature on discriminant analysis and statistical pattern recognition. To this end, an attempt has been made to provide a broad coverage of the results in these fields. Over 1200 references are given.

Concerning the coverage of the individual chapters, Chapter 1 provides a general introduction of discriminant analysis. In Chapter 2, likelihood-based approaches to discrimination are considered in a general context. This chapter also provides an account of the use of the EM algorithm in those situations where maximum likelihood estimation of the group-conditional distributions is to be carried out using unclassified feature data in conjunction with the training feature data of known group origin.

As with other multivariate statistical techniques, the assumption of multivariate normality provides a convenient way of specifying a parametric group structure. Chapter 3 concentrates on discrimination via normal theory-based models. In the latter part of this chapter, consideration is given also to reducing the dimension of the feature vector by appropriate linear projections. This process is referred to in the pattern recognition literature as linear feature selection. Chapter 4 reports available distributional results for normal-based discriminant rules. Readers interested primarily in practical applications of discriminant analysis may wish to proceed directly to Chapter 5, which discusses practical aspects and variants of normal-based discriminant rules. The aforementioned approach of regularized discriminant analysis is emphasized there.

Chapter 6 is concerned primarily with data analytic considerations with normal-based discriminant analysis. With a parametric formulation of problems in discriminant analysis, there is a number of preliminary items to be addressed. They include the detection of apparent outliers among the training sample, the question of model fit for the group-conditional distributions, the use of data-based transformations to achieve approximate normality, the assessment of typicality of the feature vector on an unclassified entity to be allocated to one of the specifed groups, and low-dimensional graphical representations of the feature data for highlighting and/or revealing the underlying group structure. Chapter 7 is devoted to parametric discrimination via non-normal models for feature variables that are either all discrete, all continuous,

or that are mixed in that they consist of both types of variables. A semiparametric approach is adopted in Chapter 8 with a study of the widely used logistic model for discrimination. Nonparametric approaches to discrimination are presented in Chapter 9. Particular attention in this chapter is given to kernel discriminant analysis, where the nonparametric kernel method is used to estimate the group-conditional densities in the formation of the posterior probabilities of group membership and the consequent discriminant rule.

Chapter 10 is devoted fully to the important but difficult problem of assessing the various error rates of a sample-based discriminant rule on the basis of the same data used in its construction. The error rates are useful in summarizing the global performance of a discriminant rule. Of course, for a specific case as, for example, in medical diagnosis, it is more appropriate to concentrate on the estimation of the posterior probabilities of group membership. Accordingly, a separate chapter (Chapter 11) is devoted to this problem.

Chapter 12 is on the selection of suitable feature variables using a variety of criteria. This is a fundamental problem in discriminant analysis, as there are many practical and theoretical reasons for not using all of the available feature variables. Finally, Chapter 13 is devoted to the statistical analysis of image data. Here the focus is on how to form contextual allocation rules that offer improved performance over the classical noncontextual rules, which ignore the spatial dependence between neighboring images.

Thanks are due to the authors and owners of copyrighted material for permission to reproduce published tables and figures. The author also wishes to thank Tuyet-Trinh Do for her assistance with the preparation of the typescript.

GEOFFREY J. MCLACHLAN

Brisbane, Queensland
January, 1991

This Page Intentionally Left Blank

CHAPTER 1

General Introduction

1.1 INTRODUCTION

Discriminant analysis as a whole is concerned with the relationship between a categorical variable and a set of interrelated variables. More precisely, suppose there is a finite number, say, g, of distinct populations, categories, classes, or groups, which we shall denote here by $G_1,\ldots,G_g$. We will henceforth refer to the G_i as groups. Note that in discriminant analysis, the existence of the groups is known *a priori*. An entity of interest is assumed to belong to one (and only one) of the groups. We let the categorical variable z denote the group membership of the entity, where $z = i$ implies that it belongs to group G_i $(i = 1,\ldots,g)$. Also, we let the p-dimensional vector $\mathbf{x} = (x_1,\ldots,x_p)'$ contain the measurements on p available features of the entity.

In this framework, the topic of discriminant analysis is concerned with the relationship between the group-membership label z and the feature vector $\mathbf{x}$. Within this broad topic there is a spectrum of problems, which corresponds to the inference-decision spectrum in statistical methodology. At the decision end of the scale, the group membership of the entity is unknown and the intent is to make an outright assignment of the entity to one of the g possible groups on the basis of its associated measurements. That is, in terms of our present notation, the problem is to estimate z solely on the basis of $\mathbf{x}$. In this situation, the general framework of decision theory can be invoked. An example in which an outright assignment is required concerns the selection of students for a special course, where the final decision to admit students is based on their answers to a questionnaire. For this decision problem, there are two groups with, say, G_1, referring to students who complete the course successfully, and G_2 to those who do not. The feature vector $\mathbf{x}$ for a student contains his/her answers to the questionnaire. A rule based on $\mathbf{x}$ for allocating

a student to either G_1 or G_2 (that is, either accepting or rejecting the student into the course) can be formed from an analysis of the feature vectors of past students from each of the two groups. The construction of suitable allocation rules is to be pursued in the subsequent sections of this chapter.

At the other extreme end of the spectrum, no assignment or allocation of the entity to one of the possible groups is intended. Rather the problem is to draw inferences about the relationship between z and the feature variables in $\mathbf{x}$. An experiment might be designed with the specific aim to provide insight and understanding into the predictive structure of the feature variables. For example, a political scientist may wish to determine the socio-economic factors that have the most influence on the voting patterns of a population of voters.

Between these extremes lie most of the everyday situations in which discriminant analysis is applied. Typically, the problem is to make a prediction or tentative allocation for an unclassified entity. For example, concerning prediction, an economist may wish to forecast on the basis of his or her most recent accounting information, those members of the corporate sector that might be expected to suffer financial losses leading to failure. For this purpose, a discriminant rule may be formed from accounting data collected on failed and surviving companies over many past years. An example where allocation, tentative or otherwise, is required is with the discrimination between an earthquake and an underground nuclear explosion on the basis of signals recorded at a seismological station (Elvers, 1977). An allocation rule is formed from signals recorded on past seismic events of known classification.

Examples where prediction or tentative allocation is to be made for an unclassified entity occur frequently in medical prognosis and diagnosis. A source for applications of discriminant analysis to medical diagnosis is the bibliography of Wagner, Tautu and Wolbler (1978) on problems in medical diagnosis. In medical diagnosis, the definitive classification of a patient often can be made only after exhaustive physical and clinical assessments or perhaps even surgery. In some instances, the true classification can be made only on evidence that emerges after the passage of time, for instance, an autopsy. Hence, frequent use is made of diagnostic tests. Where possible, the tests are based on clinical and laboratory-type observations that can be made without too much inconvenience to the patient. The financial cost of the test is also sometimes another consideration, particularly in mass screening programs. Suppose that the feature vector $\mathbf{x}$ contains the observations taken on a patient for his or her diagnosis with respect to one of g possible disease groups $G_1, \ldots, G_g$. Then the relative plausibilities of these groups for a patient with feature vector $\mathbf{x}$ as provided by a discriminant analysis may be of assistance to the clinician in making a diagnosis. This is particularly so with the diagnosis of Conn's syndrome in patients with high blood pressure, as reported in Aitchison and Dunsmore (1975, Chapter 1). The two possible groups represent the cause, which is either a benign tumor in one adrenal gland, curable by surgical removal (G_1), or a more diffuse condition affecting both adrenal glands, with the possibility of control of blood pressure by drugs (G_2). The actual cause

can be confirmed only by microscopic examination of adrenal tissue removed at an operation. However, because surgery is inadvisable for patients in G_2, a clinician is faced with a difficult treatment decision. Thus, a realistic preoperative assessment that a patient with a particular feature vector belongs to either G_1 or G_2 would be most valuable to the clinician. The available feature variables on a patient relate to age, plasma concentrations of sodium, potassium, bicarbonate, renin, and aldosterone, and systolic and diastolic blood pressures.

The relative plausibilities of group membership for a patient with an associated feature vector are also useful in medical prognosis. Here the vector is measured after the onset of some medical condition, say, an injury, and the groups represent the possible outcomes of the injury. There are several reasons why an initial prediction of the eventual outcome of the injury may be needed. For instance, in situations where the management of the patient is closely linked to the outcome, it provides a guide to the clinician as to whether his or her particular course of management is appropriate. It also provides a firmer basis for advice to relatives of an injured patient on the chances of recovery. These and other reasons are discussed by Titterington et al. (1981) and Murray et al. (1986) in the context of the prognosis for patients with severe head injuries. For these patients, the three possible outcomes were dead or vegetative, severe disability, and moderate or good recovery. The feature vector for a patient included background information such as age and cause of injury and four clinical variables (eye opening, motor response, motor response pattern, and pupil reaction).

Situations in medical diagnosis where outright rather than tentative allocations to groups are made occur in mass screening programs. Suppose that in the detection of a disease, G_1 consists of those individuals without the disease and G_2 of those with the disease. Then in a screening program for this disease, a patient is assigned outright to either G_1 or G_2, according to whether the diagnostic test is negative or positive. Usually, with a positive result, further testing is done before a final assignment is made. For example, with the enzyme-linked immunosorbent assay (ELISA) test used to screen donated blood for antibodies to the AIDS virus, a positive test would result in a more definitive test such as the Western blot being performed (Gastwirth, 1987). J. A. Anderson (1982) has given an example on patient care where an irrevocable outright assignment has to be made. It concerns the decision on whether to administer a preoperative anticoagulant therapy to a patient to reduce the risk of postoperative deep vein thrombosis.

Discriminant analysis is widely used also in the field of pattern recognition, which is concerned mainly with images. The aim of pattern recognition is to automate processes performed by humans. For example, automatic analysis and recognition of photomicrographs of tissue cells can be used in blood tests, cancer tests, and brain-tissue studies. Another example of much current interest concerns the automatic recognition of images remotely sensed from earth satellites. It is considered in Chapter 13.

The branch of pattern recognition known as statistical pattern recognition has close ties with statistical decision theory and areas of multivariate analysis, in particular discriminant analysis. In statistical pattern recognition, each pattern is considered as a single entity and is represented by a finite dimensional vector of features of the pattern. Hence, the recognition of patterns with respect to a finite number of predefined groups of patterns can be formulated within the framework of discriminant analysis. The number of features required for recognition of a pattern may become very large if the patterns under study are very complex or if, as in fingerprint identification, the number of possible pattern groups is very large. Consequently, the above approach may have to be modified; see, for example, Fu (1986) and Mantas (1987).

By now, there is an enormous literature on discriminant analysis, and so it is not possible to provide an exhaustive bibliography here. However, we have attempted to cover the main results, in particular the more recent developments. Additional references on the earlier work may be found in the books devoted to the topic as a whole or in part by Lachenbruch (1975a), Goldstein and Dillon (1978), Klecka (1980), and Hand (1981a, 1982). They have been supplemented recently by the volume edited by S. C. Choi (1986), the notes of Hjort (1986a), and the report by a panel of the Committee on Applied and Theoretical Statistics of the Board on Mathematical Sciences of the National Research Council, chaired by Professor R. Gnanadesikan (Panel on Discriminant Analysis, Classification and Clustering, 1989). Further references may be found in the symposium proceedings edited by Cacoullos (1973) and Van Ryzin (1977), the review article by Lachenbruch and Goldstein (1979), and in the encyclopedia entry by Lachenbruch (1982). There are also the relevant chapters in the rapidly growing list of textbooks on multivariate analysis. Another source of references is the pattern recognition literature. Fukunaga (1972, 1990), Patrick (1972), Duda and Hart (1973), Young and Calvert (1974), and Devijver and Kittler (1982) are examples of texts on statistical pattern recognition. A single source of references in discriminant and cluster analyses and in pattern recognition is the book edited by Krishnaiah and Kanal (1982).

1.2 BASIC NOTATION

We let $\mathbf{X}$ denote the p-dimensional random feature vector corresponding to the realization $\mathbf{x}$ as measured on the entity under consideration. The associated variable z denoting the group of origin of the entity is henceforth replaced by a g-dimensional vector $\mathbf{z}$ of zero-one indicator variables. The ith component of $\mathbf{z}$ is defined to be one or zero according as $\mathbf{x}$ (really the entity) belongs or does not belong to the ith group G_i ($i = 1,\ldots,g$); that is,

$$\begin{aligned} z_i &= 1, \qquad \mathbf{x} \in G_i, \\ &= 0, \qquad \mathbf{x} \notin G_i, \end{aligned}$$

for $i = 1,\ldots,g$. Where possible, random variables are distinguished from their realizations by the use of the corresponding uppercase letters.

The probability density function (p.d.f.) of $\mathbf{X}$ in group G_i is denoted by $f_i(\mathbf{x})$ for $i = 1,\ldots,g$. These group-conditional densities are with respect to arbitrary measure on R^p, so that $f_i(\mathbf{x})$ can be a mass function by the adoption of counting measure. Under the mixture model approach to discriminant analysis, it is assumed that the entity has been drawn from a mixture G of the g groups $G_1,\ldots,G_g$ in proportions $\pi_1,\ldots,\pi_g$, respectively, where

$$\sum_{i=1}^{g} \pi_i = 1 \qquad \text{and} \qquad \pi_i \geq 0 \qquad (i = 1,\ldots,g).$$

The p.d.f. of $\mathbf{X}$ in G can therefore be represented in the finite mixture form

$$f_X(\mathbf{x}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}). \tag{1.2.1}$$

An equivalent assumption is that the random vector $\mathbf{Z}$ of zero-one group indicator variables with $\mathbf{z}$ as its realization is distributed according to a multinomial distribution consisting of one draw on g categories with probabilities $\pi_1,\ldots,\pi_g$, respectively; that is,

$$\text{pr}\{\mathbf{Z} = \mathbf{z}\} = \pi_1^{z_1}\pi_2^{z_2}\cdots\pi_g^{z_g}. \tag{1.2.2}$$

We write

$$\mathbf{Z} \sim \text{Mult}_g(1,\boldsymbol{\pi}), \tag{1.2.3}$$

where $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)'$. Note that with a deterministic approach to the problem, $\mathbf{z}$ is taken to be a parameter rather than a random variable as here. The distribution function of $\mathbf{Y} = (\mathbf{X}',\mathbf{Z}')'$ is denoted by $F(\mathbf{y})$, where the prime denotes vector transpose. We let $F_i(\mathbf{x})$ and $F_X(\mathbf{x})$ denote the distribution functions corresponding to the densities $f_i(\mathbf{x})$ and $f_X(\mathbf{x})$, respectively.

The ith mixing proportion π_i can be viewed as the prior probability that the entity belongs to G_i $(i = 1,\ldots,g)$. With $\mathbf{X}$ having been observed as $\mathbf{x}$, the posterior probability that the entity belongs to G_i is given by

$$\begin{aligned} \tau_i(\mathbf{x}) &= \text{pr}\{\text{entity} \in G_i \mid \mathbf{x}\} \\ &= \text{pr}\{Z_i = 1 \mid \mathbf{x}\} \\ &= \pi_i f_i(\mathbf{x})/f_X(\mathbf{x}) \qquad (i = 1,\ldots,g). \end{aligned} \tag{1.2.4}$$

In the next section, we consider the formation of an optimal rule of allocation in terms of these posterior probabilities of group membership $\tau_i(\mathbf{x})$.

The term "classification" is used broadly in the literature on discriminant and cluster analyses. To avoid any possible confusion, throughout this monograph, we reserve the use of classification to describe the original definition of the underlying groups. Hence, by a classified entity, we mean an entity whose group of origin is known. A rule for the assignment of an unclassified entity

to one of the groups will be referred to as a discriminant or allocation rule. In the situation where the intention is limited to making an outright assignment of the entity to one of the possible groups, it is perhaps more appropriate to use the term allocation rather than discriminant to describe the rule. However, we will use either nomenclature regardless of the underlying situation. In the pattern recognition jargon, such a rule is referred to as a classifier.

1.3 ALLOCATION RULES

At this preliminary stage of formulating discriminant analysis, we consider the pure decision case, where the intent is to make an outright assignment of an entity with feature vector $\mathbf{x}$ to one of the g possible groups. Let $r(\mathbf{x})$ denote an allocation rule formed for this purpose, where $r(\mathbf{x}) = i$ implies that an entity with feature vector $\mathbf{x}$ is to be assigned to the ith group G_i ($i = 1,\ldots,g$). In effect, the rule divides the feature space into g mutually exclusive and exhaustive regions $R_1,\ldots,R_g$, where, if $\mathbf{x}$ falls in R_i, then the entity is allocated to group G_i ($i = 1,\ldots,g$).

The allocation rates associated with this rule $r(\mathbf{x})$ are denoted by $e_{ij}(r)$, where

$$e_{ij}(r) = \text{pr}\{r(\mathbf{X}) = j \mid \mathbf{X} \in G_i\}$$

is the probability that a randomly chosen entity from G_i is allocated to G_j ($i, j = 1,\ldots,g$). It can be expressed as

$$e_{ij}(r) = \int_{R_j} f_i(\mathbf{x})\, d\nu,$$

where ν denotes the underlying measure on $\mathbf{R}^p$ appropriate for $f_X(\mathbf{x})$. The probability that a randomly chosen member of G_i is misallocated can be expressed as

$$\begin{aligned} e_i(r) &= \sum_{j \neq i}^{g} e_{ij}(r) \\ &= \int_{\overline{R}_i} f_i(\mathbf{x})\, d\nu, \end{aligned}$$

where $\overline{R}_i$ denotes the complement of R_i ($i = 1,\ldots,g$).

For a diagnostic test using the rule $r(\mathbf{x})$ in the context where G_1 denotes the absence of a disease and G_2 its presence, the error rate $e_{12}(r)$ corresponds to the probability of a false positive, and $e_{21}(r)$ is the probability of a false negative. The correct allocation rates

$$e_{22}(r) = 1 - e_{21}(r) \qquad \text{and} \qquad e_{11}(r) = 1 - e_{12}(r)$$

are known as the sensitivity and specificity, respectively, of the diagnostic test.

1.4 DECISION-THEORETIC APPROACH

Decision theory provides a convenient framework for the construction of discriminant rules in the situation where an outright allocation of an unclassified entity is required. The present situation where the prior probabilities of the groups and the group-conditional densities are taken to be known is relatively straightforward.

Let c_{ij} denote the cost of allocation when an entity from G_i is allocated to group G_j, where $c_{ij} = 0$ for $i = j = 1,\ldots,g$; that is, there is zero cost for a correct allocation. We assume for the present that the costs of misallocation are all the same. We can then take the common value of the c_{ij} $(i \neq j)$ to be unity, because it is only their ratios that are important.

For given $\mathbf{x}$, the loss for allocation performed on the basis of the rule $r(\mathbf{x})$ is

$$l\{\mathbf{z}, r(\mathbf{x})\} = \sum_{i=1}^{g} z_i Q[i, r(\mathbf{x})], \tag{1.4.1}$$

where, for any u and v, $Q[u,v] = 0$ for $u = v$ and 1 for $u \neq v$. The expected loss or risk, conditional on $\mathbf{x}$, is given by

$$E[l\{\mathbf{Z}, r(\mathbf{x})\} \mid \mathbf{x}] = \sum_{i=1}^{g} \tau_i(\mathbf{x}) Q[i, r(\mathbf{x})], \tag{1.4.2}$$

since from (1.2.4),

$$E(Z_i \mid \mathbf{x}) = \tau_i(\mathbf{x}).$$

An optimal rule of allocation can be defined by taking it to be the one that minimizes the conditional risk (1.4.2) at each value $\mathbf{x}$ of the feature vector. In decision-theory language, any rule that so minimizes (1.4.2) for some $\pi_1,\ldots,\pi_g$ is said to be a Bayes rule. It can be seen from (1.4.2) that the conditional risk is a linear combination of the posterior probabilities, where all coefficients are zero except for one, which is unity. Hence, it is minimized by taking $r(\mathbf{x})$ to be the label of the group to which the entity has the highest posterior probability of belonging. Note that this is the "intuitive solution" to the allocation problem.

If we let $r_o(\mathbf{x})$ denote this optimal rule of allocation, then

$$r_o(\mathbf{x}) = i \quad \text{if} \quad \tau_i(\mathbf{x}) \geq \tau_j(\mathbf{x}) \qquad (j = 1,\ldots,g;\ j \neq i). \tag{1.4.3}$$

The rule $r_o(\mathbf{x})$ is not uniquely defined at $\mathbf{x}$ if the maximum of the posterior probabilities of group membership is achieved with respect to more than one group. In this case the entity can be assigned arbitrarily to one of the groups for which the corresponding posterior probabilities are equal to the maximum value. If

$$\text{pr}\{\tau_i(\mathbf{X}) = \tau_j(\mathbf{X})\} = 0 \qquad (i \neq j = 1,\ldots,g),$$

then the optimal rule is unique for almost all $\mathbf{x}$ relative to the underlying measure ν on $\mathbf{R}^p$ appropriate for $f_X(\mathbf{x})$.

As the posterior probabilities of group membership $\tau_i(\mathbf{x})$ have the same common denominator $f_X(\mathbf{x})$, $r_o(\mathbf{x})$ can be defined in terms of the relative sizes of the group-conditional densities weighted according to the group-prior probabilities; that is,

$$r_o(\mathbf{x}) = i \quad \text{if} \quad \pi_i f_i(\mathbf{x}) \geq \pi_j f_j(\mathbf{x}) \qquad (j = 1,\ldots,g;\ j \neq i). \tag{1.4.4}$$

Note that as the optimal or Bayes rule of allocation minimizes the conditional risk (1.4.2) over all rules r, it also minimizes the unconditional risk

$$\begin{aligned} e(r) &= \sum_{i=1}^{g} E\{\tau_i(\mathbf{X})Q[i, r(\mathbf{X})]\} \\ &= \sum_{i=1}^{g} \pi_i \int_{R_i} f_i(\mathbf{x})\, d\nu \\ &= \sum_{i=1}^{g} \pi_i e_i(r), \end{aligned}$$

which is the overall error rate associated with r.

Discriminant analysis in its modern guise was founded by Fisher (1936). His pioneering paper, which did not take the group-conditional distributions to be known, is to be discussed later in this chapter in the context of sample-based allocation rules. Concerning initial work in the case of known group-conditional distributions, Welch (1939) showed for $g = 2$ groups that a rule of the form (1.4.4) is deducible either from Bayes theorem if prior probabilities are specified for the groups or by the use of the Neyman–Pearson lemma if the two group-specific errors of allocation are to be minimized in any given ratio. Wald (1939, 1949) developed a general theory of decision functions, and von Mises (1945) obtained the solution to the problem of minimizing the maximum of the errors of allocation for a finite number of groups, which was in the general theme of Wald's work. Rao (1948) discussed explicit solutions of the form (1.4.4) and also the use of a doubtful region of allocation. In a subsequent series of papers, he pursued related problems and extensions; see Rao (1952, 1954) for an account. There is an extensive literature on the development of allocation rules. The reader is referred to Das Gupta (1973) for a comprehensive review.

Up to now, we have taken the costs of misallocation to be the same. For unequal costs of misallocation c_{ij}, the conditional risk of the rule $r(\mathbf{x})$ is

$$\sum_{i \neq r(\mathbf{x})}^{g} \tau_i(\mathbf{x}) c_{i,r(\mathbf{x})}. \tag{1.4.5}$$

Let $r_o(\mathbf{x})$ be the optimal or Bayes rule that minimizes (1.4.5). Then it follows that $r_o(\mathbf{x}) = i$ if

$$\sum_{h \neq i}^{g} \tau_h(\mathbf{x}) c_{hi} \leq \sum_{h \neq j}^{g} \tau_h(\mathbf{x}) c_{hj} \qquad (j = 1,\ldots,g;\ j \neq i). \tag{1.4.6}$$

For $g = 2$ groups, (1.4.6) reduces to the definition (1.4.3) or (1.4.4) for $r_o(\mathbf{x})$ in the case of equal costs of misallocation, except that π_1 is replaced now by $\pi_1 c_{12}$ and π_2 by $\pi_2 c_{21}$. As it is only the ratio of c_{12} and c_{21} that is relevant to the definition of the Bayes rule, these costs can be scaled so that

$$\pi_1 c_{12} + \pi_2 c_{21} = 1.$$

Hence, we can assume without loss of generality that $c_{12} = c_{21} = 1$, provided π_1 and π_2 are now interpreted as the group-prior probabilities adjusted by the relative importance of the costs of misallocation. Due to the rather arbitrary nature of assigning costs of misallocation in practice, they are often taken to be the same in real problems. Further, the group-prior probabilities are often specified as equal. This is not as arbitrary as it may appear at first sight. For example, consider the two-group situation, where G_1 denotes a group of individuals with a rare disease and G_2 those without it. Then, although π_1 and π_2 are disparate, the cost of misallocating an individual with this rare disease may well be much greater than the cost of misallocating a healthy individual. If this is so, then $\pi_1 c_{12}$ and $\pi_2 c_{21}$ may be comparable in magnitude and, as a consequence, the assumption of equal group-prior probabilities with unit costs of misallocation in the formation of the Bayes rule $r_o(\mathbf{x})$ is apt. Also, it would avoid in this example the occurrence of highly unbalanced group-specific error rates. The latter are obtained if $r_o(\mathbf{x})$ is formed with extremely disparate prior probabilities π_i and equal costs of misallocation. This imbalance between the group-specific error rates is a consequence of $r_o(\mathbf{x})$ being the rule that minimizes the overall error rate. In the next section, we consider the construction of rules that are optimal with respect to other criteria. In particular, it will be seen that by specifying the prior probabilities π_i in $r_o(\mathbf{x})$ so that its consequent error rates are equal, we obtain the rule that minimizes the maximum of the group-specific error rates.

1.5 UNAVAILABILITY OF GROUP-PRIOR PROBABILITIES

In some instances in practice, the prior probabilities π_i of the groups G_i are able to be assigned or reasonable estimates are available. For example, in the context of medical diagnosis where the groups represent the possible disease categories to which an individual is to be allocated, the prior probabilities can be taken to be the prevalence rates of these diseases in the population from which the individual has come. However, as to be discussed further in Section

2.3, in some instances, the very purpose of forming an allocation rule for the basis of a screening test is to estimate the prevalence rates of diseases. Also, with a deterministic approach to the construction of an allocation rule, prior probabilities are not relevant to the formulation of the problem.

We now consider the selection of suitable allocation rules where the prior probabilities of the groups are not available. We will only give a brief coverage of available results. For further details, the reader is referred to T. W. Anderson (1984, Chapter 6), who has given a comprehensive account of the decision-theoretic approach to discriminant analysis.

In the absence of prior probabilities of the groups, we cannot define the risk either unconditional or conditional on the feature vector $\mathbf{x}$. Hence, some other criterion must be used. Various other criteria have been discussed by Raiffa (1961). One approach is to focus on the group-specific unconditional losses and to look for the class of admissible rules; that is, the set of rules that cannot be improved upon. For an entity from G_i, the unconditional loss for a rule $r(\mathbf{x})$ is

$$l_i(r) = \sum_{j \neq i}^{g} c_{ij} \mathrm{pr}\{r(\mathbf{X}) = j \mid \mathbf{X} \in G_i\}$$

$$= \sum_{j \neq i}^{g} c_{ij} e_{ij}(r) \qquad (i = 1,\ldots,g).$$

A rule $r^*(\mathbf{x})$ is at least as good as $r(\mathbf{x})$ if

$$l_i(r^*) \leq l_i(r) \qquad (i = 1,\ldots,g). \tag{1.5.1}$$

If at least one inequality in (1.5.1) is strict, then $r^*(\mathbf{x})$ is better than $r(\mathbf{x})$. The rule $r(\mathbf{x})$ is said to be admissible if there is no other rule $r^*(\mathbf{x})$ that is better.

It can be shown that if $\pi_i > 0\ (i = 1,\ldots,g)$, then a Bayes rule is admissible. Also, if $c_{ij} = 1\ (i \neq j)$, and

$$\mathrm{pr}\{f_i(\mathbf{X}) = 0 \mid \mathbf{X} \in G_j\} = 0 \qquad (i,j = 1,\ldots,g),$$

then a Bayes rule is admissible. The converse is true without conditions (except that the parameter space is finite). The proofs of these and other related results can be found in T. W. Anderson (1984, Chapter 6) and in the references therein.

A principle that usually leads to the selection of a unique rule is the minimax principle. A rule is minimax if the maximum unconditional loss is a minimum. In the present context, the rule $r(\mathbf{x})$ is minimax if the maximum of $l_i(r)$ over $i = 1,\ldots,g$ is a minimum over all allocation rules. The minimax rule is the Bayes procedure for which the unconditional losses are equal (von Mises, 1945).

1.6 TRAINING DATA

We have seen in the previous section that the absence of prior probabilities for the groups introduces a complication into the process of obtaining a suitable allocation rule. A much more serious issue arises when the group-conditional densities are either partially or completely unknown.

A basic assumption in discriminant analysis is that in order to estimate the unknown group-conditional densities, there are entities of known origin on which the feature vector $\mathbf{X}$ has been recorded for each. We let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ denote these recorded feature vectors and $\mathbf{z}_1,\ldots,\mathbf{z}_n$ the corresponding vectors of zero-one indicator variables defining the known group of origin of each. We let

$$\mathbf{y}_j = (\mathbf{x}_j', \mathbf{z}_j')' \qquad (j = 1,\ldots,n).$$

The collection of data in the matrix $\mathbf{t}$ defined by

$$\mathbf{t}' = (\mathbf{y}_1,\ldots,\mathbf{y}_n) \tag{1.6.1}$$

is referred to in the literature as either the initial, reference, design, training, or learning data. The last two have arisen from their extensive use in the context of pattern recognition. Also in the latter field, the formation of an allocation rule from training data of known origin is referred to as supervised learning.

There are two major sampling designs under which the training data $\mathbf{T}$ may be realized, joint or mixture sampling and $\mathbf{z}$-conditional or separate sampling. They correspond, respectively, to sampling from the joint distribution of $\mathbf{Y} = (\mathbf{X}', \mathbf{Z}')'$ and to sampling from the distribution of $\mathbf{X}$ conditional on $\mathbf{z}$. The first design applies to the situation where the feature vector and group of origin are recorded on each of n entities drawn from a mixture of the possible groups. Mixture sampling is common in prospective studies and diagnostic situations. In a prospective study design, a sample of individuals is followed and their responses recorded.

With most applications in discriminant analysis, it is assumed that the training data are independently distributed. For a mixture sampling scheme with this assumption, $\mathbf{x}_1,\ldots,\mathbf{x}_n$ are the realized values of n independent and identically distributed (i.i.d.) random variables $\mathbf{X}_1,\ldots,\mathbf{X}_n$ with common distribution function $F_X(\mathbf{x})$. We write

$$\mathbf{X}_1,\ldots,\mathbf{X}_n \overset{iid}{\sim} F_X.$$

The associated group indicator vectors $\mathbf{z}_1,\ldots,\mathbf{z}_n$ are the realized values of the random variables $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$ distributed unconditionally as

$$\mathbf{Z}_1,\ldots,\mathbf{Z}_n \overset{iid}{\sim} \text{Mult}_g(1,\boldsymbol{\pi}). \tag{1.6.2}$$

The assumption of independence of the training data is to be relaxed in Chapter 13. Examples in remote sensing are given there where the assumption of independence is not valid.

With separate sampling in practice, the feature vectors are observed for a sample of n_i entities taken separately from each group G_i ($i = 1,\ldots,g$). Hence, it is appropriate to retrospective studies, which are common in epidemiological investigations. For example, with the simplest retrospective case-control study of a disease, one sample is taken from the cases that occurred during the study period and the other sample is taken from the group of individuals who remained free of the disease. As many diseases are rare and even a large prospective study may produce few diseased individuals, retrospective sampling can result in important economies in cost and study duration. Note that as separate sampling corresponds to sampling from the distribution of $\mathbf{X}$ conditional on $\mathbf{z}$, it does not provide estimates of the prior probabilities π_i for the groups.

1.7 SAMPLE-BASED ALLOCATION RULES

We now consider the construction of an allocation rule from available training data $\mathbf{t}$ in the situation where the group-conditional densities and perhaps also the group-prior probabilities are unknown. The initial approach to this problem, and indeed to discriminant analysis in its modern guise as remarked earlier, was by Fisher (1936). In the context of $g = 2$ groups, he proposed that an entity with feature vector $\mathbf{x}$ be assigned on the basis of the linear discriminant function $\mathbf{a}'\mathbf{x}$, where $\mathbf{a}$ maximizes an index of separation between the two groups. The index was defined to be the magnitude of the difference between the group-sample means of $\mathbf{a}'\mathbf{x}$ normalized by the pooled sample estimate of its assumed common variance within a group. The derivation of Fisher's (1936) linear discriminant function is to be discussed further in Section 3.3, where it is contrasted with normal theory-based discriminant rules.

The early development of discriminant analysis before Fisher (1936) dealt primarily with measures of differences between groups based on sample moments or frequency tables, and ignored correlations among different variates in the feature vector (Pearson, 1916; Mahalanobis, 1927, 1928). One of Fisher's first contacts with discriminant problems was in connection with M. M. Barnard's (1935) work on the secular variation of Egyptian skull characteristics. By 1940, Fisher had published four papers on discriminant analysis, including Fisher (1938) in which he reviewed his 1936 work and related it to the contributions by Hotelling (1931) on his now famous T^2 statistic and by Mahalanobis (1936) on his D^2 statistic and earlier measures of distance. Das Gupta (1980) has given an account of Fisher's research in discriminant analysis.

With the development of discriminant analysis through to the decision-theoretic stage (Wald, 1944; Rao, 1948, 1952, 1954; Hoel and Peterson, 1949), an obvious way of forming a sample-based allocation rule $r(\mathbf{x};\mathbf{t})$ is to take it to be an estimated version of the Bayes rule $r_o(\mathbf{x})$ where, in (1.4.3), the posterior probabilities of group membership $\tau_i(\mathbf{x})$ are replaced by some estimates

$\hat{\tau}_i(\mathbf{x};\mathbf{t})$ formed from the training data $\mathbf{t}$. One approach to the estimation of the posterior probabilities of group membership is to model the $\tau_i(\mathbf{x})$ directly, as with the logistic model to be presented in Chapter 8. Dawid (1976) calls this approach the diagnostic paradigm.

A more common approach, called the sampling approach by Dawid (1976), is to use the Bayes formula (1.2.4) to formulate the $\tau_i(\mathbf{x})$ through the group-conditional densities $f_i(\mathbf{x})$. With this approach, the Bayes rule is estimated by the so-called plug-in rule,

$$r(\mathbf{x};\mathbf{t}) = r_o(\mathbf{x};\hat{F}), \tag{1.7.1}$$

where we now write the optimal or Bayes rule as $r_o(\mathbf{x};F)$ to explicitly denote its dependence on the distribution function $F(\mathbf{y})$ of $\mathbf{Y} = (\mathbf{X}',\mathbf{Z}')'$. As before, $\mathbf{X}$ is the feature observation and $\mathbf{Z}$ defines its group of origin. In (1.7.1), $\hat{F}$ denotes an estimate of F that can be obtained by estimating separately each group-conditional distribution from the training data $\mathbf{t}$.

The group-prior probabilities can be estimated by the proportion of entities from each group at least under mixture sampling. Their estimation under separate sampling is considered in the next chapter, commencing in Section 2.3. Concerning the estimates of the group-conditional distribution functions, a nonparametric approach may be adopted using, say, kernel or nearest-neighbor methods. These along with other nonparametric methods are to be discussed in Chapter 9. A commonly used approach is the parametric, which is introduced in the next section in a general context. It is to be considered further in Chapter 3 for the specific choice of normal models and in Chapter 7 for nonnormal models. There is also the work, in the spirit of the empirical Bayes approach of Robbins (1951, 1964), on the allocation of a sequence of unclassified entities whose group-indicator vectors and features are independently distributed. Results under various assumptions on the available information on the underlying distributions have been obtained by Johns (1961), Samuel (1963a, 1963b), Hudimoto (1968), K. Choi (1969), Wojciechowski (1985), and Stirling and Swindlehurst (1987), among others.

1.8 PARAMETRIC ALLOCATION RULES

Under the parametric approach to the estimation of the group-conditional distributions, and hence of the Bayes rule, the group-conditional distributions are taken to be known up to a manageable number of parameters. More specifically, the ith group-conditional density is assumed to belong to a family of densities

$$\{f_i(\mathbf{x};\boldsymbol{\theta}_i) : \boldsymbol{\theta}_i \in \boldsymbol{\Theta}_i\}, \tag{1.8.1}$$

where $\boldsymbol{\theta}_i$ is an unknown parameter vector belonging to some parameter space $\boldsymbol{\Theta}_i$ ($i = 1,\ldots,g$). Often the group-conditional densities are taken to belong to the same parametric family, for example, the normal.

The density functions of $\mathbf{X}$ and $\mathbf{Y} = (\mathbf{X}', \mathbf{Z}')'$ are written now as $f_X(\mathbf{x}; \Psi)$ and $f(\mathbf{y}; \Psi)$, respectively, where

$$\Psi = (\boldsymbol{\pi}', \boldsymbol{\theta}')' \tag{1.8.2}$$

and $\boldsymbol{\theta}$ is the vector consisting of the elements of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g$ known *a priori* to be distinct. For example, if the group-conditional distributions are assumed to be multivariate normal with means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g$ and common covariance matrix $\boldsymbol{\Sigma}$, then $\boldsymbol{\theta}_i$ consists of the elements of $\boldsymbol{\mu}_i$ and of the distinct elements of $\boldsymbol{\Sigma}$, and $\boldsymbol{\theta}$ consists of the elements of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g$ and of the distinct elements of $\boldsymbol{\Sigma}$. Note that since the elements of the vector $\boldsymbol{\pi}$ of the mixing proportions π_i sum to one, one of them is redundant in Ψ, but we will not modify Ψ accordingly, at least explicitly. However, in our statements about the distribution of any estimator of Ψ, it will be implicitly assumed that one of the mixing proportions has been deleted from Ψ.

With the so-called estimative approach to the choice of a sample-based discriminant rule, unknown parameters in the adopted parametric forms for the group-conditional distributions are replaced by appropriate estimates obtained from the training data $\mathbf{t}$. Hence, if $r_o(\mathbf{x}; \Psi)$ now denotes the optimal rule, then with this approach,

$$r(\mathbf{x}; \mathbf{t}) = r_o(\mathbf{x}; \hat{\Psi}),$$

where $\hat{\Psi}$ is an estimate of Ψ formed from $\mathbf{t}$. Provided $\hat{\boldsymbol{\theta}}_i$ is a consistent estimator of $\boldsymbol{\theta}_i$ and $f_i(\mathbf{x}; \boldsymbol{\theta}_i)$ is continuous in $\boldsymbol{\theta}_i$ $(i = 1, \ldots, g)$, then $r_o(\mathbf{x}; \hat{\Psi})$ is a Bayes risk consistent rule in the sense that its risk, conditional on $\hat{\Psi}$, converges in probability to that of the Bayes rule, as n approaches infinity. This is assuming that the postulated model (1.8.1) is indeed valid and that the group-prior probabilities are estimated consistently as possible, for instance, with mixture sampling of the training data. If the conditional risk for $r_o(\mathbf{x}; \hat{\Psi})$ converges almost surely to that of the Bayes rule as n approaches infinity, then it is said to be Bayes risk strongly consistent. Consistency results for sample-based allocation rules have been obtained by Van Ryzin (1966) and Glick (1972, 1976). Initial references on the notion of consistency for sample-based allocation rules include Hoel and Peterson (1949) and Fix and Hodges (1951). The latter technical report, which also introduced several important nonparametric concepts in a discriminant analysis context, has been reprinted in full recently at the end of a commentary on it by Silverman and Jones (1989).

Given the widespread use of maximum likelihood as a statistical estimation technique, the plug-in rule $r_o(\mathbf{x}; \hat{\Psi})$ is usually formed with $\hat{\Psi}$, or at least $\hat{\boldsymbol{\theta}}$, taken to be the maximum likelihood estimate. This method of estimation in the context of discriminant analysis is to be considered further in the next section. Since their initial use by Wald (1944), Rao (1948, 1954), and T. W. Anderson (1951), among others, plug-in rules formed by maximum likelihood estimation under the assumption of normality have been extensively applied in

practice. The estimation of $r_o(\mathbf{x};\Psi)$ by $r_o(\mathbf{x};\hat{\Psi})$, where $\hat{\Psi}$ is the maximum likelihood estimate of Ψ, preserves the invariance of an allocation rule under monotone transformations.

Concerning some other parametric approaches to constructing a sample-based rule, there is the likelihood ratio criterion. The unknown vector $\mathbf{z}$ of zero-one indicator variables defining the group of origin of the unclassified entity is treated as a parameter to be estimated, along with Ψ, on the basis of $\mathbf{t}$ and also $\mathbf{x}$. It differs from the estimative approach in that it includes the unclassified observation $\mathbf{x}$ in the estimation process. Hence, in principle, there is little difference between the two approaches although, in practice, the difference may be of some consequence, in particular for disparate group-sample sizes.

Another way of proceeding with the estimation of the group-conditional densities, and, hence, of $r_o(\mathbf{x};\Psi)$, is to adopt a Bayesian approach, which is considered in Section 2.2. Among other criteria proposed for constructing allocation rules is minimum distance. With this criterion, an entity with feature vector $\mathbf{x}$ is allocated to the group whose classified data in the training set $\mathbf{t}$ is closest to $\mathbf{x}$ in some sense. Although minimum-distance rules are often advocated in the spirit of distribution-free approaches to allocation, they are predicated on some underlying assumption for the group-conditional distributions. For example, the use of Euclidean distance as a metric corresponds to multivariate normal group-conditional distributions with a common spherical covariance matrix, and Mahalanobis distance corresponds to multivariate normal distributions with a common covariance matrix. The aforementioned parametric allocation rules are discussed in more detail with others in Chapter 3 in the context of normal theory-based discrimination.

Often, in practice, the total sample size is too small relative to the number p of feature variables in $\mathbf{x}$ for a reliable estimate of $\boldsymbol{\theta}$ to be obtained from the full set $\mathbf{t}$ of training data. This is referred to as "the curse of dimensionality," a phrase due to Bellman (1961). Consideration then has to be given to which variables in $\mathbf{x}$ should be deleted in the estimation of $\boldsymbol{\theta}$ and the consequent allocation rule. Even if a satisfactory discriminant rule can be formed using all the available feature variables, consideration may still be given to the deletion of some of the variables in $\mathbf{x}$. This is because the performance of a rule fails to keep improving and starts to fall away once the number of feature variables has reached a certain threshold. This so-called peaking phenomenon of a rule is discussed further in Chapter 12, where the variable-selection problem is to be addressed. It is an important problem in its own right in discriminant analysis, as with many applications, the primary or sole aim is not one of allocation, but rather to infer which feature variables of an entity are most useful in explaining the differences between the groups. If some or all of the group-sample sizes n_i of the classified data are very small, then consideration may have to be given to using unclassified data in the estimation of $\boldsymbol{\theta}$, as discussed in Section 2.7.

1.9 ASSESSMENT OF MODEL FIT

If the postulated group-conditional densities provide a good fit and the group-prior probabilities are known or able to be estimated with some precision, then the plug-in rule $r_o(\mathbf{x};\hat{F})$ should be a good approximation to the Bayes rule $r_o(\mathbf{x};F)$. However, even if $\hat{F}$ is a poor estimate of F, $r_o(\mathbf{x};\hat{F})$ may still be a reasonable allocation rule. It can be seen from the definition (1.4.4) of $r_o(\mathbf{x};F)$ that for $r_o(\mathbf{x};\hat{F})$ to be a good approximation to $r_o(\mathbf{x};F)$, it is only necessary that the boundaries defining the allocation regions,

$$\{\mathbf{x} : \pi_i f_i(\mathbf{x}) = \pi_j f_j(\mathbf{x}), \qquad i < j = 1,\ldots,g\}, \tag{1.9.1}$$

be estimated precisely. This implies at least for well-separated groups that in consideration of the estimated group-conditional densities, it is the fit in the tails rather than in the main body of the distributions that is crucial. This is what one would expect. Any reasonable allocation rule should be able to allocate correctly an entity whose group of origin is obvious from its feature vector. Its accuracy is really determined by how well it can handle entities of doubtful origin. Their feature vectors tend to occur in the tails of the distributions.

If reliable estimates of the posterior probabilities of group membership $\tau_i(\mathbf{x})$ are sought in their own right and not just for the purposes of making an outright assignment, then the fit of the estimated density ratios $\hat{f}_i(\mathbf{x})/\hat{f}_j(\mathbf{x})$ is important for all values of $\mathbf{x}$ and not just on the boundaries (1.9.1). It can be seen in discriminant analysis that the estimates of the group-conditional densities are not of interest as an end in themselves, but rather how useful their ratios are in providing estimates of the posterior probabilities of group membership or at least an estimate of the Bayes rule. However, for convenience, the question of model fit in practice is usually approached by consideration of the individual fit of each estimated density $\hat{f}_i(\mathbf{x})$.

Many different families of distributions may be postulated for the group-conditional densities, although some may be difficult to deal with analytically or computationally. The normal assumption is commonly adopted in practice. In some cases for this to be reasonable, a suitable transformation of the feature variables is required. In many practical situations, some variables in the feature vector $\mathbf{x}$ may be discrete. Often treating the discrete variables, in particular binary variables, as if they were normal in the formation of the discriminant rule is satisfactory. However, care needs to be exercised if several of the feature variables are discrete. The use of nonnormal models, including for mixed feature vectors where some of the variables are continuous and some are discrete, is discussed in Chapter 7, and Chapter 3 is devoted entirely to discrimination via normal models. Practical aspects such as robust methods of estimating group-conditional parameters, use of transformations to achieve approximate normality, testing for normality, and detection of atypical entities are discussed in Chapters 5 and 6.

1.10 ERROR RATES OF ALLOCATION RULES

1.10.1 Types of Error Rates

The allocation rates associated with the optimal or Bayes rule are given by

$$eo_{ij}(F) = \text{pr}\{r_o(\mathbf{X};F) = j \mid \mathbf{X} \in G_i\} \qquad (i,j = 1,\ldots,g), \quad (1.10.1)$$

where $eo_{ij}(F)$ is the probability that a randomly chosen entity from G_i is allocated to G_j on the basis of $r_o(\mathbf{x};F)$. The error rate specific to the ith group G_i is

$$eo_i(F) = \sum_{j \neq i}^{g} eo_{ij}(F) \qquad (i = 1,\ldots,g),$$

and the overall error rate is

$$eo(F) = \sum_{i=1}^{g} \pi_i eo_i(F).$$

As seen in Section 1.4, $r_o(\mathbf{x};F)$ is the rule that minimizes the overall error rate in the case of unit costs of misallocation. Consequently, $eo(F)$ is referred to as the optimal (overall) error rate. The optimal overall error rate can be used as a measure of the degree of separation between the groups, as to be considered in Section 1.12.

We proceed now to define the error rates of a sample-based rule. Let $r(\mathbf{x};\mathbf{t})$ denote an allocation rule formed from the training data $\mathbf{t}$. Then the allocation rates of $r(\mathbf{x};\mathbf{t})$, conditional on $\mathbf{t}$, are defined by

$$ec_{ij}(F_i;\mathbf{t}) = \text{pr}\{r(\mathbf{X};\mathbf{t}) = j \mid \mathbf{X} \in G_i,\mathbf{t}\}, \quad (1.10.2)$$

which is the probability, conditional on $\mathbf{t}$, that a randomly chosen entity from G_i is allocated to G_j $(i,j = 1,\ldots,g)$. The group-specific conditional error rates are given by

$$ec_i(F_i;\mathbf{t}) = \sum_{j \neq i}^{g} ec_{ij}(F_i;\mathbf{t}) \qquad (i = 1,\ldots,g),$$

and the overall conditional error rate by

$$ec(F;\mathbf{t}) = \sum_{i=1}^{g} \pi_i ec_i(F_i;\mathbf{t}).$$

For equal costs of misallocation, the rule $r(\mathbf{x};\mathbf{t})$ is Bayes risk consistent (strongly consistent) if $ec(F;\mathbf{t})$ converges in probability (almost surely) to $eo(F)$, as n approaches infinity.

On averaging the conditional allocation rates over the distribution of the training data, we obtain the expected or unconditional rates defined as

$$\begin{aligned} eu_{ij}(F) &= E\{ec_{ij}(F_i;\mathbf{T})\} \\ &= \text{pr}\{r(\mathbf{X};\mathbf{T}) = j \mid \mathbf{X} \in G_i\} \qquad (i,j = 1,\ldots,g). \end{aligned} \quad (1.10.3)$$

In the case of separate sampling where $\mathbf{t}$ is based on a fixed number of entities from each group, we should, strictly speaking, denote these unconditional rates as $eu_{ij}(F_1,\ldots,F_g)$, rather than $eu_{ij}(F)$. The group-specific unconditional error rates are given by

$$eu_i(F) = \sum_{j\neq i}^{g} eu_{ij}(F),$$

and the overall unconditional error rate by

$$eu(F) = \sum_{i=1}^{g} \pi_i eu_i(F).$$

We are following Hills (1966) here in referring to the $ec_i(F_i;\mathbf{t})$ and $ec(F;\mathbf{t})$ as the conditional or actual error rates, and to the $eu_i(F)$ and $eu(F)$ as the unconditional or expected error rates. Before the introduction of his careful terminology for the various types of error rates, there had been a good deal of confusion in the literature; see the comment of Cochran (1966).

1.10.2 Relevance of Error Rates

Concerning the relevance of error rates in discriminant analysis, for allocation problems, they play a major role in providing a measure of the global performance of a discriminant rule. It has been suggested (Lindley, 1966) that more attention should be paid to the unconditional losses. However, as remarked earlier, the specification of costs in practice is often arbitrary.

On the use of error rates to measure the performance of a sample-based allocation rule, it is the conditional error rates that are of primary concern once the rule has been formed from the training data $\mathbf{t}$. If $\mathbf{t}$ denotes all the available data of known origin, then one is stuck with this training set in forming a rule. An example where these error rates enter naturally into an analysis is when the rule $r(\mathbf{x};\mathbf{t})$ forms the basis of a diagnostic test for estimating the prevalence rates of a disease, as covered in Section 2.3.

The average performance of the rule over all possible realizations of $\mathbf{t}$ is of limited interest in applications of $r(\mathbf{x};\mathbf{t})$. However, the unconditional error rates are obviously relevant in the design of a rule. They relate the average performance of the rule to the size of the training set and to the group-conditional distributions as specified. For example, consider the case of two groups G_1 and G_2 in which the feature vector $\mathbf{X}$ is taken to have a multivariate normal distribution with means μ_1 and μ_2, respectively, and common covariance matrix Σ. For separate sampling with equal sample sizes $n/2$ from each of the two groups, the sample-based analogue of the Bayes rule with equal group-prior probabilities has equal unconditional error rates. Their common value, equal to the overall error rate for equal priors, is given by

$$eu(F) \approx eo(F) + n^{-1}\{\phi(\tfrac{1}{2}\Delta)/4\}\{p\Delta + 4(p-1)\Delta^{-1}\}, \qquad (1.10.4)$$

where

$$eo(F) = \Phi(-\tfrac{1}{2}\Delta)$$

and

$$\Delta = \{(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)\}^{1/2} \tag{1.10.5}$$

is the Mahalanobis (1936) distance between G_1 and G_2. In this and subsequent work, Φ and ϕ denote the standard normal distribution and density, respectively. The error of the approximation (1.10.4) is of order $O(n^{-2})$ (Okamoto, 1963). The derivation of (1.10.4) is discussed in Section 4.2.

From (1.10.4), we can determine approximately how large n must be for a specified Δ and p in order for the unconditional error rate not to exceed too far the best obtainable, as given by the optimal rate $eo(F)$. For instance, for $\Delta = 1$ representing two groups that are close together, n on the basis of (1.10.4) has be to at least 40 with $p = 3$ for the rate to be less than 1/3 on average; that is, not more than 0.0248 in excess of the optimal rate of 0.3085. The latter value shows that it is not possible to design an accurate allocation rule in this case. Indeed, if n is small, then for $p > 1$, the error rate is not far short of 1/2, which is the error rate for a randomized rule that ignores the feature vector and makes a choice of groups according to the toss of a fair coin.

It can be seen from (1.10.2) and (1.10.3) that the conditional and unconditional allocation rates of a sample-based rule depend on the unknown group-conditional distributions and so must be estimated. In the absence of any further data of known origin, these rates must be estimated from the same data $\mathbf{t}$ from which the rule has been formed. Hence, there are difficulties in obtaining unbiased estimates of the error rates of a sample-based rule in its application to data of unknown origin, distributed independently of the training sample. Estimation of the error rates of allocation rules is thus a difficult but important problem in discriminant analysis. It is taken up in Chapter 10, which is devoted fully to it.

We have seen that the situation where some of the errors of allocation are more serious than others can be handled through the specification of unequal costs of misallocation in the definition (1.4.6) of the Bayes rule. Another approach would be to introduce regions of doubt in the feature space where no allocation is made. This approach was adopted by J. A. Anderson (1969) in his design of a rule with upper bounds specified on the errors of allocation. It was used also by Habbema, Hermans, and van der Burgt (1974b) in their development of a decision-theoretic model for allocation. Previously, Marshall and Olkin (1968) had considered situations where direct assessment of the group of origin is possible, but expensive. In these situations, after the feature vector has been observed, there is a choice between allocation and extensive group assessment. Another approach where there is an alternative to an outright allocation of the entity after its feature vector has been observed was given by Quesenberry and Gessaman (1968). Their nonparametric procedure constructs tolerance regions for each group, and an entity is allocated to the set of those groups whose tolerance regions contain the feature vector $\mathbf{x}$. If $\mathbf{x}$ falls

within all or outside all the tolerance regions, then the entity is not allocated; see also Gessaman and Gessaman (1972). Broffitt, Randles, and Hogg (1976) introduced a rank method for partial allocation with constraints imposed on the unconditional error rataes. This nonparametric approach to partial discrimination in the presence of constraints is discussed in Section 9.9.

A parametric approach to constrained discrimination with unknown group-conditional densities has been investigated by T. W. Anderson (1973a, 1973b) and McLachlan (1977b) for the sample normal-based linear discriminant rule. Their work is described in Section 4.5. Also, Gupta and Govindarajulu (1973) considered constrained discrimination in the special case of univariate normal group-conditional distributions with multiple independent measurements available on the entity to be allocated.

The error rates are not the only measure of the global accuracy of an allocation rule. Breiman et al. (1984, Section 4.6) have proposed a global measure in terms of estimates of the posterior probabilities of group membership for a rule $r(\mathbf{x};\mathbf{t})$ defined analogously to the Bayes rule $r_o(\mathbf{x};F)$. That is, $r(\mathbf{x};\mathbf{t})$ is equal to i if the estimated posterior probabilities satisfy

$$\hat{\tau}_i(\mathbf{x};\mathbf{t}) \geq \hat{\tau}_j(\mathbf{x};t) \qquad (j = 1,\ldots,g;\ j \neq i). \tag{1.10.6}$$

Their proposed measure of the accuracy (conditional here on the training data $\mathbf{t}$) of the rule $r(\mathbf{x};\mathbf{t})$ is

$$E\left[\sum_{i=1}^{g}\{\hat{\tau}_i(\mathbf{X};\mathbf{t}) - \tau_i(\mathbf{X})\}^2 \mid \mathbf{t}\right]. \tag{1.10.7}$$

They noted that if the mean-squared error (conditional on $\mathbf{t}$) of the rule $r(\mathbf{x};\mathbf{t})$ is defined as

$$\mathrm{MSE}(r) = E\left[\sum_{i=1}^{g}\{\hat{\tau}_i(\mathbf{X};\mathbf{t}) - Z_i\}^2 \mid \mathbf{t}\right], \tag{1.10.8}$$

then it can be decomposed into the two terms,

$$\mathrm{MSE}(r) = \mathrm{MSE}(r_o) + E\left[\sum_{i=1}^{g}\{\hat{\tau}_i(\mathbf{X};\mathbf{t}) - \tau_i(\mathbf{X})\}^2 \mid \mathbf{t}\right],$$

where

$$\mathrm{MSE}(r_o) = E\left[\sum_{i=1}^{g}\{\tau_i(\mathbf{X}) - Z_i\}^2\right]$$

is the mean-squared error of the Bayes rule $r_o(\mathbf{x})$. Hence, a comparison in terms of the accuracy (1.10.7) of different rules of the form (1.10.6) can be made in terms of their conditional mean-squared errors. This provides a significant advantage as, unlike (1.10.7), $\mathrm{MSE}(r)$ can be estimated directly from $\mathbf{t}$ as

$$\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{g}\{\hat{\tau}_i(\mathbf{x}_j;\mathbf{t}) - z_{ij}\}^2,$$

where $z_{ij} = (\mathbf{z}_j)_i$, and $\mathbf{z}_j$ is the vector of zero-one indicator variables defining the known group of origin of the jth feature vector $\mathbf{x}_j$ in the training data $\mathbf{t}$ ($j = 1,\ldots,n$).

Note that by virtue of their definition, error rates are concerned only with the allocatory performance of a rule. Hence, for rules of the form (1.10.6), they are concerned only with the relative sizes of the estimated posterior probabilities of group membership. By contrast, the criterion (1.10.7) attempts to measure the accuracy of a rule of the form (1.10.6) by assessing the absolute fit of the posterior probabilities of group membership.

Other ways of assessing the discriminatory performance of a fitted model have been considered by Habbema, Hilden, and Bjerregaard (1978b, 1981); Hilden, Habbema, and Bjerregaard (1978a, 1978b); and Habbema and Hilden (1981).

1.11 POSTERIOR PROBABILITIES OF GROUP MEMBERSHIP

It was shown in Section 1.8 that the posterior probabilities of group membership $\tau_i(\mathbf{x})$ or their estimates may play no role in the formation of some allocation rules in the pure decision context. On the other hand with the Bayes rule or a sample version, the relative sizes of the posterior probabilities of group membership $\tau_i(\mathbf{x})$ form the basis of the subsequent outright allocation to be made. In many real problems, only a tentative allocation is contemplated before consideration is to be given to taking an irrevocable decision as to the group of origin of an unclassified entity. For these problems, the probabilistic allocation rule implied by the $\tau_i(\mathbf{x})$ or their estimates provides a concise way of expressing the uncertainty about the group membership of an unclassified entity with an observed feature vector $\mathbf{x}$.

It has been argued (Spiegelhalter, 1986) that the provision of accurate and useful probabilistic assessments of future events should be a fundamental task for biostatisticians collaborating in clinical or experimental medicine. To this end, the posterior probabilities of group membership play a major role in patient management and clinical trials. For example, in the former context with the groups corresponding to the possible treatment decisions, the uncertainty over which decision to make is conveniently formulated in terms of the posterior probabilities of group membership. Moreover, the management of the patient may be only at a preliminary stage where an outright assignment may be premature particularly, say, if the suggested treatment decision is not clearcut and involves major surgery on the patient. The reliability of these estimates is obviously an important question to be considered, especially in applications where doubtful cases of group membership arise.

If the posterior probabilities of group membership have been estimated for the express purpose of forming an allocation rule, then their overall reliability can be assessed through the global performance of this rule as measured by its associated error rates. However, as emphasized by Critchley and Ford (1985),

even if all its error rates are low, there may still be entities about which there is great uncertainty as to their group of origin. Conversely, these global measures may be high, yet it may still be possible to allocate some entities with great certainty. Thus, in some situations, it may not be appropriate to consider an assessment in terms of the error rates. Indeed, as pointed out by Aitchison and Kay (1975), in clinical medicine, the Hippocratic oath precludes any criterion of average results over individual patients (such as error rates), so that conditioning on the feature vector $\mathbf{x}$ is an apt way to proceed. In Chapter 11, we consider methods for assessing the reliability of the estimates of the posterior probabilities of group membership from the same training data used to form these estimates in the first instance.

1.12 DISTANCES BETWEEN GROUPS

Over the years, there have been proposed many different measures of distance, divergence, or discriminatory information between two groups. Krzanowski (1983a) has put them broadly into two categories: (a) measures based on ideas from information theory and (b) measures related to Bhattacharyya's (1943) measure of affinity.

Some members of category (a) are considered first. There is the Kullback–Leibler (1951) measure of discriminatory information between two groups with distribution functions F_1 and F_2, admitting densities $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, respectively, with respect to some measure ν. This measure is given by

$$\delta_{KL}(F_1, F_2) = \int f_1(\mathbf{x}) \log\{f_1(\mathbf{x})/f_2(\mathbf{x})\}\, d\nu.$$

It is a directed divergence in that it also has a directional component, since generally, $\delta_{KL}(F_1, F_2) \neq \delta_{KL}(F_2, F_1)$; that is, it is not a metric. Jeffreys' (1948) measure is a symmetric combination of the Kullback–Leibler information,

$$\delta_J(F_1, F_2) = \delta_{KL}(F_1, F_2) + \delta_{KL}(F_2, F_1).$$

A third measure in category (a) is

$$\delta_S(F_1, F_2) = \tfrac{1}{2}[\delta_{KL}\{F_1, \tfrac{1}{2}(F_1 + F_2)\} + \delta_{KL}\{F_2, \tfrac{1}{2}(F_1 + F_2)\}],$$

which is Sibson's (1969) information radius given in its simple form.

Rényi (1961) generalized both Shannon (1948) entropy and the Jeffreys–Kullback–Leibler information by introducing a scalar parameter. Recently, Burbea and Rao (1982), Burbea (1984), and Taneja (1983) have proposed various alternative ways to generalize $\delta_J(F_1, F_2)$. The proposed measures of Burbea and Rao (1982) and Burbea (1984) involve one parameter, and the measures proposed by Taneja (1983) involve two parameters. The definitions of these generalized measures may be found in Taneja (1987). Another measure that has been proposed is the power divergence corresponding to the power-divergence family of goodness-of-fit statistics introduced by Cressie and Read (1984); see also Read and Cressie (1988, Section 7.4).

Concerning members of category (b), Bhattacharyya's original measure of affinity is

$$\rho = \int \{f_1(\mathbf{x}) f_2(\mathbf{x})\}^{1/2} d\nu. \tag{1.12.1}$$

Bhattacharyya (1946) subsequently proposed

$$\delta_B(F_1, F_2) = \cos^{-1}(\rho),$$

and, in unpublished notes, A. N. Kolmogorov used

$$\delta_{KO}(F_1, F_2) = 1 - \rho.$$

Chernoff (1952) introduced the more general distance measure

$$\delta_C(F_1, F_2) = -\log \int \{f_1(\mathbf{x})\}^{\alpha} \{f_2(\mathbf{x})\}^{1-\alpha} d\nu,$$

where $\alpha \in [0, 1]$. It reduces to $-\log\rho$ in the special case of $\alpha = 0.5$. If $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are multivariate normal densities with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively, then

$$\begin{aligned} -\log\rho = {} & \tfrac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ & + \tfrac{1}{2} \log[|\boldsymbol{\Sigma}| / \{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|\}^{1/2}], \end{aligned}$$

where

$$\boldsymbol{\Sigma} = \tfrac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2);$$

see, for example, Kailath (1967).

Matusita (1956) subsequently defined the distance measure

$$\begin{aligned} \delta_M(F_1, F_2) &= \left[\int \{\sqrt{f_1(\mathbf{x})} - \sqrt{f_2(\mathbf{x})}\}^2 d\nu \right]^{1/2} \\ &= (2 - 2\rho)^{1/2} \\ &= \{2\delta_{KO}(F_1, F_2)\}^{1/2}. \end{aligned} \tag{1.12.2}$$

The distance (1.12.2) is sometimes referred to as Hellinger's distance; see, for example, Le Cam (1970) and Beran (1977). There is little practical difference between these functionally related measures in category (b).

Additional distance measures to these defined above may be found in Ben-Bassat (1982), who summarized findings on the current measures, including their relationships with lower and upper bounds on the overall optimal error rate. A recent paper on the latter topic is Ray (1989a), who considered the maximum of the span s between the upper and lower bounds on the overall error rate $eo(F)$ of the Bayes rule, as provided by ρ. Hudimoto (1956–1957, 1957–1958) had shown that

$$\tfrac{1}{2} - \tfrac{1}{2}(1 - 4\pi_1\pi_2\rho^2)^{1/2} \leq eo(F) \leq (\pi_1\pi_2)^{1/2}\rho,$$

with span

$$s = (\pi_1\pi_2)^{1/2}\rho - \tfrac{1}{2} + \tfrac{1}{2}(1 - 4\pi_1\pi_2\rho^2)^{1/2}.$$

Ray (1989a) showed that the maximum value of s is $\frac{1}{2}(\sqrt{2}-1)$, which can be attained for values of π_1, and hence values of π_2, lying inside the interval

$$\{(2-\sqrt{2})/4,\ (2+\sqrt{2})/4\}.$$

In another related paper, Ray (1989b) considered the maximum of the span between the upper and lower bounds on $eo(F)$ as provided by the generalized distance measure of Lissack and Fu (1976). This measure is

$$\delta_{LF}(F_1, F_2) = \int |\tau_1(\mathbf{x}) - \tau_2(\mathbf{x})|^{\alpha} f_X(\mathbf{x}) d\nu \qquad 0 < \alpha < \infty,$$

where $\tau_i(\mathbf{x})$ is the posterior probability of membership of group G_i ($i = 1,2$), as defined by (1.2.4).

It is a generalization of the Kolmogorov variational distance defined by

$$\int |\pi_1 f_1(\mathbf{x}) - \pi_2 f_2(\mathbf{x})|\, d\nu.$$

For $\alpha = 1$, $\delta_{LF}(F_1, F_2)$ reduces to this distance with

$$\begin{aligned}\delta_{LF}(F_1, F_2) &= \int |\pi_1 f_1(\mathbf{x}) - \pi_2 f_2(\mathbf{x})|\, d\nu \\ &= 2\int \max\{\pi_1 f_1(\mathbf{x}), \pi_2 f_2(\mathbf{x})\}\, d\nu - 1 \\ &= 1 - 2eo(F); \end{aligned} \tag{1.12.3}$$

see Rao (1948, 1977) and Glick (1973b).

For $0 < \alpha \leq 1$, Lissack and Fu (1976) showed that

$$\tfrac{1}{2}\{1 - \delta_{LF}(F_1, F_2)\} \leq eo(F) \leq \tfrac{1}{2}[1 - \{\delta_{LF}(F_1, F_2)\}^{1/\alpha}],$$

and for $1 \leq \alpha < \infty$,

$$\tfrac{1}{2}[1 - \{\delta_{LF}(F_1, F_2)\}^{1/\alpha}] \leq eo(F) \leq \tfrac{1}{2}\{1 - \delta_{LF}(F_1, F_2)\}.$$

The lower and upper bounds coincide for $\alpha = 1$ in accordance with the result (1.12.3). For $\alpha > 1$, Ray (1989b) showed that the maximum of their difference is

$$\tfrac{1}{2}\{\alpha^{-1/(\alpha-1)} - \alpha^{-\alpha/(\alpha-1)}\},$$

and that it increases from 0 to 0.5 as α increases from 1 to infinity. He also established that the maximum difference between the upper and lower bounds increases from 0 to 0.5 as α decreases from 1 to 0.

An early study of distance measures was by Adhikari and Joshi (1956). A general class of coefficients of divergence of one distribution from another was considered by Ali and Silvey (1966), who demonstrated that the measures

above are members of this class. Furthermore, if $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are multivariate normal densities with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, and common covariance matrix $\boldsymbol{\Sigma}$, then every coefficient in their class is an increasing function of the Mahalanobis distance Δ defined by (1.10.5). For example, the affinity ρ then is given by

$$\rho = \exp(-\Delta^2/8), \tag{1.12.4}$$

as calculated previously by Matusita (1973). The Mahalanobis distance Δ has become the standard measure of distance between two groups when the feature variables are continuous.

Atkinson and Mitchell (1981) have shown how Δ arises naturally from Rao's (1945) procedure for determining distances between members of a well-behaved parametric family of distributions. The relevant family in this case is the multivariate normal with common shape but varying location. Concerning alternative models for multivariate data, a rich source of models is provided by the class of elliptic distributions whose densities have elliptical contours and which include the multivariate normal, multivariate Student's t, and Cauchy. The p-dimensional random variable $\mathbf{X}$ is said to have an elliptic distribution with parameters $\boldsymbol{\mu}$ ($p \times 1$ vector) and $\boldsymbol{\Sigma}$ ($p \times p$ positive-definite matrix) if its density is of the form

$$|\boldsymbol{\Sigma}|^{-1/2} f_S[\{\delta(\mathbf{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma})\}^{1/2}], \tag{1.12.5}$$

where $f_S(\cdot)$ is any function such that $f_S(\|\mathbf{x}\|)$ is a density on R^p and

$$\delta(\mathbf{x}, \boldsymbol{\mu}; \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}).$$

The class of elliptic densities defined by (1.12.5) can be generated by a non-singular transformation of $\mathbf{x}$ from the class of spherically symmetric densities $f_S(\|\mathbf{x}\|)$, where $\|\mathbf{x}\| = (\mathbf{x}'\mathbf{x})^{1/2}$ denotes the Euclidean norm of $\mathbf{x}$. It follows in (1.12.5) that $\boldsymbol{\mu}$ is the mean of $\mathbf{X}$ and $\boldsymbol{\Sigma}$ is a scalar multiple of the covariance matrix of $\mathbf{X}$. Mitchell and Krzanowski (1985) have shown that the Mahalanobis distance Δ remains appropriate when the family of distributions under consideration is one of the elliptic class having fixed shape but varying location. One implication of this result noted by Mitchell and Krzanowski (1985) is that the sample analogue of Δ is the appropriate measure of the distance between the estimates of $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ fitted using either the estimative or the predictive approaches under the assumption of multivariate normal densities with a common covariance matrix. As discussed in the next chapter, the fitted densities are multivariate normal in the estimative case and multivariate Student's t in the predictive case.

Bhattacharyya's (1946) measure $\delta_B(F_1, F_2)$, which has become known as the angular separation between the groups, was the first measure proposed for discrete feature data. An alternative approach was adopted by Balakrishnan and Sanghvi (1968) and Kurezynski (1970), who attempted to create Mahalanobis-like distance measures for discrete feature data. These subsumed some earlier measures based on chi-squared statistics (Sanghvi, 1953). Bhattacharyya's

(1946) measure in the context of multinomial data has received strong support from Edwards (1971), who provided a further measure through the stereographic projection as an approximation to the angular separation.

The affinity-based measure (1.12.1) has been developed by Matusita (1964, 1967a, 1967b, 1971, 1973). For $g > 2$ groups, the affinity of the group-conditional distributions $F_1,\ldots,F_g$ is defined as

$$\rho_g(F_1,\ldots,F_g) = \int \{f_1(\mathbf{x})\cdots f_g(\mathbf{x})\}^{1/g}\, d\nu.$$

The affinity ρ_g is connected with the distance

$$\left|\int [\{f_i(\mathbf{x})\}^{1/g} - \{f_j(\mathbf{x})\}^{1/g}]^g\, d\nu\right|^{1/g}$$

between any pair of distributions F_i and F_j, $i \neq j = 1,\ldots,g$. In the case of two groups, there is a complete duality between the distance measure and the affinity measure of Matusita (1956). However, this is not clear where there are more than two groups (Ahmad, 1982).

Toussaint (1974b) has extended the definition of $\rho_g(F_1,\ldots,F_g)$ to

$$\int [\{f_1(\mathbf{x})\}^{c_1}\cdots\{f_g(\mathbf{x})\}^{c_g}]\, d\nu,$$

where

$$\sum_{i=1}^{g} c_i = 1,$$

and $c_i \geq 0\,(i = 1,\ldots,g)$. It reduces to $\rho_g(F_1,\ldots,F_g)$ when $c_i = 1/g$ for $i = 1,\ldots,g$. As noted by Glick (1973b), a measure of separation between $F_1,\ldots,F_g$ is provided by

$$1 - \rho_g^{g/2}(F_1,\ldots,F_g). \tag{1.12.6}$$

Since it can be shown that

$$\rho_g^g(F_1,\ldots,F_g) \leq \min_{i\neq j} \rho_2^2(F_i,F_j),$$

it implies that the separation among the g groups according to (1.12.6) is not less than the separation between any two of them. This measure also has the other desirable properties of a separation measure in that it is symmetric in its arguments and has a minimum value of zero at $F_1 = F_2 = \cdots = F_g$. Glick (1973b) also generalized the two-group result (1.12.3) by showing that $1 - 2eo(F)$ can be viewed as a separation measure for an arbitrary number g of groups when in equal proportions; see also Cleveland and Lachenbruch (1974).

In a series of papers, Krzanowski (1983a, 1984a, 1987a) has studied the use of Matusita's (1956) measure of distance in situations where the feature vector consists of continuous and discrete variables. His work is discussed in Section 7.4, where discriminant analysis in the case of mixed features is presented.

CHAPTER 2

Likelihood-Based Approaches to Discrimination

2.1 MAXIMUM LIKELIHOOD ESTIMATION OF GROUP PARAMETERS

We consider maximum likelihood estimation of the vector $\boldsymbol{\theta}$ containing all the unknown parameters in the parametric families (1.8.1) postulated for the g group-conditional densities of the feature vector $\mathbf{X}$. As in the previous work, we let $\mathbf{t}' = (\mathbf{y}_1,\ldots,\mathbf{y}_n)$ denote the observed training data, where $\mathbf{y}_j = (\mathbf{x}_j', \mathbf{z}_j')'$ for $j = 1,\ldots,n$. For training data $\mathbf{t}$ obtained by a separate sampling scheme, the likelihood function $L(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ is formed by evaluating at their observed values $\mathbf{x}_1,\ldots,\mathbf{x}_n$ the joint density of the feature vectors conditional on their known group-indicator vectors $\mathbf{z}_1,\ldots,\mathbf{z}_n$. We proceed here under the assumption that $\mathbf{y}_1,\ldots,\mathbf{y}_n$ denote the realized values of n independent training observations. Then the log likelihood function for $\boldsymbol{\theta}$ is given under (1.8.1) by

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} \log f_i(\mathbf{x}_j;\boldsymbol{\theta}_i), \tag{2.1.1}$$

where log denotes the natural logarithm. An estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained as a solution of the likelihood equation

$$\partial L(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0},$$

or, equivalently,

$$\partial \log L(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}. \tag{2.1.2}$$

Briefly, the aim of maximum likelihood estimation (Lehmann, 1980, 1983) is to determine an estimate for each n ($\hat{\boldsymbol{\theta}}$ in the present context) so that it defines a sequence of roots of the likelihood equation that is consistent and asymptotically efficient. Such a sequence is known to exist under suitable regularity conditions (Cramér, 1946). With probability tending to one, these roots correspond to local maxima in the interior of the parameter space. For estimation models in general, the likelihood usually has a global maximum in the interior of the parameter space. Then typically a sequence of roots of the likelihood equation with the desired asymptotic properties is provided by taking $\hat{\boldsymbol{\theta}}$ for each n to be the root that globally maximizes the likelihood; that is, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate. We will henceforth refer to $\hat{\boldsymbol{\theta}}$ as the maximum likelihood estimate, even though it may not globally maximize the likelihood. Indeed, in some of the examples on mixture models to be presented, the likelihood is unbounded. However, for these models, there may still exist under the usual regularity conditions a sequence of roots of the likelihood equation with the properties of consistency, efficiency, and asymptotic normality; see McLachlan and Basford (1988, Chapter 1).

For $\mathbf{t}$ obtained under mixture sampling, the log likelihood function for $\Psi = (\boldsymbol{\pi}', \boldsymbol{\theta}')'$ is given by

$$\log L(\boldsymbol{\theta}) + \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \log \pi_i.$$

It follows from consideration of the likelihood equation for Ψ that it is estimated by

$$\hat{\Psi} = (\hat{\boldsymbol{\pi}}', \hat{\boldsymbol{\theta}}')',$$

where $\hat{\boldsymbol{\theta}}$ is defined as before and $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_g)'$, and where

$$\begin{aligned} \hat{\pi}_i &= \sum_{j=1}^{n} z_{ij}/n \\ &= n_i/n \qquad (i = 1, \ldots, g). \end{aligned}$$

Given that a statistical model is at best an approximation to reality, it is worth considering here the behavior of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ if the postulated parametric structure for the group-conditional densities is not valid. Suppose now that the group-conditional densities $f_i(\mathbf{x})$ do not belong to the parametric families postulated by (1.8.1). Working with a mixture sampling scheme, the true mixture density of $\mathbf{X}$ can be expressed as

$$f_X(\mathbf{x}) = \sum_{i=1}^{g} \pi_{i,o} f_i(\mathbf{x}),$$

where $\pi_{i,o}$ denotes the true value of π_i $(i = 1, \ldots, g)$.

As seen previously, regardless of whether a mixture or separate sampling scheme applies, $\hat{\boldsymbol{\theta}}$ is obtained by consideration of the same function $\log L(\boldsymbol{\theta})$

given by (2.1.1). Following Hjort (1986a, 1986b), we have that as $n \to \infty$, $1/n$ times $\log L(\boldsymbol{\theta})$ tends almost surely to

$$\sum_{i=1}^{g} \pi_{i,o} \left\{ \int f_i(\mathbf{x}) \log f_i(\mathbf{x};\boldsymbol{\theta}_i) \right\} d\nu. \tag{2.1.3}$$

Suppose there is a unique value of $\boldsymbol{\theta}, \boldsymbol{\theta}_o$, that maximizes (2.1.3) with respect to $\boldsymbol{\theta}$. Then this value also minimizes the quantity

$$\sum_{i=1}^{g} \pi_{i,o} \left[\int f_i(\mathbf{x}) \log\{f_i(\mathbf{x})/f_i(\mathbf{x};\boldsymbol{\theta}_i\} \, d\nu \right],$$

which is a mixture in the true proportions $\pi_{1,o},\ldots,\pi_{g,o}$ of the Kullback–Leibler distances between the true and the postulated group-conditional densities of $\mathbf{X}$.

Under mild regularity conditions, it follows that if $\hat{\boldsymbol{\theta}}$ is chosen by maximization of $\log L(\boldsymbol{\theta})$, it tends almost surely to $\boldsymbol{\theta}_o$. Hence, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ under the invalid model (1.8.1) is still a meaningful estimator in that it is a consistent estimator of $\boldsymbol{\theta}_o$, the value of $\boldsymbol{\theta}$ that minimizes the Kullback–Leibler distances between the actual group-conditional densities of $\mathbf{X}$ and the postulated parametric families, mixed in the proportions in which the groups truly occur.

2.2 A BAYESIAN APPROACH

A review of the Bayesian approach to discriminant analysis has been given by Geisser (1966, 1982). This approach is based on the concept of the predictive density of the feature vector $\mathbf{X}$. The predictive density of $\mathbf{X}$ within group G_i is defined by

$$\hat{f}_i^{(P)}(\mathbf{x};\mathbf{t}) = \int f_i(\mathbf{x};\boldsymbol{\theta}_i) p(\boldsymbol{\theta} \mid \mathbf{t}) \, d\boldsymbol{\theta} \qquad (i = 1,\ldots,g), \tag{2.2.1}$$

where $p(\boldsymbol{\theta} \mid \mathbf{t})$ can be regarded either as some weighting function based on $\mathbf{t}$ or as a full Bayesian posterior density function for $\boldsymbol{\theta}$ based on $\mathbf{t}$ and a prior density $p(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$. In the latter case,

$$p(\boldsymbol{\theta} \mid \mathbf{t}) \propto p(\boldsymbol{\theta}) L(\boldsymbol{\theta};\mathbf{t}),$$

where $L(\boldsymbol{\theta};\mathbf{t})$ denotes the likelihood function for $\boldsymbol{\theta}$ formed from the training data $\mathbf{t}$. Note that for economy of notation, we are using $p(\cdot)$ here as a generic symbol for a density function. In the subsequent discussion, the vector $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)'$ defining the group-prior probabilities is taken to be specified, so that the vector $\boldsymbol{\Psi}$ of parameters to be estimated is reduced to $\boldsymbol{\theta}$. We therefore write the posterior probability of membership of the ith group $\tau_i(\mathbf{x};\boldsymbol{\Psi})$ as $\tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta})$.

The predictive estimate of $\tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta})$ is obtained by substituting the predictive estimates of the group-conditional densities in its defining expression (1.2.4) for $\tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta})$ to give

$$\hat{\tau}_i^{(P)}(\mathbf{x};\mathbf{t}) = \pi_i \hat{f}_i^{(P)}(\mathbf{x};\mathbf{t})/\hat{f}_X^{(P)}(\mathbf{x};\mathbf{t}) \qquad (i = 1,\ldots,g), \tag{2.2.2}$$

where

$$\hat{f}_X^{(P)}(\mathbf{x};\mathbf{t}) = \sum_{j=1}^{g} \pi_j \hat{f}_j^{(P)}(\mathbf{x};\mathbf{t}).$$

According to Aitchison and Dunsmore (1975, Chapter 11), the predictive approach was first presented explicitly by Geisser (1964) for multivariate normal group-conditional distributions, and in Dunsmore (1966). For moderately large or large-size training samples, the predictive and estimative approaches give similar results for the assessment of the posterior probabilities of group membership. However, for small sample sizes, there can be dramatic differences. This appears to have been first noted by Aitchison and Kay (1975). These two approaches are to be compared under normal models in Section 3.8. It will be seen there that if the estimates produced by the estimative approach are corrected for bias, then the differences between the two approaches are considerably reduced (Moran and Murphy, 1979). For further discussion of the estimation of $\tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta})$ in a Bayesian framework, the reader is referred to Critchley, Ford, and Rijal (1987).

We now consider the semi-Bayesian approach as adopted by Geisser (1967) and Enis and Geisser (1970) in the estimation of the log odds under normal models for $g = 2$ groups. With the semi-Bayesian approach to the estimation of $\tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta})$, its posterior distribution is calculated on the basis of the information in $\mathbf{t}$ but not $\mathbf{x}$. The mean of this posterior distribution so calculated is given by

$$\int \tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{t}) d\boldsymbol{\theta} \qquad (i = 1,\ldots,g). \tag{2.2.3}$$

Corresponding to a squared-error loss function, (2.2.3) can be used as an estimate of $\tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta})$. By using different loss functions, other estimates of $\tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta})$ can be obtained, for example, the median or the mode of the posterior distribution of the probability of membership of G_i.

It is of interest to contrast the estimate (2.2.3) of $\tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta})$ with the predictive estimate (2.2.2). Following Rigby (1982), we have from (2.2.1) and (2.2.2) that

$$\begin{aligned}\hat{\tau}_i^{(P)}(\mathbf{x};\mathbf{t}) &= \frac{\pi_i \hat{f}_i^{(P)}(\mathbf{x};\mathbf{t})}{\hat{f}_X^{(P)}(\mathbf{x};\mathbf{t})} = \int \frac{\pi_i f_i(\mathbf{x};\boldsymbol{\theta}_i) p(\boldsymbol{\theta} \mid \mathbf{t})}{p(\mathbf{x} \mid \mathbf{t})} d\boldsymbol{\theta} \\ &= \int \frac{\pi_i f_i(\mathbf{x};\boldsymbol{\theta}_i) p(\mathbf{x},\boldsymbol{\theta} \mid \mathbf{t})}{p(\mathbf{x} \mid \boldsymbol{\theta},\mathbf{t}) p(\mathbf{x} \mid \mathbf{t})} d\boldsymbol{\theta} = \int \tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x};\mathbf{t}) d\boldsymbol{\theta}.\end{aligned} \tag{2.2.4}$$

It can be seen from (2.2.4) that the predictive estimate of $\tau_i(\mathbf{x},\boldsymbol{\pi};\boldsymbol{\theta})$ corresponds to a fully Bayesian approach, as it averages $\tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta})$ over the posterior distribution of $\boldsymbol{\theta}$ given both $\mathbf{t}$ and $\mathbf{x}$. On comparing it with the semi-Bayesian estimate of $\tau_i(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\theta})$ given by (2.2.3), it follows that these two estimates will be practically the same if the information provided by $\mathbf{x}$ about $\boldsymbol{\theta}$ is negligible compared to that provided by $\mathbf{t}$.

2.3 ESTIMATION OF GROUP PROPORTIONS

We consider now the problem where the aim is to estimate the proportions $\pi_1,\ldots,\pi_g$ in which a mixture G of g distinct groups $G_1,\ldots,G_g$ occur. McLachlan and Basford (1988, Chapter 4) have given several examples where this problem arises. One example concerns crop-acreage estimation on the basis of remotely sensed observations on a mixture of several crops; see, for instance, Chhikara (1986) and the references therein. The problem is to estimate the acreage of a particular crop as a proportion of the total acreage. Training data are available on each of the crops to provide estimates of the unknown parameters in the distribution for an individual crop. Another example concerns the case study of Do and McLachlan (1984), where the aim was to assess the rat diet of owls through the estimation of the proportion of each of seven species of rats consumed.

If the training data $\mathbf{t}$ have been obtained by sampling from the mixture of interest G, then the proportion π_i can be estimated simply by its maximum likelihood estimate n_i/n, where n_i denotes the number of entities known to belong to G_i $(i = 1,\ldots,g)$. Therefore, we consider the problem in the context where the n_i provide no information on the proportions π_i. This would be the case if the training data were obtained by sampling separately from each of the groups or from some other mixture of these groups. In order to obtain information on the desired proportions π_i, it is supposed that there is available a random sample of size m, albeit unclassified, from the relevant mixture G. We let $\mathbf{x}_j$ $(j = n+1,\ldots,n+m)$ denote the observed feature vectors on these m unclassified entities having unknown group-indicator vectors $\mathbf{z}_j$ $(j = n+1,\ldots,n+m)$.

An obvious and computationally straightforward way of proceeding is to form a discriminant rule $r(\mathbf{x};\mathbf{t})$ from the classified training data $\mathbf{t}$, and then to apply it to the m unclassified entities with feature vectors $\mathbf{x}_j$ $(j = n+1,\ldots,n+m)$ to find the proportion assigned to the ith group G_i $(i = 1,\ldots,g)$. That is, if m_i denotes the number of the m unclassified entities assigned to G_i, then a rough estimate of π_i is provided by m_i/m $(i = 1,\ldots,g)$. Unless $r(\mathbf{x};\mathbf{t})$ is an infallible rule, m_i/m will be a biased estimator of π_i. For $g = 2$, it can be easily seen that, conditional on $\mathbf{t}$,

$$E(m_1/m) = \pi_1 ec_{11} + \pi_2 ec_{21} \tag{2.3.1}$$

and

$$E(m_2/m) = \pi_1 ec_{12} + \pi_2 ec_{22}, \tag{2.3.2}$$

with either equation giving the so-called discriminant analysis estimator of π_1,

$$\hat{\pi}_{1D} = (m_1/m - ec_{21})/(ec_{11} - ec_{21}),$$

as an unbiased estimator of π_1. In the equations above, the conditional allocation rates of $r(\mathbf{x};\mathbf{t})$, $ec_{ij}(F_i;\mathbf{t})$, are written simply as ec_{ij} for convenience $(i,j = 1,2)$. If $\hat{\pi}_D$ is outside [0,1], then it is assigned the appropriate value zero or one.

On considering (2.3.1) and (2.3.2) simultaneously, $\hat{\pi}_{1D}$ and $\hat{\pi}_{2D} = 1 - \hat{\pi}_{1D}$ can be expressed equivalently as

$$\hat{\boldsymbol{\pi}}_D = \mathbf{J}^{-1}(m_1/m,\ m_2/m)', \tag{2.3.3}$$

where $\hat{\boldsymbol{\pi}}_D = (\hat{\pi}_{1D}, \hat{\pi}_{2D})'$ and

$$\mathbf{J} = \begin{pmatrix} ec_{11} & ec_{21} \\ ec_{12} & ec_{22} \end{pmatrix}.$$

The result (2.3.3) can be generalized to $g > 2$ groups to give

$$\hat{\boldsymbol{\pi}}_D = \mathbf{J}^{-1}(m_1/m,\ldots,m_g/m)', \tag{2.3.4}$$

where the (i,j)th element of the confusion matrix $\mathbf{J}$ is equal to

$$(\mathbf{J})_{ij} = ec_{ji} \qquad (i,j = 1,\ldots,g).$$

According to Macdonald (1975), $\hat{\boldsymbol{\pi}}_D$ seems to have been first suggested by Worlund and Fredin (1962). For known conditional allocation rates $ec_{ij}(F_i;\mathbf{t})$, $\hat{\boldsymbol{\pi}}_D$ is the maximum likelihood estimate of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)'$ based on the proportions m_i/m $(i = 1,\ldots,g)$, and is unbiased. However, in practice, these conditional allocation rates are unknown and must be estimated before $\hat{\boldsymbol{\pi}}_D$ can be calculated from (2.3.4).

It can be seen that if a nonparametric rule $r(\mathbf{x};\mathbf{t})$ is used and the conditional allocation rates are estimated nonparametrically, then the discriminant analysis estimator $\hat{\boldsymbol{\pi}}_D$ can be made distribution-free. In this sense, it should be more robust than a parametric estimator of $\boldsymbol{\pi}$, for example, the maximum likelihood estimator $\hat{\boldsymbol{\pi}}$ whose computation is to be described in Section 2.7. The latter, which is based on the classified training data $\mathbf{t}$ in conjunction with the unclassified data $\mathbf{x}_j$ $(j = n+1,\ldots,n+m)$, is of course fully efficient if the assumed parametric structure holds. Ganesalingam and McLachlan (1981) have investigated the efficiency of $\hat{\boldsymbol{\pi}}_D$ relative to $\hat{\boldsymbol{\pi}}$ in the case of two groups in which the feature vector has a multivariate normal distribution with a common covariance matrix. They concluded that the relative efficiency of $\hat{\boldsymbol{\pi}}_D$ can be quite high provided the mixing proportions are not too disparate and n is not too small relative to m. More recently, for the same normal model, Lawoko and McLachlan (1989) have studied the bias of $\hat{\pi}_{1D}$ as a consequence of using estimates of the conditional allocation rates in its formation. They also considered the case where the classified training observations are correlated.

There are other methods of estimation of the mixing proportions. For a mixture of $g = 2$ groups, π_1 can be estimated also by

$$\hat{\pi}_{1M} = \{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_u^{-1}(\bar{\mathbf{x}}_u - \bar{\mathbf{x}}_2)\}/\{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_u^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\},$$

where

$$\bar{\mathbf{x}}_i = \sum_{j=1}^{n} z_{ij}\mathbf{x}_j/n_i \qquad (i = 1,2),$$

$$\bar{\mathbf{x}}_u = \sum_{j=n+1}^{n+m} \mathbf{x}_j/m,$$

and

$$\mathbf{S}_u = \sum_{j=n+1}^{n+m} (\mathbf{x}_j - \bar{\mathbf{x}}_u)(\mathbf{x}_j - \bar{\mathbf{x}}_u)'/(m-1).$$

The estimator $\hat{\pi}_{1M}$ can be viewed as the moment estimator of π_1 after transformation of the original feature data $\mathbf{x}_j$ from $\mathbf{R}^p$ to $\mathbf{R}$ by

$$\mathbf{x}_j'\mathbf{S}_u^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \qquad (j = 1,\ldots,n+m).$$

The asymptotic relative efficiency of $\hat{\pi}_{1M}$ has been derived by McLachlan (1982) for a mixture of two multivariate normal distributions with a common covariance matrix.

There are also minimum-distance estimators. The discriminant analysis, maximum likelihood, and moment estimators of the mixing proportions can all be obtained by using the method of minimum distance through an appropriate choice of the distance measure. A review of these various estimators of the mixing proportions can be found in McLachlan and Basford (1988, Chapter 4).

2.4 ESTIMATING DISEASE PREVALENCE

A situation where the model in the previous section is appropriate occurs in epidemiological studies, where an important aim is to estimate disease prevalence within a population. The groups represent the possible disease categories. Without some idea of prevalence, it is very difficult to plan prospective studies, to interpret retrospective studies, or to make rational health planning decisions; see Rogan and Gladen (1978). It is often impracticable to examine the entire population and so a random sample is taken. Further, it is usually too expensive and perhaps too arduous an experience for the individual being tested for a definitive classification of the disease to be made. Also, even if exhaustive physical and clinical tests were carried out, a true diagnosis may still not be possible. Hence, typically, the sample drawn from the population is allocated to the various groups on the basis of some straightforward but fallible diagnostic test, whose error rates are assessed by applying it to patients with

known disease classification. Screening programs often use tests in this way. Even if the prime purpose of the program is finding cases of disease rather than estimating prevalence, the performance of the test is still of interest.

The performance of a screening test designed to detect the presence of a single disease can be evaluated in terms of the sensitivity and specificity of the diagnostic rule $r(\mathbf{x};\mathbf{t})$, which are given conditional on $\mathbf{t}$ by $ec_{22}(F_2;\mathbf{t})$ and $ec_{11}(F_1;\mathbf{t})$, respectively. Here the two groups G_1 and G_2 refer to the absence or presence of the disease. An individual is assigned to G_1 or G_2 according as to whether the test is negative or positive; that is, according as $r(\mathbf{x};\mathbf{t})$ equals 1 or 2. We can write the sensitivity of the test (conditional on $\mathbf{t}$) as

$$\text{pr}\{r(\mathbf{X};\mathbf{t}) = 2 \mid Z_2 = 1\}, \tag{2.4.1}$$

and the specificity as

$$\text{pr}\{r(\mathbf{X};\mathbf{t}) = 1 \mid Z_1 = 1\},$$

where Z_i is one or zero according as the individual with feature vector $\mathbf{x}$ belongs to G_i or not ($i = 1,2$).

It is sometimes mistakenly assumed in the confusion over conditional probabilities that, because a test has a high sensitivity as given by the conditional probability (2.4.1), the reverse conditional probability

$$\text{pr}\{Z_2 = 1 \mid r(\mathbf{x};\mathbf{t}) = 2\} \tag{2.4.2}$$

must also be high. This is what Diaconis and Freedman (1981) have referred to in a general context as the "fallacy of the transposed conditional." Although a test may have high sensitivity, the conditional probability (2.4.2), called the predictive value of a positive test (PVP), may be small. Hence, the usefulness of a test is often evaluated in terms of the PVP and the PVN, where the latter denotes the predictive value of a negative test, defined by

$$\text{pr}\{Z_1 = 1 \mid r(\mathbf{x};\mathbf{t}) = 1\}.$$

By Bayes theorem, the PVP can be expressed as

$$\text{pr}\{Z_2 = 1 \mid r(\mathbf{x};\mathbf{t}) = 2\} = \pi_2 ec_{22}/\{\pi_2 ec_{22} + (1-\pi_2)(1-ec_{11})\}, \tag{2.4.3}$$

with a similar expression for the PVN. In (2.4.3), the sensitivity and specificity of the test are abbreviated to ec_{22} and ec_{11}, respectively. It can be seen from (2.4.3) that in order to evaluate the PVP and PVN of a test, the disease prevalence rate π_2, as well as the sensitivity and specificity, must be assessed.

Recently, Gastwirth (1987) established the asymptotic normality of the estimator of the PVP as given by (2.4.3), where π_2 is estimated according to (2.3.3) and where the sensitivity and specificity are replaced by independent binomial estimators formed by applying the test to subsequent data of known origin. A nonmedical example where the abovementioned methodology is applicable is with the use of lie detectors and the associated issues of veracity and admissibility of polygraph evidence in judicial and preemployment screening of applicants.

Where not explicitly stated, it is implicitly assumed in the above that the results are conditional on the training data t. In some screening applications, the diagnostic test may be based on a rather sophisticated discriminant rule $r(\mathbf{x};\mathbf{t})$ formed from t. For example, for the screening of keratoconjunctivitis sicca in rheumatoid arthritic patients, J. A. Anderson (1972) proposed a diagnostic test based on ten symptoms. However, with many screening tests, in particular presymptomatic ones, the diagnostic test is based on some ad hoc rule, generally using only one feature variable. In the latter situation, the variable is usually measured on a continuous scale and a threshold is imposed to define a positive test. The training data t are used then solely to assess the performance of the test for a given threshold. In an example considered by Boys and Dunsmore (1987), patients were designated as either malnourished or nonmalnourished according as their plasma cholesterol levels were less or greater than some threshold.

The choice of threshold in such tests depends on the role in which they are applied. With the ELISA test applied in the context of routine screening of blood donations for AIDS, the threshold for declaring the ELISA assay to be positive is set so that the test is highly sensitive at the expense of having rather low specificity. A high specificity is achieved subsequently by following a positive ELISA with a confirmatory Western blot test (Weiss et al., 1985).

Hand (1986c, 1987a) cautioned that, as the aims of screening and estimation of disease prevalence are somewhat different, the threshold should be chosen with the particular aim in mind. For prevalence estimation, the aim is to minimize the variance of $\hat{\pi}_2$, whereas with screening, the aim is to maximize the accuracy of the rule, that is, some function of the sensitivity and specificity, such as their sum. He investigated the choice of threshold separately for each aim, but using a different sampling scheme to that taken above. In his scheme, the entities of known origin were part of the unclassified data. They were obtained by sampling from each lot of m_i entities assigned to G_i by the fallible diagnostic test and then using an infallible rule to classify them correctly.

2.5 MISCLASSIFIED TRAINING DATA

It is usually assumed in applications of discriminant analysis that the training entities are correctly classified. The classification of a training set is often expensive and difficult, as noted in the previously presented examples on discriminant analysis in the context of medical diagnosis. Another applicable example concerns the classification of remotely sensed imagery data. An important consideration besides the expense and difficulty in procuring classified training observations is that the actual classification may well be subject to error. Indeed, the concept of a true diagnosis is probably inappropriate in some medical fields, and certainly in some such as psychiatry. In the remote-sensing example, say, of crop patterns, the classification of the training pixels may be undertaken visually and hence be prone to error.

In considering now the possibility of misclassification in the training set, we let $\alpha_i(\mathbf{x})$ denote the probability that an entity from group G_i and with feature vector $\mathbf{x}$ is misclassified in the formation of the training set ($i = 1,\ldots,g$). The misclassification is said to be nonrandom or random depending on whether the $\alpha_i(\mathbf{x})$ do or do not depend on the feature vector $\mathbf{x}$. The simple error structure of random misclassification,

$$\alpha_i(\mathbf{x}) \equiv \alpha_i \qquad (i = 1,\ldots,g), \tag{2.5.1}$$

may rise, for example, where the classification of the training data is made on the basis of machine output (for example, X-ray interpretation or blood test results) and either the output or interpretation for each entity is inaccurate in a way that is independent of its feature vector $\mathbf{x}$. In this same context of medical diagnosis, nonrandom misclassification should be more applicable than random if the classification of the patients in the training set was carried out by clinicians using symptoms closely related to the feature variables.

In his initial work on this problem, Lachenbruch (1966) considered the effect of random misclassification of training entities on the error rates of the sample linear discriminant rule obtained by plugging in the usual maximum likelihood estimates of the parameters in the case of two multivariate normal group-conditional distributions with a common covariance matrix. McLachlan (1972a) derived the asymptotic theory for this model under the additional assumption that one group is not misclassified. For instance, it is often reasonable to assume in the formation of a diagnostic rule that the training sample of healthy patients are all correctly classified. Lachenbruch (1979) subsequently showed that whereas random misclassification is not a serious issue for the sample normal-based linear discriminant rule if α_1 and α_2 are similar, it is for its quadratic counterpart. Lachenbruch (1974) and Chhikara and McKeon (1984) considered the problem under more general misclassification models to allow for nonrandom misclassification. The general conclusion is that ignoring errors in the classification of the training set can be quite harmful for random misclassification. For nonrandom misclassification, the error rates of the sample discriminant rule appear to be only slightly affected, although the optimism of its apparent error rates is considerably increased.

Random misclassification of training entities has been investigated further by Chittineni (1980, 1981) and by Michalek and Tripathi (1980), who also considered the effect of measurement error in the feature variables. More recently, Grayson (1987) has considered nonrandom misclassification of training entities in the context of two groups G_1 and G_2, representing healthy and unhealthy patients, respectively. He supposed that the health status of a patient as specified by the group-indicator vector $\mathbf{z}$ can only be ascertained unreliably. Recall that $z_i = (\mathbf{z})_i$ is one or zero, according as the entity belongs or does not belong to G_i ($i = 1,2$). We let $\tilde{\mathbf{z}}$ denote the value of $\mathbf{z}$ under the uncertain (noisy) classification. Corresponding to the posterior probability that a patient with feature vector $\mathbf{x}$ belongs to G_i, we let $\tau_i(\mathbf{x})$ be the probability that the ith element $\tilde{Z}_i$ of $\tilde{\mathbf{Z}}$ is one for a patient with feature vector $\mathbf{x}$ ($i = 1,2$). Under very

general conditions on the misclassification errors,

$$\alpha_1(\mathbf{x}) = \text{pr}\{\tilde{Z}_1 = 0 \mid Z_1 = 1, \mathbf{x}\}$$

and

$$\alpha_2(\mathbf{x}) = \text{pr}\{\tilde{Z}_2 = 0 \mid Z_2 = 1, \mathbf{x}\},$$

Grayson (1987) showed that the likelihood ratio for arbitrary group-conditional densities is not ordinally affected. That is, $\text{logit}\{\tau_i(\mathbf{x})\}$ is a monotonic function of $\text{logit}\{\tilde{\tau}_i(\mathbf{x})\}$. Thus, the class of admissible decision rules is unaffected by this error structure in the classification of the training data. As an illustration of a consequence of the monotonicity of the scales, Grayson (1987) gave an example where there is a need to select the 40 most ill patients (only 40 hospital beds being available). The same patients would be chosen regardless of whether the $\tau_i(\mathbf{x})$ or the $\tilde{\tau}_i(\mathbf{x})$ were used.

2.6 PARTIALLY CLASSIFIED TRAINING DATA

In this section, we consider the problem of forming a discriminant rule, using classified data in conjunction with data on entities unclassified with respect to the g underlying groups $G_1, \ldots, G_g$. We will see there is a number of separate problems in discriminant analysis that fall within this context.

Consistent with our previous notation, we let $\mathbf{t}' = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ contain the information on the classified entities, where $\mathbf{y}_j = (\mathbf{x}_j', \mathbf{z}_j')'$, and $\mathbf{z}_j$ denotes the known group-indicator vector for the jth entity with observed feature vector $\mathbf{x}_j$ $(j = 1, \ldots, n)$. It is supposed that in addition to the classified training data $\mathbf{t}$, there are available the feature vectors $\mathbf{x}_j$ $(j = n+1, \ldots, n+m)$ observed on m entities of unknown origin. The latter are assumed to have been drawn from a mixture G of $G_1, \ldots, G_g$ in some unknown proportions $\pi_1, \ldots, \pi_g$. We let

$$\mathbf{t}_u' = (\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m}).$$

The unknown group-indicator vector associated with the unclassified feature vector $\mathbf{x}_j$ is denoted by $\mathbf{z}_j$ $(j = n+1, \ldots, n+m)$. If the aim is solely to make a tentative or outright allocation of the m unclassified entities to $G_1, \ldots, G_g$, then we can proceed as discussed in the previous sections. A discriminant rule $r(\mathbf{x}; \mathbf{t})$ can be formed from the classified training data $\mathbf{t}$ and then applied in turn to each of the m unclassified entities with feature vector $\mathbf{x}_j$ $(j = n+1, \ldots, n+m)$ to produce an assessment of its posterior probabilities of group membership and, if required, an outright allocation.

However, in some situations, it is desired to construct a discriminant rule using the unclassified data $\mathbf{t}_u$ in conjunction with the classified training set $\mathbf{t}$. The updating problem falls within this framework. The observing of feature vectors and the subsequent allocation of the corresponding unclassified entities is an ongoing process and, after a certain number of unclassified observations has been obtained, the discriminant rule is updated on the basis of

all the observed data. In most updating problems, it is the allocation of unclassified entities subsequent to those whose features have been observed that is of prime concern. Those unclassified entities whose features have been observed may be reallocated as part of the updating process, but their new allocations are generally of no practical consequence in their own right. The latter would be the case if irrevocable decisions had to be made on the basis of the original allocations. For instance, in medical diagnosis, one does not always have the luxury of being able to wait until further information becomes available before making a decision.

If there are sufficiently many classified observations available from each of the groups, then updating may not be a worthwhile exercise: However, often in practice, there are impediments to the procurement of entities of known origin, which limit the number of classified entities available. We have seen in the examples on medical diagnosis given in the previous sections that it may be physically inconvenient to the patient, as well as very expensive, to attempt to make a true diagnosis of the diseased status. In such situations where there is an adequate number of unclassified entities whose features can be measured easily, updating provides a way of improving the performance of the discriminant rule formed solely from the limited classified training data. Of course, as an unclassified observation contains less information than a classified one, many unclassified entities may be needed to achieve an improvement of practical consequence; see Ganesalingam and McLachlan (1978, 1979), O'Neill (1978), and McLachlan and Ganesalingam (1982). Their work was obtained under the assumption of multivariate normality with a common covariance matrix for the two group-conditional distributions. Amoh (1985) later considered the case of inverse normal group-conditional distributions.

Updating procedures appropriate for nonnormal group-conditional densities have been suggested by Murray and Titterington (1978), who expounded various approaches using nonparametric kernel methods, and J. A. Anderson (1979), who gave a method for the logistic discriminant rule. A Bayesian approach to the problem was considered by Titterington (1976), who also considered sequential updating. The more recent work of Smith and Makov (1978) and their other papers on the Bayesian approach to the finite mixture problem, where the observations are obtained sequentially, are covered in Titterington, Smith, and Makov (1985, Chapter 6).

Another problem where estimation on the basis of both classified and unclassified data arises is in the assessment of the proportions in which g groups occur in a mixture G. This problem was discussed in Section 2.3. It is supposed that the classified data have been obtained by a scheme, such as separate sampling, under which they provide no information on the mixing proportions. For the purposes of estimation of the latter, a random but unclassified sample is available from the mixture G. The mixing proportions can be estimated then by the discriminant analysis estimator (2.3.4). However, this estimator is not fully efficient in cases where a parametric family is postulated for the group-conditional distributions. Hence, in such cases, maximum likelihood es-