STATISTICAL DISCRIMINATION ANALYSIS

by

Farouk Elbishti

A thesis submitted for the degree of M.Sc. in the University of Glasgow.

December, 1967.

CONTENTS

ACKNOWLEDGEMENTS

I should like to express my gratitude to Professor J. Aitchison for suggesting the subject of this thesis, and for his general supervision and encouragement throughout the preparation of this thesis. I should like also to express my thanks to Professor S. D. Silvey for his help and encouragement.

Chapter 1

INTRODUCTION

1.1. The problem of discrimination

The problem of discrimination, also known as the problem of classification or allocation, may be described in the following general terms. Previous work has separated a number of individual experimental units (individuals) into k classes or categories, labelled 1, ..., k, these being n_i individuals in class i ($i = 1, \ldots, k$). On each of the $n = n_1 + \ldots + n_k$ individuals the same m characteristics have been measured. The characteristics may be qualitative, such as the presence or absence of a headache or quantitative, such as diastolic blood pressure. The data available are therefore n×m-dimensional vectors, $\underset{\sim}{x}_{ij}$ being the vector associated with the j^{th} individual in the i^{th} class ($j = 1, \ldots, n_i$; $i = 1, \ldots, k$). For easy reference we shall denote the complete set of data by $z = \{\underset{\sim}{x}_{ij} : j = 1, \ldots, n_i;\ i = 1, \ldots, k\}$. Almost invariably there is appreciable variability in the vectors, even within the vectors of a single class. A new unclassified individual with vector observation $\underset{\sim}{x}$ is now under scrutiny. The problem of discrimination is : to which of the classes 1, ..., k does this individual belong?

Fig. 1a gives a graphical illustration of a typical set z of data for the case of two-dimensional vectors. More generally the picture is one of k clusters of points in an m-dimensional record or sample space, with possible overlaps of the clusters. A discriminant procedure can then be defined as a partition of this m-dimensional record space into k regions $A_1, \ldots, A_k$ with the classification rule : if $x \in A_i$ classify the unit as from class i ($i = 1, \ldots, k$). See Fig. 1b.

There are two requirements for a satisfactory solution of such a problem. First, we must be able to postulate a suitable model for the generation of the data. This model must be probabilistic in order to explain the variability in the data. Secondly, we must be able to define what is meant by good discrimination; this will involve consideration of the consequences of misclassification of various kinds. The particular type of model and the criterion of good discrimination must depend on the particular practical situation under investigation.

The problem of discrimination can be formally set out as one of multiple hypothesis testing, the k hypotheses involved being: $\underline{x}$ arises from the distribution associated with the i^{th} class ($i = 1, \ldots, k$). The theory of testing many hypotheses is, however, one of the areas of statistics where there is no general consensus of opinion and so there appears to be no great advantage in such a formulation. It seems better to exploit whatever particular aspects the discrimination problem holds.

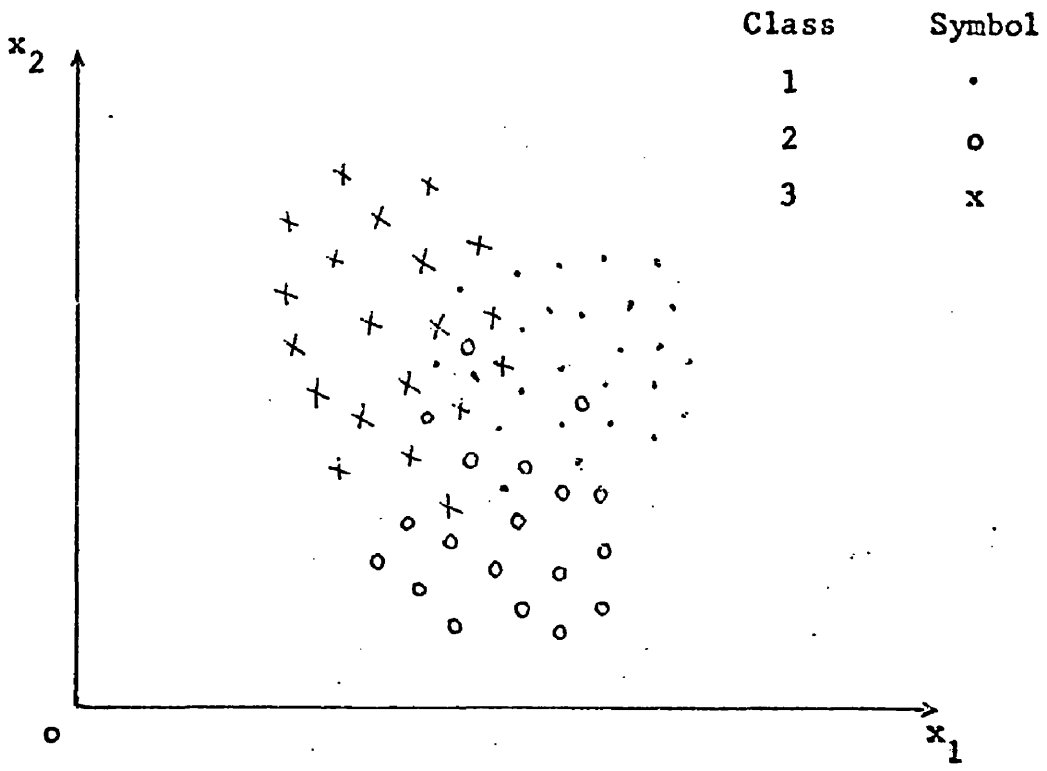


Fig. 1a. A typical picture of the record space in two-dimension.

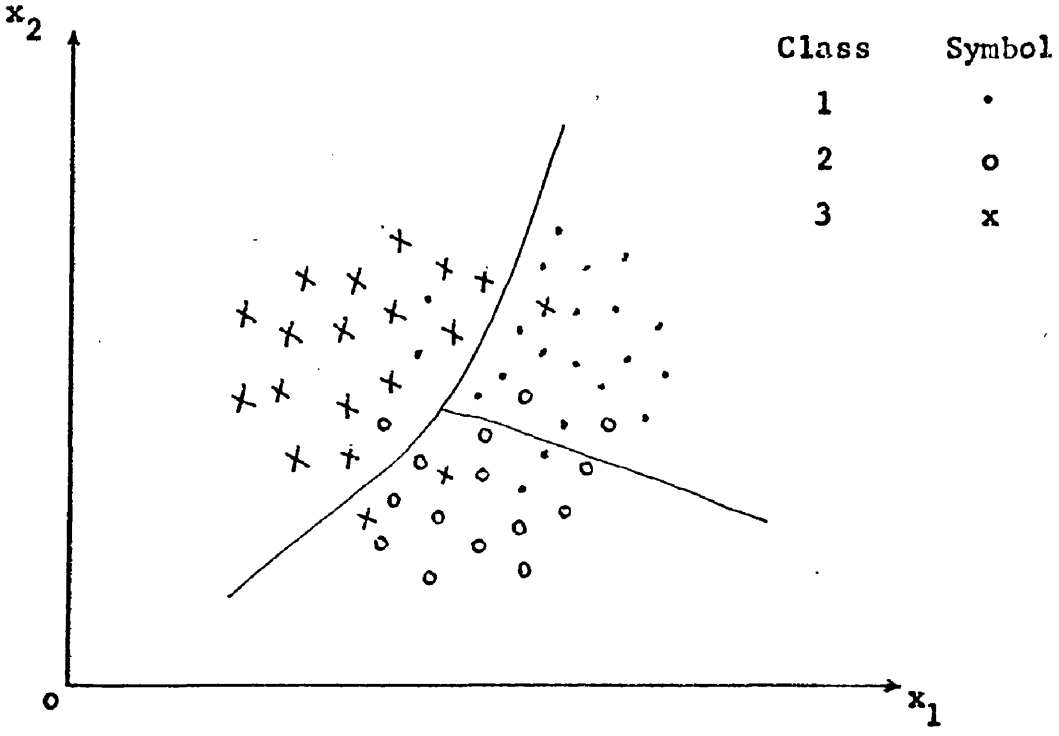


Fig. 1b. A typical partition of the record space for discrimination purposes.

1.2. Examples of fields of application

(i) Taxonomy. The first effective statistical treatment of discrimination was presented by Fisher (1936) in a problem of botanical taxonomy. This was the problem of discriminating between the two plants Iris setosa and Iris versicolor, on the basis of four measurements: sepal width, sepal length, petal width and petal length. Similar taxonomic problems occur in other branches of science. Another familiar early study is one in anthropology by Martin (1936); this involves an investigation of a series of Egyptian mandibles, the vector of characteristics being 6-dimensional.

(ii) Medical diagnosis. One of the more promising recent applications of statistical discriminant analysis and one which seems likely to grow in importance, is that of diagnosis of diseases; see Bailey (1965), Boyle (1965), Ledley and Lusted (1962), Radhakrishna (1964), and Warner (1961). Here the observation vector $\underset{\sim}{x}$ describes the state (signs, symptoms, results of clinical trials, medical history, etc.) of a patient as yet undiagnosed, and that data z are the set of such state vectors recorded in the past for diagnosed patients within a relevant class of diseases.

The difficulties in this field of application are obvious and considerable. The state vector required for careful diagnosis appears in most cases to be of high dimension. The problem can be/...

/be complicated by the cost and time of observing some of the elements of the vector. It is certainly a problem which when eventually properly formulated will involve the use of large automatic computers for its operation ; see Baron (1965), Boyle (1965)

(iii) Locating faults. An interesting and less familiar application is in the location of a fault in a machine. Here we have a number of sources of fault which give rise to various symptoms. The problem is to try to locate the fault efficiently on the basis of the symptom vector $\underline{x}$ presented, and the past history z of located faults.

1.3. Outline of the thesis

Chapter 2 deals with so-called classical discrimination procedures. The main idea underlying these is the attempt to construct some linear combination of the elements of the state vector to form a linear discriminant. The magnitude of this discriminant for the state vector $\underline{x}$ of the new individual is then used as a means of allocating the individual to his class. The theory is entirely based on the assumption of multivariate normal distributions for the description of the variability of the data.

In Chapter 3 we discuss Bayesian discrimination procedures. For the operation of these procedures some a priori information about the relative plausibilities of the various classes is required. This prior information is subsequently converted, after the observation of x, into a posterior appraisal of the various classes. The allocation is based on this posterior appraisal, /.....

/appraisal, possibly taking into account the relative seriousness of the various types of misclassification. In the literature on Bayesian discrimination the data z are usually assumed to be extensive and this assumption is used to make plausible the use of known distributions in the basic Bayesian model. At the end of Chapter 3 we suggest how this assumption might be relaxed.

In the next two chapters procedures of more recent origin are discussed. In Chapter 4 we consider two procedures - an order statistic procedure and a convex-hull procedure, both recently presented by Kendall (1965). The subject of Chapter 5 is a review of methods suggested by Sebestyen (1962), in particular a distance or similarity index procedure and a non-linear procedure.

Where possible, the procedures are illustrated by examples and in Chapter 6 we present some general comments and conclusions.

Chapter 2

CLASSICAL DISCRIMINATION PROCEDURES

2.1. Introduction

The first discrimination procedure was presented by Fisher (1936). This was for two classes and based on the assumption of multivariate normality. The main ideas underlying his procedure were as follows. If two clusters of points in m-dimensional space are roughly ellipsoidal in shape then an appropriate way to attempt to separate them is by means of a hyperplane. The equation of such a hyperplane involves a linear combination $\beta'\underline{x}$ of the elements of the state vector $\underline{x}$ of an individual. This combination will be normally distributed whether the individual is class 1 or class 2. The choice of β is at our disposal. Each β gives two one-dimensional normal distributions for $\beta'\underline{x}$, one on a class 1 assumption and one on a class 2 assumption. Now the amount of separation of two one-dimensional normal distributions $N(\lambda_1,\sigma^2)$ and $N(\lambda_2, \sigma^2)$ increases as $|\lambda_1-\lambda_2|/\sigma$ increases, as shown in Fig. 2a, b and c. It therefore seems plausible to take as appropriate values of β in the discriminant $\beta'x$ that β which maximises

$$\frac{|\beta' x_{1.} - \beta' x_{2.}|}{\sqrt{(\beta' S \beta)}}$$

where $x_{1.}$, $x_{2.}$ are the means of the sets of class 1 and class 2 vectors, and S is the pooled sample variance-covariance matrix.

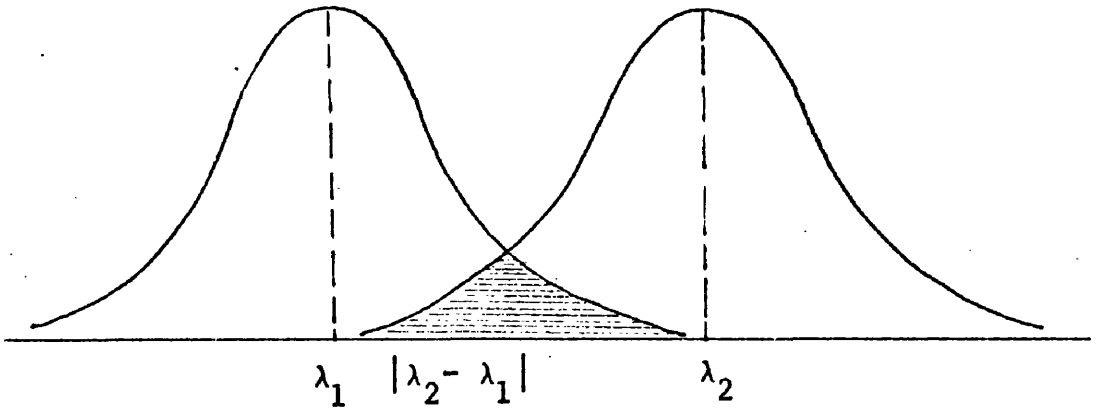


Fig. 2a.

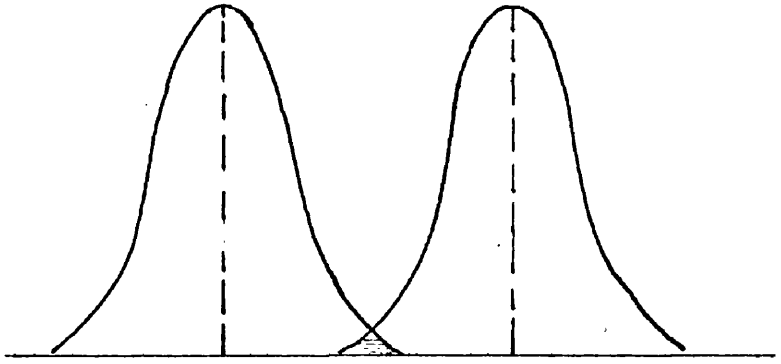


Fig. 2b. where σ is smaller

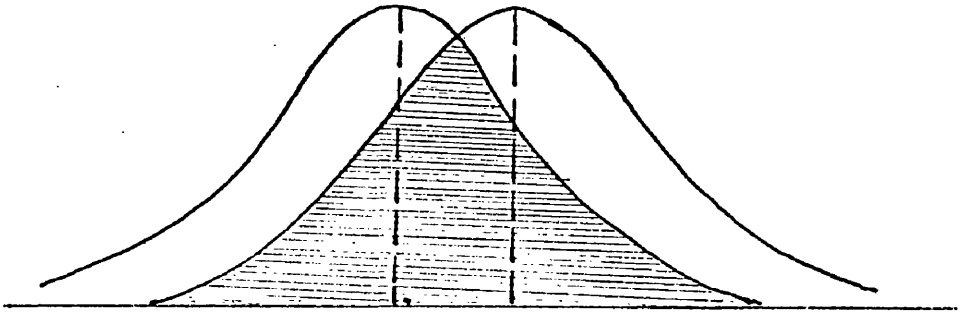


Fig. 2c. where $|\lambda_2 - \lambda_1|$ is smaller

In this chapter we first present Fisher's procedure, and then various other versions of it based on likelihood-ratio arguments by Wald (1944), Smith (1945) and Anderson (1957).

2.2. Notation

We make the assumptions of multivariate normality and of independence of the state vectors. Thus our model is the following.

The state vectors in $z = \{\underset{\sim}{x}_{ij} : j = 1, \ldots, n_i;\ i = 1, 2\}$ are independent,

$$x_{1j} \ (j = 1, \ldots, n_1) \text{ being } N(\mu_1, \Sigma),$$
$$x_{2j} \ (j = 1, \ldots, n_2) \text{ being } N(\mu_2, \Sigma),$$

Note that there is an assumption of equality of variance-covariance matrices. We write

$$x_{1\cdot} = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1j} ,$$

$$x_{2\cdot} = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j} ,$$

$$S = \frac{1}{\Sigma n_i - k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - x_{i\cdot})(x_{ij} - x_{i\cdot})' ,$$

for the usual estimates of μ_1, μ_2 and Σ.

2.3. Fisher's approach

If x is $N(\mu, \Sigma)$ then $\beta'x$ is $N(\beta'x, \beta'\Sigma\beta)$ and the mean and variance of the distribution of $\beta'x$ are estimated by

(i) $\beta'x_{1.}$ and $\beta'S\beta$ if x is from class 1 ,

(ii) $\beta'x_{2.}$ and $\beta'S\beta$ if x is from class 2.

The use of $\beta'x$ as a discriminant gives an estimated "measure of separability" of these two distributions :

$$|\beta'(x_{1.} - x_{2.})| \,/\sqrt{(\beta'S\beta)}.$$

The choice of β to maximize this "measure of separability" is given by,

$$\frac{\partial}{\partial \beta} \quad \frac{\beta'(x_{1.} - x_{2.})}{\sqrt{(\beta'S\beta)}} = 0 ,$$

i.e.

$$(x_{1.} - x_{2.})\,\beta'S\beta - \beta'(x_{1.} - x_{2.})\,S\beta = 0 .$$

By noting that $\beta'S\beta / \beta'(x_{1.} - x_{2.})$ is a factor constant for all the unknown coefficients, we see that the required coefficients are proportional to the solutions of the equations.

$$(x_{1.} - x_{2.}) = S\beta.$$

Thus Fisher's discrimination rule chooses a constant c and then allocates x to class 1 if

$$(x_{1.} - x_{2.})'\, S^{-1}\, x > c ;$$

otherwise x is allocated to class 2.

2.4. Wald's contribution

Wald's contribution to the problem of discrimination was essentially to point out that a likelihood-ratio criterion leads, for the case of known normal distributions, to the use of a linear discriminant.

The likelihood-ratio criterion is the following. If the density functions associated with class 1 and 2 are $p_1(x)$ and $p_2(x)$ then the likelihood-ratio is defined as

$$\Lambda(x) = p_1(x)/p_2(x) ,$$

and the rule of classification is : allocate x to class 1 if this ratio is greater than a given number c; otherwise allocate it to class 2, Welch (1939). For normal density functions,

$$p_i(x) = \frac{1}{(2\pi)^{\frac{1}{2}n_i}|\Sigma|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\} \quad (i = 1, 2)$$

so that we are interested in an equality such as

$$\Lambda(x) = \frac{\exp\{-\tfrac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\}}{\exp\{-\tfrac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\}} > c$$

Since the logarithmic function is a monotonic increasing function an equivalent inequality can be obtained in terms of logarithm as

$$-\tfrac{1}{2}\{(x-\mu_1)'\Sigma^{-1}(x-\mu_1) - (x-\mu_2)'\Sigma^{-1}(x-\mu_2)\} > \log c = c'$$

i.e. $x'\Sigma^{-1}(\mu_1-\mu_2) - \tfrac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) > c'$.

The first term of the above inequality is a linear function of the components of the observation vector; it is a linear discriminant. The second term is, of course, a known constant and can be absorbed into the constant on the right-hand side of the inequality to give procedures of the following form : if

$$x' \Sigma^{-1}(\mu_1 - \mu_2) > c'' ,$$

a constant, allocate to class 1 ; otherwise allocated to class 2.

Note that the data z are not relevant to this case since the assumption is that the class distributions are known.

2.5. Smith's likelihood-ratio approach

The difficulty of Wald's analysis of the problem of discrimination is that we seldom know the distributions associated with the two classes. Smith's solution to this difficulty was to assume that the class variabilities are described by classes of density functions

$$\{p_i(x|\theta_i) \ : \ \theta_i \in \Theta_i\} \qquad (i = 1, 2)$$

of known (in his treatment, normal) forms, to obtain from the data z, estimates $\hat{\theta}_1(z)$ and $\hat{\theta}_2(z)$ of the unknown parameters θ_1 and θ_2 and to apply the Wald criterion with

$$\Lambda(x, z) = p_1\left(x|\hat{\theta}_1(z)\right) \Big/ p_2\left(x|\hat{\theta}_2(z)\right)$$

For the case where $\underset{\sim}{x}_{ij}$ $(j = 1, \ldots, n_i \ ; \ i = 1, 2)$ are independent $N(\mu_i, \Sigma)$ $(i = 1, 2)$ we have

$$\Lambda(x, z) = \frac{\exp\left[-\frac{1}{2}(x-x_{1.})' S^{-1}(x-x_{1.})\right]}{\exp\left[-\frac{1}{2}(x-x_{2.})' S^{-1}(x-x_{2.})\right]} > c$$

i.e. $$-\frac{1}{2}\left[(x-x_{1.})' S^{-1}(x-x_{1.}) - (x-x_{2.})' S^{-1}(x-x_{2.})\right] > \log c = c'$$

i.e. $$x' S^{-1}(x_{1.}-x_{2.}) - \frac{1}{2}(x_{1.}-x_{2.})' S^{-1}(x_{1.}-x_{2.}) > c'$$

Since the second term is constant, we can absorb it in the right hand side with the constant c'. Then the rule of discrimination is: allocate x to class 1 if $x' S^{-1}(x_{1.}-x_{2.}) > c''$ otherwise allocate it to class 2. Thus Smith's approach gives exactly the same linear discriminant as Fisher's original approach.

Smith also considers the case where the variance-covariance matrices could be different, that is where $\underline{x}_{ij}$ $(j = 1, \ldots, n_i)$ are independent $N(\mu_i, \Sigma_i)$ $(i = 1, 2)$. Writing S_1 and S_2 for the estimators of Σ_1 & Σ_2 so that

$$S_i = \frac{1}{n_i - 1} \Sigma(x_{ij}-x_{i.})(x_{ij}-x_{i.})' \qquad (i = 1, 2)$$

we have

$$\Lambda(x, z) = \frac{\exp\left[-\tfrac{1}{2}(x-x_{1.})' S_1^{-1}(x-x_{1.})\right]}{\exp\left[-\tfrac{1}{2}(x-x_{2.})' S_2^{-1}(x-x_{2.})\right]} > c'$$

leading to

$$\tfrac{1}{2}x'(S_1^{-1} - S_2^{-1})x - 2x'(x_{2.}S_2^{-1} - x_{1.}S_1^{-1}) > c'.$$

Here the rule of discrimination is : allocate x to class 1 if x satisfies the above inequality: otherwise allocate to class 2.

2.6. Anderson's likelihood-ratio criterion

Smith had overcome the limitations of Wald's assumption of known class distributions by substituting estimates for unknown parameters in the Wald likelihood-ratio criterion. One point made by Anderson (1957) is that Smith's substitution is in a sense incomplete in that he absorbs the second term $(x_{1.} - x_{2.})' S^{-1}(x_{1.} - x_{2.})$ of the inequality into the constant of the inequality. Anderson suggests that it may be more reasonable to carry the substitution into this second term and so use a discrimination procedure as follows ; if

$$x'S^{-1}(x_{1.} - x_{2.}) - \tfrac{1}{2}(x_{1.} - x_{2.})' S^{-1}(x_{1.} - x_{2.}) > c$$

allocate to class 1 ; otherwise allocate to class 2.

But Anderson goes further and shows that it is possible to apply a generalised likelihood-ratio criterion to the problem, the likelihood function being formed for the complete set of data, z and x. More specifically, if the two classes of density functions are

$$\{P_i(\cdot|\theta_i) : \theta_i \in \Theta_i\} \quad (i = 1, 2)$$

the generalised-likelihood ratio is defined as

$$\Lambda(x,z) = \frac{\sup_{\theta_1 \times \theta_2} p_1(x|\theta_1) \prod_{i=1}^{2} \prod_{j=1}^{n_i} p_i(x_{ij}|\theta_i)}{\sup_{\theta_1 \times \theta_2} p_2(x|\theta_2) \prod_{i=1}^{2} \prod_{j=1}^{n_i} p_i(x_{ij}|\theta_i)}$$

A discrimination procedure is then of the form : if $\Lambda(x,z) > c$ allocate to class 1; otherwise allocate to class 2.

For the multivariate normal case this means that he considers $x, x_{11}, \ldots, x_{1n_1}$ as being independent $N(\mu_1, \Sigma)$ and $x_{21}, \ldots, x_{2n_2}$ as being independent $N(\mu_2, \Sigma)$ as class 1 hypothesis against the class 2 hypothesis that $x_{11}, \ldots, x_{1n_1}$ are drawn from $N(\mu_1, \Sigma)$ and $x, x_{21}, \ldots, x_{2n_2}$ are drawn from $N(\mu_2, \Sigma)$ with μ_1, μ_2 and Σ unspecified.

Under the class 1 hypothesis the maximum likelihood estimates of μ_1, μ_2 and Σ are

$$\hat{\mu}_{11} = \frac{(n_1 x_{1\cdot} + x)}{n_1 + 1} , \qquad \hat{\mu}_{12} = x_{2\cdot} ,$$

$$\hat{\Sigma}_1 = \frac{1}{n_1 + n_2 + 1} \{ \sum_{j=1}^{n_1} (x_{1j} - \hat{\mu}_{11})(x_{1j} - \hat{\mu}_{11})' + (x - \hat{\mu}_{11})(x - \hat{\mu}_{11})' + \sum_{j=1}^{n_2} (x_{2j} - \hat{\mu}_{12})(x_{2j} - \hat{\mu}_{12}) \} .$$

Since

$$\sum_{j=1}^{n_1} (x_{1j} - \hat{\mu}_{11})(x_{1j} - \hat{\mu}_{11})' + (x - \hat{\mu}_{11})(x - \hat{\mu}_{11})'$$

$$= \Sigma (x_{1j} - x_{1\cdot})(x_{1j} - x_{1\cdot})' + n_1 (x_{1\cdot} - \hat{\mu}_{11})(x_{1\cdot} - \hat{\mu}_{11})' + (x - \hat{\mu}_{11})(x - \hat{\mu}_{11})'$$

$$= \Sigma (x_{1j} - x_{1\cdot})(x_{1j} - x_{1\cdot})' + \frac{n_1}{n_1 + 1} (x - x_{1\cdot})(x - x_{1\cdot})'$$

We can write $\hat{\Sigma}_1$ as

$$\hat{\Sigma}_1 = \frac{1}{n_1 + n_2 + 1} \left[\frac{n_1}{n_1 + 1} (x - x_{1\cdot})(x - x_{1\cdot})' + H \right] ,$$

where $$H = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - x_{i\cdot})(x_{ij} - x_{i\cdot})' .$$

Under the assumption of the class 2 hypothesis we find that the maximum likelihood estimates of the parameters are

$$\hat{\mu}_{21} = x_{1\cdot} , \qquad \hat{\mu}_{22} = \frac{n_2 x_{2\cdot} + x}{n_2 + 1}$$

and $$\hat{\Sigma}_2 = \frac{1}{n_1 + n_2 + 1} \left[\frac{n_2}{n_2 + 1} (x - x_{2\cdot})(x - x_{2\cdot})' + H \right]$$

The likelihood ratio criterion $\Lambda(x, z)$ is then given by

$$\left[\Lambda(x, z)\right]^{2/(n_1+n_2+1)} = \frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|} = \frac{\left| H + \frac{n_2}{n_2+1} (x - x_{2.})(x - x_{2.})' \right|}{\left| H + \frac{n_1}{n_2+1} (x - x_{1.})(x - x_{1.})' \right|}$$

$$\text{i.e.} \quad \frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|} = \frac{1 + \frac{n_2}{n_2+1} (x - x_{2.})' H^{-1} (x - x_{2.})}{1 + \frac{n_1}{n_1+1} (x - x_{1.})' H^{-1} (x - x_{1.})} .$$

2.7. Illustration

To illustrate these different methods we give this example [Kendall (1957)].

A group of 25 normal and 25 psychotic individuals were given certain tests, and for each individual a size and shape variable x and y were determined. The results are shown in the following table.

Normals		Psychotics		Normals		Psychotics	
x	y	x	y	x	y	x	y
22	6	24	38	13	13	3	12
20	14	19	36	20	14	10	51
23	9	11	43	19	13	22	22
23	1	6	60	20	11	11	30
17	8	9	32	18	17	6	30
24	9	10	17	20	7	20	61
23	13	3	17	23	6	20	43
18	18	15	56	23	23	15	43
22	16	14	43	25	4	5	53
19	18	20	8	25	5	10	43
20	17	8	46	21	12	13	19
20	31	20	62	23	7	12	4
21	9	14	36				

	Normals	Psychotics
Mean of x	$x_{1.} = 20.8$	$x_{2.} = 12.8$
Mean of y	$y_{1.} = 12.32$	$y_{2.} = 36.4$
Variance of x	6.92	36.75
Covariance	- 5.27	13.92
Variance of y	40.89	287.92
Correlation	- 0.3	0.14

We can see from the last table that the correlation is not significant but the variances are significantly different. So it seems that a quadratic discriminant function will be better than a linear one.

(i) Fisher's technique

To apply Fisher's linear discriminant we have to assume that the variances and covariances are the same, so it is necessary to get a common value for the two classes. The most natural one to take is Fisher's variance within groups, which is simply the weighted mean of the variances for normals and psychotics, the weights being the number of degrees of freedom.

$$\text{Common Variance of } x = \frac{24 \times 6.92 + 24 \times 36.75}{48} = 21.83$$

$$\text{Common covariance} = \frac{24 \times -5.27 + 24 \times 13.72}{48} = 4.33$$

$$\text{Common Variance of } y = \frac{24 \times 40.89 + 24 \times 287.92}{48} = 164.4$$

∴ The linear discriminant function is

$$\Lambda(x, y) = 5x - 2y - 36.$$

By applying this linear function we found that Λ is positive for all the normals giving a zero empirical error of misclassification of normals, and negative for all but 4 of psychotics giving a 16% empirical error of misclassifying psychotics.

(ii) Smith's technique

By using Smith's method, we get the quadratic discriminant function

$$\Lambda(x, y) = \left(\frac{x - 23}{2}\right)^2 + \left(\frac{y - 8}{5}\right)^2 - 16$$

By applying this function we found that $\Lambda(x, y)$ is negative for all but 2 of the normals giving an error of 8% and positive for all but 2 of the psychotics giving an error of 8%.

(iii) Anderson's technique

By applying Anderson's method to this example, where

$$H^{-1} = \begin{bmatrix} 0.046 & -0.001 \\ -0.001 & 0.006 \end{bmatrix},$$

$$\bar{x}_1 = (20.8 \quad 12.32)$$

$$\bar{x}_2 = (12.8 \quad 36.4)$$

We found that for all the normals the ratio is greater than one giving zero error, while it is less than one for all psychotics except 4 giving 16% error. We notice that the same persons who were misclassified by Fisher's method were misclassified by this method. So this technique gives for this example the same result as Fisher's discriminant function.

Chapter 3

BAYESIAN DISCRIMINATION PROCEDURES

3.1. Introduction

The pattern of the development of Bayesian discriminant procedures is very similar to that of classical or frequentist discrimination procedures from Wald to Anderson - first, the assumption of known class distributions, secondly the substitution of estimated parameter values within the procedure developed for the first case, and then a full Bayesian analysis.

We first recall the result in conditional probability theory, commonly known as Bayes's theorem. If a class of density functions $\{p(\cdot|\theta) : \theta \in \Theta\}$ on a record space X form the possible descriptions of an experiment and if a density function $\pi(\theta)$ can be assumed to express the assessment of the uncertainty about the parameter prior to experimentation, then the assessment posterior to the observation of the record x in the experiment is given by the density function $\pi(\theta|x)$, where

$$\pi(\theta|x) = \frac{p(x|\theta)\ \pi(\theta)}{p(x)} \qquad (\theta \in \Theta)$$

where, of course,

$$p(x) = \int_{\Theta} p(x|\theta)\ \pi(\theta)\ d\theta \qquad (x \in X)$$

3.2. Case of known class distributions

On the assumption of known class distributions the data z are irrelevant, and the parametric space Θ can be identified with the set $I = \{1, 2, \ldots, k\}$, the set of class labels. We suppose that the i^{th} class distribution has known density function $p(\cdot|i)$ on X and that the a priori probability that an individual is from class i is $\pi(i)$. If x is the observed state vector of the individual awaiting classification, then the posterior probability that the individual is in the i^{th} class is, by $\pi(i|x)$

$$\text{where} \quad \pi(i|x) = \frac{\pi(i)p(x|i)}{p(x)}$$

3.2.1. Bayesian discriminant function (Birnbaum (1960), Bailey (1965))

We can define a discrimination procedure in terms of a partition $\{A_1, \ldots, A_k\}$ of the record set X in the following way.

Let $$A_i = \{ x : \pi(i|x) = \max_{j \in I} \pi(j|x) \}$$

and define the discriminant function $\delta_\pi : X \to I$

by

$$\delta_\pi(x) = i \qquad (x \in A_i) \quad (i = 1, \ldots, k).$$

In words, we assign x to the class for which the posterior probability $\pi(i|x)$, or equivalently $\pi(i)\, p(x|i)$, is largest. Note that we have explicitly shown in the notation the dependence of the discriminant function and hence the procedure on π.

3.2.2. Misclassification probabilities

Probabilities of misclassification can be of two types (i) conditional and (ii) unconditional.

(i) The probability that an individual of class i is misclassified as of class j is a conditional probability; we denote it by $q_{\delta_\pi}(j|i)$.

(ii) The probability that an individual (whose class is assumed selected by the prior probability structure) is misclassified in some way is an unconditional probability; we denote it by q_{δ_π}.

These misclassification probabilities can be evaluated in terms of the class density functions.

$$q_{\delta_\pi}(j|i) = \int_{A_j} p(x|i)\, dx \quad ,$$

$$q_{\delta_\pi} = \sum_{i=1}^{k} \pi(i) \sum_{j \neq i} q_{\delta_\pi}(j|i)$$

3.2.3. Admissibility of δ_π.

We say δ_π is admissille if there is no other δ such that $q_\delta(j|i) < q_{\delta_\pi}(j|i)$ for every j and i and $i \neq j$.

The Bayesian discriminant function δ_π which maximizes the quantity $\pi(i)\ p(x|i)$ is the same as the one which minimizes the probability of misclassification q_{δ_π}.

δ_π which minimizes q_{i_π} is admissible for, if not, there exists another discriminant function having none of its error-probabilities larger, and one or more smaller than the $q_{\delta_\pi}(j|i)$'s of δ_π in q_{δ_π} . But this would give a smaller value to q_{δ_π}, thus contradicting the fact that δ_π minimizes q_{δ_π} . Thus δ_π is admissible.

We have seen that when the $p(x|i)$ are known density functions we can find a number of admissible Bayesian discriminant functions δ_π in terms of the relative magnitude of the quantities $\pi(i)\ p(x|i)$; each arbitrary choice of the hypothetical probabilities $\pi(i)$ defines onesuch function.

In a given classification problem it will be of interest to consider for possible use at least several different such functions δ_π and to compare them on the basis of their respective sets of error-probabilities $q_{\delta_\pi}(j|i)$.

The first problem which must be met in using these methods in practice is the choice of one or several sets of hypothetical probabilities $\pi(i)$ to use in defining the first Bayesian discriminant function to be examined. Unfortunately there are no general rules available which give very useful quantitative information to govern this choice, except for the case where a priori estimates are available. However, much qualitative information becomes available as to the directions in which the $\pi(i)$'s should be varied to get desired modifications of successive discriminant rules considered.

Investigations of robustness are usually advisable in most statistical applications. In this context such an investigation would take the following line. Construct the (A_i) corresponding to the suggested π and then ask: Is there some subset of the set $\{\pi(i) : \pi(i) \geq 0, \sum_i^k \pi(i) = 1\}$ of all possible priors which leads to the same (A_i).

The basic concept here is the property of admissibility of each of the discriminant functions obtained. Also a useful fact is that if a given set of $\pi(i)$'s gives error-probabilities which include too small a probability of correct classification for individuals from a certain class i, then, by increasing only the corresponding $\pi(i)$, and decreasing some or all of the /.....

/the other $\pi(i)$'s one will generate a new disciminant function with smaller probabilities of error of each kind possible for individuals from the given class.

3.3. Generalised Bayesian Procedure, (Birnbaum (1960)).

The Bayesian procedure discussed up to this point is a simple Bayesian procedure. Sometimes the error-probabilities found by this simple procedure are unsatisfactory. In this case we consider some generalised Bayesian procedure. The basic method for constructing a generalised Bayesian procedure is as follows.

Let (ℓ_{ji}) be a k×k array where $\ell_{ii} = 0$ for each i. For every discriminant function δ we can calculate the weighted sum of its error-probabilities

$$W = \sum_{i=1}^{k} \sum_{j=1}^{k} \ell_{ji} q_{\delta}(j|i) .$$

The procedure which minimizes W for any given array (ℓ_{ji}) is a generalised Bayesian procedure which is admissible. W is minimized by the discriminant function which takes for each x the value j for which

$$W_j(x) = \sum_{i=1}^{k} \ell_{ji} p(x|i)$$

is minimized. Then, knowing the distribution $p(x|i)$ and the /.....

/the array (ℓ_{ji}), to construct the generalised Bayesian discriminant function we have to compare for each x the k quantities $W_j(x)$, as contrasted with the k quantities $\pi(i)\ p(x|i)$ required for a simple Bayesian discriminant function.

It is useful to note that by increasing one ℓ_{ji} $j \neq i$ while leaving the others unchanged tends to reduce the corresponding $q_\delta\ (j|i)$ and of course to increase one or more of other error-probabilities.

While the method is not restricted to a special class of distributions (e.g. normal) it does however involve the assumption that the distributions are known.

We note that we do not usually know the distribution in practical work. Even if the $p(x|i)$'s are known to be of a given parametric form we need a large sample in order to use the above method in an approximate way by substituting estimates for parameter values. If we use a small sample in this case the method will not be very satisfactory since the estimates will typically be subject to fairly large sampling error.

It would be possible to apply Bayesian procedure in the case where the data are not sufficient to allow the assumption of known distribution. For example, suppose that the distributions are of known forms, say $p_1(\cdot|\theta_1)$, ..., $p_k(\cdot|\theta_k)$.

Write $\theta = (\theta_1, \ldots, \theta_k)$ and $Z = (x_{11}, \ldots, x_{kn_k})$ and

$$p(z|\theta) = \prod_{i=1}^{k} \prod_{\nu=1}^{n_i} p_i(x_{i\nu} | \theta_i)$$

The parameter space is now $I \times \theta$ and an assumption about a prior on this, say $\pi(i, \theta)$ would lead, by a straightforward application of Bayes's theorem to a posterior evaluation

$$\pi(i, \theta|z, x) = \frac{\pi(i,\theta)\, p_i(x, \theta_i)\, p(z|\theta)}{\sum_i \pi(i,\theta)\, p_i(x|\theta_i)\, p(z|\theta)}$$

We might then consider evaluating the marginal posterior densities, say

$$\pi(i|z, x) = \int_\theta \pi(i,\theta \,|\, z, x)\, d\theta$$

and choosing the class which gives maximum $\pi(i|z, x)$.

3.4. Illustration

There are two locations 1 and 2, at which a fault may occur in a machine, and there are four mutually exclusive symptoms which may be displayed when a fault occurs. The previous history of machines of this type has shown proportions of location × symptoms combinations as in the table below. The problem is to devise a discrimination procedure, which tells us which location should be examined first for each given symptom.

		Symptom			
		1	2	3	4
Machine fault at	1	0.05	0.16	0.10	0.03
	2	0.09	0.13	0.33	0.11

We shall find the following discrimination procedures,

(i) the simple Bayesian discrimination procedure;

(ii) the generalised Bayesian discrimination procedures, based on the following cost structures :

(a) costs of inspecting locations 1 and 2 are 1 and 2 respectively.

(b) costs of inspecting locations 1 and 2 are 1 and 4 respectively.

(i) From the table, we have

$$\pi(1) = 0.34 \qquad \pi(2) = 0.66 \ ;$$

also the table gives $p(x, i)$ $(i = 1, 2)$ on each symptom from which we can obtain $p(x|i)$ from the formula $p(x, i) = p(x|i)\pi(i)$.

Knowing the $p(x|i)$'s we can get the posterior probabilities $\pi(i|x)$'s, where

$$\pi(i|x) = \frac{p(x|i)\pi(i)}{p(x)} \ ,$$

$$p(x) = \pi(1)p(x|1) + \pi(2)p(x|2).$$

The following table shows the posterior probability, one on each symptom.

$\pi(1\|x)$	$\pi(2\|x)$	Symptom
0.357	0.642	1
0.552	0.448	2
0.232	0.767	3
0.214	0.785	4

From this table we see that,

$\pi(1\|x) < \pi(2\|x)$	for the symptom	1
$\pi(1\|x) > \pi(2\|x)$	" " "	2
$\pi(1\|x) < \pi(2\|x)$	" " "	3
$\pi(1\|x) < \pi(2\|x)$	" " "	4

which tells us that for symptoms 1, 3 and 4 location 2 should/.....

/should be examined first, for symptom 2 location 1 should be examined first.

(ii) If the costs of inspection are a_1 and a_2 for 1 and 2 respectively then expected cost of inspecting location 1 first is

$$a_1 \pi(1|x) + (a_1 + a_2) \pi(2|x) ,$$

and of inspecting location 2 first is

$$(a_1 + a_2) \pi(1|x) + a_2 \pi(2|x)$$

The following table shows these values for each symptom in the two different cases.

The rule of discrimination is : we examine the location which has minimum expected cost of inspection.

	Case a		Case b	
Symptom	Expected cost of inspecting location 1 first	Expected cost of inspecting location 2 first	Expected cost of inspecting location 1 first	Expected cost of inspecting location 2 first
1	2.283	1.713	3.567	4.351
2	1.896	2.104	2.792	4.552
3	2.533	1.463	4.067	4.228
4	2.569	1.427	4.139	4.210

According to these results we examine location 1 first for the symptom 2 only in case a, but in case b it is better to examine location 1 for all the symptoms.

Chapter 4

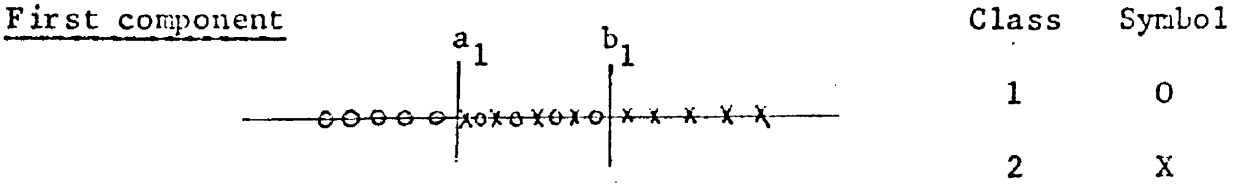
RECENT DISCRIMINATION PROCEDURES - THE ORDER STATISTIC AND CONVEX HULL PROCEDURES

4.1. Introduction

Two discriminant procedures recently suggested by Kendall (1965) are described in this chapter. The first - the order statistic procedure - has the advantage of being distribution-free but suffers from the handicap that consideration of components of the state vectors one at a time may miss some effective means of discrimination depending on combinations of components. The second - the convex hull procedure - consists of constructing the convex hull of each class cluster and allocating a new state vector only if it falls within one and only one of the convex hulls. It is unfortunately difficult computationally.

4.2. The order-statistic discrimination procedure.

Consider the case of two classes. We can represent the ordering of the first components of the set z of state vectors as in the diagram below, and we can say that the vectors in class 1 to the left of a_1 and the vectors in class 2 to the right of b_1 are "separated" by this component.



We can obtain such a separation for each of the components. Suppose that i_1 is the component which separates most vectors, say class 1 having components below a_{i_1} and class 2 having components above b_{i_1}. Kendall suggests that we now reprocess the vectors unseparated by component i_1, examining what is the best additional separation by another component. Suppose that it is i_2, and that for this component the separation is for class 1 components falling above b_{i_2} and class 2 components falling below a_{i_2}. The still unseparated vectors are then reprocessed, and the processing continues until a satisfactory separation is obtained.

The discriminant procedure then takes the following form:

(i) If the i_1th component is below a_{i_1} allocate to class 1, if above b_{i_1} allocate to class 2; otherwise proceed to (ii).

(ii) If the i_2th component is above b_{i_2} allocate to class 1, if below a_{i_2} allocate to class 2; otherwise proceed to (iii).

And so on.

4.2.1. Illustration

To illustrate this method consider again problem (2.7) of classifying individuals as normal or psychotic, based on two measurements x (size) and y (shape).

The ordering process can be conveniently set out here as two frequency tables.

Frequency table of x and y for normals and psychotics

x	Normals	Psychotics	y	Normals	Psychotics
≤10	-	10	≤ 6	5	1
11	-	2	7-10	6	1
12	-	1	11-14	6	1
13	1	1	15-18	6	2
14	-	1	19-22	-	2
15	-	2	23-26	1	-
16	-	2	27-30	-	2
17	1	-	31-34	1	1
18	2	-	35-38	-	3
19	2	1	39-42	-	-
20	6	4	43-46	-	6
21	2	-	47-over	-	6
22	2	1			
23	7	-			
24	1	1			
25	1	-			

We observe that the y-component gives the greater separation, there being a common range $0 \leq y \leq 34$ with 35 vectors in it, and one non-overlaping range $y > 34$ with 15 vectors in it and so separated. For the x-component the overlap range is $13 \leq x \leq 24$, containing 36 vectors, so that only 14 vectors are separated by x. Thus we take as our first discriminating variable y with the following first part of the discrimination rule.

(i) If $y \geq 35$ allocate to psychotic; if $y < 35$ proceed to step (ii).

By doing so we have 35 cases for which y lies in the common range.

We now take the 35 unseparated vectors and construct a frequency table for them in respect of the x component.

Frequency Table for 35 unseparated vectors.

x	Normals	Psychotics
≤ 10	-	5
11	-	1
12	-	1
13	1	1
14	-	-
15	-	-
16	-	-
17	1	-
18	2	-
19	2	-
20	6	1
21	2	-
22	2	1
23	7	-
≥ 24	2	-

According to this table we see that there is a common range $13 \leqslant x \leqslant 22$, 19 cases lying inside this range. We can thus add to step (i) the following step:

(ii) If $x \leqslant 12$ allocate to psychotics; if $x \geqslant 23$ allocate to normals; otherwise proceed to step (iii).

Since this step exhausts the components available we add the final step.

(iii) No reasonable allocation can be made.

4.3. The convex hull discrimination procedure.

Any cluster of points $z_1 = \{x_{11}, \ldots, x_{1n_1}\}$ in m-dimensional space has a convex hull

$$C(z_1) = \{ \sum_{i=1}^{n_1} a_i x_{1i} : a_i \geqslant 0 \ (i = 1, \ldots, n_1), \sum_{i=1}^{n_1} a_i = 1 \},$$

the smallest convex set containing all the points. Thus, for the data of each class cluster we can construct a corresponding convex hull. If $z_2 = \{x_{21}, \ldots, x_{2n_2}\}$ is the cluster associated with class 2 then we denote by $C(z_2)$ its convex hull. Kendall (1965) then suggests the following discrimination procedure for the case of two classes.

If $x \in C(z_1) - C(z_2)$ allocate to class 1, if $x \in C(z_2) - C(z_1)$ allocate to class 2; otherwise regard the vector as unclassified on the information available.

Note that by this procedure x remains unclassified if it belongs either to the intersection $C(z_1) \cap C(z_2)$ or to the exterior of both convex hulls, $\{C(z_1) \cup C(z_2)\}'$. See Fig. 3, where /.....

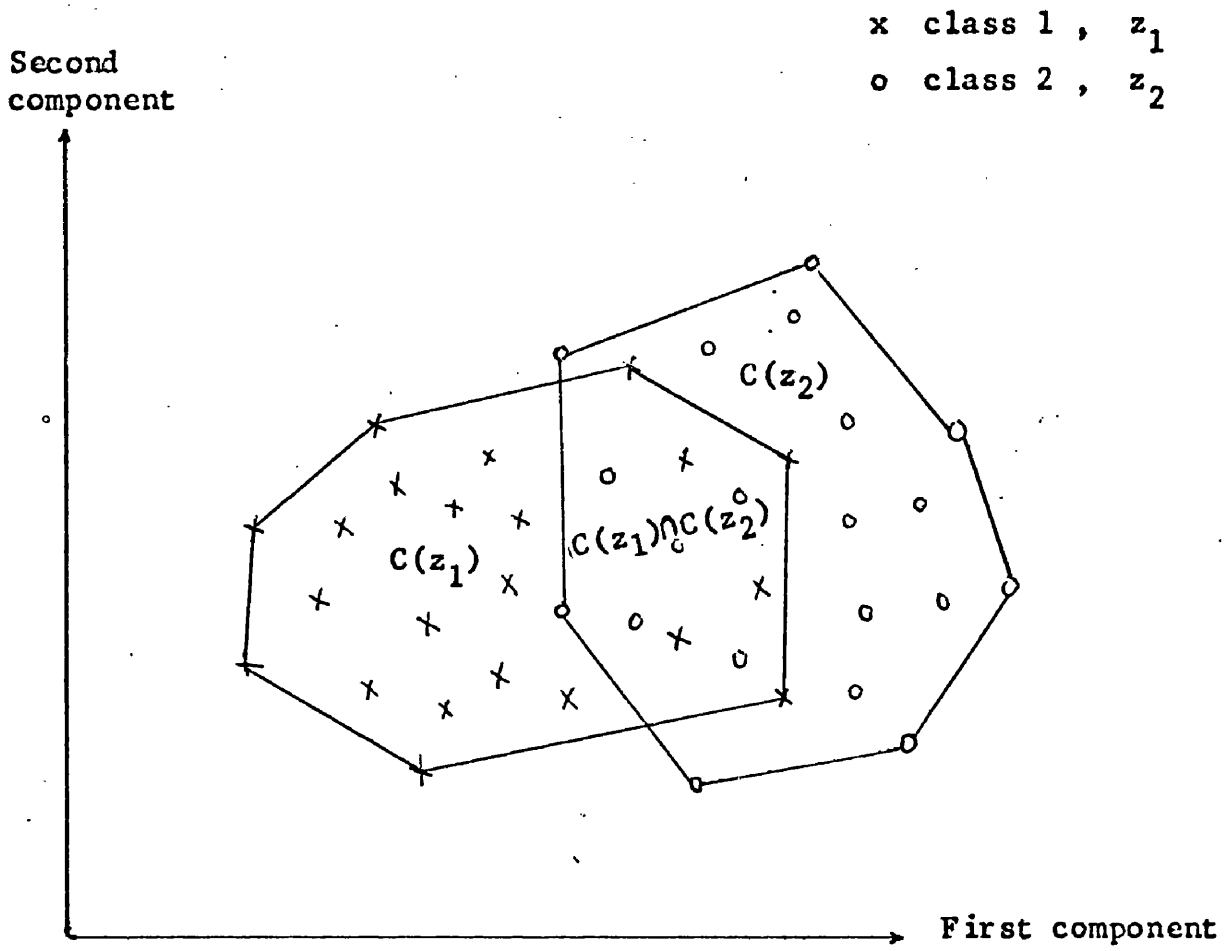

x class 1 , z_1
o class 2 , z_2
Second component
$C(z_2)$
$C(z_1)$
$C(z_1) \cap C(z_2)$
First component

Fig. 3.

/where the case of two-dimensional vectors is illustrated.

For more than two classes the extension is straightforward. If $x \in C(z_j) - \sum_{i \neq j} C(z_i)$ allocate to class j; otherwise regard the vector as unclassifiable. Presumably in the latter case some partial classification may be possible; for example, if $x \in C(z_{j_1}) \cap C(z_{j_2}) - \sum_{i \neq j_1, j_2} C(z_i)$ then we might conclude that x is likely to belong to one or other of classes j_1 and j_2 but not to any other class.

In most applications the vectors will be multi-dimensional and the main problem is how to determine, without the possibility of a graphical representation, whether x belongs to a given convex hull. Kendall expresses this as mathematical programming problem as follows.

Minimize, with respect to (a_i) and subject to $a_i \geq 0$ $(i = 1, \ldots, n)$ and $\Sigma a_i = 1$, the sum of the absolute values of the components of $\sum_{i=1}^{n_i} a_i x_{1i} - x$. If this minimum is zero then $x \in C(z_1)$.

We illustrate the method by applying it to the normal-psychotic problem. Fig. 4 shows the clusters of points for the two classes and their convex hulls. For this example we see that 12 cases are left unclassified.

On comparing the result of the two methods we find that all 12 vectors unclassified by the convex hull procedure are unseparated by the order-statistic procedure, and the 7 further vectors unclassified by the order-statistic procedure are in fact classified by the convex-hull procedure.

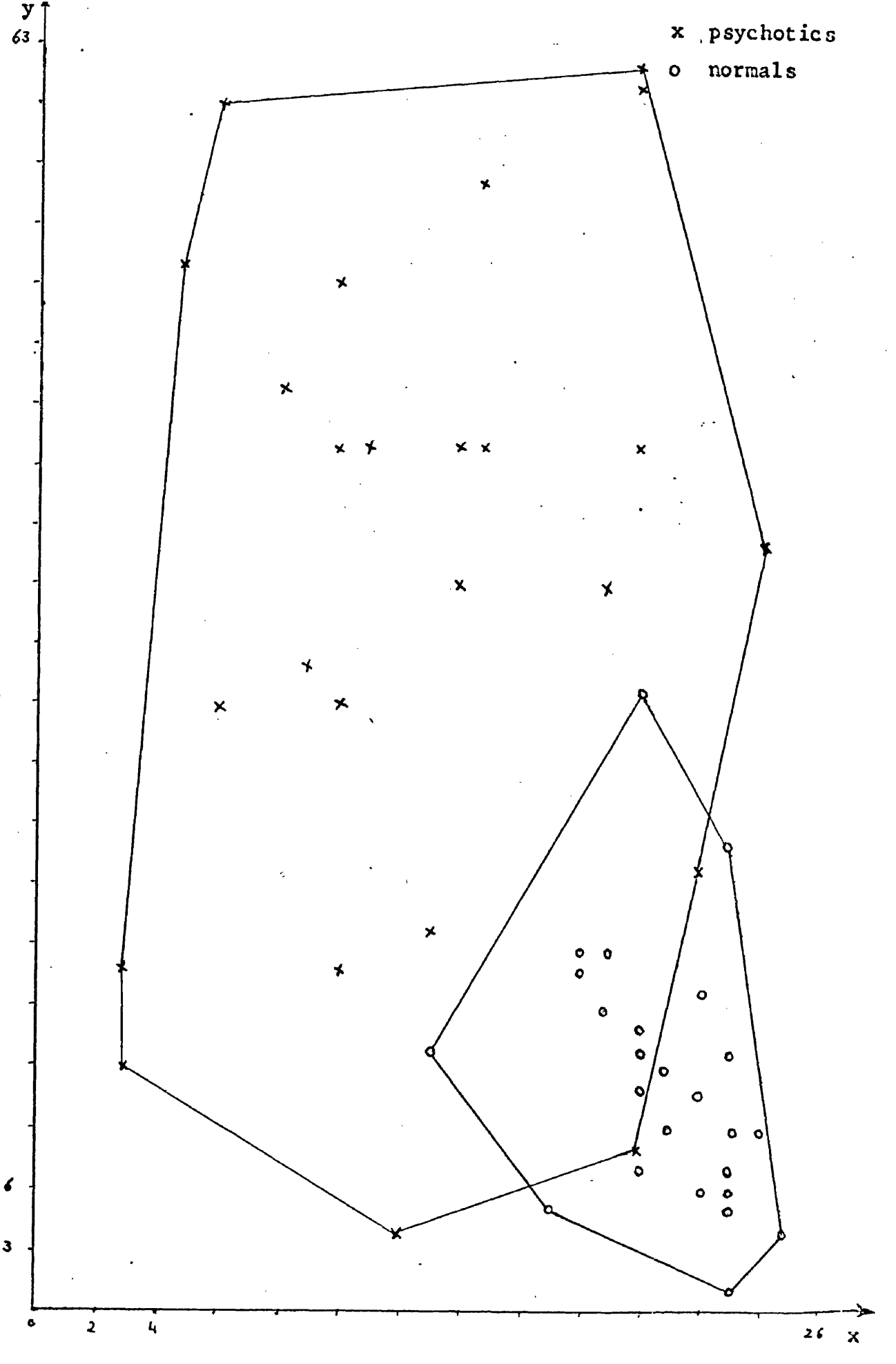


Fig. 4.

Chapter 5

RECENT DISCRIMINATION PROCEDURES BASED ON DISTANCE AND SIMILARITY INDICES

5.1. Similarity S.

Let $z = \{x_1, \ldots, x_n\}$ be a cluster of n points in m-dimensional space and x some point in this space. Suppose that a distance function or metric d is defined for this space. The similarity $S(x, z)$ of x and the cluster z is defined as the mean-square distance between x and the n members of the cluster, so that

$$S(x, z) = \frac{1}{n} \sum_{i=1}^{n} d^2(x, x_i) \qquad (5.1.1)$$

It is, of course, possible to define similarity indices without the use of a distance function (see, for example Sokal and Sneath (1963)), but the use of S does provide one convenient way of ordering points by their closeness to z.

5.2. Similarity discrimination procedure

Suppose that there are two classes, with given clusters z_1 and z_2 of data. The similarity discrimination procedure, as defined by Sebestyen (1962), can then be stated in the following way.

If $S(x, z_1) < S(x, z_2)$ allocate x to class 1, if $S(x, z_2) < S(x, z_1)$ allocate x to class 2. The extension to more than two classes is obvious : allocate x to class j, where/.....

/where $S(x, z_j) = \min_i S(x, z_i)$.

The metric d to be used has not been specified. There will usually be many possible metrics on the space, and the question now arises as to how the metric should be chosen.

Clearly any discrimination procedure depends not only on the "nearness" of x to the various clusters, but also on the relative measures of concentration of the clusters. One way of taking account of this aspect is to attempt to choose a metric which in some sense gives small distances between points of the same class cluster and large distance between points of different clusters.

To achieve this we define the following two measures. The intra-cluster similarity index $A(z_1)$ of cluster z_1 is defined as the mean-square distance between all pairs of points in z_1, so that

$$A_d(z_1) = \frac{1}{n_1(n_1 - 1)} \sum_{i \neq j} \sum d^2(x_{1i}, x_{1j}) \qquad (5.2.1.)$$

The inter-or between-cluster similarity index of z_1 and z_2 is defined as the mean-square distance between all pairs of points, one from z_1 and one from z_2, so that

$$B_d(z_1, z_2) = \frac{1}{n_1 n_2} \sum_i \sum_j d^2(x_{1i}, x_{2j}) \qquad (5.2.2.)$$

The objective of "separating" the class clusters while concentrating the points within clusters is then achieved by choosing a metric to maximise $B(z_1, z_2)$ while holding constant $A(z_1 \cup z_2)$, the intra-cluster index for the two clusters z_1 and z_2 regarded as one cluster. That is, we try to choose d^* so that

$$A_{d^*}(z_1 \cup z_2) = k \qquad (5.2.3.)$$

and

$$B_{d^*}(z_1, z_2) = \max \{B_d(z_1, z_2) : A_d(z_1 \cup z_2) = k \} \qquad (5.2.4.)$$

Sebestyen modifies this approach in the following way. He first simplifies the problem by a restriction of the class of possible metrics to those of the form

$$d^2(x, y) = \sum_{\nu=1}^{m} w_\nu^2 (x^\nu - y^\nu)^2 \qquad (5.2.5.)$$

The justification of this choice is considered in 5.3.

5.3. Non-linear procedures

Sebestyen describes his similarity index approach of § 5.2. in terms of linear transformations and Euclidean metrics. He points out that if the metric is the square root of a positive definite form then the problem of determining the optimum metric is equivalent to asking the question : what linear transformation is such that the transformed clusters have maximum mean square/.....

/square inter-cluster Euclidean distance for given specified mean square intra-set Euclidean distance.

Consider the linear transformation

$$y = Wx \tag{5.3.1.}$$

so that

$$y_{1i} = Wx_{1i} (i = 1, \ldots, n_1) \ , \ y_{2j} = Wx_{2j} (j = 1, \ldots, n_2) \tag{5.3.2.}$$

form the transformed clusters. If $B_W(z_1, z_2)$ denotes the mean-square inter-cluster Euclidean distance for the W-transformed data then

$$B_W(z_1, z_2) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} || y_{1i} - y_{2j} ||^2 \tag{5.3.3a}$$

$$= \frac{1}{n_1 n_2} \sum_{i} \sum_{j} (x_{1i} - x_{2j})' W' W (x_{1i} - x_{2j}) \tag{5.3.3b}$$

$$= \sum_{k=1}^{m} w_k' V w_k \ , \tag{5.3.3c}$$

where $w_k' = \{w_{k1} \ldots w_{km}\}$ is the k^{th} row of W and

$$V = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (x_{1i} - x_{2j})(x_{1i} - x_{2j})' \tag{5.3.3d}$$

has the structure of a variance covariance matrix. In exactly the same way the mean-square intra-cluster Euclidean distance can be expressed in the form /.....

/form

$$C_W(z_1, z_2) = \Sigma\, v_k' T w_k \qquad (5.3.4.)$$

where T is again a kind of variance covariance matrix constructed from $z = (z_1, z_2)$.

The problem is thus seen to be that of choosing $w_k (k = 1, \ldots, m)$ to maximise $\Sigma\, w_k' V w_k$ subject to the condition that $\Sigma\, w_k' T w_k = c$, a constant. Using the method of Lagrange multipliers we thus maximise

$$\Sigma\, w_k' V w_k - \lambda(\Sigma w_k' T w_k - c)$$

with respect to w_k $(k = 1, \ldots, m)$ and λ. Thederivative equations for the maximising $\hat{w}_k$ and $\hat{\lambda}$ are thus

$$(V - \hat{\lambda} T)\, \hat{w}_k = 0 \qquad (5.3.5.)$$

$$\Sigma\, \hat{w}_k' T \hat{w}_k = c \qquad (5.3.6.)$$

From (5.3.5.) and (5.3.6.) we have that

$$B_{\hat{W}}(z_1, z_2) = \hat{\lambda}\, \Sigma\, \hat{w}_k' T \hat{w}_k = \hat{\lambda}\, c \qquad (5.3.7.)$$

Thus we must take $\hat{\lambda}$ to be the largest root of $|V - \lambda T| = 0$, i.e. the largest eigenvalue of $T^{-1}V$. Then $\hat{w}_1$, say may be the eigenvector corresponding to this largest eigenvalue and satisfying

$$\hat{w}_1' T \hat{w}_1 = c, \text{ and } \hat{w}_k = 0 \ (k = 2, \ldots, m) \qquad (5.3.8.)$$

The fact that

$$W = \begin{bmatrix} w_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

justifies Sebestyen's special form of metric.

This technique of dealing with linear transformations may be extended to the case of non-linear transformations. The class of all continuous transformations is too general to yield a practical solution, and the class considered by Sebestyen is restricted to polynomial transformations, in the following sense. The k^{th} component $y^{(k)}$ of y, the result of transforming x is given by

$$y^{(k)} = \sum_{\nu=1}^{p} w_k \left[x^{(k)} \right]^{\nu} \qquad (5.3.9.)$$

Thus the matrix $W = \left[w_{k\nu} \right]$ defines the transformation. Again Sebestyen succeeds in determining a W which gives an optimum metric in the sense already defined. The computations are naturally more complex and we do not reproduce the complicated algebra here.

5.4. An illustration

The computations involved in the applications of § 5.2. and 5.3. are heavy though not prohibitive on an automatic computer. We illustrate the technique here for another, computationally simpler version suggested by Sebestyen. This in effect uses two metrics, one for measuring distances from z_1, the other for measuring distances from z_2. These metrics are selected from the class (5.2.5.) and are such that the mean-square distance between vectors of a cluster is minimised subject to the volume of the cluster being held constant.

Applying this technique to the normals-psychotics problem we find

$$w_1 = .93 \ , \ w_2 = 0.07 \quad \text{for normals}$$

$$w_1 = .97 \ , \ w_2 = 0.03 \quad \text{for psychotics}$$

and that for all the 25 normals except two $S(x, N) < S(x, P)$ and for all 25 psychotics except 6 $S(x, P) < S(x, N)$. We note that 8 cases are misclassified by this method, 2 from normals and 6 from psychotics.

Chapter 6

GENERAL COMMENTS ON THE DISCRIMINANT PROCEDURE

6.1. General remarks

In this chapter we set out briefly some of the advantages and disadvantages of the various discrimination procedures we have discussed. There is certainly no procedure which is generally applicable. It is possible to envisage for each procedure a simple practical situation where it will fail. The problem of statistical discrimination is clearly only at a very early stage of its development. It could be argued that it remains a difficult problem because of its multi-dimensional character. For state vectors of one, two or three dimensions it is easy to obtain a pictorial representation of the clusters of classes and, from the patterns we see, be guided to sensible discrimination procedures. For higher dimensional vectors no such picture is available to us and this prevents us from easily detecting patterns or clusters. The main hope at the moment for off-setting this human deficiency seems to lie in the possibility of organising computers to engage in some sophisticated form of pattern recognition.

6.2. The advantages and disadvantages of the procedures

(i) Classical procedures. These are wholly based on the assumption of multivariate normality of the state vectors, and largely on the assumption of equality of covariance matrices. The question of how useful such a procedure is when these assumptions are not justifiable is the question of robustness. Little work on robustness has been done in this area, and only conjectures can be made. The procedure leads to a linear discriminant, that is a division of the state vector space by a hyperplane, or hyperplanes. It might be conjectured that the procedure would give reasonable discrimination for most situations where the clusters are roughly ellipsoidal (see Fig. 5). It does not follow that any calculations of misclassification probabilities based on normality assumptions would necessarily be reliable.

When the covariance matrices are unequal classical procedures lead to "quadratic discriminants". For example with data of Fig. 5 some curved dividing line would probably be more reliable than the straight line division because of differing concentration pattern of the classes. The classical procedure would probably again give sensible discrimination even although normality could not be assumed.

Class	Symbol
1	o
2	x

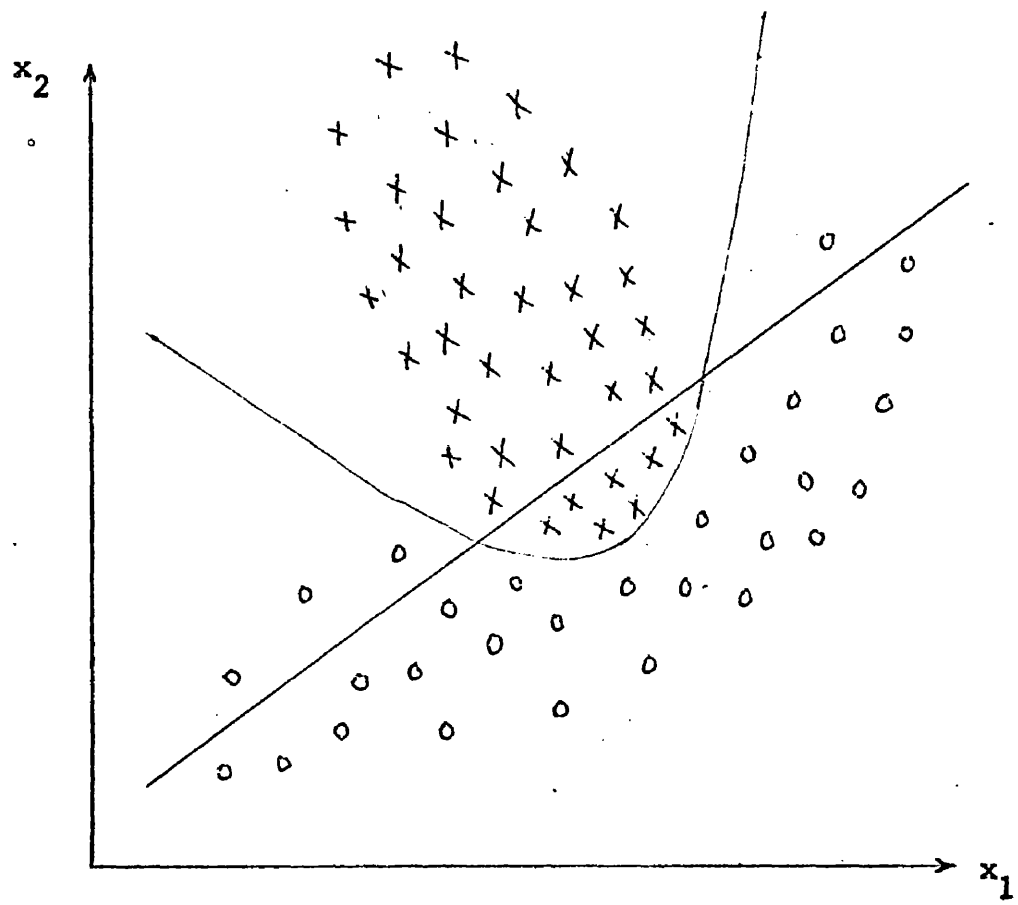


Fig. 5.

(ii) Bayesian procedures. These depend on the ability of the experimenter to provide prior information. If this is available then Bayesian procedures are versatile enough to deal with most parametric models. They are very readily applied to situations where the state vector consists of qualitative components. Because of the greater difficulty of making prior assessments for distributions of quantitative factors they are clearly less attractive in this field.

(iii) Order-statistic procedure. The classical and Bayesian procedures are both based on assumptions about the parametric form of the distributions of state vectors. The main attraction of the order-statistic procedure is that it is distribution-free and is computationally simple. Its great disadvantage is that it examines the components of the state vector one at a time. The examination of the components of the state vector one at a time is a mixed blessing. In some situations it will as claimed by Kendall give some clue about which components are most effective in discriminating. For example, an application to Fisher's example (see Table 6.1) of discriminating between Iris setosa and Irs versicolor, the state vector being of four components, showed that the classes could be completely discriminated with the use of a single component petal length or petal width, and the rule of discriminating is : allocate setosa if petal length less than/.....

able 6.1. Measurements of the flowers of 50 plants each of the two species Iris setosa and Iris versicolor.

Iris Setosa				Iris Versicolor			
Sepal Length	Sepal Width	Petal Length	Petal Width	Sepal Length	Sepal Width	Petal Length	Petal Width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4
4.9	3.0	1.4	0.2	6.4	5.1	4.5	1.5
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5
5.4	3.6	1.7	0.4	5.7	2.8	4.5	1.3
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6
5.0	3.4	1.5	0.2	4.9	2.9	3.3	1.0
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0
4.8	3.4	1.6	0.2	5.6	3.0	4.2	1.5
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8
5.1	3.4	1.5	0.4	6.1	2.8	4.0	1.3
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2

Iris Setosa				Iris Versicolor			
Sepal Length	Sepal Width	Petal Length	Petal Width	Sepal Length	Sepal Width	Petal Length	Petal Width
4.8	3.4	1.1	0.2	6.4	2.9	4.3	1.3
5.0	3.0	1.6	0.2	6.6	3.0	4.7	1.4
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5
4.8	3.1	1.5	0.2	5.5	2.4	3.8	1.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0
5.2	4.1	1.5	0.1	5.8	2.9	3.9	1.2
5.5	4.2	1.4	0.2	6.0	3.0	5.1	1.6
4.9	3.1	1.5	0.2	5.4	3.4	4.5	1.5
5.0	3.2	1.2	0.2	6.0	3.1	4.5	1.0
5.5	3.5	1.3	0.2	6.7	2.3	4.7	1.5
4.9	3.6	1.4	0.1	6.3	3.0	4.4	1.3
4.4	3.0	1.3	0.2	5.6	2.5	4.1	1.3
5.1	3.4	1.5	0.2	5.5	2.6	4.0	1.3
5.0	3.5	1.3	0.3	5.5	3.0	4.4	1.2
4.5	2.3	1.3	0.3	6.1	2.6	4.6	1.4
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2
5.0	3.5	1.3	0.6	5.0	2.3	3.3	1.0
5.1	3.8	1.6	0.4	5.8	2.7	4.2	1.3
4.8	3.0	1.9	0.3	5.7	3.0	4.2	1.2
4.1	3.8	1.4	0.2	5.7	2.9	4.2	1.3
4.6	3.2	1.6	0.2	6.2	2.9	4.3	1.3
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3

/than 2, otherwise allocate to versicolor. But on petal width allocate to setosa if petal width less than 0.7, otherwise allocate to versicolor. We note that in two cases there is no range of overlaping. But if we consider sepal length getting a range of overlaping which is $4.9 \leq$ sepal length ≤ 5.8 with 57 vectors in it. And on sepal width the common range is $2.3 \leq$ sepal width ≤ 4.4 with 96 vectors in it. While the procedure thus gives the appearance of placing the components in order of discriminatory importance it may fail completely even in a simple case where discrimination can be satisfactorily carried out in terms of a combination of components. For example, in Fig. 6 order-statistic procedure gives no discrimination while it is clear that the straight line shown separates the two classes effectively.

The procedure differs from classical and Bayesian procedures in that it may leave some state vectors unclassified. This may well be a more realistic conclusion than that of forcing a complete classification.

(iv) Convex-hull procedure. The idea underlying this procedure is attractive but it has a number of serious practical disadvantages. It involves a considerable computation when the state vectors are of high dimension, and it cannot be applied to qualitative components. Moreover, we can envisage situations/.....

Class	Symbol
1	o
2	x

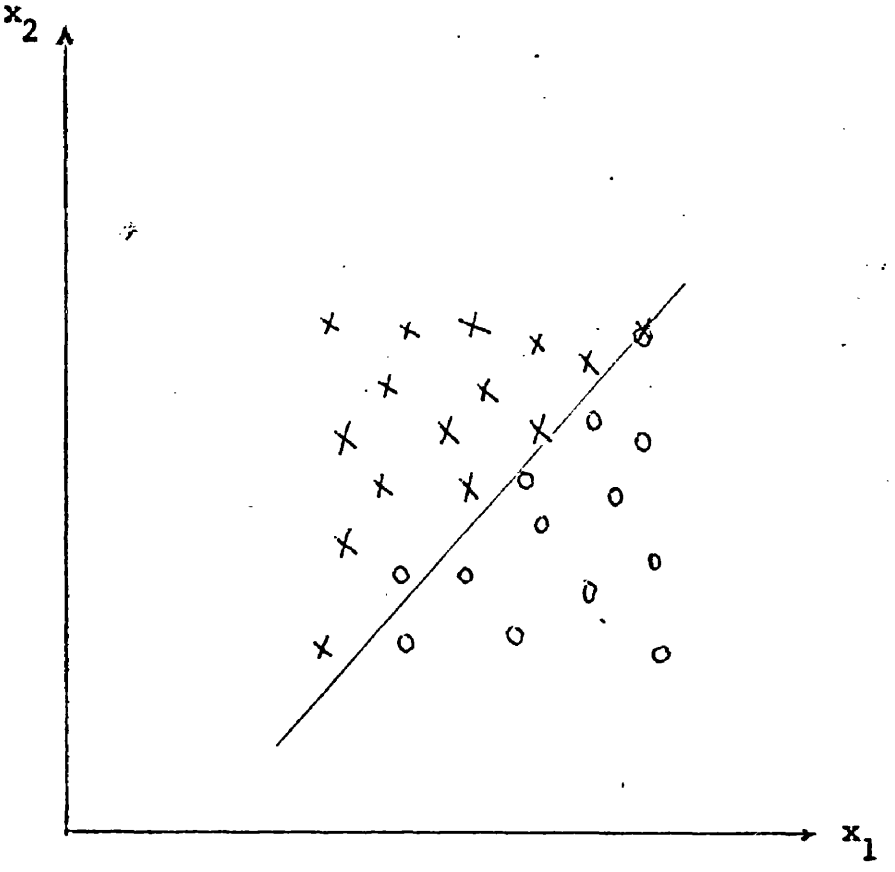


Fig. 6.

/situations where it will fail completely; see, for example Fig. 7.

(v) Linear and non-linear transformation and similarity. For the different ways of defining and measuring similarity it is left to the investigator to choose the one which is applicable under the given circumstances. The simplest concept of measuring similarity mean-square distance was discussed by Sebesteyn.

The success of linear transformation depends on the choice of the model and the nature of classes to be discriminated. It gives a good discrimination in the situation where the probability densities of classes are unimodal i.e. they possess only a single hump. The linear transformations operate on one dimensional information and ignoring all the other directions. When the number of classes increase linear transformations give very poor results and using non-linear transformations yield good results. The difficulties of these techniques involve a complicated calculation and the use of digital computers is necessary.

Class	Symbol
1	o
2	x

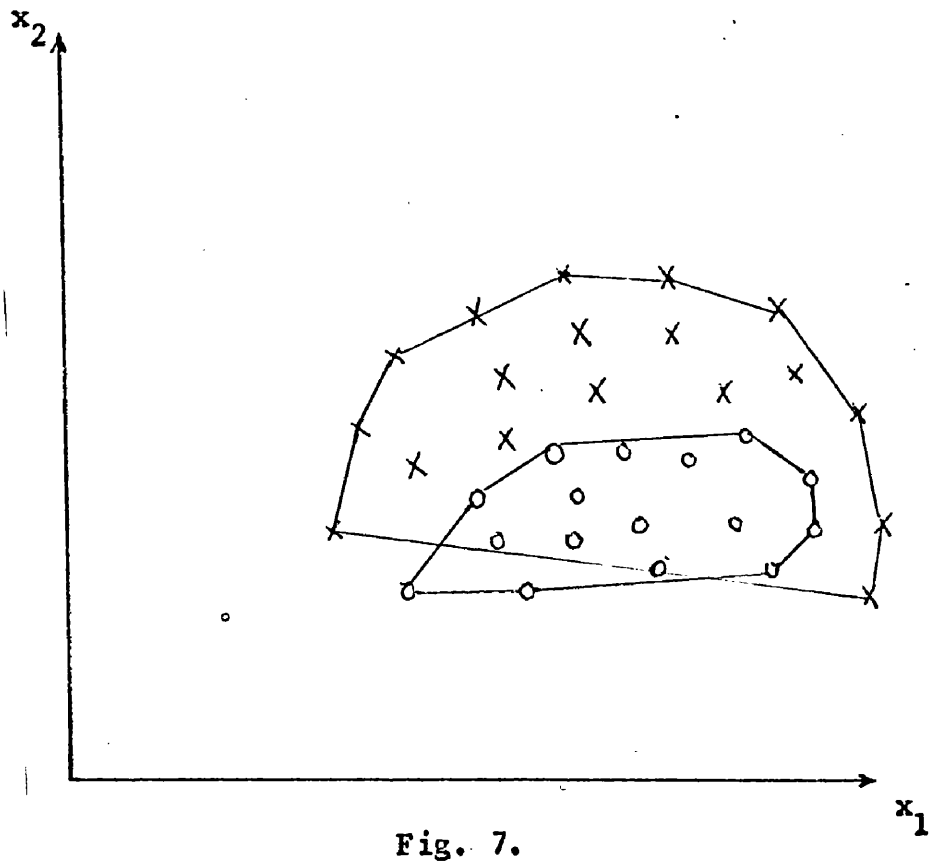


Fig. 7.

REFERENCES

ANDERSON, T.W. (1957) "Introduction to Multivariate Statistical Analysis." John Wiley.

BAILEY, N.J.S. (1965) "Probability Methods of Diagnosis Based on Small Sample." Mathematics and Computer Science in Biology and Medicine.

BARON, D.N. and FRASER, P.M. (1965) "Theory and Application of Computer in Classification and Diagnosis of Liver Disease." Computers in the Hospital Service. Proceedings of a Symposium, p. 46.

BIRNBAUM, A. and MAXWELL. (1960) "Classification Procedure Based on Bayes's Formula". Applied Statistics, 9, 152.

BOYLE, J.A. (1965) "An Experimental Approach to the Calculation of Diagnosis using an Automatic Digital Computer." Computers in Hospital Service, p. 54.

FISHER, R.A. (1936-37) "The Use of Multiple Measurements in Taxonomic Problems." Annals of Eugenics, 7, 179.

KENDALL, M.G. (1957) "A Course in Multivariate Analysis." Griffin.

KENDALL, M.G. (1965) "Discrimination and Classification" in Multivariate Analysis by Paruchuri R. Krishnaiah. Academic Press.

LEDLEY, R.S. and LUSTED, L.B. (1962) "Mathematical diagnosis and modern decision-making" in Mathematical problems in biological sciences. Proceedings of a Symposium in Applied Mathematics, Vol. 14. Providence, R.I., American Mathematical Society p. 117.

MARTIN, E.S. (1936) "A study of an Egyptian series of Mandibles with Special Reference to Mathematical Methods of Sexing". Biometrika 28, 149.

RADHAKRISHNA, S. (1964) "Discrimination Analysis in Medicine". The Statistician 14 No. 2.

SEBESTYEN, S.G. (1962) "Decision-making Processes in Pattern Recognition." Macmillan.

SMITH, C.A.B. (1946-47) "Some Examples of Discrimination." Annals of Eugenics 13, 272.

SOKAL, R. and SNEATH, P.H.A. (1963) "Principles of Numerical Taxonomy." W. H. Freedman and Co.; San Francisco.

WALD, A. (1944) "On a statistical problem arising in the classification of an individual into one of two groups." Ann. Math. Stat. 15, 145.

WARNER, H.R. (1961) "A Mathematical Approach to Medical Diagnosis". J. Amer. med. Ass. 177, 177.

WELCH, B.C. (1939). "Note on Discriminant Function." Biometrika 31, 218.

SUMMARY

In this thesis a critical survey of techniques of statistical discrimination is undertaken. The problem of statistical discrimination arises where previous work has separated a number of individuals into k distinct classes, there being available on each individual a vector of m-measurements. The problem is to assign a new unclassified individual for which the vector of m-measurements is available, to one of the k classes.

Many different techniques for solving this problem have been suggested and these are considered in Chapters 2-5 of the thesis.

The idea of classical techniques (Chapter 2) is to find a linear combination of the m-measurements and use its value for the allocation of the new individual. This idea is derived from the assumption of normality of the variability of the data for the different classes. The technique was introduced by Fisher (1936).

Another technique, called Bayesian discrimination (Chapter 3) requires some prior information about the relative frequenci the different classes which, after the observation of the new individual, can be converted into a posterior information by the use of Bayes's theorem. The main developments of this theory to data require knowledge of the distributions of the different classes.

Some techniques - order-statistic and convex-hull methods (Chapter 4) have recently been introduced by Kendall (1965). For the first method the discrimination procedure is built up in stages. A first step towards discrimination is taken by considering the measurements one at a time, and using that measurement which separates into classes the most individuals. At subsequent stages, only previously unclassified individuals and unused measurements are considered in the search for the further refinement of the procedure. The second method consists of constructing the convex-hull of each class and allocating the new individual if and only if it falls in one of the convex-hulls.

Other recent techniques have been introduced by Sebestyen (1962 and are termed similarity index procedures (Chapter 5). The idea underlying this theory is to calculate the similarity of the individual to each class and allocate it to the class which is most similar. The concept of measuring similarity which is suggested by Sebestyen is the calculation of the mean-square distance between a point and a class of points.

The main conclusion (Chapter 6) is that there is no general procedure which can be followed in every situation. The application of any technique depends on the nature of the practical problem. The hope of obtaining improved procedures seems to lie in the use of large scale computers to provide in some convenient form a geometric picture of the high-dimensional data involved in most practical problems.