**Title:** SOME OBSERVATIONS IN PATTERN RECOGNITION

**Author(s):** Kishan G. Mehrotra

**Performing Organization:** Syracuse University

**Contract or Grant Number:** F30602-78-C-0148

**Report Date:** February 1982

**Number of Pages:** 66

**Security Class.:** UNCLASSIFIED

**Distribution Statement:** Approved for public release; distribution unlimited

**Key Words:** Pattern Recognition, Feature Selection

**Abstract:** In this manuscript some very basic ideas of important consequence are discussed. These ideas are important for any practicing engineer in pattern recognition. The topics include equivalent classifier, dimensionality reduction, fusion of classifier, time varying statistics, etc. Throughout this presentation, it is assumed that the reader is familiar with the mechanics of constructing discriminant, selecting features and other related properties. Therefore no attempt is made to make this (over)
presentation are considered "obvious" in standard books written on the subject of pattern recognition. It is our belief that the readers of this manuscript will benefit considerably by giving some time to these "obvious" results; mainly because the obvious results are sometimes very confusing results.
INTRODUCTION

In this manuscript some very basic ideas of important consequence are discussed. These ideas are important for any practicing engineer in pattern recognition. The topics include equivalent classifier, dimensionality reduction, fusion of classifier, time varying statistics etc.

Throughout, this presentation, it is assumed that the reader is familiar with the mechanics of constructing discriminant, selecting features and other related properties. Therefore no attempt is made to make this presentation comprehensive. Most of the subjects, discussed in this presentation are considered 'obvious' in standard books written on the subject of pattern recognition. It is our belief that the readers of this manuscript will benefit considerably by giving some time to these "obvious" results; mainly because the obvious results are sometimes very confusing results.
1. OPTIMAL AND EQUIVALENT DISCRIMINANTS

Many important applications of pattern recognition can be characterized as either waveform classification or geometric figures classification. In order to perform this type of classification, typically one measures some observable characteristic of the object. This collection of measurements is called the features, and the process of deriving the features is called feature extraction. Typically a classifier is developed using these features.

If any classification problem, one of the basic assumptions is that there exists some difference between the populations from which the objects are sampled. Thus, there is always a classifier which can be used to differentiate between the populations. We will call it "the natural classifier" and denote it by C. Existence of such a natural classifier is of fundamental importance in pattern recognition. This will also be useful in the following discussion.

To fix the ideas, we consider the example of character recognition between letters A and B. Note that there exists a natural classifier [which human mind employs] to distinguish between A and B. For mechanical or computerized discrimination one would select features to construct a classifier. This feature selection can be done in many ways and success of the corresponding classifier depends very heavily on these features. For the hand written characters (A and B) two possible feature extraction pro-
cedures are:

(a) put a standard grid on each letter and measure the shaded area in each cell, [see attached figure].

(b) Record the presence and absence of a portion of the letter in each cell by 1 and 0 respectively and obtain the feature vector consisting of 0's and 1's. The collection of these features can be employed to construct two respective classifiers.

Mathematically, the feature extraction is equivalent to transformation from the natural space, \( S \), to euclidean space \( \mathbb{R}^n \), i.e.

\[
[F] = T[S]
\]

where \( F \) denotes the new features and \( T \) is the transformation. Typically the transformation \( T \) is nonlinear and some time one may not be able to express it in terms of mathematical equations.

Let \( C \) and \( C' \) denote the classifiers in the natural space and the feature space respectively and \( T^{-1} \) the inverse transformation of \( T \). Then \( C \) and \( C' \) will be equivalent.

This equivalence is obvious because the existence of \( T^{-1} \) implies
that there is a one to one transformation from $S$ to space of features and conversely.

Extending this idea, suppose that $T_1$ and $T_2$ are two transformations, $\{F_1\}$ and $\{F_2\}$ the corresponding feature then the classifiers $C_1$ and $C_2$ would be equivalent to each other and to $C$ if $T_1^{-1}$ and $T_2^{-1}$ exist. Obviously $C_1$ and $C_2$ could be equivalent to each other, if there exists a one to one transformation from $\{F_1\}$ to $\{F_2\}$, without being equivalent to $C$.

Thus, the optimal classifier is the 'natural' classifier and generally, it is not possible to explain how it works. On the other hand to obtain a classification procedure a set of features is obtained. Given these features one can attempt to obtain optimal classification procedures. But it must be remembered that this optimality is conditional upon the given feature set. In other words, if a new feature set is given than another 'optimal' classifier will be obtained. The two 'optimal' classifiers will be equivalent if and only if it is possible to obtain a one to one transformation from one feature set to the other.

To summarize, a classifier is optimal only after a set of features have been selected. This optimality should not be confused with 'global' optimality.
2. ON MEMBER OF FEATURES

In a typical pattern recognition problem there are two stages:
(a) the feature selection stage (b) design of a classifier based
on the selected features. Classifier design is relatively easier
in the sense that if the features and their class dependent
joint distributions are available then one can apply Bayes proce-
dure to obtain optimal classifier. In case the class dependent
distributions are partially known, or even if they are completely
unknown, modifications of the optimal classifier or nonparametric
classifiers are applicable. On the otherhand the problem of fea-
ture selection is quite difficult because no standard procedures
can be applied and moreover the features are specific to the pro-
blem under consideration.

The problem of feature selection arises generally because the
data collected in the natural space is not suitable for mathemati-
cal manipulation. For example, consider computerized classifica-
tion of ECG curves to one of the several disease classes. In
this case mathematical manipulations with these random ECG curves are
almost impossible, therefore the need for alternative ways of
storing the information in a curve. For this particular problem,
one possible procedure is to apply Karhunen-Loeve expansion.

Feature selection also plays an important role as a method of
data reduction. For example, although the data may be available
in a vector form, suitable for mathematical manipulations, yet
its dimensionality may be very large. In such situation it is
desired to compress the dimensionality without sacrificing in the
performance.
Let $C_n$ be a classifier based on $n$ features. Let $C_m$ be a classifier based on a subset of the original $n$ features. Then it appears to be a well known property that the performance of $C_m$ cannot be superior than $C_n$. A proof of this property is easily obtained in the case of two class classification problem with underlying normal distribution with common covariance matrix.

In this case the performance of the optimal classifier is measured in terms of Mahalonobis distance $\delta' \Sigma^{-1} \delta$ where $\delta = \mu_1 - \mu_2$ and $\Sigma$ is the common covariance matrix, $\mu_1$ is the mean vector $i=1,2$.

The error probability decreases as the Mahalonobis distance $\delta' \Sigma^{-1} \delta$ increases because the error probability is given by $\phi(-\frac{1}{2}(\delta' \Sigma^{-1} \delta)^{1/2})$.

Since

$$\delta' \Sigma^{-1} \delta = \delta_1' \Sigma_{11}^{-1} \delta_1 + \delta_{2.1}' \Sigma_{22.1}^{-1} \delta_{2.1}$$

where

$$\delta = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

$$\delta_{2.1} = \delta_2 - \Sigma_{21} \Sigma_{11}^{-1} \delta_1, \quad \Sigma_{22.1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$$

and

$$\delta_{2.1}' \Sigma_{22.1}^{-1} \delta_{2.1} \geq 0.$$

it follows immediately that $\phi(-\frac{1}{2}(\delta_1' \Sigma_{11}^{-1} \delta_1)^{1/2}) \geq \phi(-\frac{1}{2}(\delta_{2.1}' \Sigma_{22.1}^{-1} \delta_{2.1})^{1/2})$.

Thus, a subset selection may not lead to a better classifier.

But this does not imply that if $n > m$ and the first classifier is based on $n$ features and the second classifier is based on $m$ features then the first classifier is necessarily better than the second.

In fact, in some cases a classifier based on $n$ features may do worse than another classifier based on $m$ features ($m < n$).
To demonstrate this property, consider the following trivial example. Consider two populations in which the underlying random vector is \((k+1)\) dimensional, \(k > 1\). Suppose the marginal distributions of the first \(k\) components are identical in the two populations, thus the first \(k\) components have no discriminatory capability. On the other hand the \((k+1)\)th component has different distributions in the two population. Two researchers, who are unaware of this property, select feature sets consisting of the first \(k\) components and the last \((k+1)\)th component only respectively. It is obvious that the first researcher will obtain poorer discriminant although his feature set contains a larger number of components than the second researcher.

In general, for every classifier based on \(m\) features, one can produce an equivalent classifier with \(n\) features where \(n \geq m\) because all we need to do is to add \(n-m\) non-informative independent features to the set of \(n\) features. On the other hand given a classifier based on \(m\) features one can produce an equivalent classifier based on one feature alone, as seen below.

The existence of a 1 dimensional equivalent criterion is easily seen in the case of two class problem when the underlying distributions are normal with common covariance \(\Sigma\). In this case, the classification rule, based on \(n\) features \(x\) is given by:

\[
\text{classify } x \text{ to class 1 iff } (x - \frac{\mu_1 + \mu_2}{2}) \Sigma^{-1} (\mu_1 - \mu_2) > 0
\]

where standard notations are employed. Choosing the one dimensional feature \(Y\), where

\[
Y = (x - \frac{\mu_1 + \mu_2}{2}) \Sigma^{-1} (\mu_1 - \mu_2)
\]

we obtain an equivalent classifier. The result is now obvious for
m ≥ 2. In general for any arbitrary distributions, let

\[ y = \ln f_1(x) - \ln f_2(x) \]

where \( f_i(x) \) is the probability distribution function corresponding to the \( i \)th class.

In summary, it cannot be said that a discriminant based on larger number of features is necessarily better than another discriminant which uses a smaller number of features, unless the second set of features is a subset of the first set. Additional features will improve the performance of a discriminant only if they are informative. Finally, the performance of a discriminant depends not on the number of features but on the choice of features themselves.
3. DISCRIMINATION VERSUS CLASSIFICATION

There are two main goals in pattern recognition. The first goal could be called "Separating distinct sets of objects" and the second goal is to "allocate new items to previously defined groups". Fisher (1938) used the term "discrimination" to refer to the first goal. A more descriptive term is "separation".

The second goal is referred to as "classification" which is also called "allocation" see Johnson and Wichran (1980) and "identification" see Rao (1974). These concepts are further explained below.

By discrimination or separation we understand how to describe either graphically or algebraically, the differential features of objects (observations) from several known collections (populations). We try to find "discriminants" whose values are such that the collections are separated as much as possible.

By classification or allocation we understand how to sort objects (observations) into 2 or more leveled classes. The emphasis is on deriving a rule which can be used to "optimally" assign a new object to the leveled classes.

The difference, just pointed out, between discrimination and classification is generally not explained in standard texts on pattern recognition. Inconsistent use of the terminology by statisticians and pattern recognizers has also caused confusion. Moreover, a function which separates may also be used for allocation and conversely, an allocatory rule may suggest a discriminatory procedure. Thus in practice the two goals may overlap and the distinction between separation and allocation becomes blurred.

Allocation or classification rules are usually developed from
"learning" samples. Observations are randomly selected and are known to come from specified populations. These samples, also known as training set, are then examined for differences and based on the results of this examination, the entire sample space is partitioned in as many regions as the number of populations. If we denote these disjoint and exhaustive regions by $R_1, R_2, \ldots, R_p$ where $p$ = number of populations and if new observation falls in the region $R_i$, it is allocated to the $i$th population.

Fisher's idea, in discriminating between two populations $\pi_1$ and $\pi_2$ on the basis of observed values of presumably relevant variables $\chi$ was to transform the multivariate observations $\chi$ to univariate observations $y$ such that the $y$'s derived from populations $\pi_1$ and $\pi_2$ were separated as much as possible. For simplcity, Fisher suggested the use of linear combinations of $\chi$ to create the $y$'s. This idea can be extended to several classes and also to several discriminants $y_1, y_2, \ldots$ where $y_1$ provides the best separation, $y_2$ the next best separation and so on. It is well known that these discriminants have also been used for classification and have "optimum" properties for the normal distributions.
4. **SAMPLE SIZE CONSIDERATIONS OF CLASSIFIERS AND TESTING OF CLASSIFIERS:**

One of the most important issues, after a classifier has been designed, based on a training set, is to find how well it performs. Considerable attention has been paid to this problem. To study this problem, most attention has been given to the two class problem assuming the underlying distributions are normal.

Denoting the probability of error of misclassification by $\rho$ there are several types of error probabilities which should be distinguished.

- $\rho$: When the discriminant uses the known population parameters and this discriminant is applied to independent observations from the population,
- $\hat{\rho}$: when the discriminant is based on a training set and its performance is measured using the given training set.
- $\rho^*$: when the discriminant is based on the training set and its performance is measured on another independent set called the test set.
- $\tilde{\rho}$: when the discriminant is based on a training set and its performance is measured on the independent samples of the population.

A general result

$$\hat{\rho} < \rho < \tilde{\rho}$$

was established by Mills in 1965. The dependence of $\hat{\rho}$ and $\tilde{\rho}$ on the ratio $n/k$, where $n$ is the size of the training set and $k$ is a dimension of the underlying normal random variable, was studied by Foley (1972), when $\Sigma$, the common covariance matrix is
assumed known. Foley observed that the difference between $E(\hat{\rho})$ and $\rho$ is very large if $n/k << 3$. Only if $n/k > 3$, $E(\hat{\rho}) - \rho$ is small and therefore $\hat{\rho}$ can be considered a reasonable estimator of $\rho$. Mehrotra (1973) observed that if $E$ is also estimated, and if $n/k > 5$ then only $\hat{\rho}$ can be considered as a good estimator.

However, obtaining the estimate $\hat{\rho}$ is the most important problem, but its distribution is very complex. Asymptotic results have been obtained by several investigators. Lachenbruch and Mickey (1965) studied several possible estimators of $\rho$ for the normal distribution and concluded that the leave-one-out method, which is equivalent to jackknifing the estimator $\hat{\rho}$, provides a good estimator of $\rho$. This work was further studied by Cochran (1968). Due to space considerations, it is prohibitive to go into details of work in this area. Toussaints (1974) bibliography provides useful references related to this problem.

Several studies have also been performed to study the performance of the Fisher's linear discriminant. These include the study of its performance when their common covariance assumption is not applicable, when the underlying distributions are not normal. Most of these studies are empirical. Overall performance of the Fisher's linear is found to be satisfactory.

In the study of the Fisher Linear discriminant, other problems of interest are: (i) study of the coefficients of the Fisher linear discriminant and (ii) the problem of testing the significance of the obtained discriminant function. Sitgreaves (1961) observed that the estimates of the coefficients in the linear discriminant are biased and obtained the bias. Nanda (1949)
has shown that as the sample size increases the standard errors of the estimates of the coefficients decrease but do not converge to zero. Using these and other similar results one can construct confidence intervals and test the hypotheses regarding these estimates. Of particular interest is the hypotheses whether or not a certain coefficient is zero.

The second problem, namely the testing of the significance of the obtained discriminant function, was considered by Fisher by means of developing a test for $D^2$, the Mahalonobis distance. Rao (1946, 1948) obtained a test based on the distributional property of

$$F = \frac{(n_1 + n_2 - k - 1) n_1 n_2}{(n_1 + n_2) (n_1 + n_2 - 2) k} \frac{(x_1 - x_2)' S^{-1} (x_1 - x_2)}{(x_1 - x_2)' S^{-1} (x_1 - x_2)}$$

which is distributed as $F(k, n_1 + n_2 - k - 1)$. In the above expression $n_1, n_2$ are sample sizes, $x_1, x_2$ are sample means, $S^{-1}$ is the common covariance matrix and $k$ is the dimensionality of the underlying random variable.
5. **SEQUENTIAL VS. NONSEQUENTIAL CLASSIFICATION PROCEDURES WITH SEVERAL POPULATIONS**

For simplicity of presentation, we consider the case of 3 populations denoted by \( \pi_1, \pi_2 \) and \( \pi_3 \). Given an observation we wish to classify it to one of these three populations.

Let \( f_i(\xi) \) be the density associated with population \( \pi_i, i=1,2,3 \), \( p_i \) the prior probability of population \( \pi_i \), and \( C(k|i) \) the cost of allocating an item to \( \pi_k \) when it belongs to \( \pi_i \), for \( i,k=1,2,3 \). The Bayes classification rule, which minimizes the expected cost of misclassification is given as follows.

The observation \( \xi \) is classified to population \( \pi_k, k=1,2,3 \) for which

\[
\sum_{i=1}^{3} p_i C(k|i) f_i(\xi) \quad \text{is smallest.}
\]

If all the misclassification costs are equal, then the term in (5.1) will be smallest when the omitted term is largest. Thus, for equal cost of misclassification, the observation \( \xi \) is classified to

\[
\begin{align*}
\pi_1 & \text{ if } p_1 f_1(\xi) > p_2 f_2(\xi) &> p_3 f_3(\xi) \\
\pi_2 & \text{ if } p_2 f_2(\xi) > p_3 f_3(\xi) &> p_1 f_1(\xi) \\
\pi_3 & \text{ if } p_3 f_3(\xi) > p_1 f_1(\xi) &> p_2 f_2(\xi)
\end{align*}
\]
A second classification rule is obtained by comparing two populations at a time. If the costs of misclassification are equal then this rule is: classify $\xi$ to $\pi_1$ if $p_1 f_1(\xi) > p_2 f_2(\xi)$ and $p_1 f_1(\xi) > p_3 f_3(\xi)$ or equivalently if $p_1 f_1(x) - p_2 f_2(x) > 0$ and $p_1 f_1(x) - p_3 f_3(x) > 0$. The other two cases can be described in a similar manner. This procedure, which is alternatively written as (5.3) below is equivalent to (5.2) described earlier.

Allocate $\xi$ to $\pi_k$ if

$$\frac{f_k(\xi)}{f_1(\xi)} \geq \frac{p_i}{p_k} \quad \text{for all } i=1,2,3. \quad (5.3)$$

Note that in (5.3) one obtains the same inequality which is obtained in the case of two population classification, with the major difference that the desired inequality should be satisfied for all three possible values of $i$.

One may alternatively decide to follow a third procedure described below, which is sequential in nature. First allocate $\xi$ to $\pi_1$ or ($\pi_2$ or $\pi_3$) by Bayes rule. If the decision is to allocate $\xi$ to $\pi_2$ or $\pi_3$, then in this second stage allocate it to one of the two populations by again using the Bayes rule. Note that this third procedure is not equivalent to the Bayes rule described above.

It can be easily seen by means of an example. Consider the case of three univariate normal populations with common variance 1 and respective means 3, 5 and 6. Let the a priori probabilities be all equal to 1/3 and the costs of misclassification be also all equal. In this case, using procedure (5.2) the boundaries are obtained at 4 and 5.5. That is, if $x < 3$ classify it to population $\pi_1$ (with mean 3), if $4 \leq x < 5.5$ classify it to population $\pi_2$ (with mean 5)
and if $x > 5.5$ classify it to population $\pi_3$. On the otherhand, the third procedure described above, will allocate $x$ to $\pi_1$ if

$$\frac{\frac{1}{3} f_1(x)}{\frac{1}{3} f_2(x) + \frac{1}{3} f_3(x)} > 1$$

and to ($\pi_2$ or $\pi_3$) otherwise. In this case the boundary is given by 3.9075. In otherwords, if $x < 3.9075$ then it is classified to $\pi_1$ otherwise to $\pi_2$ or $\pi_3$. As before the boundary between $\pi_2$ and $\pi_3$ is 5.5.

In summary, the two alternative sequential procedures are different and clearly the first procedure of comparing two at a time is optimum.
6. **ON THE POSSIBILITY OF AN UNKNOWN GROUP**

Typically, in a classification problem it is assumed that there are a specified number $k$ of classes and the objective is to classify a new observation into one of these $k$ classes. In a more realistic situation it is possible that a given observation may not belong to any one of the given $k$ classes. In other words, there exists the possibility of the new object belonging to an unknown class, a class not previously specified. Thus, in such situation we encounter two problems (a) classify the given object in one of the $k$ given classes (b) show that there exists another class to which the new object belongs.

This problem has not been considered in great detail in the literature. This is because the class conditional density, the prior probability etc. are all unknown for this unknown class. However, in one particular situation the problem can be considered as seen below, (Rao, 1974).

Let $N(\mu, \Sigma)$ denote the normal density of a $p$-dimensional random vector with mean vector $\mu$ and covariance matrix $\Sigma$. Let $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ be two class conditional densities and let $\mu_1$, $\mu_2$ and $\Sigma$ be known (or estimated from very large training sets). Consider the well known Fisher linear classifier. For classifying a new observation $x$, it is given by

\[ x' \Sigma^{-1} (\mu_1 - \mu_2) - \frac{1}{2} (\mu_1 + \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \geq 0 \]

or equivalently by
Let us now consider normal densities with covariance matrix $\Sigma$ and mean vector given by $\lambda \mu_1 + (1-\lambda) \mu_2$. This density is given by

$$
\frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \lambda \mu_1 - (1-\lambda) \mu_2)' \Sigma^{-1} (x - \lambda \mu_1 - (1-\lambda) \mu_2) \right\}
$$

In the above representation, using the Neyman-Fisher factorization theorem, it is clear that $(x - \mu_1)' \Sigma^{-1} \mu_1 - \mu_2)$ is a sufficient statistic for the unknown parameter $\lambda$. As a consequence of this sufficient property, it suffices to know $Y = (x - \mu_1)' \Sigma^{-1} \mu_1 - \mu_2$ to draw statistical inference regarding all normal populations with means lying on the straight line joining $\mu_1$ and $\mu_2$.

Now consider the following problem. Given a new observation $x$ and two classes with densities $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$, the problem is to classify $x$ into one of these two classes. But our above result implies that $Y$ is sufficient for $\lambda$ and therefore, it should be possible to test whether $x$ belongs to a class which has the normal distribution with mean $\lambda \mu_1 + (1-\lambda) \mu_2$ and covariance $\Sigma$. The idea is that if $x$ does not belong to this class
of populations, it makes little sense to classify $x$ to one of the two specified classes. In short, first we wish to test the hypothesis

$$H_0: \text{mean} = \lambda \mu_1 + (1-\lambda)\mu_2, \lambda \text{ unknown}$$

versus

$$H_1: H_0 \text{ is not true.}$$

A test for this hypothesis is given by $T>C$ where

$$T = (x-\mu_1)' \Sigma^{-1} (x-\mu_1) - \frac{[(x-\mu_1)' \Sigma^{-1} (\mu_1-\mu_2)]^2}{(\mu_2-\mu_1)' \Sigma^{-1} (\mu_2-\mu_1)}$$

and $T$ is distributed as a chi-square random variable with $(p-1)$ degrees of freedom. Thus, we have the following result.

**Result:** Let $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ be two normal densities. Given $x$, classify it to one of the two populations. However, it may be possible that $x$ belongs to another class which is neither of the above two classes or any other normal population with mean lying on a straight line joining these means. Then one can follow these steps:

**Step 1:** First test the hypothesis $H_0$ vs. $H_1$ using $T$. If $T > C$, where $C$ is obtained by using the chi-square property of $T$, then we conclude that $x$ does not belong to any one of the two specified classes.

**Step 2:** If $T < C$, then use the usual Fisher linear classifier to classify $x$ to one of the two specified class.

**Example:** Suppose $\mu_1 = (2,6), \mu_2 = (4,9)$ with common covariance matrix $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Then, a given observation $x' = (13,18)$ should be classified to one of these two populations only if there is evid-
ence that it does not belong to any other class. For the above problem

\[ T = \frac{1107}{20} - \frac{261^2}{63/20} = 8.57 \]

With 1 d.f. the acceptable value of a chi-square random variable is 3.841 at level of significance 0.05 and 7.879 at level of significance 0.005. Hence, one would conclude that this observation does not belong to any one of the two specified classes.

It is worthwhile to note that the above procedure is developed for the case when the parameters \( u \)'s and \( E \) are all known. If they are unknown and the training set is large, then the above procedure applies in the asymptotic sense. For small training set a satisfactory procedure has not been developed.

In a recent publication Lin (1978) suggests a variation of the idea used in Neyman-Pearson theory of hypothesis testing. To understand his approach, consider the problem of testing a simple hypothesis \( H_0 \) versus a simple hypothesis \( H_1 \) where

\[ H_0: f(x) = f_0(x) \]

and

\[ H_1: f(x) = f_1(x) \]

and \( f(x) \) denotes the probability density function of the random variable \( x \). According to the Neyman-Pearson theory a test for the above problem is obtained by minimizing the probability \( P[x \text{ is classified as having the pdf } f_0(x) | \text{true pdf is } f_1(x)] \) keeping the probability, \( P[x \text{ is classified as having the pdf } f_1(x) | \text{the true pdf if } f_0(x)] \) fixed. Equivalently, if the entire sample space \( S \) is partitioned in two sets \( R \) and \( R^C = S-R \), and \( x \)
is classified as having the pdf $f_0(x)$ if $x \in R$, then according to the Neyman Pearson theory $R$ is chosen such that:

$$\alpha = \int_{R^C} f_0(x) \, dx$$

is fixed, and

$$\beta = \int_{R} f_1(x) \, dx$$

is minimized.

In the absence of the knowledge of $f_1(x)$, Lin suggests the following test: Choose $R$ such that

$$\alpha = \int_{R^C} f_0(x) \, dx$$

is fixed and

$$V(R) = \int_{R} dx$$

is minimized.

An adaptation of this concept to the classification problem is suggested as follows. Suppose the problem is to classify $x$ to one of the two classes with respective pdf's $h_1(x)$ and $h_2(x)$ and the prior probabilities $p_1$ and $p_2$ respectively. However, let there exists a possibility that the object may not belong to any one of these two classes. In this situation Lin suggests that one can use a two step procedure

**Step (a) Test for the hypothesis**

$$H_0: f(x) = f_0(x) \equiv p_1 h_1(x) + p_2 h_2(x)$$

**vs**

$$H_1: \text{density is unknown}$$

**Step (b).** If in Step (a) the null hypothesis is accepted, then apply the conventional classification rule.
Remark: It can be easily seen that Lin's proposal, in the framework of testing, is to test the null hypothesis

\[ H_0: f(x) = f(x) \]

vs

\[ H_1: f(x) = \text{uniform over a certain unknown interval.} \]

Thus, Lin is replacing the unknown density by the uniform density.
7. ON SELECTION OF THE BEST k OUT OF n MEASUREMENTS IN GAUSSIAN DISTRIBUTIONS

7.0 ABSTRACT:

The purpose of this note is to show that it is not possible to obtain a subset of k measurements out of a set of n measurements, which provides the best discrimination between two populations by extending the set of (k-1) best measurements. This result is demonstrated for the Gaussian distribution.

7.1 INTRODUCTION:

Cover and Campenhoul ( ) considered the problem of selecting the best k out of n measurements, for the purpose of discriminating between two populations. They showed that it is not possible to extend the set of best (k-1) measurements to obtain the set of best k measurements. In fact, they proved that there does not exist any systematic method of obtaining a subset of k measurements for the purpose of discriminating. Only the exactive search provides the desired answer.

In order to prove the above mentioned property Cover and Campenhoul first related n measurements of the distribution under consideration to n^{n-1} Gaussian random variables, then established the result for these new Gaussian variables. This author has not been able to follow their method of relating n variables of some distribution to n^{n-1} Gaussian random variables. Moreover, it is not clear how can one obtain any desired ordering in the subsets of n measurement, in terms of probability of error of misclassification and also obtain the specified magnitude of these error probabilities.
The purpose of this note is to show that any extension of the best \((k-1)\) measurements to a set of \(k\) measurements \((k<n)\) need not give us the set of best \(k\) measurements. This is demonstrated in the case of Gaussian random variables and by means of two simple examples. It is also demonstrated, by means of an example, that although it may be possible to obtain a desired ordering of subsets of \(n\) measurement (subject to a natural constraint given below and also in Cover and Campenholt), it may not be possible to obtain the desired magnitudes of the error probabilities. This later property is also obtained in the case of Gaussian random variables.

Before proceeding further we wish to recall that this phenomenon, of not being able to select best subset of \(k\) out of \(n\), also occurs in the context of regression analysis [see Draper and Smith (1966)].
7.2 TWO EXAMPLES TO SHOW THAT ANY EXTENSION OF "BEST" k TO "k+1"
MEASUREMENTS MAY NOT BE THE "BEST" SET OF (k+1) MEASUREMENTS:

Before presenting the examples we wish to recall two basic results.

**Result (A):** Let \( \mathbf{x} \) be a \( n \)-dimensional normal random vector with mean \( \mu_i \) and covariance matrix \( \Sigma_i \) where \( i=1 \) or \( 2 \) depending upon whether \( \mathbf{x} \) was drawn from population \( \pi_1 \) or \( \pi_2 \). Let the prior probabilities of \( \pi_1 \) or \( \pi_2 \) be equal and the costs of misclassification be equal. Then, the minimum probability of misclassification, is given by \( \Phi(-\frac{A}{2}) \) where \( \Phi \) denotes the distribution function of the standard normal random variable and

\[
A^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)
\]

**Result (B):** Let \( A \) be a \((p+q) \times (p+q)\) positive definite symmetric matrix and \( \mathbf{x} \) be a \((p+q)\) dimensional column vector. Let \( A \) and \( \mathbf{x} \) be partitioned as follows.

\[
A = \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix}
\quad \text{and} \quad
\mathbf{x} = \begin{bmatrix}
\mathbf{x}_1 \\
\mathbf{x}_2
\end{bmatrix}
\]

In terms of these partitioned matrices we have the following well known equality.

\[
\mathbf{x}' A^{-1} \mathbf{x} = \mathbf{x}_1' A_{11}^{-1} \mathbf{x}_1 + \mathbf{x}_2' A_{22.1}^{-1} \mathbf{x}_2
\]

where \( \mathbf{x}_1 \) and \( \mathbf{x}_2 \) are \( p \) and \( q \) dimensional vectors, \( A_{11}, A_{22} \) and \( A_{12} = \mathbf{A}_{21} \) are \( p \times p, q \times q \) and \( p \times q \) dimensional matrices and

\[
A_{22.1} = A_{22} - A_{21} A_{11}^{-1} A_{12}
\]

\[
A_{2.1} = A_2 - A_{21} A_{11}^{-1} A_1
\]

In particular if \( A_{12} \) is a matrix of all zero elements, then,
from result (B) the following equality is obtained.

\[
\mathbf{c}' \mathbf{A}^{-1} \mathbf{c} = \mathbf{c}_1' \mathbf{A}_{11}^{-1} \mathbf{c}_1 + \mathbf{c}_2' \mathbf{A}_{22}^{-1} \mathbf{c}_2
\]

Let \( \mathbf{c} = \mathbf{u}_1 - \mathbf{u}_2 = (\mathbf{c}_1', \ldots, \mathbf{c}_n')' \). From result (A) it follows that if one wishes to choose only one component out of \( n \), then the optimum choice is to select the \( i \)th component such that

\[
\frac{\delta_i^2}{\sigma_{ji}} = \max_{i \leq j \leq n} \frac{\delta_i^2}{\sigma_{ii}}
\]

Example 1: In this example we consider the special case \( n=3 \). It is observed that the best singleton set is \( \{X_1\} \) but neither \( \{X_1, X_2\} \) nor \( \{X_1, X_3\} \) is the best set of two components.

Let \( \mathbf{c} \) be a 3-dimensional normal random variable such that

\[
\left( \frac{\delta_1}{\sigma_1}, \frac{\delta_2}{\sigma_2}, \frac{\delta_3}{\sigma_3} \right) = (2, 1.5, 1)
\]

and the correlation matrix is given by

\[
P = (\rho_{ij}) = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & .96 \\ \rho & .96 & 1 \end{bmatrix}, \rho = \sqrt{\frac{1}{2}} \tag{7.1}
\]

Clearly, the best singleton set of measurements is given by \( \{X_1\} \), because \( (\delta_1^2/\sigma_1^2) \) is largest among all \( (\delta_i^2/\sigma_i^2) \) for \( i=1,2,3 \).

Next, we calculate

\[
\left( \left( \frac{\delta_1}{\sigma_1}, \frac{\delta_2}{\sigma_2}, \frac{\delta_3}{\sigma_3} \right) \right)' \begin{bmatrix} \sigma_1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3 \end{bmatrix}^{-1} \left( \left( \frac{\delta_1}{\sigma_1}, \frac{\delta_2}{\sigma_2}, \frac{\delta_3}{\sigma_3} \right) \right) = \begin{bmatrix} \frac{\delta_1^2}{\sigma_1^2} - 2\rho_{12} \frac{\delta_1 \delta_2}{\sigma_1 \sigma_2} + \frac{\delta_2^2}{\sigma_2^2} \end{bmatrix} = 4.31
\]

and finally

\[
\left( \left( \frac{\delta_2}{\sigma_2}, \frac{\delta_3}{\sigma_3} \right) \right)' \begin{bmatrix} \sigma_2 & \sigma_{23} & \sigma_{23} \\ \sigma_{23} & \sigma_3 & \sigma_{33} \end{bmatrix}^{-1} \left( \left( \frac{\delta_2}{\sigma_2}, \frac{\delta_3}{\sigma_3} \right) \right) = \begin{bmatrix} \frac{\delta_2^2}{\sigma_2^2} - 2\rho_{23} \frac{\delta_2 \delta_3}{\sigma_2 \sigma_3} + \frac{\delta_3^2}{\sigma_3^2} \end{bmatrix} = 4.719
\]
It is obvious that the set \( \{X_2, X_3\} \) is the best set of two components, which it is not an extension of \( \{X_1\} \).

**Example 2:** The above example is extended to arbitrary \( n \). In other words we show that there exists a normal distribution such that the best set of \( k \) components, when extended to \( (k+1) \) components does not provide the best set of \( (k+1) \) components. Without loss of generality let \( \{X_1, \ldots, X_{k-1}\} \) be the best set of \( (k-1) \) components and let \( n=k+2 \). Let \( \mathbf{X} = (X_1, \ldots, X_{k+2})' \) be normally distributed with mean vectors \( \mu_1 \) and \( \mu_2 \) and common correlation matrix \( \Sigma \), given that it belongs to population \( \pi_1 \) and \( \pi_2 \) respectively. Let \( \xi = \mu_1 - \mu_2 \) and \( \Sigma \) be such that

\[
\begin{pmatrix}
\frac{\delta_1}{\sigma_1} & \ldots & \frac{\delta_{k+2}}{\sigma_{k+2}} \\
\end{pmatrix} = \begin{pmatrix}
\frac{\delta_1}{\sigma_1} & \ldots & \frac{\delta_{k-1}}{\sigma_{k-1}} \\
\end{pmatrix}, 2, 1.5, 1
\]

and the correlation matrix, \( R \), corresponding to \( \Sigma \) satisfies

\[
R = \begin{bmatrix}
R_{11} & 0 \\
\vdots & \ddots & \ddots \\
0 & \ddots & 0 & P \\
\end{bmatrix}
\]

where \( P \) is given by (7.1) and \( R_{11} \) is a \((k-1) \times (k-1)\) dimensional matrix. From result (B), our assumption that the best set of \( (k-1) \) components is \( \{X_1, \ldots, X_{k-1}\} \), and Example 1 it is obvious that the best set of \( k \) components is given by \( \{X_1, \ldots, X_k\} \). But, when we search for the best set of \( (k+1) \) components, it turns out to be \( \{X_1, \ldots, X_{k-1}, X_{k+1}, X_{k+2}\} \) which is not an extension of \( \{X_1, \ldots, X_k\} \).

**Remark:** The above result is of a negative nature. A more useful result would be to discover conditions such that it would be possible to obtain the best set of \( k \) components by extending the best set of \( k \) components by extending the best set of \( (k-1) \) components.
This work is under investigation.

In the remaining part of this note we consider one other aspect of Cover and Campenhout's result. This example appears to contradict their basic theorem.

Suppose \( M_1, M_2, \ldots, M_n \) are \( n \)-measurements (Scalar) and \( S_i \) denotes an arbitrary, non-empty, subset of these measurements and \( P_e(S_i) \) denotes the minimal probability of error when the elements of \( S_i \) are used for the purpose of classification (between two classes). It is well known that if \( S_i \subset S_j \) then \( P_e(S_i) \geq P_e(S_j) \). Probability of error \( P_e(S_i) \) can be used to establish an ordering among all \( (2^n - 1) \) possible selections of measurements. Cover and Campenhout state the following theorem.

**Theorem:** Given an arbitrary ordering on the subsets of measurements \( M_1, M_2, \ldots, M_n \), subject to the monotonicity constraints, there exists a jointly normally distributed random vector \( \mathbf{x} \) of \( n \) dimension which has exactly the same ordering and the same probability of error.

**Example 3:** The following example shows that the above result is not correct. Suppose \( M_1, M_2, M_3 \) are three measurements. The 7 possible non empty measurement selections are ordered below (subject to the monotonicity criterion mentioned above). The corresponding error probabilities are also specified. Let the ordering be \( \{M_2\} \geq \{M_3\} \geq \{M_1\} \geq \{M_1, M_2\} \geq \{M_1, M_3\} \geq \{M_2, M_3\} \geq \{M_1, M_2, M_3\} \), and the corresponding error probabilities be \( .4, .38, .35, .3, .29, .28 \) and \( .22 \) respectively. At this stage we are interested in the following question.

Is it possible to find \( \mathbf{x}' = (X_1, X_2, X_3) \) which is normally dis-
tributed with mean \( \mu_i \), \( i=1,2 \) depending on the class membership and common covariance matrix \( \Sigma \) such that \( X \)'s have exactly the above ordering and error magnitudes? In this simple example, we observe that it is possible to have the same order but it is not possible to match the error probabilities.

For a normal random vector, the error probability is \( \Phi(-\frac{A}{\sigma}) \) where \( A^2 \) is the Mahalanobis distance. Thus, it suffices to obtain the mean vectors \( \mu_1 \) and \( \mu_2 \) and then common covariance matrix \( \Sigma=(\sigma_{ij}) \) such that if \( \delta = \mu_1 - \mu_2 = (\delta_1, \delta_2, \delta_3)' \) then,

\[
\Phi\left(-\frac{\delta_1}{\sigma_1}\right) = .35, \quad \Phi\left(-\frac{\delta_2}{\sigma_2}\right) = .8, \quad \Phi\left(-\frac{\delta_3}{\sigma_3}\right) = .3
\]

\[
\Phi\left[-\frac{1}{2}\left( \frac{\delta_2^2}{\sigma_1^2} - 2\rho_{12} \frac{\delta_1 \delta_2}{\sigma_1 \sigma_2} + \frac{\delta_2^2}{\sigma_2^2} \right) \right] = .3
\]

\[
\Phi\left[-\frac{1}{2}\left( \frac{\delta_3^2}{\sigma_1^2} - 2\rho_{13} \frac{\delta_1 \delta_3}{\sigma_1 \sigma_3} + \frac{\delta_3^2}{\sigma_3^2} \right) \right] = .29
\]

\[
\Phi\left[-\frac{1}{2}\left( \frac{\delta_3^2}{\sigma_2^2} - 2\rho_{23} \frac{\delta_2 \delta_3}{\sigma_2 \sigma_3} + \frac{\delta_3^2}{\sigma_3^2} \right) \right] = .28
\]

and

\[
\Phi\left[-\frac{1}{2}(\delta' \Sigma^{-1} \delta)^{1/2}\right] = .22
\]

where, as before, \( \sigma_i = \sigma_{ii}^{1/2} \) and \( \rho_{ij} = \sigma_{ij}/\sigma_i \sigma_j \), \( i,j = 1,2,3 \).

Solving the first three equations in (6) we get

\[
\frac{\delta_1}{\sigma_1} = .7706, \quad \frac{\delta_2}{\sigma_2} = .5066 \quad \text{and} \quad \frac{\delta_3}{\sigma_3} = .611
\]

The next three equations, along with the above values of \( (\delta_i/\sigma_i)'s \), specify \( \rho_{ij} \). These must be \( \rho_{12} = -.23966 \) or \( .9489 \), \( \rho_{13} = -.2135 \) or \( .9828 \) and \( \rho_{23} = -.5392 \) or \( .9948 \).
Although, it may appear that are 8 possible correlation matrices such that the first six equalities in (6) are satisfied, however only one of these P matrix is positive definite. This matrix is given below. The remaining seven matrices are all negative definite.

\[
P = \begin{bmatrix}
1 & -.23966 & -.2135 \\
-.23966 & 1 & -.5392 \\
-.2135 & -.5392 & 1
\end{bmatrix}
\]

Given the above values of $\rho$ and $P$, the remaining probabilities of error are fixed. Consequently, the last equality in (7.2) will not be satisfied. In this particular example

\[
\phi\left\{\frac{1}{2}(\delta'_1\Sigma^{-1}\delta_1)^{1/2}\right\} = \phi\left\{\frac{1}{2}(\delta_2', \rho^{-1}(\delta_2')^{1/2}\right\} = .1714
\]

which is different from the desired value of .22.

In general, suppose $M_1, M_2, \ldots, M_n$, $n \geq 3$ are $n$ measurements. A certain order among subsets of $\{M_1, M_2, \ldots, M_n\}$ and the associated error probabilities are given. The order satisfies the natural constraints. Using the error magnitudes corresponding to the singletons $\{M_i\}$ we obtain $(\delta_i/\sigma_i)$'s. Using the error magnitudes corresponding to $\{M_i, M_j\}$'s we get $P_{ij}$'s. At this stage all of the free parameters are fixed and consequently all of the other error probabilities are also fixed. It is unlikely that a specified order among the components will be satisfied.

Remark: Cover and Campenhout have generated $n^2^{n-1}$ Gaussian random variables for the $n$ original measurements. This gives them enough freedom of selection to match the error probabilities. But, their
process of arriving at $n$ normal random variables from these $n^{2^n-1}$ variables is not at all clear. This is the major source of the disagreement between their theorem and our counter example.
8.1. **INTRODUCTION**

In discriminant analysis, we try to allocate objects into one of the given classes based on some measurements of the objects. There exists instances where several independent attempts have been made to classify a population of objects, each time a different set of measurements are chosen and a new decision rule is constructed. Suppose we are presented with the results of these decision rules, and we are asked to utilize these results to classify the population of objects with a performance better than each of the decision rules, whenever possible. In this paper we investigate such methods for two class and three class problems. We confine to the case when only three independent sets of measurements are taken. The results can be generalized in a similar manner for other cases.

Let $X = (X_1, X_2, X_3)^t$ be a vector constructed by the juxtaposition of the 3 sets of measurements $X_i$ of an object. The dimension for each measurement vector $X_i$ is $p_i$. Thus $X$ is of dimension $p$, where $p = p_1 + p_2 + p_3$. Let $X' = (D_1(X_1), D_2(X_2), D_3(X_3))$, where $D_i$'s are the three discriminants given, and $D_1(X_1)$ is the decision of the discriminant $D_1$ based on the set of measurements $X_1$ of the object.

The purpose of this paper is to develop some schemes such that an object represented by the vector $X$ can be classified in one of the two classes $w_1$ and $w_2$. 

32
In some cases, the set of measurements on an object are naturally grouped in subsets such as the $X_i$'s. For instance if the classification was performed on three different aspects of the same individual. Since the complexity in constructing a decision rule for an object vector $X$ increases with the dimensionality $p$ of $X$, it may be beneficial to develop a decision rule for each of the subsets of measurements $X_i$, and combine the results of the different decision rules in a sensible fashion to obtain a decision rule for $X$. This avoids the complications caused by high dimensionality, yet all the features of the object are considered in the classification. In other cases it may be difficult to get the complete data at one place.

In section 8.2, we consider the intuitively appealing majority logic and in Section 8.3 the optimum method of combining the results of three discriminants in the two class problem. Section 8.4 contains some general remarks on the two class problem. In Section 8.5, it is demonstrated that there would always be some loss in terms of probability of correct classification when the three discriminants are combined as opposed to using the best possible discriminant for the entire $X$. In Section 8.6 the case of multivariate normal is further investigated. We finish this paper with a quick review of these ideas as they apply to the three class problem.
2. THE MAJORITY RULE

Without loss of generality the function $D_1(X_1)$ can be defined as

$$D_1(X_1) = \begin{cases} 1 & \text{if } X_1 \text{ is classified into } w_1 \text{ by } D_1 \\ 0 & \text{if } X_1 \text{ is classified into } w_2 \text{ by } D_1 \end{cases}$$

Clearly, the $D_i$'s are independent from one another provided the $X_i$'s are mutually independent. In the following derivations, we assume that the $X_i$'s, for $i = 1, 2, 3$ are mutually independent.

A random vector $X$ in the object space is thus mapped by the function $D_i$'s into the 3-dimensional binomial space $S'$, where $S' = \{(000)^t, (001)^t, (010)^t, (011)^t, (100)^t, (101)^t, (110)^t, (111)^t\}$.

In this section we consider the majority rule of combining the results from $D_i$'s which belong to the sample space $S'$.

The majority rule is an intuitive approach to classify the samples in $S'$. The rule is: whenever more than one $X_i$'s are classified by the $D_i$'s to $w_1$, then $X$ should be classified to $w_1$.

Let $D_m$ denote the majority decision function i.e. let $D_m(X) = 1$ represent that $X$ is classified to $w_1$. Thus $D_m(X) = 1$ iff $\sum_{i=1}^{3} D_i(X_i) \geq 2$, else $D_m(X) = 0$ and $X$ is classified to $w_2$.

Let $Q = Pr(D_m(X) = 1|w_1)$, the probability of correctly recognizing an object from $w_1$. Let $a_i = Pr(D_i(X) = 1|w_1)$. We assume that $a_i \geq \frac{1}{2}$ for all $i=1, 2$ and $3$, such that $a_1 \leq a_2 \leq a_3$. 


The above theorem tells us that the intuitively appealing majority rule is sometimes inferior in performance to the best of the $D_i$'s. It may be possible to improve the performance by, instead of using the majority rule, using a linear combination of the $D_i(X_1)$'s, $\frac{1}{3} \sum_{i=1}^{3} C_i D_i(X_1)$, where the weight $C_i$ is chosen to reflect
the magnitude of the $a_1$'s. However, we will not disagress into this. Instead, we will derive a nonlinear procedure which provides the best possible discriminant based on the $D_1$'s.
8.3. THE LIKELIHOOD RATIO DECISION RULE

Let $\alpha_i = \Pr(D_i(X_1) = 1|\omega_1)$, $\beta_i = \Pr(D_i(X_1) = 0|\omega_2)$, $i = 1, 2, 3$.

The $\alpha_i$'s and $\beta_i$'s are the conditional probabilities of correct recognition for observations from $\omega_1$ and $\omega_2$, respectively. $1-\alpha_i$ and $1-\beta_i$ will be probabilities of misclassification. Let

$$f \equiv f(\alpha_i) = \prod_{i=1}^{3} \Pr(D_i(X_1)|\omega_1), \quad g \equiv g(\beta_i) = \prod_{i=1}^{3} \Pr(D_i(X_1)|\omega_2).$$

The likelihood ratio decision rule $D_1$ is such that if $f/g > 1$, decide $X_{\omega_1}$, if $f/g < 1$ decide $X_{\omega_2}$, and if $f/g = 1$ then decide $X_{\omega_1}$ with probability $0.5$.

**Theorem 2** Given the decisions of the $D_i$'s, the decision rule $D_1$ is the optimal decision rule.

**Proof** Since $D_i(X_1)$'s are mutually independent, the following equality holds

$$\prod_{i=1}^{3} \Pr(D_i(X_1)|\omega_j) = \Pr(D_1 \cap D_2 \cap D_3|\omega_j).$$

Therefore the decision rule $D_1$ is the same as the most powerful test for a simple hypothesis $X_{\omega_1}$ versus a simple alternative $X_{\omega_2}$, given by the Neyman-Pearson lemma. Consequently, with the obvious choices of the constants of the Neyman-Pearson lemma the discriminant $D_1$ is optimum.

It remains to find the performance of the above, optimal decision rule $D_1$. The proposition 3 given below not only answers this question but also relates the two decision rules $D_1$ and $D_2$ discussed in Theorem 1. In the proposition 3 we also assume $\alpha_i = \beta_i$, $i = 1, 2, 3$.

**Proposition 3** Let $\alpha_i = \beta_i$, $i = 1, 2, 3$, and $\alpha_1 \leq \alpha_2 \leq \alpha_3$. The probability of the correct decision by $D_2$ is $Q^*$ where

$$Q^* = a_1 a_2 a_3 + a_1 a_2 (1-a_3) + a_1 (1-a_2) a_3 + \max((1-a_1) a_2 a_3, a_1 (1-a_2) (1-a_3))$$

$$= \max(\text{probability of correct decision by } D_m, \text{probability of correct decision by } D_3).$$
Proof. By $D$, any observation $(D_1(X_1), D_2(X_2), D_3(X_3))$ will be classified as belonging to $w_1$ when it actually belongs to $w_1$ provided

$$\sum_{i=1}^{3} D_i(X_i) \left(1-D_i(X_i) \right) > \sum_{i=1}^{3} (1-D_i(X_i)) a_i$$

or equivalently,

$$\sum_{i=1}^{3} \frac{a_i}{\left(\frac{1-a_i}{1-a_1}\right)} > \sum_{i=1}^{3} \frac{a_i}{\left(\frac{1-a_i}{1-a_1}\right)}$$

(with probability .5). Let $(D_1(X_1), D_2(X_2), D_3(X_3))$ takes values in \{(1,1,1), (0,1,1), (1,0,1)\}. Clearly the decision rule $D$ will classify $x \in w_1$ because $a_1(1-a_1)^{-1} \leq a_2(1-a_2)^{-1} \leq a_3(1-a_3)^{-1}$. Similarly if $(D_1(X_1), D_2(X_2), D_3(X_3))$ takes values in \{(0,0,0), (1,0,0), (0,1,0)\} then the decision rule $D$ will classify them as coming from $w_2$. The only other cases are when $(D_1(X_1), D_2(X_2), D_3(X_3)) = (1,1,0)$ or $(0,0,1)$. Obviously, we could classify $(1,1,0)$ as coming from $w_1(w_2)$ if

$$\frac{a_1 a_2 (1-a_3)}{(1-a_1)(1-a_2)a_3} > (\frac{<}{<}) 1$$

and we could classify $(0,0,1)$ as coming from $w_1(w_2)$ provided

$$\frac{(1-a_1)(1-a_2)a_3}{a_1 a_2 (1-a_3)} > (\frac{<}{<}) 1$$

By symmetry of the above two cases one and only one of these two points will result in the acceptance of $w_1$. Since the above two conditions are the same as in the second part of Theorem 1, the $G^*$ is given by the first equation. To show that the second equality holds, we observe that if

$$G^* = a_1 a_2 a_3 + a_1 a_2 (1-a_3) + a_1 (1-a_2) a_3 + a_1 (1-a_2) (1-a_3)$$

then the right hand side simplifies to $a_3$. 38
8.4 SOME REMARKS

Since both of the above suggested decision rules are based on the samples in $S'$ only, $(S' = \{(000)^t, \ldots, (111)^t\})$, once the segments $X_1$ of an observations $X$, are classified by the $D_1$'s, we can use template matching (checking the content of the $3 \times 3$ vector $X'$ against all the possible elements in $S'$) to decide the classification of $X$. Thus, classification using either the majority or optimum decision rule involves the same amount of work.

For the majority rule to work properly, the number of subgroups $k$ of the elements of $X$ should be an odd number. For the likelihood ratio decision rule, there is no such restriction on the number of subgroups of the elements of $X$, and we can be sure that this rule is always the best based on the available information on elements of $S'$. 
Earlier in the introduction it was said if $p$, the dimension of $X = (X_1, X_2, X_3)$, is very large we may prefer to construct first the $D_1, D_2, D_3$ and then combine their results to form $D_8$. The discriminants $D_1, D_2, D_3$ along with $D_8$ are assumed to be optimum and $D_8$'s are assumed independent. Suppose, on the other hand we construct the optimum discriminant $D$ using the whole set $X$. Then a natural question is: Are $D$ and $D_8$ different from each other?

There is no doubt that $D$, being optimum, must perform at least as well as $D_8$ in terms of probability of correct decision. Thus all we need to verify is that: is $D_8$ inferior in its performance? In the following discussion we have tried to answer this question under some assumptions, which are reasonable for two class problems.

Suppose apriori-probabilities of the two classes are the same. Let $f_j^{(k)}(X_j)$ be the p.d.f. of $X_j$ when $w_k$ is the true class, $k = 1, 2, j = 1, 2, 3$. Then, by rule of optimum decision (we assume all of the parameters are known)

$$
D_j(X) = \begin{cases} 
1 & \text{i.e. } X_j \text{ is classified as from } w_1 \text{ if } \frac{f_j^{(1)}(X_j)}{f_j^{(2)}(X_j)} > 1 \\
0 & \text{i.e. } X_j \text{ is classified as from } w_2 \text{ otherwise.}
\end{cases}
$$

$j = 1, 2, 3$. Similarly

$$
D(X) = \begin{cases} 
1 & \text{i.e. } X \text{ is classified as from } w_1 \text{ if } \frac{f_j^{(1)}(X)}{f_j^{(2)}(X)} > 1 \\
0 & \text{i.e. } X \text{ is classified as from } w_2 \text{ if } \frac{f_j^{(1)}(X)}{f_j^{(2)}(X)} < 1
\end{cases}
$$

[for sake of simplicity we assume that $\frac{f_j^{(1)}(X)}{f_j^{(2)}(X)} = 1$ with probability zero].
First we consider the case where

(a) \( a_j = \Pr(D_j(X_j) = 1|w_1) = \Pr(D_j(X_j) = 0|w_2) \)

(b) \( \frac{1}{2} \leq a_1 \leq a_2 \leq a_3 \) such that \( a_1 a_2 (1-a_3) > (1-a_1)(1-a_2)a_3 \)

so that \( D_2 \) is the majority rule: 2 out of 3 rule.

Under the above assumptions,

\[
\Pr(D(X) = 1|w_1) = \left[ \prod_{j=1}^{3} \frac{f_j^{(1)}(X_j)}{f_j^{(2)}(X_j)} > 1 \right] w_1
\]

\[
= \int_A \prod_{j=1}^{3} f_j^{(1)}(X_j) \, dx_j
\]

where \( A = \left\{ X : \prod_{j=1}^{3} \frac{f_j^{(1)}(X_j)}{f_j^{(2)}(X_j)} > 1 \right\} \).

Let \( \beta_j(X_j) \equiv \beta_j = \frac{f_j^{(1)}(X_j)}{f_j^{(2)}(X_j)} \), \( j = 1, 2, 3 \). Then

\[
A = \left\{ X : \beta_1 \beta_2 \beta_3 > 1 \right\}.
\]

Clearly, the set \( A \) contains \( X \) such that at least one of the three \( \beta_j \)'s is larger than 1 and, of course, \( \beta_1 \beta_2 \beta_3 > 1 \). We can easily see that \( A \) can be written in terms of the union of 7 disjoint sets \( A_1 \) through \( A_7 \) where

\[
A_1 = \left\{ X : \beta_1 > 1, \beta_2 > 1, \beta_3 > 1 \right\}
\]

\[
A_2 = \left\{ X : \beta_1 > 1, \beta_2 > 1, \beta_3 < 1, \beta_1 \beta_2^{-1} < \beta_3 \right\} ; A_3 \text{ and } A_4 \text{ are defined similarly}
\]

with \( \beta_3 \) replaced by \( \beta_2 \) and \( \beta_1 \) respectively.

\[
A_5 = \left\{ X : \beta_1 > 1, \beta_2 < 1, \beta_3 < 1 \text{ such that } \beta_1 (\beta_2 \beta_3)^{-1} \right\}
\]

\( A_6 \) and \( A_7 \) defined in a similar manner. Thus

\[
\Pr(D(X) = 1|w_1) = \int_{\bigcup_{A_1}^{A_7}} \prod_{j=1}^{3} f_j^{(1)}(X_j) \, dx_j
\]  

(8.1)
Next,
\[
Pr(D_2(X) = 1|a_1) = Pr(\text{At least two out of three } D_i\text{'s are } 1|a_1)
\]
\[
= a_1 a_2 a_3 + \frac{1}{3} a_1 (1-a_2) a_3 + \frac{1}{3} (1-a_1) a_2 a_3
\]
\[
= \int_0^1 f_j^{(1)}(x_j) \, dx_j
\]
which follows from independence of \(X_j\)'s \(j = 1,2,3\); the set
\[
B = A_1 \cup B^e
\]
where \(B^e\) consists of all those \(X\)'s for which exactly two \(D_i\)'s are 1. Consider the case obtained from \(a_1 a_2 (1-a_3)\). This contributes the set
\[
\begin{align*}
\{x : & \beta_1 > 1, \beta_2 > 1, \beta_3 < 1\} \\
= & \{x : \beta_1 > 1, \beta_2 > 1, (\beta_1 \beta_2)^{-1} < \beta_3 < 1\} \cup \{x : \beta_1 > 1, \beta_2 > 1, \beta_3 < (\beta_1 \beta_2)^{-1}\}
\end{align*}
\]
\[
= A_2 \cup B_2
\]
to the set \(B^e\). In a manner similar to this, one can easily show that
\[
B^e = A_2 \cup A_4 \cup A_5 \cup B_2 \cup B_3 \cup B_4
\]
Thus, the two integrals given by (61) and (62) differ from each other only over the sets \(A_5 \cup A_6 \cup A_7\) and \(B_2 \cup B_3 \cup B_4\). Thus, it can be easily concluded that the two discriminants \(D\) and \(D_4\) are not identical because for any \(x \in A_1\), \(D(x) = 1\) whereas \(D(x) = 0\) and conversely for any \(x \in B_j, D_2(x) = 1\) whereas \(D_2(x) = 0\).

One could still get equal probabilities of errors. To see if these probabilities are same or not we consider
\[
Pr(D(x) = 1|a_1) - Pr(D_4(x) = 1|a_1)
\]
\[
= \left[f_2 - f_4\right] \int_{A_j} \int_{B_j} f_j^{(1)}(x_k) \, dx_k = \theta_2 - \frac{1}{2}
\]
In a manner similar to the above it can be shown that

$$\Pr(D(x) = 0|\omega_1) - \Pr(D_k(x) = 0|\omega_2) = \left(\int_{\mathcal{A}_j} - \int_{\mathcal{A}_j}\right) \prod_{\mathcal{A}_j} f_k(x_k) dx_k = \phi_1 - \phi_2$$

By assumption the above two differences $\theta_1 - \theta_2$ and $\phi_1 - \phi_2$ are equal. But

$$\theta_1 = \int_{\mathcal{A}_j} \prod_{\mathcal{A}_j} f_k^{(2)}(x_k) dx_k > \int_{\mathcal{A}_j} \prod_{\mathcal{A}_j} f_k^{(2)}(x_k) dx_k = \phi_2$$

and

$$\theta_2 = \int_{\mathcal{A}_j} \prod_{\mathcal{A}_j} f_k^{(2)}(x_k) dx_k < \int_{\mathcal{A}_j} \prod_{\mathcal{A}_j} f_k^{(2)}(x_k) dx_k = \phi_1$$

$$. \therefore \theta_1 - \theta_2 > 0 \text{ and } \theta_2 - \theta_1 > 0.$

Now, we observe that $\theta_1 \geq \theta_2$ because $D$ is better than $D_k$. If $\theta_1 = \theta_2$ then $\phi_1 = \phi_2$ and on the other hand

$$\phi_1 > \theta_2 = \theta_1 > \phi_2$$

would imply $\phi_1 > \phi_2$, a contradiction. Thus we must have $\theta_1 > \theta_2$. This implies that $D$ is always better than $D_k$ except in the case when probabilities of the sets $\omega_2$ through $\omega_7$ and $D_2$ to $D_4$ are zero under $\omega_2$ as well as $\omega_1$.

In the remaining part of this section we consider the special case of the normal population which plays a significant role in discriminant analysis.

Consider, once again, the simplest form of the discrimination problem. Let
follows multivariate normal distribution with mean vector \( \mu_i^{(1)} \) and covariance matrix \( \Sigma_i \); \( i=1,2,3 \) and \( j=1,2 \) associated with two classes. We assume that \( X_i \)'s are all stochastically independent. Under the above assumptions, the probabilities of correct classification are given by \( a_i = \phi(\delta_i \mid 2); \) \( i=1,2,3 \) when optimum discriminants \( D_i(X_1) \) are used, \( i=1,2,3 \). Here:

\[
\delta_i^2 = \left\{ \begin{array}{l}
\left[ \mu_i^{(2)} - \mu_i^{(1)} \right] \Sigma_i^{-1} \left[ \mu_i^{(2)} - \mu_i^{(1)} \right]^\top, i=1,2,3 \end{array} \right.
\]

and

\[
D_i(X_1) = \begin{cases} 
1 & \text{if } \left\{ X_1 \Sigma_i^{-1} \left[ \mu_i^{(2)} - \mu_i^{(1)} \right] \right\} > 0 \\
0 & \text{if } < 0.
\end{cases}
\]

On the other hand, if the whole vector \( X = (X_1, X_2, X_3)^\top \) is used then the optimum discriminant rule \( D(X) \) is given by

\[
D(X) = \begin{cases} 
1 & \text{if } \left\{ X \Sigma^{-1} \left[ \mu^{(2)} - \mu^{(1)} \right] \right\} > 0 \\
0 & \text{if } < 0,
\end{cases}
\]

where

\[
\mu^{(1)} = \left( \mu_1^{(1)}, \mu_2^{(1)}, \mu_3^{(1)} \right)^\top
\]

and

\[
\Sigma = \begin{pmatrix} 
\Sigma_1 & 0 & 0 \\
0 & \Sigma_2 & 0 \\
0 & 0 & \Sigma_3
\end{pmatrix}.
\]

The probability of correct classification by this rule is \( \phi(\frac{1}{2} \delta) \),

\[
\delta^2 = \delta_1^2 + \delta_2^2 + \delta_3^2.
\]

Thus, if \( \delta_i = 1 \) for all \( i \), then the probability of correct classification by this rule is \( \phi \left( \frac{\sqrt{3}}{2} \right) = .806 \); whereas the probability of correct decision by the majority rule is given by

\[
\phi^3(.5) + 3 \phi_2(.5) [1-\phi(.5)] = .7726.
\]
Thus, there is a loss in not using the whole set together. The following is a small table consisting of some of these probabilities of correct classification and the associated differences.

| $\sigma^2$ | $\sigma_1^2$ | $\sigma_2^2$ | $\sigma_3^2$ | Prob. of correct classification $D$ | Prob. of correct classification by $D_\hat{D} = D_\hat{D}$ | Difference in Probabilities |
|---|---|---|---|---|---|---|}
| 1| .50 | .17 | .33 | .69146 | .66387 | .02758 |
| 2| 1.00 | .33 | .67 | .76024 | .72490 | .03533 |
| 3| 1.50 | .50 | 1.00 | .80675 | .76757 | .03917 |
In this section we evaluate the difference between the probabilities of correct classification by the two methods discussed in this paper. We consider the case when the class conditional distributions are normal with common, known, covariance matrix I and unknown mean vectors. That is, in the notations of the previous section, for j=1,2; X follows multivariate normal distribution with mean vector \( \mu_j \) which is unknown, and covariance matrix \( \Sigma \), assumed to be known, i=1,2,3. Clearly \( X = (X_1, X_2, X_3)' \) also follows multivariate normal distribution with mean vector \( \mu = (\mu_1, \mu_2, \mu_3) \) and covariance matrix

\[
\Sigma = \begin{bmatrix}
\Sigma_1 & 0 & 0 \\
0 & \Sigma_2 & 0 \\
0 & 0 & \Sigma_3
\end{bmatrix}
\]

Obviously independence of \( X_1, X_2, \) and \( X_3 \) is implied by this covariance structure.

Under these assumptions, the discriminant \( D(\cdot) \) is used where

\[
D(Y) = \begin{cases}
1 & \text{if } \left( Y - \frac{1}{2}(\overline{X}^{(1)} + \overline{X}^{(2)}) \right)' \Sigma^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)}) > 0 \\
0 & \text{if } \left( Y - \frac{1}{2}(\overline{X}^{(1)} + \overline{X}^{(2)}) \right)' \Sigma^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)}) \leq 0
\end{cases}
\]

where \( \overline{X}^{(j)} \) are sample means of \( n_j \) observations from the jth class, j=1,2 and \( Y \) is an observation to be classified in one of the two classes. A similar expression for the discriminant \( D_i(X) \) will hold if the ith subset is used for this purpose, i=1,2,3.

Clearly, the probability of correct decision, when \( D(Y) \) is employed is given by

\[
\Pr[\text{correct decision by } D(X) | \mu^{(1)}]
\]
\[
\begin{align*}
&\Pr\left[\left(1 - \frac{X^{(1)} + X^{(2)}}{2}\right)^2 \geq \frac{1}{\mu^{(1)}} \left(\bar{X}^{(1)} - \bar{X}^{(2)}\right) > 0\right] \\
&= \mathbb{E}_Y\left[\Pr\left(1 - \frac{X^{(1)} + X^{(2)}}{2}\right)^2 \geq \frac{1}{\mu^{(1)}} \left(\bar{X}^{(1)} - \bar{X}^{(2)}\right) > 0\right] \\
&= \mathbb{E}_Y \left(\phi\left(\bar{X}^{(1)}, \bar{X}^{(2)}\right)\right)
\end{align*}
\]

where
\[
\phi\left(\bar{X}^{(1)}, \bar{X}^{(2)}\right) = \frac{1}{\sqrt{2\pi\mu^{(1)}}} \left(1 - \frac{\left(\bar{X}^{(1)} + \bar{X}^{(2)}\right)^2}{\mu^{(1)}} \left(\bar{X}^{(1)} - \bar{X}^{(2)}\right)\right) \left(\frac{1}{\sqrt{2\pi\mu^{(1)}}}\right) \left(\frac{1}{\sqrt{2\pi\mu^{(1)}}}\right)^{1/2}
\]

and \(\phi(\cdot)\) denotes the distribution function of standard normal random variables.

In a similar manner, given \(\bar{X}^{(1)}\) and \(\bar{X}^{(2)}\) the conditional probability of correct classification by \(D(Y)\) will be given by \(\phi\left(\varphi_{\phi}\left(\bar{X}^{(1)}, \bar{X}^{(2)}\right)\right)\) where \(\varphi_{\phi}\) is also defined with appropriate changes. Denote \(\phi\left(\varphi_{\phi}\left(\bar{X}^{(1)}, \bar{X}^{(2)}\right)\right)\) by \(\phi_1\).

Conditional on the event that \(\bar{X}^{(1)}\) and \(\bar{X}^{(2)}\) are given, the probability of correct classification after combining the results of \(D_1, D_2, D_3\) is given by

\[
\begin{align*}
\phi_1 \quad &\text{if } x \in A_1 = \left\{x : \phi_1 (1-\phi_2) (1-\phi_3) > (1-\phi_1) \phi_2 \phi_3\right\} \\
\phi_2 \quad &\text{if } x \in A_2 = \left\{x : (1-\phi_1) \phi_2 (1-\phi_3) > \phi_1 (1-\phi_2) \phi_3\right\} \\
\phi_3 \quad &\text{if } x \in A_3 = \left\{x : (1-\phi_1) (1-\phi_2) \phi_3 > \phi_1 \phi_2 (1-\phi_3)\right\}
\end{align*}
\]

\[
\phi_1 \phi_2 \phi_3 + \phi_1 \phi_2 (1-\phi_3) + \phi_1 (1-\phi_2) \phi_3 + (1-\phi_1) \phi_2 \phi_3.
\]

otherwise

Thus the unconditional probability of correct classification is given by the integral of the above probabilities over \(A_1, A_2, A_3\) and the remaining region with respect to the joint density of \(\bar{X}^{(1)}, \bar{X}^{(2)}\). This being a difficult problem of integration, we obtain a lower bound by integrating the last expression over the entire range. Thus, the probability of correct classification, when results of \(D_1, D_2, D_3\) are employed is greater than
\[
E \left[ \theta_1 \theta_2 \theta_3 + \theta_1 \theta_2 (1-\theta_3) + \theta_1 (1-\theta_2) \theta_3 + (1-\theta_1) \theta_2 \theta_3 \right]
\]

using the independence of \(X_i\)'s the above reduces considerably because, for instance we can replace \(E(\theta_1 \theta_2 \theta_3)\) by \(E[\theta_1] E[\theta_2] E[\theta_3]\).

The expression for \(E[1-\theta]\) is given by Equation 77 of John (1961). The \(E[1-\theta_i]\) can also be obtained similarly. Thus an upper bound for the difference in the probabilities of correct classification can be evaluated. The following table gives these upper bounds for few choices of number of observations in the training sample. The number of training samples, \(N\), are equal in both classes. In the table \(\delta_i^2\), \(i=1,2,3\) denotes the Mahalonobis distance between the two populations, measured for the \(i\)th subset only.
The following table gives an upper bound of the difference between the probabilities of correct classification by the two methods.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>$P_4$</th>
<th>$P_5$</th>
<th>$P_6$</th>
<th>$P_7$</th>
<th>$P_8$</th>
<th>$P_9$</th>
<th>$P_{10}$</th>
<th>$P_{11}$</th>
<th>$P_{12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>7</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>11</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>12</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

The table shows the probabilities of correct classification for each method along with the difference between them.
From the table it is clear that the difference increases as N increases provided other parameters are fixed. For all other parameters fixed, the difference increases with $\sigma^2$. The most significant observation, from the table, is that if the actual probability of correct classification is large then the difference is also large. One also concludes that the actual difference of probability of correct decision between using the linear discriminant with the whole set of observation and in parts decreases if the parameters of the population are unknown. Thus, in otherwords, one would be less concerned about the loss in using the alternative method $D_\lambda$ of discrimination when the parameters are unknown.
8.7. EXTENSION TO THE THREE CLASS PROBLEM

In the previous sections we have considered the case when there are only two classes. In this section the ideas of the previous sections are extended to three class problem. Extension to more than three classes will be straightforward. It will be seen that due to a large number of parameters, there are some difficulties in getting results in the most general form, but the basic concepts remain unchanged. To formalise the concepts we denote the three classes by \( w_1, w_2, \text{ and } w_3 \). As before, we confine our attention to the case when there are three independent subsets of measurements on each object. Furthermore, as in the previous section, we have the results of the three discriminants operating at each subset. Our aim is to combine these results to decide to which class the given object belongs.

The sample space of the results of the three discriminants is given by

\[ S' = \left\{(i_1, i_2, i_3) : 1 \leq i_1, i_2, i_3 \leq 3\right\} \]

where the triplet \((i_1, i_2, i_3)\) means that the first discriminant using the first segment of the measurement on the given object classifies it to class \( i_1 \), the second discriminant, using the second segment, to class \( i_2 \) and the third discriminant, using the last segment, to the class \( i_3 \). Our object, as pointed out earlier also, is to combine the result \((i_1, i_2, i_3)\) and classify the object to one of the three classes. Given the sample space \( S' \) and associated probability measures, the optimum criteria of classification would use the Bayesian approach. Under the assumption of equally probable classes and equal costs the Bayes approach would employ the likelihood functions only, but the unequal costs and unequal a priori probabilities can be accommodated using the standard procedures, Anderson (1968).

The probability structure associated with the sample space \( S' \) is given by the following 27 parameters.
\[ P_{i,j}^{(k)} \] = Probability that the discriminant \( k \) will classify \( X \) into class \( i \) when it actually comes from class \( j \), for \( 1 \leq i, j, k \leq 3 \). Obviously these parameters are not all independent because

\[ \sum_{i=1}^{3} P_{i,j}^{(k)} = 1 \quad \text{for each } j \text{ and } k \]

Furthermore, \( P_{j,j}^{(k)} \) denotes the probability of correct classification for each \( j \) and \( k \), whereas \( P_{i,j}^{(k)} \) denotes probability of misclassification for \( i \neq j \) and \( k \). Given these parameters one can easily calculate the probability of observing any sample point of \( S \)'s conditional on \( X \) belongs to a specified class.

For instance, the probability that the result of the three discriminants will be \((1, 1, 1)\) when \( X \) comes from class 1 is given by

\[
\begin{pmatrix}
P_{1,1}^{(1)} & P_{1,2}^{(2)} & P_{1,3}^{(3)}
\end{pmatrix}.
\]

The optimum discriminant method which is based upon the elements of \( S \)'s then given by \( D(X) \) where \( D(X) = D(1, 1, 1) = 1 \) means that \( X \) is classified to class 1.

This discriminant [See Anderson [1]], is given by

\[ D(X) = 1 \text{ iff } P_{1,1}^{(1)} P_{2,1}^{(2)} P_{3,1}^{(3)} > P_{1,1}^{(1)} P_{2,1}^{(2)} P_{3,1}^{(3)} \text{ for } \sigma \cdot 1 = 1 \]

Based on the available information \((1, 1, 1)\) only \( D \) will be optimum which can be seen from the arguments given in Anderson [1].

In the remaining section we study the above discriminant when \( P_{i,j}^{(k)} \) satisfies additional conditions.

**Special Cases I:**

Suppose \( P_{i,j}^{(k)} \) are such that all of the probabilities of misclassifications are equal for each class and each discriminant. Denote this probability by \( p \).

That is

\[
P_{i,j}^{(k)} \begin{cases} = p & \text{if } i \neq j \\ = 1-2p & \text{if } i = j \end{cases}
\]

52
It is easily seen that, for $p < 1/3$, $D$ is equivalent to the majority rule except that the 6 permutations of $(1,2,3)$ are classified arbitrarily with equal probabilities. Further, the probability of correct classification will be given by

$$(1-2p)^3 + 3(1-2p)^2 p + \frac{1}{2}(1-2p)p^2.$$  

**Case 2**

Suppose $P_{i,j}(k)$ satisfies the following conditions. Let $P_{i',k}(k) = p(k)$ for $i' = i$, $i = 1,2,3$, $k = 1,2,3$. Clearly $P_{i,j}(k) = 1 - 2p(k)$. This implies that the two probabilities of misclassification of an object from class $w_i$ into $w_{i'}$ are the same for any specified $k$, and equal for each $i$. In this case, without loss of generality, we can assume that

$$\frac{1}{3} > p(1) > p(2) > p(3).$$

These inequalities imply that, compared in terms of probabilities of correct classification, the first discriminant is worst and the third is the best. As would be expected, the majority rule will not necessarily be the best and for instance $D$ classifies all of the following 5 sample points to class $w_1$.

$$\{(1,1,1), (2,1,1), (3,1,1), (1,2,1), (1,3,1)\}$$

As for the points $(1,1,2)$ or $(1,1,3)$ they would be classified to class 1 if

$$\frac{1-2p(1)}{p(1)} \cdot \frac{1-2p(2)}{p(2)} > \frac{1-2p(3)}{p(3)}$$

and to class $w_2$ and $w_3$, respectively, if the direction of the inequality is reversed. Recall that this is exactly the same condition that we have seen in an earlier section. Similarly we can find points which will be classified to $w_2$ and $w_3$. The remaining 6 permutations of $(1,2,3)$ are classified as follows:

$(1,2,3)$ and $(2,1,3)$ are classified such that the associated $X$ belongs to $w_3$;

$(1,3,2)$ and $(3,1,2)$ to $w_2$ and $(2,3,1)$ and $(3,2,1)$ to class $w_1$. 53
Probabilities of correct classifications can be computed easily by summing over all points which lead to the acceptance of \( x \in u_1 \) when it is truly the case. For instance, given

\[
\frac{1-2p^{(1)}}{p^{(1)}} \cdot \frac{1-2p^{(2)}}{p^{(2)}} \cdot \frac{1-2p^{(3)}}{p^{(3)}}
\]

the probability of correct classification of \( x \in u_1 \) is given by

\[
(1-2p^{(1)}) (1-2p^{(2)}) (1-2p^{(3)}) + 2p^{(1)} (1-2p^{(2)}) (1-2p^{(3)})
\]

\[+ 2(1-2p^{(1)})p^{(2)} (1-2p^{(3)}) + 4p^{(1)}p^{(2)} (1-2p^{(3)}),
\]

which, as expected, simplifies to \( 1-2p^{(3)} \), implying that the decision is equivalent to the third discriminant. The probabilities can be evaluated in a similar manner. The other special cases can also be studied.

The extension of the above idea to more than 3 classes is straightforward. The above concepts can also be extended in exactly the same manner to the case when there are more than 3 independent segments of measurement on each object. The case when these segments of observations are stochastically dependent is relatively hard to study although the similar concepts apply in that situation also.

Finally, we present the formulas which can be employed to evaluate the differences between the probabilities of the correct decisions for the two methods under investigation, in the case of multivariate normal distributions. We consider the simple situation when all of the parameters are assumed known and the covariance matrix is same. Suppose for \( \bar{x} \in u_1 \), \( i=1,2,3 \) the mean vector is \( \mu_1 \) and (without loss of generality) the covariance matrix is \( I \). As before \( \bar{x} \) (and therefore \( \mu_1 \)) is \( p \) dimensional which consists of three segments \( \bar{x}_1, \bar{x}_2, \bar{x}_3 \) of \( p_1, p_2 \) and \( p_3 \) dimensions. In the following we give the formulas for the probabilities of correct and incorrect classification when \( \bar{x} \in u_1 \) is used. The same formulas can be used to calculate the associated probabilities when any one
of the segment is used, of course, with obvious modifications. Given this structure it would be straightforward to calculate the desired difference.

In this case of three class problem, the decision to classify \( x \) in class \( w_1 \) will be taken if

\[
\exp - \frac{1}{2} \sum_{i=1}^{P} \left[ x_i - \mu_{1i} \right]^2 > \left\{ \begin{array}{ll}
\exp - \frac{1}{2} \sum_{i=1}^{P} \left[ x_i - \mu_{2i} \right]^2 \\
\exp - \frac{1}{2} \sum_{i=1}^{P} \left[ x_i - \mu_{3i} \right]^2
\end{array} \right.
\]

or equivalently when \( (U_1 > 0, V_1 > 0) \) where

\[
U_1 = \sum_{i=1}^{P} \left( \mu_{1i} - \mu_{2i} \right) \left[ x_i - \frac{\mu_{1i} + \mu_{2i}}{2} \right]
\]

\[
V_1 = \sum_{i=1}^{P} \left( \mu_{1i} - \mu_{3i} \right) \left[ x_i - \frac{\mu_{1i} + \mu_{3i}}{2} \right]
\]

Define \( (U_2, V_2) \) and \( (U_3, V_3) \) by

\[
U_2 = \sum_{i=1}^{P} \left[ \mu_{2i} - \mu_{1i} \right] \left[ x_i - \frac{\mu_{2i} + \mu_{1i}}{2} \right]
\]

\[
V_2 = \sum_{i=1}^{P} \left[ \mu_{2i} - \mu_{3i} \right] \left[ x_i - \frac{\mu_{2i} + \mu_{3i}}{2} \right]
\]

\[
U_3 = \sum_{i=1}^{P} \left[ \mu_{3i} - \mu_{1i} \right] \left[ x_i - \frac{\mu_{3i} + \mu_{1i}}{2} \right]
\]

\[
V_3 = \sum_{i=1}^{P} \left[ \mu_{3i} - \mu_{2i} \right] \left[ x_i - \frac{\mu_{3i} + \mu_{2i}}{2} \right]
\]

It can be easily shown that \( x \) will be classified to class \( w_2 \) if \( (U_2 > 0, V_2 > 0) \) and to class \( w_3 \) if \( (U_3 > 0, V_3 > 0) \). The distribution of all of these pairs is bivariate normal with means, variances and covariances given below. It is assumed, as mentioned earlier also, that \( x \) is assumed to be from \( w_1 \) in the following calculations.
2 \text{E} \left( U_1 \right) = \text{Var} \left( U_1 \right) = \frac{p}{i=1} \left( u_{1i} - \mu_{11} \right)^2

2 \text{E} \left( V_1 \right) = \text{Var} \left( V_1 \right) = \frac{p}{i=1} \left( u_{1i} - \mu_{31} \right)^2

and the correlation coefficient between $U_1$ and $V_1$ is given by

$$\rho_{U_1V_1} = \frac{\sum_{i=1}^{p} \left( u_{1i} - \mu_{11} \right) \left( u_{1i} - \mu_{31} \right)}{\sqrt{\sum_{i=1}^{p} \left( u_{1i} - \mu_{11} \right)^2 \sum_{i=1}^{p} \left( u_{1i} - \mu_{31} \right)^2}}$$

-2 \text{E} \left( U_2 \right) = \text{Var} \left( U_2 \right) = \frac{p}{i=1} \left( u_{2i} - \mu_{11} \right)^2

$$\text{E} \left( V_2 \right) = \frac{p}{i=1} \left( u_{2i} - \mu_{31} \right) \left( u_{1i} - \frac{u_{2i} + u_{3i}}{2} \right), \text{Var} \left( V_2 \right) = \frac{p}{i=1} \left( u_{2i} - \mu_{31} \right)^2$$

the correlation coefficient between $U_2$ and $V_2$ is given by

$$\rho_{U_2V_2} = \frac{\sum_{i=1}^{p} \left( u_{2i} - \mu_{11} \right) \left( u_{2i} - \mu_{31} \right)}{\sqrt{\sum_{i=1}^{p} \left( u_{2i} - \mu_{11} \right)^2 \sum_{i=1}^{p} \left( u_{2i} - \mu_{31} \right)^2}}$$

and finally

-2 \text{E} \left( U_3 \right) = \text{Var} \left( U_3 \right) = \frac{p}{i=1} \left( u_{3i} - \mu_{11} \right)^2

$$\text{E} \left( V_3 \right) = \frac{p}{i=1} \left( u_{3i} - \mu_{2i} \right) \left( u_{1i} - \frac{u_{3i} + u_{2i}}{2} \right), \text{Var} \left( V_3 \right) = \frac{p}{i=1} \left( u_{3i} - \mu_{2i} \right)^2$$

56
and correlation coefficient between $U_3$ and $V_3$ is given by

$$
\rho_{U_3, V_3} = \frac{\sum_{i=1}^{P} \left( \mu_{3_i} - \mu_{2_i} \right) \left( \mu_{3_i} - \mu_{2_i} \right)}{\sqrt{\sum_{i=1}^{P} \left( \mu_{3_i} - \mu_{2_i} \right)^2 \sum_{i=1}^{P} \left( \mu_{3_i} - \mu_{2_i} \right)^2}}
$$

Given the above distributions, it is easy to calculate that the probability of classifying $X \in M_1$ to $M_1$ is

$$
P(U_1 > 0, V_1 > 0|u_1) = \int_{-h_1}^{h_2} \int_{-h_2}^{h_1} \frac{1}{2\pi \sqrt{1-P_1}} \exp \left( -\frac{1}{2} \left( u^2 + v^2 - 2\rho_{U_1, V_1} uv \right) \right) \, du \, dv
$$

$$
= L(-h_1, -h_2; P_{U_1, V_1})
$$

following the notations of Johnson and Kotz [See equation 19, page 94,[3]]. Here

$$
h_1 = \frac{1}{2} \left\{ \sum \left( \mu_{1_i} - \mu_{2_i} \right)^2 \right\}^{1/2}, \quad h_2 = \frac{1}{2} \left\{ \sum \left( \mu_{1_i} - \mu_{3_i} \right)^2 \right\}^{1/2}.
$$

Since the arguments $-h_1, -h_2$ in $L(-h_1, -h_2; P_{U_1, V_1})$ are negative, we have to use equation 22.3 of page 95 of Johnson and Kotz. At this stage to calculate $P(U_1 > 0, V_1 > 0|u_1)$ it remains to look up $L(-h_1, -h_2; P_{U_1, V_1})$ from a table, for example in National Bureau of standard publication (1939). In exactly a similar manner one could calculate

$$
P(U_2 > 0, V_2 > 0|u_2) \text{ and } P(U_3 > 0, V_3 > 0|u_3),
$$

which provide the probabilities of misclassifications. It is significant to observe that the above probabilities not only depend upon the Mahalanobis distance

$$
\left( \frac{D}{\sum \left( \mu_{2_i} - \mu_{3_i} \right)^2} \right)
$$

but also upon the angles between $(\mu_1 - \mu_2)$, $(\mu_1 - \mu_3)$, $(\mu_2 - \mu_3)$ etc.
REFERENCES


References Continued


