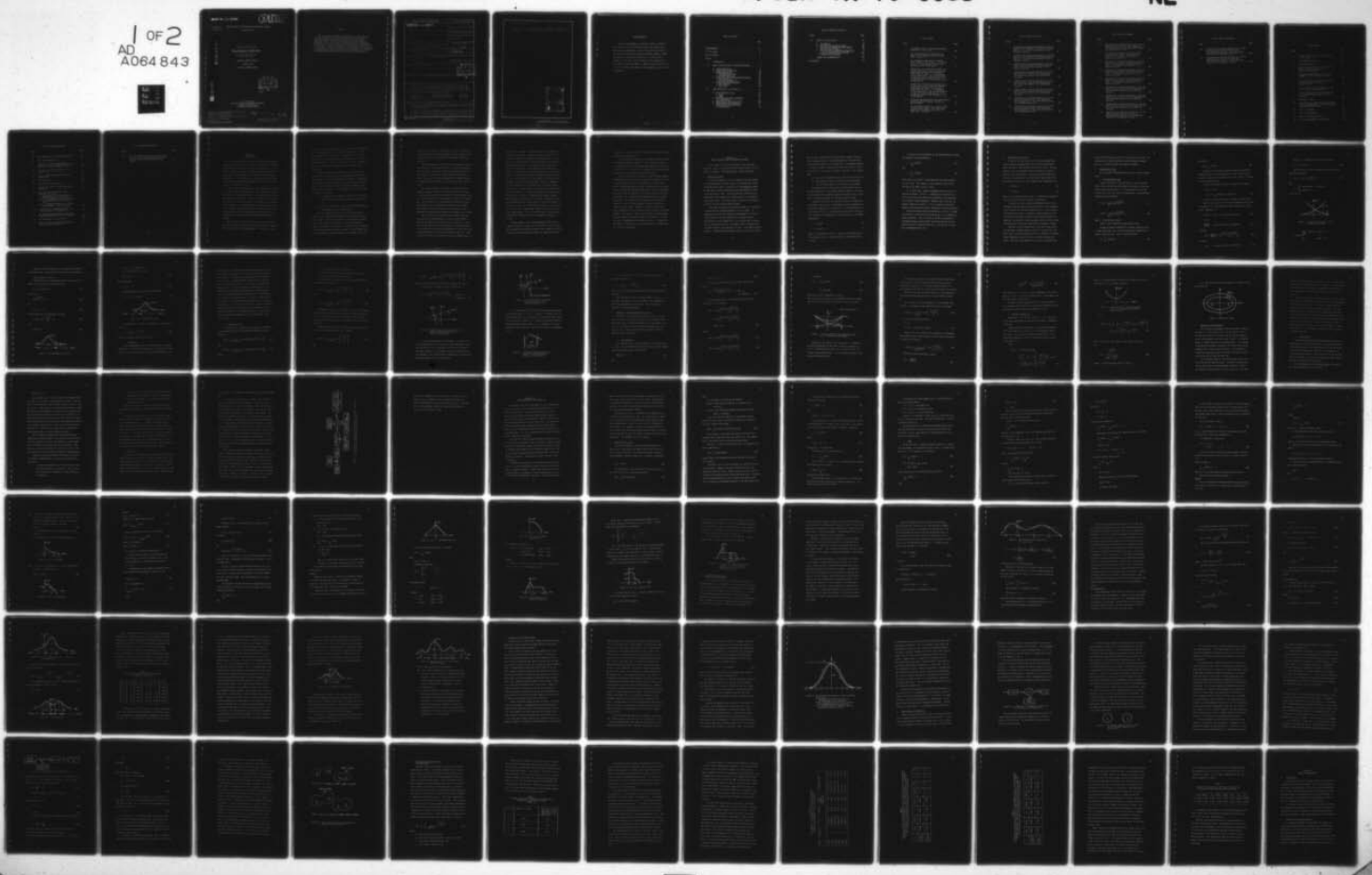
AD-A064 843 OHIO STATE UNIV COLUMBUS ELECTROSCIENCE LAB F/G 12/2
IDENTIFICATION OF CATALOGUED AND UNCATALOGUED CLASSES.(U)
DEC 78 H LIN AFOSR-74-2611
UNCLASSIFIED ESL-783815-7 AFOSR-TR-79-0055 NL

1 OF 2
AD
A064 843



OSU

IDENTIFICATION OF CATALOGUED AND UNCATALOGUED CLASSES

Heng-Cheng Lin

The Ohio State University

ElectroScience Laboratory

Department of Electrical Engineering
Columbus, Ohio 43212

TECHNICAL REPORT 783815-7

December 1978

Grant No. AFOSR-74-2611

U. S. Air Force
Air Force Office of Scientific Research
1400 Wilson Boulevard
Arlington, Virginia 22209

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AFOSR-TR-79-0055	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): IDENTIFICATION OF CATALOGUED AND UNCATALOGUED CLASSES.	5. TYPE OF REPORT & PERIOD COVERED: Interim 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Heng-Cheng Lin	8. CONTRACT OR GRANT NUMBER(s): AFOSR 74-2611
9. PERFORMING ORGANIZATION NAME AND ADDRESS: The Ohio St. Univ. Electro Science Laboratory, Department of Electrical Engineering, Columbus, Ohio 43212	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304/A2
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, DC 20332	12. REPORT DATE: December 1978 13. NUMBER OF PAGES: 124
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): ESL-78381-F	15. SECURITY CLASS. (of this report): UNCLASSIFIED 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The material contained in this report was also used as a dissertation submitted to the Department of Electrical Engineering, The Ohio State University as partial fulfillment for the degree Doctor of Philosophy.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Classification
Uncatalogued class
Identification
Minimum volume classifier
Monte Carlo simulation
Hypersphere approximation
Unknown target detection
Null class
Subgrouping
Clustering

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A technique to discriminate listed objects from unlisted ones is described. It is based on the principle of minimizing the probability of error in an identification process.

The devised classifier is implemented as a threshold test. The classifier was applied to an aircraft identification problem. It was shown that the error of misclassifying catalogued targets as uncatalugued and vice versa can be made very small, while keeping a high probability of correct

DD FORM 1 JAN 73 1473

402 251

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract continued.

identification when the presence of a listed object is detected. The implementation of the developed scheme was shown to be simple and efficient.

ACKNOWLEDGMENTS

Special acknowledgment is given to my advisor, Professor A.A. Ksienski, for his continued guidance, many helpful suggestions and encouragement, not only during the work covered in this dissertation but also through my graduate program. A debt of gratitude is also owed to Professors L.J. White and R.T. Compton, Jr. for their critical reading of the manuscript and many valuable ideas and comments. Thanks are also due to Dr. C.D. Liu for his reading and comments on the mathematical part of this dissertation.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

CHAPTER I
INTRODUCTION

This study deals with the problem of determining whether an observed object is or is not a member of a listed class of objects. The technique developed classifies the observed object as listed or unlisted based on a priori information regarding the listed class alone. The technique assures that the misclassification probability of a listed object as an unlisted one is kept under a specified threshold while attempting to minimize the reverse, "type II" probability of error.

In most identification processes the signatures of an object, e.g., light, sound or temperature etc., are assumed to be measurable and can be converted into electric signals, which are called the responses of the object. The basic assumption in statistical classification is that the response of each object has a probability distribution. Some algorithms[1,2,3] were developed to identify each object by using the available information, which ranges from complete statistical knowledge of the distribution to no knowledge except that which can be deduced from the measured response.

More specifically, when the statistics of all the listed objects are known, the response of an observed object is compared with the responses of these known objects. The object is then

classified as one of the known objects by a predetermined criterion. The set of known objects is catalogued beforehand and is called the catalogued (or listed) class. Conventional classification procedures assume that the object to be identified is in the list of the known objects. However, since the list of the catalogued objects is very often not exhaustive, this assumption is not always valid.

Another often encountered problem is that the functional forms of the distributions of the listed objects are known, but some parameters of these distributions are unknown. A great deal of effort[4,5] has been devoted to solve these kinds of problems, namely, classification of the objects without knowing the parameters of the distributions. Two procedures are usually employed in solving this problem:

(1) Estimation of the parameters and their use to do the classification as ordinary parametric classification, and

(2) Classification of the object directly without knowing the parameters of all distributions.

The first procedure turns out to be a learning process, supervised or unsupervised depending on the labeling of the observed samples. Both learning processes are themselves another form of the statistical estimation. The supervised process labels each sample "to be learned" as an object in the catalogued class and the unsupervised process assigns a priori probabilities for all the objects in the catalogued class although it does not

label any of the samples. Both processes possess an underlying assumption that all the learned samples originate from the listed class, which is, as stated before, not always complete and exhaustive.

The second procedure is a classification which does not use the parameters of the distributions and is called non-parametric classification. It does not require any information about the distribution of each object in the catalogued class but the number of the objects and their deterministic responses have to be known before an algorithm can be devised to do the classification.

The problem studied in this work is different from the *ordinary parametric classification* problem in that it includes an additional alternative, the presence of an uncatalogued class. It is also different from a non-parametric problem in an obvious way. The alternatives in a non-parametric problem are clearly defined while they are incompletely defined in this problem. The parameters of the distributions are not known in a non-parametric problem but are not necessarily unknown in our problem.

The fact that all the classification schemes ignore the existence of the unlisted objects is due to the difficulty in handling the problem, not because of its insignificance. Since the unlisted objects are usually unknown to a system designer, it is very difficult to predict the performance of a system classifying the unlisted class as distinct from the listed class. Also since the statistics of the unlisted class are unknown, there is

no way that an "optimum" criterion can be set up to guide the design of such a system. However, for some classifiers it is important to identify an unexpected or unknown object. For instance, the radar detectors in air surveillance are developed to detect airborne objects. There are many schemes developed to identify an object after a radar detects some response from a target. A conventional aircraft classifier detects the electromagnetic scattering return from an airplane and compares it with the list of the responses of all the known aircraft targets and attempts to classify an observed target into one of them[6,7]. Yet, the stored data of the responses are by no means complete and exhaustive. When a target is present and its response looks "odd" on the radar, it is extremely important not to exclude the possibility of its being an unknown airborne target, perhaps newly developed. Therefore, it is essential to determine at the very beginning stage of classification if the observed target is in the list. Once it is decided that the target is in the listed class, one can use a conventional scheme to classify it as one of the known targets. If it is not, the target is designated as a new one and a learning process is employed to estimate its response.

Another example is the Electrocardiogram (EKG) used to diagnose a disease. A person can be classified into nine different states[8] which include the normal state. However, a peculiar disease not described by any of these nine states may never be

recognized. It would be desirable to develop a scheme to identify unexpected or unknown diseases.

It is therefore the purpose of this work to develop a method to distinguish whether an object is in the listed or unlisted class before one can use any conventional method to do the classification. This is referred to as a problem of identifying uncatalogued objects as distinct from catalogued objects.

Our main effort will be devoted to solve the above problem. The method devised is based on the idea that an optimum classifier is the one which minimizes the probability of classification error. The developed approach tends to minimize the error probability of identifying an uncatalogued object into the catalogued one while fixing the probability of the other kind of classification error, viz., identifying a catalogued object as an uncatalogued one. The reasoning for this approach is described in Chapter II, and a criterion proposed to guide the design of a classifier is described. Chapter III discusses some of the special properties of this classifier which is applied to several examples. The scheme is shown to be effective in distinguishing the listed objects from unlisted ones. The algorithm is simple in terms of implementation and computation time. In Chapter IV, the scheme is applied to an aircraft identification problem and the results are presented.

CHAPTER II
MODEL DISCUSSION AND THE PROPOSED CRITERION

In this chapter, the basic philosophy of the algorithm is discussed. Due to its similarity to the Neyman-Pearson rule, the latter is reviewed. The proposed model is then introduced.

A. General Discussion

A classifier measures a set of N features from the response of an observed object to form a point in an N-dimensional space. The N-dimensional space is referred to as the observation space. In other words, every observation can be represented as a point in the observation space. Every observation is assumed to be a measurement of the response of an object, which is corrupted by some kind of noise. This noisy signal is then used to determine to which class the observed object belongs.

In building an optimum classification scheme, it is necessary to define what an optimum performance is for a system. An appropriate criterion for judging the performance of a classification system is the error probability in making decisions. Minimizing this error probability is our ultimate goal in designing a classification system. An optimum system is therefore the one which minimizes the probability of error in the identification process. In the problem considered, there are two kinds of errors

one can make in performing the classification, namely, classifying a catalogued object as an uncatalogued one, and identifying an uncatalogued object as a catalogued one. A reasonable approach is to make the overall probability of misclassification as small as possible, as is the case for a Bayes classifier. This requires that

(1) the a priori probabilities for each class be given, and

(2) the probability density functions of the observation vector be known when any of the classes are present.

The first one can be easily met by assigning each class an a priori probability. The second condition, however, will never be satisfied since no knowledge about the unlisted class is available. A different approach is therefore taken as follows.

In making a decision, the observation space is divided into two regions, say Z_N and Z_X, corresponding to the catalogued class N and uncatalogued class X, respectively. The observed object is classified to N (or X) if the measured response falls into Z_N (or Z_X). Therefore, the probabilities of making errors in the identification are

$$\beta = P\{\bar{x} \in Z_N | X\} \quad , \tag{1}$$

$$\alpha = P\{\bar{x} \in Z_X | N\} \quad , \tag{2}$$

where $\bar{x}$ is the observed vector, α represents the probability of error when N is true, β represents the error probability when X is true.

In terms of the distributions of the observed vector $\bar{x}$, these two equations can be expressed as

$$\beta = \int_{Z_N} P(\bar{x}|X)d\bar{x} \quad , \tag{3}$$

and

$$\alpha = \int_{Z_X} P(\bar{x}|N)d\bar{x} \quad , \tag{4}$$

where $P(\bar{x}|X)$ (or $P(\bar{x}|N)$) is the probability distribution when X (or N) is true. The integral is the probability that $\bar{x}$ falls into Z_N (or Z_X) when X (or N) is true.

It is clear that β cannot be computed since the class X is unknown and neither is $P(\bar{x}|X)$. In designing a classifier, one can try to minimize the second error probability α by minimizing Z_X if $P(\bar{x}|N)$ is known beforehand. However, when Z_X shrinks, Z_N expands, tending to increase β. This is contrary to the main purpose of minimizing the overall probability of misclassification.

One way of solving the above difficulty is to fix α instead of minimizing it while trying to minimize Z_N, i.e., to make Z_X as large as possible. This is similar to the basic assumption in the Neyman-Pearson classification: minimizing one kine of error probability while fixing the other one. This leads us to introduce the Neyman-Pearson rule.

B. Conventional Classifiers

In the case of two classes, both of the two conventional criteria, Bayes' and Neyman-Pearson's, classify the observed object into one of the two known classes, say L and K. The two classes L and K are sometimes called a learning class and an exterior class respectively. We refer to the error made when L is true as a type one error (or false alarm) and the other as a type two error (or miss). And we express their probabilities as

$$\alpha = P\{\bar{x} \in K | L\} \quad , \tag{5}$$

$$\beta = P\{\bar{x} \in L | K\} \quad , \tag{6}$$

where $\bar{x}$ is the observed vector, and α, β represent the probabilities of type one and type two errors, respectively.

The Bayes classifier, using the knowledge of the a priori information and the statistical characteristic of the corrupting noise, minimizes the average weighted error (or risk) by adjusting the decision process. An ordinary system governed by this rule sets up a threshold to which all the received signals are compared. The observed is then assigned to one of the two classes. The probability of misclassification of this scheme is minimum.

When the a priori probabilities of the two listed classes are not known, a Neyman-Pearson criterion is usually used. The basic idea is to specify one kind of error, whichever is considered more important, say the first one α, while minimizing the other. Very often, the method turns out to be a threshold test

and the threshold depends upon the first kind of error only. The characteristic, that the detector functions without knowledge about K, is called uniformly most powerful (UMP)[5].

C. Neyman-Pearson Rule

We first apply the Neyman-Pearson rule to a special example below.

1. One dimensional case

Suppose that there is only one object in each of the two known classes and that the noise added to the signal is Gaussian with zero mean and variance σ^2. In the absence of noise, the signals from L and K are S_1 and S_2 respectively. The conditional probabilities are therefore

$$P(x|L) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-S_1)^2}{2\sigma^2}\right) \quad , \tag{7}$$

$$P(x|K) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-S_2)^2}{2\sigma^2}\right) \quad , \tag{8}$$

where x is the observed signal.

Again, L and K represent the two known classes.

We want to keep the probability of making a type one error at some fixed value, say α, while minimizing the probability of making a type two error. That is, we want to have

$$P_F \triangleq \int_{Z_L} P(x|L)dx = \alpha \quad , \tag{9}$$

and minimize

$$P_M \triangleq \int_{Z_L} P(x|K)dx = \beta \quad . \tag{10}$$

Z_L and Z_K are two disjoint regions occupying the whole observation space, and are associated with L and K respectively. The object whose measured response lies in any of them is assigned to the corresponding class.

To solve the problem, construct a function F as follows:

$$F \triangleq P_M + \lambda [P_F - \alpha]$$

$$= \lambda(1-\alpha) + \int_{Z_L} [P(x|K) - \lambda P(x|L)]dx \quad . \tag{11}$$

Since $P_F - \alpha = 0$, minimizing P_M is equivalent to minimizing F.

Now, for any positive λ we like the integrand to be as small as possible (i.e., to be negative or zero) such that F is minimized. Therefore, if

$$P(x|K) - \lambda P(x|L) < 0, \text{ assign point x to } Z_L \tag{12}$$

Or, if

$$\frac{P(x|K)}{P(x|L)} < \lambda \text{ assign to L, and to K otherwise.} \tag{13}$$

Defining

$$\Lambda(x) = \frac{P(x|K)}{P(x|L)} = \exp\left[\frac{2x(S_2-S_1)-(S_2^2-S_1^2)}{2\sigma^2}\right] \quad , \tag{14}$$

we have

$$\Lambda(x) > \lambda \quad \text{assign to L, and to K otherwise.} \tag{15}$$

Since $P_F = \alpha$, Equation (9) can also be written as

$$P_F = \int_{\lambda}^{\infty} P(\Lambda|L)d\Lambda = \alpha \quad . \tag{16}$$

The function $P(\Lambda|L)$ can be evaluated by using the transformation equation[9]

$$P(\Lambda|L) = P(x(\Lambda)|L) \left|\frac{dx(\Lambda)}{d\Lambda}\right| I_{\Lambda} \quad , \tag{17}$$

where

$$I_{\Lambda} = \begin{cases} 1, & \text{in the region } \Lambda \text{ is defined} \\ 0, & \text{otherwise} \end{cases}$$

(see Figure 1).

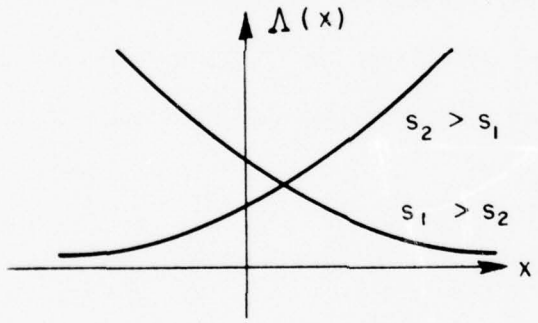


Figure 1. I_{Λ}=0 for Λ<0 since $\Lambda(x)$ is not defined in this range.

$$P(\Lambda|L) = \begin{cases} \dfrac{1}{\left|\dfrac{d\Lambda}{dx}\right|} P(x(\Lambda)|L), & \text{where } \Lambda > 0 \\ 0, & \text{where } \Lambda < 0 \end{cases} .$$

Therefore, we find a solution λ^* for Equation (16), which, substituting back to Equation (14), will give us the threshold x^*.

Case a. $S_2 > S_1$ (Figure 2)

Equation (16) can be transformed back to the form of $P(x|L)$ when $S_2 > S_1$ (see Figure 1 and Equation (15)),

$$P_F = \int_{x^*}^{\infty} P(x|L)dx = \alpha \quad . \tag{18}$$

This leads to

$$\text{erfc}\left(\frac{x_1^* - S_1}{\sigma}\right) = \alpha \tag{19}$$

or

$$x_1^* = \sigma z_\alpha + S_1 \quad , \tag{20}$$

where $\text{erfc}(x)$ and z_α are defined as follows:

$$\text{erfc}(x) \triangleq \int_x^{\infty} \frac{e^{-y^2/2}}{\sqrt{2\pi}} \, dy \quad , \tag{21}$$

and

$$\text{erfc}(z_\alpha) \triangleq \alpha \quad . \tag{22}$$

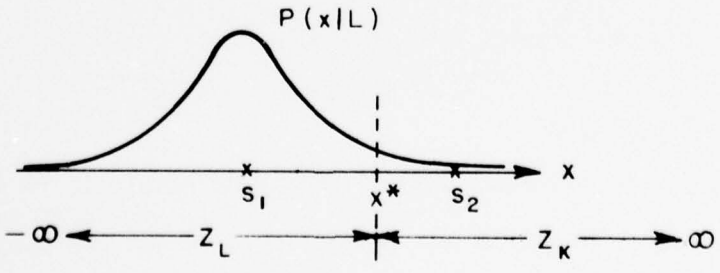


Figure 2. The threshold x^* for $S_2 > S_1$.

Case b. $S_2<S_1$ (Figure 3)

It is clear that for this case

$$\int_{-\infty}^{x_2^*} P(x|L)dx = \alpha \quad , \tag{23}$$

and the threshold

$$x_2^* = S_1 - \sigma z_\alpha \quad . \tag{24}$$

Combining Equations (20) and (24), one obtains

$$x^* - S_1 = \sigma z_\alpha \, \mathrm{sgn}(S_2-S_1) \quad . \tag{25}$$

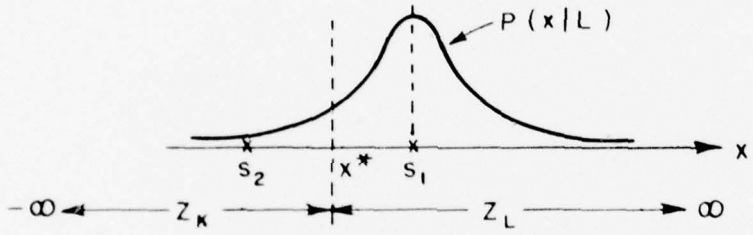


Figure 3. The threshold for $S_2<S_1$.

The probability of making a type two error β for both cases a and b is

$$\beta = \mathrm{erfc}\left(z_\alpha - \frac{|S_2-S_1|}{\sigma}\right) \quad , \tag{26}$$

which is minimized according to the rule.

Discussion

A very common application of the above result is in radar detection. L is set to be the absence of a target and K its presence. The false alarm is kept to some prefixed constant, say α.

A CW radar detects the backscattered returns and assigns the object to one of the above cases. If S_2, the noise free response of the target, is assumed to be greater than zero, a threshold is set at $x^*=\sigma z_\alpha$, provided that the injected noise is Gaussian with zero near and variance of σ^2. The noise is assumed to be Gaussian since it is the most commonly encountered type. The threshold x^* is completely determined by the system parameters σ and α even without knowing the amplitude of S_2. Nevertheless, it is based on the assumption that $S_2>0$ and $S_1=0$. When a noise free signal S_2 is not greater than or equal to zero, the result is different. It is clear that if we have no a priori information about the response of the external class K, we have no way of finding the threshold x^* according to the rule. A more complicated case with some knowledge of K will be brought up in the next section.

2. n-dimensional case

If the response of an object is an n-tuple $\bar{x}$, corrupted by zero mean Gaussian noise, the conditional probability density function is

$$P(\bar{x}|L) = \frac{1}{(2\pi)^{n/2}|\bar{\Gamma}|^{1/2}} \exp\left[-\frac{1}{2}(\bar{x}-\bar{S}_1)^T\bar{\Gamma}^{-1}(\bar{x}-\bar{S}_1)\right] , \quad (27)$$

and

$$P(\bar{x}|K) = \frac{1}{(2\pi)^{n/2}|\bar{\Gamma}|^{1/2}} \exp\left[-\frac{1}{2}(\bar{x}-\bar{S}_2)^T\bar{\Gamma}^{-1}(\bar{x}-\bar{S}_2)\right] \quad (28)$$

where

$\bar{x}$ is the observed signal (column) vector,

$\bar{S}_1$ and $\bar{S}_2$ are noise free signal (column) vectors, and

$\bar{\Gamma}$ is an $n \times n$ noise covariance matrix.

The above can be simplified by a linear transformation as described in [10] such that

$$P(\bar{x}'|L) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left[-\frac{\sum_i (x'_i - S'_{1i})^2}{2\sigma^2}\right], \tag{29}$$

and

$$P(\bar{x}'|K) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left[-\frac{\sum_i (x'_i - S'_{2i})^2}{2\sigma^2}\right]. \tag{30}$$

The summation above is carried over i=1 to i=n. The diagonization of the noise covariance matrix and the equalization of the variance do not require the uncorrelatedness of the individual components and the above process can be applied for any Gaussian process[11].

Equations (29) and (30) can be further simplified by rotating and shifting the coordinate axis to obtain

$$P(\bar{x}''|L) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{\sum_{i=1}^{n} x''^2_i}{2\sigma^2}\right] \tag{31}$$

and

$$P(\bar{x}''|K) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{(x_1''-|\bar{S}_1'-\bar{S}_2'|)^2}{2\sigma^2} - \frac{\sum_{i=2}^{n} x_i''^2}{2\sigma^2}\right] \quad (32)$$

where the absolute value represents the length of the vector.

Using the same technique as before, one obtains

$$\Lambda(\bar{x}') = \exp\left(\frac{2x''|\bar{S}_2'-\bar{S}_1'|-|\bar{S}_2'-\bar{S}_1'|^2}{2\sigma^2}\right) > \lambda \text{ assign to K,}$$

$$\text{to L otherwise} \quad (33)$$

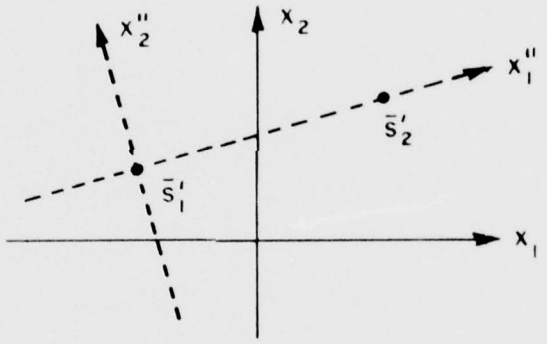


Figure 4. Rotating and shifting coordinate axis to obtain a simpler expression for Equations (29) and (30).

It is clear that the decision boundary is a straight line in a two-dimensional space, a plane in a three-dimensional space and a hyperplane in an n-dimensional space in the transformed coordinate system $\bar{x}''$. This boundary surface is then used, with the knowledge of the previous linear transformation to obtain the threshold boundary in the original system.

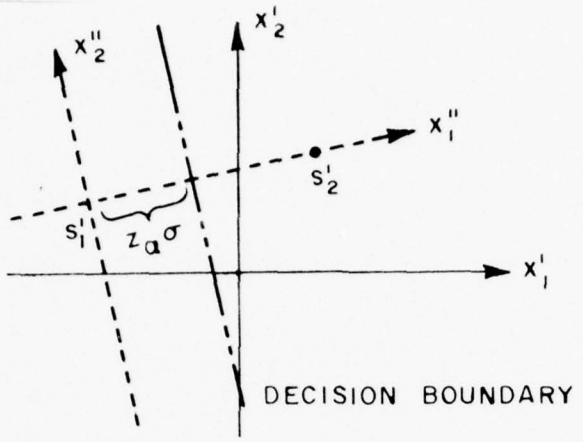


Figure 5. The distance between $\bar{S}_1'$ and the decision boundary in the $\bar{x}'$ coordinate system is $z_\alpha\sigma$.

It is interesting to note that the distance between the decision boundary and S_1' is still $z_\alpha\sigma$ (Figure 5). Obviously, the decision plane depends on the orientation of $\bar{S}_1'$ and $\bar{S}_2'$ and their relative positions too. However, it is always a hyperplane tangential to the hypersphere, whose radius is $z_\alpha\sigma$, centered at $\bar{S}_1'$ in the $\bar{x}'$ coordinate system (Figure 6). Note that the radius is independent of $\bar{S}_2'$.

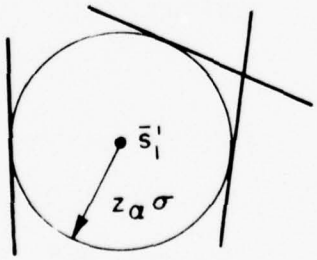


Figure 6. All possible decision planes are tangent to the hypersphere of radius $z_\alpha\sigma$, centered at $\bar{S}_1'$.

The error probability β for this n-dimensional case is obtained similarily as

$$\beta = \text{erfc}\left(z_\alpha - \frac{|\bar{S}_2' - \bar{S}_1'|}{\sigma}\right) \quad , \tag{35}$$

where $\bar{S}_2'$ and $\bar{S}_1'$ are the transformed vectors of $\bar{S}_2$ and $\bar{S}_1$ discussed before.

Note that when a train of observed signals, $x_1, x_2, \cdots, x_n$ are taken independently, the noises added to each component are very often independent of one another. One can simply use Equations (29) and (30) directly.

D. Middleton's Modified Neyman-Pearson Rule

When there is more than one element in either of the two classes, or when $\bar{S}_1$ and $\bar{S}_2$ are statistically distributed in the observation space Z, Middleton proposed to fix the average of P_F, i.e., $\langle P_F \rangle$, instead of P_F and minimize the average of P_M, i.e., $\langle P_M \rangle$[12]. This requires the conditional probability density functions $P(\bar{S}_1|L)$ and $P(\bar{S}_2|K)$.

1. One element in L

(a) Suppose there is only one element in the learning class L, say S_1, in a one-dimensional observation space, and two elements S_{21} and S_{22} in the exterior class K with

$$P(S_{21}|K) = p_1 \quad , \tag{35}$$

and

$$P(S_{22}|K) = p_2 = 1 - p_1 \quad . \tag{36}$$

By using the similar method, we can obtain the decision rule as:

$$\Lambda(x) = \frac{p_1 P(x|S_{21}) + p_2 P(x|S_{22})}{P(x|L)} > \lambda \text{ assign to K,}$$

$$\text{to L otherwise} \quad . \tag{37}$$

If the corrupting noise is Gaussian with zero mean and variance of σ^2, $\Lambda(x)$ becomes

$$\Lambda(x) = p_1 \exp\left[\frac{2x(S_{21}-S_1)-(S_{21}^2-S_1^2)}{2\sigma^2}\right]$$

$$+ p_2 \exp\left[\frac{2x(S_{22}-S_1)-(S_{22}^2-S_1^2)}{2\sigma^2}\right]$$

$$= A(x) + B(x) \quad , \tag{38}$$

where

$$A(x) = p_1 \exp\left[\frac{2x(S_{21}-S_1)-(S_{21}^2-S_1^2)}{2\sigma^2}\right] \tag{39}$$

$$B(x) = p_2 \exp\left[\frac{2x(S_{22}-S_1)-(S_{22}^2-S_1^2)}{2\sigma^2}\right] \tag{40}$$

.

As before

$$P_F = \alpha = \int_{\Lambda^*}^{\infty} P(\Lambda|L)d\Lambda \tag{16}$$

$$= \int_{Z_K} P(x|L)dx \quad . \tag{41}$$

Note that $P(x|L)$ is independent of S_{21}, S_{22}.

When $S_{21}, S_{22} > S_1$ or $S_{21}, S_{22} < S_1$, the threshold turns out to be the same as in the one element case obtained in the last section.

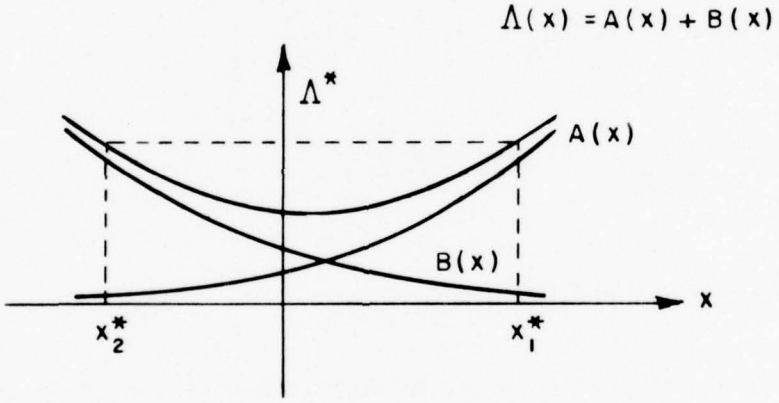


Figure 7. A special case for $\Lambda(x)$, where $S_{22} < S_1 < S_{21}$. Z_L is a region covering $[x_2^*, x_1^*]$.

When $S_{22} < S_1 < S_{21}$ (Figure 7) or $S_{22} > S_1 > S_{21}$, $Z_L = [x_2^*, x_1^*]$ can be found by Equations (41) and (38). A table is constructed of $\alpha(\Lambda)$ vs Λ from Equation (16) and a Λ^* is chosen such that $\alpha(\Lambda^*)$ is equal to the prefixed value. The corresponding x_1^*, x_2^* are then obtained from Equation (38).

The region Z_L depends not only upon the relative positions of S_{21},S_{22} to S_1, but also upon the values of S_{21}-S_1, S_{22}-S_1, $p_1P(S_{21}|K)$ and $p_2P(S_{22}|K)$. This above conclusion holds in a multi-dimension case with the scalors being replaced by the vectors.

When there are more than two elements in K, the computation is even more complicated and results in a more intricate region Z_L.

(b) If there is only one element $\bar{S}_1$ in the learning class L and $\bar{S}_2$ is distributed in the observation space Z as follows:

$$P(\bar{S}_2|K) = \left(\frac{1}{\sqrt{2\pi\sigma_s^2}}\right)^n \exp\left[-\frac{|\bar{S}_2-\bar{S}_1|^2}{2\sigma_s^2}\right] \tag{42}$$

i.e., normally distributed around $\bar{S}_1$.

We have

$$\langle P(\bar{x}|K)\rangle = \int P(\bar{S}_2|K)\, P(\bar{x}|\bar{S}_2)d\bar{S}_2 \quad . \tag{43}$$

Suppose, the noises added to each component are independent of one another and are again Gaussian with zero mean and variance σ^2, Equation (43) becomes

$$\langle P(\bar{x}|K)\rangle = \left(\frac{1}{\sqrt{2\pi(\sigma_s^2+\sigma^2)}}\right)^n \exp\left[-\frac{|\bar{x}-\bar{S}_1|^2}{2(\sigma_s^2+\sigma^2)}\right] \tag{44}$$

Applying the same technique, we obtain

$$\Lambda(\bar{x}) = \frac{\langle P(\bar{x}|K)\rangle}{\langle P(\bar{x}|L)\rangle} \tag{45}$$

$$= \left(\frac{\sigma^2}{\sigma_s^2+\sigma^2}\right)^{n/2} \exp\left[\frac{\sigma^2|\bar{x}-\bar{S}_1|^2}{2\sigma_s^2(\sigma^2+\sigma_s^2)}\right] \tag{46}$$

indicating that Λ^* or $|\bar{x}^*-\bar{S}_1|$ is again independent of $\bar{S}_2$, and depends on α only. The decision boundary here is a hypersphere with the radius being a function of α.

The above example demonstrates that if the distribution of $\bar{S}_2$ is symmetric with respect to $\bar{S}_1$, the decision boundary will depend upon the preset false alarm probability α only.

2. Multiple elements in L

When there is more than one element in L, the computation becomes more complicated but the rules still hold. An example is given here.

Suppose there are two elements in L, whose corresponding noise free points are $\bar{S}_{11}$ and $\bar{S}_{12}$ in a two-dimensional observation space. Also, $\bar{S}_{11}=-\bar{S}_{12}$ and $P(\bar{S}_{11}|L)=P(\bar{S}_{12}|L)=1/2$. $\bar{S}_2$ is uniformly distributed in a circle, centered at the origin with the radius of r_s.

Then

$$\langle P(x|K)\rangle = \int P(\bar{x}|\bar{S}_2)P(\bar{S}_2|K)d\bar{W}_2$$

$$= \int_0^{2\pi}\int_0^{r_s} \frac{1}{2\pi\sigma^2} \exp\left[-\frac{|\bar{x}-\bar{S}_2|^2}{2\sigma^2}\right]\frac{|\bar{S}_2|}{\pi r_s^2}\, d|\bar{S}_2|d\theta$$

$$= I_0\left(\frac{xr_s}{\sigma^2}\right)\left(\frac{1}{2\pi\sigma^2}\right)\exp\left[-\frac{(x^2+r_s^2)}{2\sigma^2}\right] \tag{47}$$

where I_o is the modified Bessel function of order zero and is an even function of its argument (Figure 8).

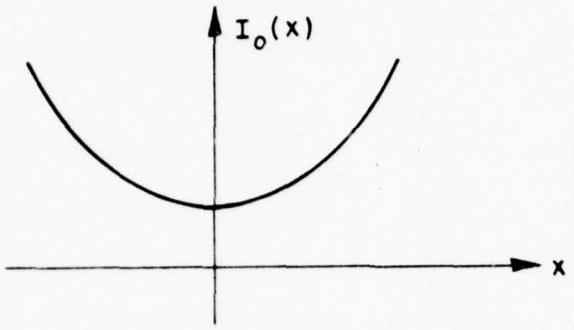


Figure 8. A modified Bessel function of degree zero is a two-dimensional hyperbolic cosine function.

Also

$$\langle P(x|L)\rangle = \frac{1}{2}\frac{1}{2\pi\sigma^2}\left[\exp\left(-\frac{|\bar{x}-\bar{S}_{11}|^2}{2\sigma^2}\right)+\exp\left(-\frac{|\bar{x}-\bar{S}_{12}|^2}{2\sigma^2}\right)\right]$$

$$= \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+S_{11}^2}{2\sigma^2}\right)\cosh\left(\frac{\bar{x}\cdot\bar{S}_{11}}{\sigma^2}\right) \qquad (48)$$

where $\bar{x}\cdot\bar{S}_{11}$ is the inner product of two vectors $\bar{x}$ and $\bar{S}_{11}$, and

$$\Lambda(\bar{x}) = \frac{I_o\left(\frac{x\, r_s}{\sigma^2}\right)}{\cosh\left(\frac{x\, S_{11}\cos\theta}{\sigma^2}\right)} \qquad (49)$$

where θ is the directional angle of $\bar{x}$ and $\bar{S}_{11}$.

It is clear that the decision boundary is no longer a circle here (Figure 9).

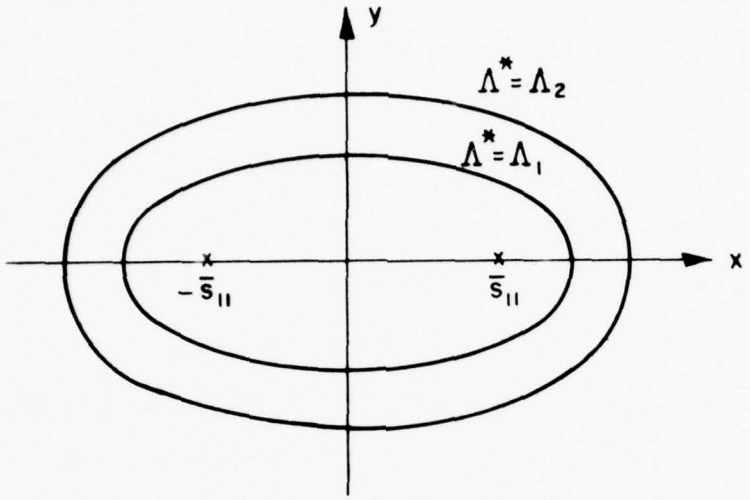


Figure 9. Constant Λ^* curves.

E. Identifying Unlisted Objects

So far, we have catalogued all the known objects into two classes; viz., the learning class L and the exterior class K. The observation space Z is divided into two disjoint regions Z_L and Z_K, corresponding to the two classes L and K. The observed object is assumed to be one of the known objects. Its response is measured and assigned, according to the region where the measurement falls, to one of the two classes. The classifier is claimed to be optimum in that it minimizes one kind of classification error while fixing the other one.

In our problem, there are also two disjoint classes: viz., the listed and the unlisted classes. To make a classification, or more specifically, to decide whether an object is known or not, there are also two kinds of errors, namely, the error made

when N is true and the error made when X is true. However, since the objects in the unlisted class are unknown, no a priori information about them is available. Consequently, minimizing the error made when X is true or fixing this kind of error is not possible.

We note that in the case of Neyman-Pearson rule the observation space Z is divided into two regions, Z_L and Z_K corresponding to the two classes L and K. Similarily, in dealing with our problem, we can also divide the observation space Z into two regions Z_N and Z_X, associated with the listed class and the unlisted class respectively. Any observed measurement that falls into the region Z_N (or Z_X) will be identified as belonging to the listed (or unlisted) class. The error made when N is true is also called a type one error and the error made when X is true is a type two error. The following criterion is used to determine an optimum classifier.

1. The criterion

In discriminating between uncatalogued objects from catalogued objects, the observation space Z is divided into two disjoint regions Z_N and Z_X, associated with the listed and the unlisted classes respectively. When the error probability of classifying a known object in the listed class to the unlisted class is prefixed to a specific value other than zero, the optimum rule is the one which minimizes Z_N.

By minimizing Z_N we mean minimizing $V(Z_N)$, the volume of the region Z_N, where we use the notation

$$V(Z_a) \triangleq \int_{Z_a} d\bar{x} \quad .$$

The main idea here is similar to that of the Neyman-Pearson classifier. In the latter, the probability of making a type two error is minimized while that of making a type one error is fixed. In our case, a type two error is impossible to determine; instead, the type one error probability is fixed, while the volume of Z_N, the region associated with the listed class, is minimized. By minimizing $V(Z_N)$, the possibility of an unlisted object response falling into Z_N will also be minimized, thus reducing the possibility of making a type two error. This can be clearly seen from Equation (3). Since $P(\bar{x}|X)$ is always positive, whenever the region Z_N decreases, β will be reduced accordingly.

This is indeed an optumum classifier which we need in identifying an unknown object. The criterion prefixes the type one error probability such that the region Z_N is large enough not to exclude noisy responses from the known objects. Meanwhile Z_N is kept by the criterion as small as possible so that the classifier can most effectively identify an unknown object.

Therefore, the main philosophy in designing a classifier to identify an unknown object is based on the following two principles:

1. Keeping the probability of misclassifying a listed object as unlisted to a fixed value. This is done by fixing the error probability α when constructing the region Z_N.

2. Minimizing the likelihood of classifying an uncatalogued object as catalogued. This is accomplished by minimizing the volume of Z_N, the region associated with the known class.

Note that the last rule implies minimizing the type two error probability without any information regarding the unknown class.

Z_N can be a compact region or a set of several disjoint regions, depending upon the distribution of the observed vector $\bar{x}$ in the observation space Z. However, the volume of Z_N is always finite when the error probability is greater than zero (when α equals zero, the problem becomes trivial and is not considered here). This will be proved later. Since the volume of the region Z_N is minimized, the criterion is referred to as a minimum volume criterion. The two regions Z_N and Z_X are disjoint and occupy the whole observation space. Some of their properties will be further discussed in the next chapter.

2. Null class

We classify an observed object to be one of the two classes; namely, N and X, as we get some responses from an object. This actually implies that the classifier may make an identification even without the presence of any object, so long as a response is shown in the measurement. The case of no object present is independent of the two aforementioned classes and is another kind of

class itself. It is therefore called a null class and is denoted by ζ.

In a practical system, noise is added to the original signal and is reflected in all measurements. Therefore, when a detector detects a signal, the possibility that a tested vector is originating from a null class cannot be ruled out and should be taken into consideration in an identification problem. The noise free signal of a null class is obviously zero, invariant with the features selected and may be considered a known object.

Figure 10 shows an application of this concept. The null class is combined with the listed class N to form a new class N'. N' is then used, with the preset error probability for a type one error as *defined before, to construct a region* $Z_{N'}$ for the new listed class N'. A measured vector is tested whether it is in $Z_{N'}$. If it is in $Z_{N'}$, a conventional scheme is used to identify the object as one of the catalogued objects, including the case of no object.

Note that the relative frequency of occurrence of each object in the new listed class, including that of the null class, has to be known to construct the region $Z_{N'}$. Therefore, the a priori probability of the null class when both ζ and N are present has to be determined beforehand. Since it can range from a small number close to zero to some number near one, depending on the system, we will not include the null class in our subsequent discussions. However, the process is similar to the N case and is implicitely included in the general discussion. Unless

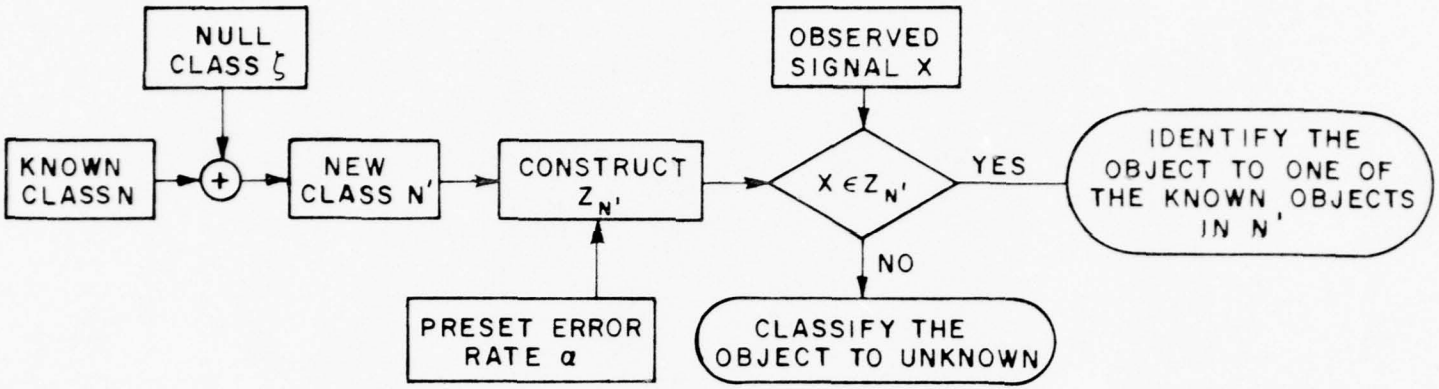


Figure 10. A classifier of identifying unknown objects with the introduction of a null class before forming a decision region.

the a priori probability of the null class is close to one, in which case Z_N is enlarged in the vicinity of the origin, no special consideration is necessary for the null class. Nevertheless, a design engineer should not ignore the existence of a null class in a practical system.

CHAPTER III
SOME PROPERTIES OF THE REGION Z_N

The previous discussion illustrated the rules governing the design of a classifier for identifying unlisted objects as distinct from listed ones. The goal of the design is to find a region Z_N satisfying the constraint that the volume $V(Z_N)$ be minimized while fixing the probability of making a decision error when the observed object is in the listed class. The problem of finding a boundary surface separating the observation space Z into Z_N and Z_X is equivalent to that of finding a threshold in a transformed space. It will be shown that utilizing the threshold greatly simplifies the problem.

We begin this chapter by demonstrating some characteristics of the region Z_N introduced in the last chapter. With the assumption that the probability density function of the observed vector is continuous and bounded, we first show that there exists some Z_N satisfying the proposed criterion and that under certain conditions Z_N is unique. The proof also reveals a method of constructing Z_N.

A method of transformation is then deduced from the above results. A few one-dimensional examples are worked out analytically. In general, however, a closed form solution is difficult to obtain, due to the complexity of the transformation for most of the functions encountered in practice. Yet, with the help of

Monte Carlo simulations, all the problems are numberically solvable except in the cases where Z_N is not unique. Two algorithms are proposed and applied to a practical aircraft identification problem. The results show that the criterion leads to an effective and easily implementable scheme.

The proposed criterion is compared with the Neyman-Pearson criterion in a one-dimensional case. It is shown that our scheme is less sensitive to a priori information of the unlisted class than a Neyman-Pearson classifier to the a priori information of the exterior class. However, in terms of the probability of misclassification, the latter does perform better because it minimizes the error probability while the former minimizes the volume of the region only. The tradeoffs are also discussed.

A. Some Properties of Z_N

A classifier measures a noise corrupted response of an object, say $\bar{x}$, and assigns the observed object into the listed class or the unlisted one. The probability density function of the measurement $\bar{x}$ when the listed class is present is denoted by

$$g(\bar{x}) \triangleq P(\bar{x}|N) \quad . \tag{50}$$

In case there are several subclasses (or objects) $C_1, C_2, \cdots, C_n$ in the listed class, Equation (50) becomes

$$g(\bar{x}) = \sum_{i=1}^{n} P(\bar{x}|C_i)P(C_i|N) \tag{51}$$

where

n is the number of subclasses (or objects)

$P(C_i|N)$ is the relative frequency of occurrence of C_i in N, and

$P(\bar{x}|C_i)$ is the conditional probability density function when C_i is present.

If the ith noise free signal $\bar{S}_i$ is distributed over the observation space Z and its distribution function is denoted by $f_i(\bar{S}_i|C_i)$, Equation (51) becomes

$$g(\bar{x}) = \sum_{i=1}^{n} P(C_i|N) \int P(\bar{x}|\bar{S}_i) f_i(\bar{S}_i|C_i) d\bar{S}_i \quad . \tag{52}$$

Here, $P(\bar{x}|\bar{S}_i)$ is the probability density function of the measured vector $\bar{x}$ when the noise free signal is $\bar{S}_i$. The integration is carried out over the entire observation space.

Obviously, if there is only one subclass in N, Equation (52) can be simplified as

$$g(\bar{x}) = \int P(\bar{x}|\bar{S}) f(\bar{S}|N) d\bar{S} \quad , \tag{53}$$

again, $f(\bar{S}|N)$ is the probability density function of the noise free signal $\bar{S}$.

In general, $g(\bar{x})$ is either continuous or discrete in the observation space. However, for most of practical cases, like those where Gaussian noise is added to noise free signal, $g(\bar{x})$ is continuous and defined over the whole observation space. Therefore, in the following discussions, we will make the assumption that $g(\bar{x})$ is continuous and bounded everywhere in the observation space.

From the basic characteristics of a probability density function we also have

$$\int_Z g(\bar{x})d\bar{x} = 1 \quad , \tag{54}$$

and

$$g(\bar{x}) \geq 0 \quad \text{for all } \bar{x} \text{ in } Z \quad . \tag{55}$$

By definition, Z_N is a region of observation space in which an observed signal is assigned to the listed class. The probability of misclassifying known objects as unknown is therefore

$$P_F = \int_{\bar{Z}_N} g(\bar{x})d\bar{x} \tag{56}$$

where

$$\bar{Z}_N \triangleq Z - Z_N = Z_X \quad , \tag{57}$$

and $Z_N \cup \bar{Z}_N = Z$ by definition.

In other words, P_F can be expressed as

$$P_F = 1 - \int_{Z_N} g(\bar{x})d\bar{x} \quad . \tag{58}$$

Our criterion is to hold P_F at a fixed value and to minimize the region Z_N, viz., minimize

$$V(Z_N) \triangleq \int_{Z_N} d\bar{x} \quad , \tag{59}$$

provided that $V(Z_N)$ exists.

The observation space Z, as stated before, is divided into two disjoint regions Z_N and Z_X by a threshold surface S_T. The classification rules are as follows:

Let N denote the listed (known) class, X the unlisted class and C_X the observed object.

if $\bar{x} \in Z_N$ C_X is assigned to N,

if $\bar{x} \in Z_X$ C_X is assigned to X, and

if $\bar{x}$ is on S_T, C_X can go either way.

Usually S_T is contained in either Z_N or Z_X, indicating that Z_N is a closed or an open set. Our main objective here is to find a threshold surface S_T.

We now show that for a continuous and bounded density function $g(\bar{x})$ there exists at least one region Z_N described by the criterion. We first state and prove a lemma below and use it to prove the existence of Z_N.

1. Lemma

Assume that $g(\bar{x})$, a probability density function, is continuous and bounded in the observation space Z and α is a number such that $0<\alpha<1$. If a region Z_N in Z satisfies

$$(1) \quad \int_{Z_N} g(\bar{x})d\bar{x} = 1 - \alpha \quad , \tag{60}$$

(2) for any $\bar{x} \in Z_N$, $\bar{y} \notin Z_N$

$$g(\bar{x}) \geq g(\bar{y}) \quad ; \tag{61}$$

and a region $Z_{N'}$ in Z, different from Z_N, satisfies

$$\int_{Z_{N'}} g(\bar{x})d\bar{x} = 1 - \alpha \quad , \tag{62}$$

then

$$V(Z_N) \leq V(Z_{N'}) \tag{63}$$

2. Proof

We first show that the region Z_N specified by Equations (60) and (61) is finite and then show that for any $Z_{N'}$ specified by Equation (62), Equation (63) holds.

Since $1>\alpha>0$, from Equation (60) we have

$$\int_{\bar{\bar{Z}}_N} g(\bar{y})d\bar{y} = \alpha > 0$$

where $\bar{\bar{Z}}_N$ is the complement of Z_N. This equation shows that not all $g(\bar{y})\equiv 0$ for $\bar{y} \in \bar{\bar{Z}}_N$.

Therefore, there must exist a $\bar{y}_1 \notin Z_N$, such that $g(\bar{y}_1)=a>0$. From the given condition Equation (61), we have

$$g(\bar{x}) \geq g(\bar{y}_1) = a > 0 \qquad \text{for any } \bar{x} \in Z_N .$$

Also, by Equation (60) and the above

$$1 > \int_{Z_N} g(\bar{x})d\bar{x} \geq a \int_{Z_N} d\bar{x}$$

we get

$$\int_{Z_N} d\bar{x} = V(Z_N) < \frac{1}{a} \qquad .$$

Since $a>0$, $V(Z_N)$ is finite.

If the region $Z_{N'}$ specified by Equation (62) is infinite, then Equation (63) is always true.

If $Z_{N'}$ is finite and different from Z_N, then let

$$Z_C \triangleq Z_{N'} \cap Z_N$$

and define

$$Z_{f'} \triangleq Z_{N'} - Z_C$$

$$Z_f \triangleq Z_N - Z_C \quad .$$

From the given conditions

$$\int_{Z_N} g(\bar{x})d\bar{x} = \int_{Z_{N'}} g(\bar{x}')d\bar{x}' = 1 - \alpha \quad .$$

Subtracting the above by the integration over Z_C, one gets

$$\int_{Z_f} g(\bar{x})d\bar{x} = \int_{Z_{f'}} g(\bar{x}')d\bar{x}' \quad .$$

However, since

$$g(\bar{x}) \geq g(\bar{x}') \qquad \text{for any } \bar{x} \in Z_f$$
$$\bar{x}' \in Z_{f'} \quad ,$$

the above equality holds only if

$$\int_{Z_f} d\bar{x} \leq \int_{Z_{f'}} d\bar{x} \quad .$$

That is

$$V(Z_f) \leq V(Z_{f'}) \quad .$$

Adding the volume of Z_C on both sides, we have

$$V(Z_N) \leq V(Z_{N'}) \quad .$$

This proves the lemma.

We have shown that among all the regions satisfying Equation (62), the one specified by Equation (61) is finite and has the minimum volume. Now we show that, for any $g(\bar{x})$ being continuous and bounded, there exists at least a minimum region Z_N satisfying Equation (62).

We first construct a region

$$S(\xi) \triangleq \{\bar{x} | g(\bar{x}) \geq \xi\} \quad , \text{ for } \xi \geq 0 \tag{64}$$

i.e., a region formed by all points for which the value of the function $g(\bar{x})$ is larger than or equal to ξ.

Its complement is denoted by

$$\bar{S}(\xi) \triangleq Z - S(\xi)$$
$$= \{\bar{x} | g(\bar{x}) < \xi\} \quad , \quad \text{for } \xi > 0 \quad . \tag{65}$$

It can be proved, following the same reasoning line which proved that $V(Z_N)$ was finite, that $V(S(\xi))$ is always finitie if $\xi > 0$.

From the lemma, if

$$\int_{S(\xi)} g(\bar{x}) d\bar{x} = 1 - \alpha \tag{66}$$

then $S(\xi)$ is a minimum region constrained by the criterion.

Now we prove the following theorem.

Theorem

For any continuous and bounded probability density function $g(\bar{x})$, and an α such that $0<\alpha<1$, there exists a minimum Z_N in Z satisfying

$$\int_{Z_N} g(\bar{x})d\bar{x} = 1 - \alpha \quad .$$

Proof

Defining

$$G(\xi) \triangleq \int_{S(\xi)} g(\bar{x})d\bar{x} \tag{67}$$

where $S(\xi)$ is defined by Equation (64).

By the definition of $S(\xi)$ and continuity of $g(\bar{x})$,

$$S(\xi_2) \supset S(\xi_1) \text{ if } \xi_m > \xi_1 > \xi_2 > 0 \quad ,$$

where ξ_m is the maximum of $g(\bar{x})$ in the observation space Z.

Note that $S(\xi_2)$ strictly contains $S(\xi_1)$. Therefore we can obtain

$$G(\xi_2) > G(\xi_1) \quad \text{if } \xi_m > \xi_1 > \xi_2 > 0$$

indicating that $G(\xi)$ is a strictly decreasing function.

Also, from the basic characteristics of a probability density function, we have

$$G(0) = 1 \quad ,$$

and

$$G(\xi_n) = 0 \quad , \quad \text{for any } \xi_n > \xi_m \quad .$$

(a) If $G(\xi)$ is continuous over $[0,\xi_m]$ (Figure 11), then since $G(\xi)$ is strictly decreasing in ξ, there is a one-to-one correspondence between ξ and $G(\xi)$. For every $\alpha \ni 0<\alpha<1$, we find a unique ξ_T such that

$$G(\xi_T) = 1 - \alpha \quad . \tag{68}$$

ξ_T is between 0 and ξ_m and is not equal to 0 or ξ_m since $0<\alpha<1$.

As shown before, $S(\xi_T)$ is the minimum region Z_N.

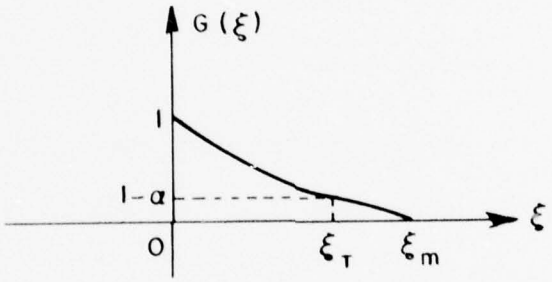


Figure 11. $G(\xi)$ is continuous.

(b) If $G(\xi)$ is not continuous in ξ, say discontinuous at a point $\xi=\xi_d$ (Figure 12), let

$$S_1(\xi) \triangleq \{\bar{x} | g(\bar{x}) > \xi\} \quad . \tag{69}$$

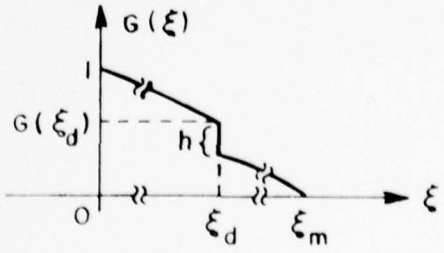


Figure 12. $G(\xi)$ is discontinuous.

We have

$$G(\xi_d) = G_T(\xi_d^+) + h \quad , \tag{70}$$

where $h>0$ is a jump of $G(\xi)$ at ξ_d and

$$G(\xi_d^+) \triangleq \int_{S_1(\xi_d)} g(\bar{x})d\bar{x} \quad . \tag{71}$$

From the definition of $G(\xi)$, it is clear that

$$G(\xi_d) - G(\xi_d^+) = \int_{S_o(\xi_d)} g(\bar{x})d\bar{x} \quad , \tag{72}$$

where S_o is defined as

$$S_o(\xi) \triangleq \{\bar{x} | g(\bar{x}) = \xi\} \quad , \tag{73}$$

viz., the set of all the points where $g(\bar{x}) = \xi$.

This indicates that there is a finitie volume for the region $S_o(\xi_d)$ in which the function $g(\bar{x})$ is constant and equals to ξ_d, where the discontinuity of $G(\xi)$ occurs.

Consequently, for any number b such that $0<b<h/\xi_d$ we can obtain a region $S_{o_b}(\xi_d)$ in $S_o(\xi_d)$ so that

$$V(S_{o_b}(\xi_d)) = b \quad . \tag{74}$$

This is equivalent to

$$\int_{S_{o_b}(\xi_d)} g(\bar{x})d\bar{x} = \xi_d b \tag{75}$$

with

$$S_{o_b}(\xi_d) \subset S_o(\xi_d) \quad . \tag{76}$$

Therefore for any α such that $G(\xi_d^+)<(1-\alpha)<G(\xi_d)$, we can choose a region

$$S_\alpha(\xi_d) = S_1(\xi_d) \cup S_{o\alpha}(\xi_d) \tag{77}$$

such that

$$S_{o\alpha}(\xi_d) \subset S_o(\xi_d) \tag{78}$$

and

$$V(S_{o\alpha}(\xi_d)) = \frac{(1-\alpha)-G(\xi_d)}{\xi_d} < \frac{h}{\xi_d} \quad . \tag{79}$$

$S_\alpha(\xi_d)$ is not unique since $S_{o\alpha}(\xi_d)$ defined by Equations (78) and (79) is not. Nevertheless, from the previous lemma, $S_\alpha(\xi_d)$ is a minimum region.

For any $(1-\alpha)$ lying in the continuous portion of $G(\xi)$, we can, as before, obtain a unique ξ_T for each α such that $G(\xi_T)=1-\alpha$ since there is a one-to-one correspondence relationship between $G(\xi)$ and ξ over this range. The corresponding $S(\xi_T)$ is again a minimum region.

In conclusion, for any continuous and bounded density function $g(\bar{x})$ and any α such that $0<\alpha<1$, we can always find a minimum region Z_N satisfying

$$\int_{Z_N} g(\bar{x})d\bar{x} = 1 - \alpha \quad .$$

QED.

From the above discussion, we can also have the following

(1) If Z_N satisfies the criterion, then for any $\bar{x} \in Z_N$ and any $\bar{y} \notin Z_N$

$$g(\bar{x}) \geq g(\bar{y}) \quad .$$

(2) If for any $\bar{x} \in Z_N$ and any $\bar{y} \notin Z_N$

$$g(\bar{x}) \geq g(\bar{y}) \quad ,$$

Z_N is a minimum region among all the $\bar{Z}_N$'s satisfying

$$\int_{\bar{Z}_N} g(\bar{x})d\bar{x} = \int_{Z_N} g(\bar{x})d\bar{x} \quad .$$

(3) If Z_N is a region which satisfies the criterion and is closed and if

$$\xi_T = \min_{x \in Z_N} g(\bar{x})$$

then Z_N is unique unless $V(S_o(\xi_T)) \neq 0$, in other words, the volume of the region over which $g(\bar{x})=\xi_T$ is not zero.

A few examples are given here to demonstrate how the above theorem works.

Example 1

When the listed class is present, the probability density function of the observed vector $\bar{x}$ is triangular as shown in Figure 13. We would like to find the region Z_N associated with the listed class when α=0.05, 0.01 and 0.001.

Obviously, g(x) is continuous and its integral over the entire x-axis is one. Also, g(x) is bounded and ranges from 0 to 1.

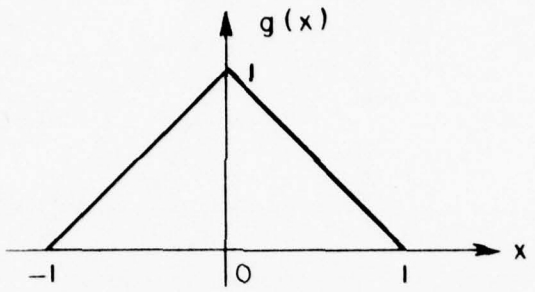


Figure 13. g(x) is a triangular function.

We first find the function $G(\xi)$. As before

$$G(\xi) = \int_{S(\xi)} g(x)dx$$

where

$$S(\xi) = \{x | g(x) \geq \xi\} \quad .$$

We obtain (Figure 14)

$$G(\xi) = \begin{cases} 1-\xi^2 & 0 \leq \xi \leq 1 \quad , \\ 1 & \xi < 0 \quad , \\ 0 & \xi > 1 \quad . \end{cases}$$

From Equation (68)

$$\xi_T = \sqrt{\alpha} \qquad \text{when } 0 < \alpha < 1 \quad .$$

Therefore

$$\begin{aligned} \xi_T &= 0.224 && \text{when } \alpha = 0.05 \quad , \\ &= 0.1 && \text{when } \alpha = 0.01 \quad , \\ &= 0.0316 && \text{when } \alpha = 0.001 \quad . \end{aligned}$$

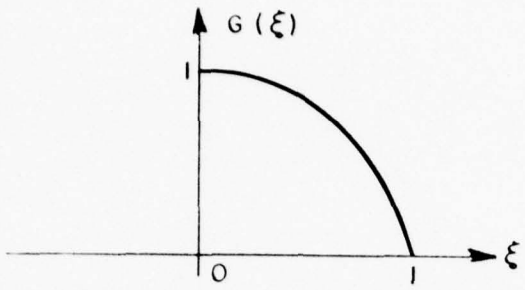


Figure 14. $G(\xi)$ for a triangular function $g(x)$.

Corresponding to the above ξ_T's

$$Z_N = S(\xi_T) = [-0.776, 0.776] \quad \text{when } \alpha = 0.05 ,$$

$$= [-0.9, 0.9] \quad \text{when } \alpha = 0.01 ,$$

$$= [-0.9684, 0.9684] \quad \text{when } \alpha = 0.001 .$$

Example 2

If $g(x)$ is a function as shown in Figure 15, find a corresponding Z_N when α is a fixed number in $(0,1)$.

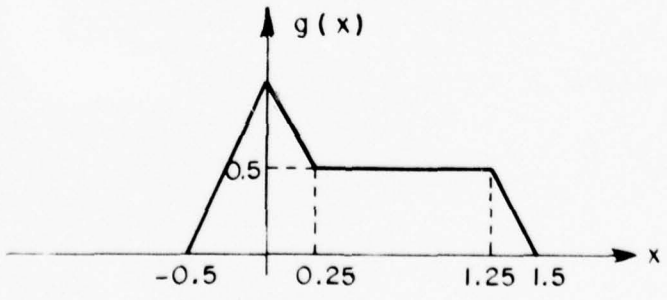


Figure 15. $g(\bar{x})$ including a flat probability density.

Again, g(x) is continuous and bounded everywhere on the x-axis and its integral over the entire x-axis is unity. Like before, we first find $G(\xi)$ by inspection

$$G(\xi) = \begin{cases} 1 & \xi \leq 0 \quad , \\ 1 - \frac{1}{2}\xi^2 & 0 < \xi < 0.5 \quad , \\ \frac{1}{2} - \frac{1}{2}\xi^2 & 0.5 \leq \xi \leq 1 \quad , \\ 0 & \xi > 1 \quad . \end{cases}$$

It is clear that there is a one-to-one correspondence between ξ and $G(\xi)$ when ξ is in (0,1), except at ξ=0.5, or equivalently, when $G(\xi)$ is between 0.375 and 0.875. We can, as before, obtain a unique Z_N for any α whose value is not in (0.125,0.625) and $0<\alpha<1$.

When $0.125<\alpha<0.625$, we can set ξ_T=0.5 and form a region Z_N according to the method described in the theorem.

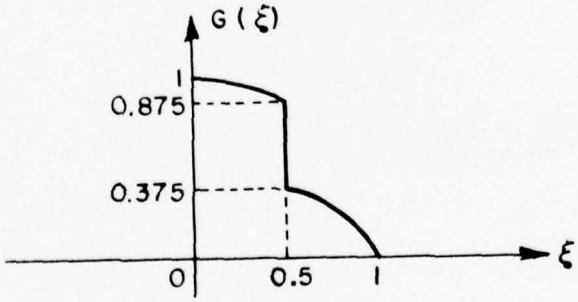


Figure 16. $G(\xi)$ has a jump at ξ=0.5.

For instance, when α=0.5, we can form a region Z_N in Z (i.e., x-axis in this example) such that

$$Z_N = [-0.25,0.25] \cup S_{\alpha 0}(0.5)$$

when $S_{\alpha o}(0.5)$ is a region consisting of a segment or a set of segments whose total length is 0.25 and lies between x=0.25 and x=1.25. However, to have a higher stability in the classification process, a subregion of x=0.25 to x=0.5, right adjacent to x=0.25, is preferred in forming the region Z_N.

This illustrates that for any α such that $0<\alpha<1$, there always exists a region Z_N satisfying the criterion.

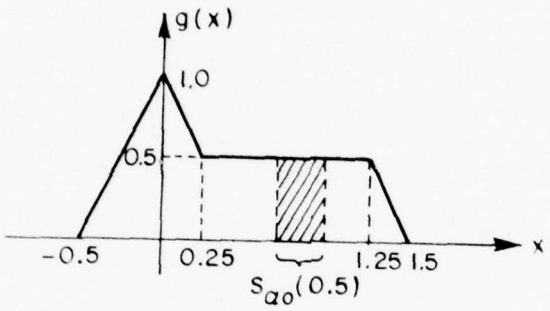


Figure 17. $S_{\alpha o}(0.5)$ is a region consisting of a segment or a set of segments between x=0.25 and x=1.25; its total length is 0.25.

B. One-dimensional Case and Method of Transformation

It is clear that if $V(S_o(\xi_T)) \neq 0$, the region Z_N is not unique and there is an instability in constructing the classifier. This can be more closely related to the function $G(\xi)$ in Equation (67). If the rate of change of $G(\xi)$ with respect to ξ is large, an instability would become noticable. The rate of change is strongly associated with the behavior of $g(\bar{x})$ in the observational space. In most of practical situations, the error probability α is set to be a small number. Under these circumstances, the

instability becomes almost irrelevant to the overall performance because the effect is definitely small when ξ_T is small. This will be shown later in Chapter III-E that even when $g(\bar{x})$ is almost flat, the instability is still not noticeable. The derivative of $G(\xi)$ at ξ_T can serve as an indicator to this instability.

However, if necessary, the instability can be eliminated completely if the classification process is carried out in the $\bar{x}$ domain where a region Z_N can always be found and consists of some compact regions. This essentially guarantees that the problem is solvable as far as $g(\bar{x})$ is continuous and bounded in the observation space.

From the above discussions, it is seen that the solution for our problem is unique when $V(S_o(\xi))\equiv 0$ for all $\xi>0$. This occurs in many practical situations and represents most of the distribution functions that one will encounter. Any monotonic function or the summation of a finite number of monotonic functions are in this category, particularly Gaussian distributions that occur in most communications or measuring systems. The function $g(\bar{x})$ consists of a finite number of such monotones if the number of subclasses in the listed class is assumed to be finite, which is a legitimate postulate in all practical cases. Therefore, in this study, the function $G(\xi)$ defined in Equation (67) will be considered continuous and its corresponding region $S(\xi)$ unique.

With the assumptions made to construct the region Z_N, associated with the listed class for a prefixed error probability α, we can carry out the integration in Equation (67) from the portion where the function $g(\bar{x})$ has the maximum value, in the descending order of $g(\bar{x})$, until $G(\xi)$ reaches the value $1-\alpha$. In doing so, we are actually carrying out the integration of $g(\bar{x})$ over $S(\xi)$ in the descending magnitude of ξ. Therefore, we can define a new variable $P_\varepsilon(\xi)$ such that

$$\begin{aligned} P_\varepsilon(\xi)d\xi &= g(\bar{x})dS(\xi) \\ &= \xi\, dS(\xi) \end{aligned} \tag{80}$$

at every ξ.

In a one-dimensional space, the right side of Equation (80) is equivalent to

$$g(x_1)|dx_1| + g(x_2)|dx_2| + \text{---} + g(x_n)|dx_n$$

where (Figure 18)

$$\xi = g(x_1) = g(x_2) = \text{---} = g(x_n)$$

Using the above two expressions, we obtain

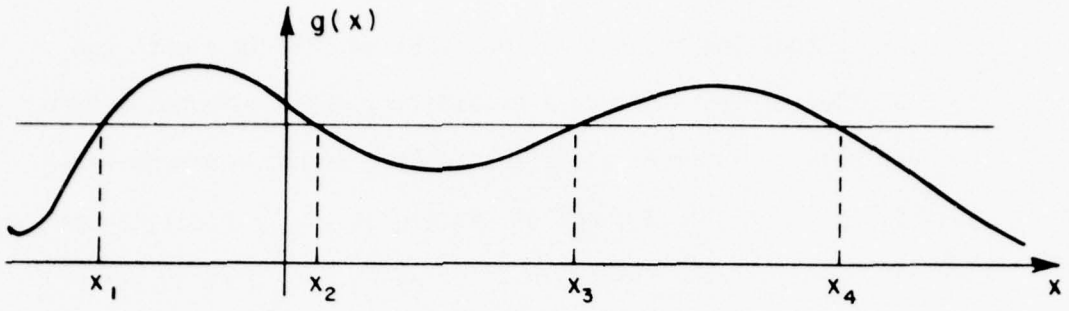


Figure 18. The function $y=g(x)$ intersects $y=\xi$ at $x_1, x_2, ---, x_n$.

$$P_\varepsilon(\xi) = g(x_1)\left|\frac{dx_1}{dg(x_1)}\right| + g(x_2)\left|\frac{dx_2}{dg(x_2)}\right| + --- + g(x_n)\left|\frac{dx_n}{dg(x_n)}\right|$$

$$= \xi \sum_i \frac{1}{\left|\frac{dg(x_i)}{dx_i}\right|} \qquad (81)$$

provided that $g(x)$ is continuous[9,13].

Obviously, if the curve $y=\xi$ intersects $y=g(x)$ at a finite number of points, the function $P_\varepsilon(\xi)$ can be always obtained by the above transformation. Also, $P_\varepsilon(\xi)$ is positive and defined over $(0,\xi_m)$ and

$$\int_0^\infty P_\varepsilon(\xi)d\xi = \int_0^{\xi_m} P_\varepsilon(\xi)d\xi = 1 \quad . \qquad (82)$$

The threshold ξ_T is obtained by solving

$$\int_{\xi_T}^{\xi_m} P_\varepsilon(\xi)d\xi = 1 - \alpha \qquad (83)$$

which can be used to determine the boundary surface S_T.

Solving Equation (83) is a straightforward step if $P_\varepsilon(\xi)$ can be analytically determined from Equation (81).

Note that the transformation Equation (81) is almost the same as that of transforming a probability density function from one variable to another except that ξ here is not a random variable[9,13]. $P_\xi(\xi)$ is a function containing the probability contributions of all the points in Z for which the function $g(x)=\xi$. Reference [13] has a very thorough discussion about the one-to-one and non one-to-one transformation for a continuous function $g(x)$. It also describes the transformation from Z space into ξ domain for a multiple-variable function $g(x_1,x_2,---,x_n)$, which is denoted as $g(\bar{x})$ in our notation. As the number of dimensions increases, the transformation becomes very complicated even when $\xi=g(\bar{x})$ is a very simple function. However, $g(\bar{x})$ is usually not very simple and the transformation is not an easy task in most situations. We will not go into any detailed discussion of the transformation in multi-dimensional space in this work. Instead, a method of Monte Carlo simulation is presented later to ease this computational difficulty, which enables us to bypass all the problems numerically without having to use the transformation process.

One Element Case

A one-dimensional problem given in Section II-C is used here to demonstrate the above approach. Suppose C_1 is the one element in N and its noise free signal is S_1. No information about the unlisted class is available. We would like to find a region Z_N associated with the listed class N when the misclassification probability for the listed class P_F is set to be α.

The probability density function of $\bar{x}$ when the listed class is present, is, as before, assumed to be

$$g(x) = P(x|N) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-S_1)^2}{2\sigma^2}} \quad . \tag{3}$$

We first transform the above probability density function to

$$P_\varepsilon(\xi) = \xi \left|\frac{dx_1}{d\xi}\right| + \xi \left|\frac{dx_2}{d\xi}\right| \tag{84}$$

where ξ ranges from 0 to $1/\sqrt{2\pi\sigma^2}$.

From the symmetry of ξ w.r.t. S_1, it can be shown that

$$S_1 - x_1(\xi_T) = x_2(\xi_T) - S_1,$$

and Equation (83) becomes

$$\int_{-\infty}^{x_1(\xi_T)} g(x)dx + \int_{x_2(\xi_T)}^{\infty} g(x)dx = \alpha$$

or

$$\int_{x_2(\xi_T)-S_1}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx = \frac{\alpha}{2}$$

i.e.,

$$\operatorname{erfc}\left(\frac{x_2(\xi_T)-S_1}{\sigma}\right) = \frac{\alpha}{2} \quad . \tag{85}$$

We have

$$x_2(\xi_T) - S_1 = S_1 - x_1(\xi_T) = z_{\alpha/2}\sigma \tag{86}$$

where z_α is defined in Equation (22), or

$$|X_i(\xi_T) - S_1| = z_{\alpha/2}\sigma \qquad \text{for } i=1,2 \quad . \tag{87}$$

Corresponding to this

$$Z_N = [-z_{\alpha/2}\sigma + S_1,\ z_{\alpha/2}\sigma + S_1] \quad , \tag{88}$$

and

$$\xi_T = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{z^2_{\alpha/2}}{2}} \quad . \tag{89}$$

Comparing this with Equation (25), we see that the boundaries are no longer dependent on S_2, which is not known in our problem anyway.

Two Element Case

If there are two elements S_1 and S_2 in N, the problem becomes more complicated. Now ξ becomes

$$g(x) = p_1\ P(x|S_1) + p_2\ P(x|S_2) \tag{90}$$

where

$$p_1 \triangleq P(S_1|N) \tag{91}$$

$$p_2 \triangleq P(S_2|N) = 1 - p_1 \text{ as defined before.} \tag{92}$$

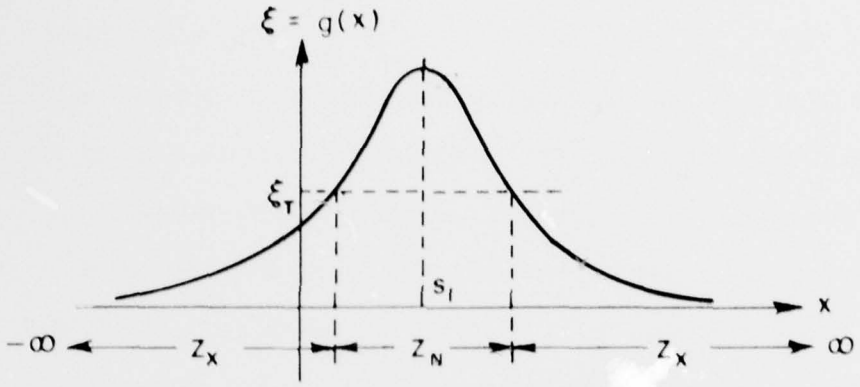


Figure 19. Z_N is a region covering both sides of S_1 symmetrically.

Again, assume the noise added to the signal to be Gaussian with zero mean, then

$$g(x) = p_1 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-S_1)^2}{2\sigma^2}} + p_2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-S_2)^2}{2\sigma^2}} . \qquad (93)$$

By transforming the above into the ξ domain and employing the same technique as before, one gets

$$\int_0^{\xi_T} p_\xi(\xi)d\xi = \alpha \qquad . \qquad (94)$$

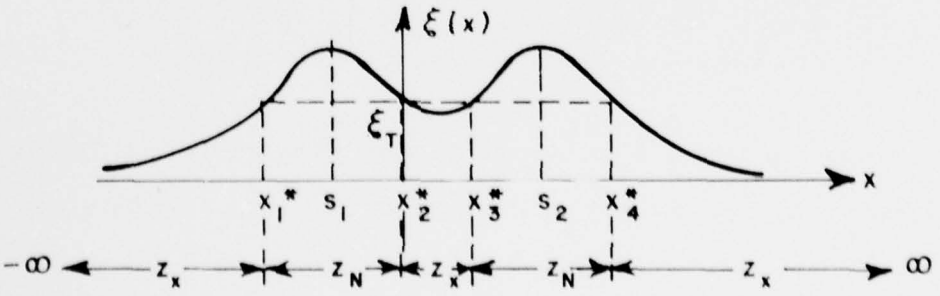


Figure 20. Two element case.

And ξ_T can therefore be solved if $P_\varepsilon(\xi)$ can be obtained by Equation (81). The points x_1^*, x_2^*, x_3^*, x_4^* (Figure 20) corresponding to ξ_T are then obtained by substituting ξ back to $\xi=g(x)$. Note that analytically, $P_\varepsilon(\xi)$ is difficult to obtain because the points for which $g(x)$ is constant are difficult to compute for Equation (93), and even more so the derivatives at these points. Nevertheless, it was carried out and the results listed in Table 1. The new region Z_N, in general, is the union of two disjoint zones $[x_1^*,x_2^*]$ and $[x_3^*,x_4^*]$ since $\xi=g(x)$ is a multi-valued function. Some results are shown in Table 1 for $p_1=p_2=1/2$, $\alpha=0.05$ and $S_1=0$.

TABLE 1
Two Element Case for the Misclassification Probability $\alpha=0.05$, $S_1=0$

S_1/σ	S_2/σ	x_1^*/σ	x_2^*/σ	x_3^*/σ	x_4^*/σ
0	1	-1.68	---	---	2.68
0	2	-1.645	---	---	3.645
0	5	-1.927	1.96	3.04	6.927
0	10	-1.96	1.96	8.04	11.96
0	50	-1.96	1.96	48.04	51.96

In the table, S_1 is set to be zero to simplify the computation. This does not affect the generality of the results since it is considered as a reference point; consequently, the numbers S_2, x_1^*, x_2^* --- in the table are actually S_2-S_1, $x_1^*-S_1$, $x_2^*-S_1$

--- etc. The signal to noise standard deviation (σ) ratios are associated with the SNR of the system. The interval $[x_3^*, x_4^*]$ is a symmetrical reflection of $[x_1^*, x_2^*]$ with respect to the mid-point of S_1 and S_2. For $(S_2-S_1)<3\sigma$, the threshold line only intersects $\xi=g(x)$ twice when $\alpha=0.05$; therefore x_2^*, x_3^* do not exist in that case. Note that under these conditions, $|S_i-x_j^*|$ where $j=2(i-1)$ or $2(i-1)+1$, is very close to zero. As $|S_2-S_1|$ becomes larger in terms of σ, the size of the subregion surrounding S_j, i.e., $|S_j-x_j^*|$ becomes closer to $\sigma z_{\alpha/2}$, the same as that of Z_N found in the last example. This is so because when all the subclasses are equally weighted, the probability density contributed from other subclasses around each element is small as long as the distances between elements are large; consequently, the threshold boundaries around each element are close to those of single element systems. Thus, under the same constraint of α, the sizes and shapes of the subregions of Z_N are like those of Z_N generated by single elements. This indicates that if the noise free signals of each element in the known class are far apart in terms of σ, Z_N will be a joint region of several individual ones which are obtained as if there were only one element in the listed class, provided that each element is equally weighted. This conclusion holds no matter how many subclasses there are in the listed class. We can approximate Z_N by combining the subregions associated with each subclass, obtained as if it were the only one in the listed class, and reduce the complicated computation substantially.

In the last example, if $p_1 \neq p_2$, the function $\xi=g(x)$ will no longer be symmetric with respect to the midpoint of (S_1,S_2) namely, $(S_1+S_2)/2$ (Figure 21). Still, we can transform the original function $\xi=g(x)$ into the ξ domain and compute ξ_T by Equation (94). The required region Z_N is again obtained by substituting ξ_T back into Equation (94) by setting $\xi_T=g(x^*)$. If there are two subregions of Z_N, each one, associated with its noise free signal S_i, will be different in size and shape. The two subregions will merge into one as α decreases.

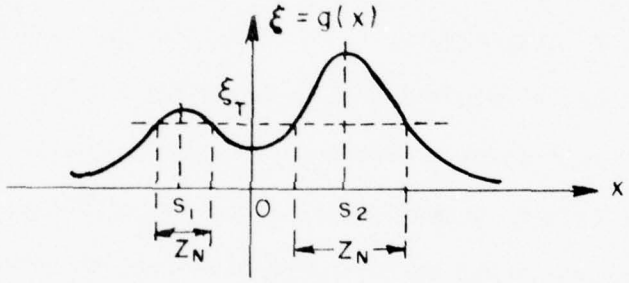


Figure 21. Two element case when $p_1 \neq p_2$.

Another common situation similar to this occurs when the amount of noise added to each signal is not the same. For instance, the noise injected to each signal is proportional to its signal amplitude in some systems. This multiplicative noise will also result in different sizes and shapes of Z_N for different subregions than those described in the above example.

In case that α is large, very often some of the noise free points originating from the subclasses of the listed class are not contained in Z_N at all (Figure 22).

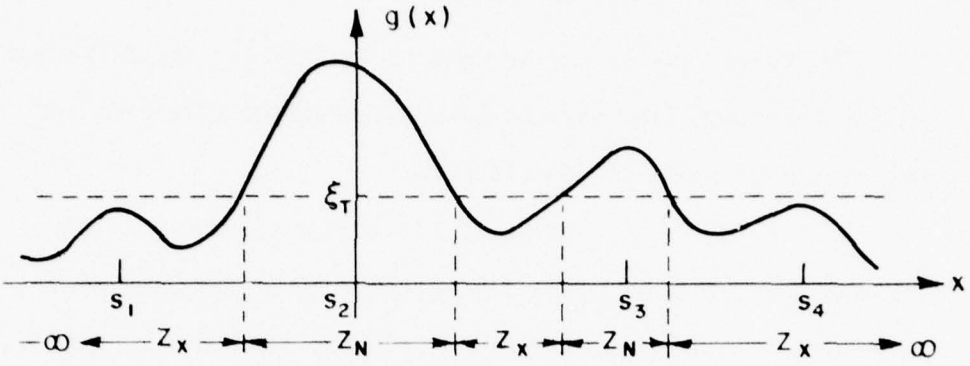


Figure 22. Z_N does not consists of the signal points S_1 and S_4.

This is most likely due to the following:

(1) The a priori probabilities of some subclasses in the listed class are very low as compared with those of others. When α is large, Z_N consists of only the highly weighted regions, leaving out the signal points of low densities. Consequently, they are excluded from Z_N.

(2) The noises added to some signal points are larger than those added to others, making the probability contributions from these points smaller than those from others. For the noise distribution being Gaussian or decreasingly unimodal, a classifier is therefore more likely to exclude some signal points corrupted by larger noise from Z_N when the detection probability $1-\alpha$ is not very high.

C. Comparisons of the Two Criteria

The above results can be used to demonstrate the differences between the classifier devised by Neyman-Pearson criterion and that by the minimum volume criterion.

Both criteria are aiming at the same objective of minimizing one kind of error while fixing the other. The lack of a priori information about one of the classes leads one to introduce the idea of minimizing the region associated with the known class, which is really the only class we know and can do something about. A Neyman-Pearson classifier can fix either of the two error probabilities and minimize the other while the classifier devised by the proposed criterion can only minimize the specified region and fix the error probability associated with the known class. This indicates that the former is more flexible in choosing a class whose associated error is to be fixed. However, in almost all practical situations, a system very often gives a strong preference in fixing one of the two error probabilities, thus making the two classifiers almost identical in terms of the flexibility with respect to this kind of selection.

Another difference between the two classifiers is the quantity to be minimized. A Neyman-Pearson classifier minimizes a type two error probability which is associated with the exterior class (see Chapter II-C), while the classifier devised in this work minimizes the region associated with the exterior class. It can be shown that the two classifiers are identical (see the examples in Chapter II-D) when the corrupting noise is Gaussian with

equal variances in all dimensions, the noise free signal of the first class occupies only a single point in the observation space, and the noise free signal of the exterior class is normally and symmetrically distributed around the noise free signal of the learning class. Extending this to the limiting case of $\sigma \to \infty$, we can assert that the two classifiers are identical when the noise free signal of the second class is uniformly distributed over the whole observation space and the noise is Gaussian of equal variance in every dimension. This kind of "equal distribution of the noise free signal" of the second class, which happens to be the unknown class in the problem defined here, is an implicit assumption in the minimum volume criterion, but the criterion can be generalized by weighing the observation space according to the distribution of noise free signal of the second class. However, it minimizes the region instead of the error probability since the former is strongly linked with the latter. Since the noise free signal of the unknown class is uniformly distributed, it is a reasonable approach not to weight any region more heavily than others. In other words, the criterion should weight equally every point in the space. This leads to the insensitivity of the proposed classifier to the distribution of the noise free signal of the second class.

To illustrate the above behavior, an example is given below. Suppose there are two classes, Class I and Class II. There is one corresponding noise free signal for each class in a one-dimensional observation space. The noise free signal of Class I

is known to be zero and that of Class II is unknown. The noise added to each class is Gaussian with zero mean and variance of one. When the error probability of identifying the known class (Class I) as the unknown class is fixed to be 0.05, according to the proposed criterion the region associated with the known class Z_N is [-1.645,1.645] and the probability of misclassifying the unknown class as the known class is

$$P_e = \text{erfc}(1.645-x) + \text{erfc}(1.645+x) \tag{95}$$

where x is the noise free signal of the unknown class and the function erfc(x) is defined in Equation (21).

The above is, of course, a fictitious one because we do not have any information about x, much less the error probability associated with it. Its result is plotted in curve A in Figure 23. The probability of error decreases as the absolute value of x increases. This is quite reasonable since the classifier better distinguishes the two classes as their responses are further separated.

If we use the Neyman-Pearson classifier to distinguish the two classes and the probability of a type one error is again fixed to be 0.05, the threshold is found to be either 1.96 or -1.96, depending on the position of x. Since we do not know x, the value 1.96 is arbitrarily chosen as the threshold (as if x were positive) and the probability of misclassifying the unknown class to the known class is shown in Figure 23 as curves B_2 (when x is negative) and B_1 (when x is positive). When x is positive,

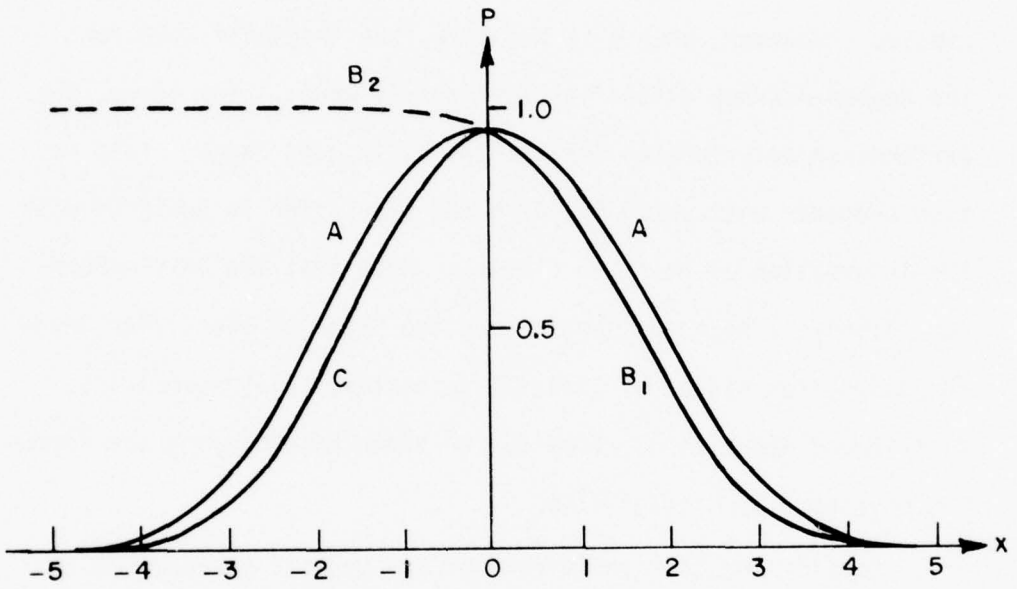


Figure 23. The fictitious performances of three different classifiers.
A. The proposed classifier without any knowledge about the second class.
B_1-B_2. The Neyman-Pearson classifier without any knowledge about the second class.
C-B_1. The Neyman-Pearson classifier with the knowledge of both two classes.

the Neyman-Pearson classifier here yields better performance than the proposed classifier. This is so because in this range, the Neyman-Pearson classifier fully utilizes the available information (both x and zero, response of the first class) while the proposed classifier only uses the information of one of the two classes. However, when x is negative, the threshold 1.96 for the Neyman-Pearson classifier is actually wrong. Therefore, the performance deteriorates drastically as -x gets large. This is also compared with curve C, where the classifier is designed with the information of both the classes. Note that the Neyman-Pearson classifier becomes identical to the proposed classifier when the noise free signal of Class II is statistically symmetrical distributed about zero. This can be shown by employing the formulas developed in Chapter II-D.

Considering the sensitivity of the overall performance to the position of x, it is apparent that the proposed classifier is superior to that of the Neyman-Pearson. In case that x is known, the latter yields the best performance. Nevertheless, the proposed cliassifier does yield a comparable performance. The performance is insensitive to x, the response of the unknown class.

D. Approximation and Simulation

So far, we have been trying to find a region Z_N, specified by the proposed criterion, to provide decision surfaces. A close look at the theorem of Chapter III-B shows that if Z_N is unique, it is not necessary to find Z_N in order to make a classification.

Whenever we want to identify an observed object, all we need is to obtain a ξ_T corresponding to the specified α. The measurement $\bar{x}$ is then substituted into $\xi=g(x)$ and compared with ξ_T. If ξ is larger or equal to ξ_T, the test vector $\bar{x}$ is identified as being in Z_N and otherwise, in Z_X.

This process demonstrates that the implementation of the classifier can be very simple (Figure 24) no matter how complicated Z_N is. Instead of finding the required region Z_N, the computation of ξ_T becomes the main task in designing the classifier for identifying unknown objects. Therefore, it is obvious that the complexity of the technique is not in implementing the classifier but in computing the error probability and the threshold ξ_T.

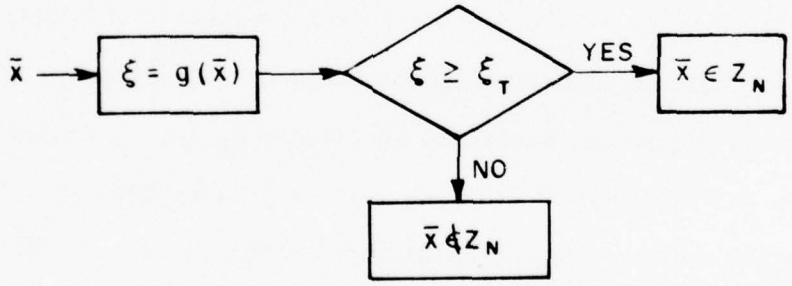


Figure 24. A simple way of implementation makes the classification as a threshold test.

The few examples worked out in the last few sections illustrate, however, that analytically computing ξ_T is extremely difficult for most cases. Two algorithms are devised here to alleviate this difficulty.

The first one was sketched out in Chapter III-B right after the two examples. When the a priori probabilities for all the known subclasses are equal, i.e., the occurrence of each class is equally likely, the subregion associated with each subclass is approximated by a hypersphere centered at the noise free point of each subclass in a multi-dimensional observation space. The approximation also implicitly assumes that the mean of noise added to each individual noise free signal is zero so that the centers of hyperspheres are coincident with the noise free points of the subclasses. This is not a necessary constraint in applying this kind of approximation since we can shift the centers of these hyperspheres by the means of the contaminating noise and approximate the region Z_N by the hyperspheres centered at these new points.

When the distances between any two of the noise free signals are large in terms of the noise standard deviation, this approximation is close to the exact Z_N specified by the criterion. This is shown in Figure 25, where two subclasses S_1 and S_2 are equally probable and embedded in Gaussian noise with zero mean in a two-dimensional space. The approximation ignores the mutual influence among the known objects and hence simplifies the construction of Z_N rather significantly.

Figure 25. The optimum region Z_N (dotted lines) and the two approximated circles (solid lines).

This can also be used as another independent way of identifying unknown objects. The optimum region Z_N is formed by constructing hyperspheres around each signal point and any noisy response falling into Z_N can then be considered as one of the listed objects.

The approximation is especially good when it applies to a higher dimensional space. This is so because the influence from each noise free point S_i is smaller in a higher dimensional space. A hyperspherical shell of a specified thickness at a fixed distance from the noise free point in a higher dimensional space gets less probability contribution as compared with that in a lower dimensional space, when both of them are in the same noise environment. For instance, suppose the contaminating noise in every dimension is Gaussian and independent of one another with variance σ^2 and zero mean, the probability contribution to a region bounded by radii 0.1 and 0.2 from the origin is 0.07886 in a one dimensional space, 0.01481 in a two dimensional space, 0.00183 in a three dimensional space and so on. The contribution to any spherical volume surrounding a noise free point lying in this region ($.1<r\leq.2$) is further reduced by the higher dimensionality since the ratio of such a spherical volume to the spherical shell ($.1<r<.2$) decreases as the dimensionality increases.

With the assumption of additive noise to each noise free signal $\overline{S}_i$, the radius of each hypersphere is assumed to be the same and is computed for each designated α. The result, in turn, can be used to obtain the threshold ξ_T. Some examples employing

this scheme in the detection of unknown aircraft targets are worked out in Chapter III-E.

Diagrammatically, the scheme is easy to implement and gives good approximation at a high SNR system. The computation of the Type I error probability, an integration over the joint regions of several hyperspheres, subject to a specified radius r is difficult, however, in a high-dimensional space. A Monte Carlo simulation technique is employed to circumvent this later.

The second scheme is to employ a Monte Carlo simulation directly to the computation of ξ_T. First, a train of random vectors $\bar{n}_1, \bar{n}_2, \cdots \bar{n}_N$ are generated according to the distribution of the noise added to each signal. Each random vector is added to its noise free signal $\bar{S}_i$ to form a test vector

$$\bar{x}_k = \bar{S}_i + \bar{n}_k \quad . \tag{96}$$

The test vector $\bar{x}_k$ is substituted into $\xi = g(\bar{x})$ to obtain a scalar ξ_k, which is then stored into the data bank D_ξ. The number of random vectors added to each noise free signal $\bar{S}_i$ is proportional to its corresponding a priori probability $P(\bar{S}_i|N)$. After all the transformed ξ_k's are stored into D_ξ, ξ_k's are then lined up in the order of magnitude and we designate the $[\alpha \cdot m]$th smallest ξ to be ξ_T, where m is the total number of ξ_k's in D and $[\_\_\_]$ is the symbol for the largest integer less than or equal to its argument.

It is clear that ξ_T splits the ξ_k's into two groups, those less than ξ_T and those that are not. Corresponding to this, the ξ space is divided into two sections $Z_X = [0, \xi_T)$ and $Z_N = [\xi_T, \xi_m]$.

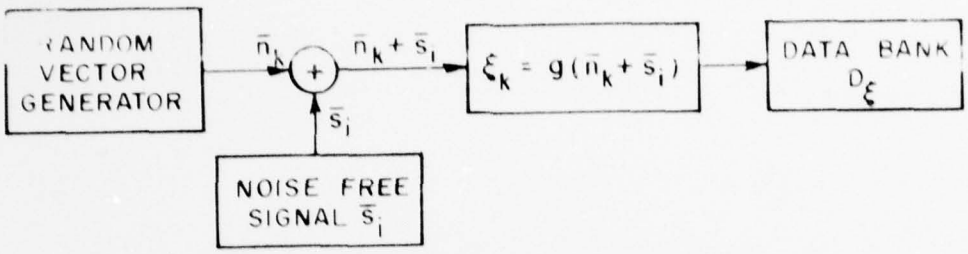


Figure 26. The generation of a data bank D_ξ.

The probability of each ξ_k falling into Z_X is α and that falling into Z_N is $\gamma=1-\alpha$. The probability of exactly ℓ points falling into Z_X in m trials is therefore given by

$$P(\ell) = \binom{m}{\ell} \alpha^{\ell} \gamma^{m-\ell} \tag{97}$$

The expected number of $\bar{x}_k$ falling into Z_X is

$$\langle \ell \rangle = m\alpha \quad , \tag{98}$$

and the variance of ℓ is

$$\sigma_\ell^2 = m \alpha \gamma \quad , \tag{99}$$

or

$$\sigma_\ell = \sqrt{m\alpha\gamma} \quad . \tag{100}$$

The ratio of the spread or standard deviation to the mean is

$$\frac{\sigma_\ell}{\langle \ell \rangle} = \frac{\sqrt{m\alpha\gamma}}{m\alpha} = \sqrt{\frac{\gamma}{m\alpha}} \tag{101}$$

which goes to zero as m goes to infinity, showing that the concentration around the mean increases with m.

We may then use the sample mean as an estimate of α, denoting it

$$\hat{\alpha} = \frac{\ell}{m} \tag{102}$$

and obtain

$$\langle\hat{\alpha}\rangle = \frac{\langle\ell\rangle}{m} = \alpha \tag{103}$$

indicating that $\hat{\alpha}$ is unbiased.

The variance of the estimate

$$\sigma^2_{\hat{\alpha}} = \langle(\hat{\alpha}-\alpha)^2\rangle = \frac{\alpha\gamma}{m} \quad . \tag{104}$$

The relative spread is

$$\frac{\sigma_{\hat{\alpha}}}{\langle\hat{\alpha}\rangle} = \sqrt{\frac{\gamma}{m\alpha}} \quad . \tag{105}$$

This represents the error spread of $\hat{\alpha}$ as a function of α. Therefore, if we like to have the estimate to be in error within some specific range, say ε, we can just have the number of trials larger than or equal to

$$m = \frac{1}{\varepsilon^2}\,\frac{\gamma}{\alpha} \quad . \tag{106}$$

For instance if ε=10%, α=0.05, then γ=0.95 and we obtain from Equation (106) that m should be 1900. Any estimate using more than 1900 trials will yield better accuracy. Even for α=0.01, it only requires 9900 random vectors to have a 90% accuracy, which makes the implementation of this simulation feasible in terms of computer time.

The value ε_T can be used to obtain Z_N. Yet, as indicated before, the criterion does not require the system to construct Z_N

in order to build the classifier. Instead, utilizing ξ_T as a threshold is good enough for the purpose of classification.

The proposed criterion classifies the object into one of the two catagories, namely, the one consisting of the listed class and the other consisting of everything unlisted. The region Z_N associated with the former constructed by the criterion actually provides a way of clustering different listed subclasses. This is especially useful in a multi-feature space where the criterion yields a way of getting those most likely together. By decreasing the value of α, we expand the region Z_N and hence have some subregions associated with the known subclasses merged into a new cluster, reducing the number of subregions. This is indeed a generalized single linkage hierarchical clustering[4] except that it operates in a multi-dimensional space. Figure 27 shows an example of forming clustering regions when employing four points embedded in Gaussian noise. Each point is from one of the four subclasses. The classes C_3 and C_4 are grouped to form a new region associated with both of them when $\alpha=a_1$. The classes C_1 and C_2 are then grouped into another new one as α decreases to another value. All the subregions are eventually grouped into one single region as α decreases to a specific value α_g, depending on the distribution of $\bar{S}_i$'s. This illustrates a method of clustering in terms of "influence" of each individual subclass.

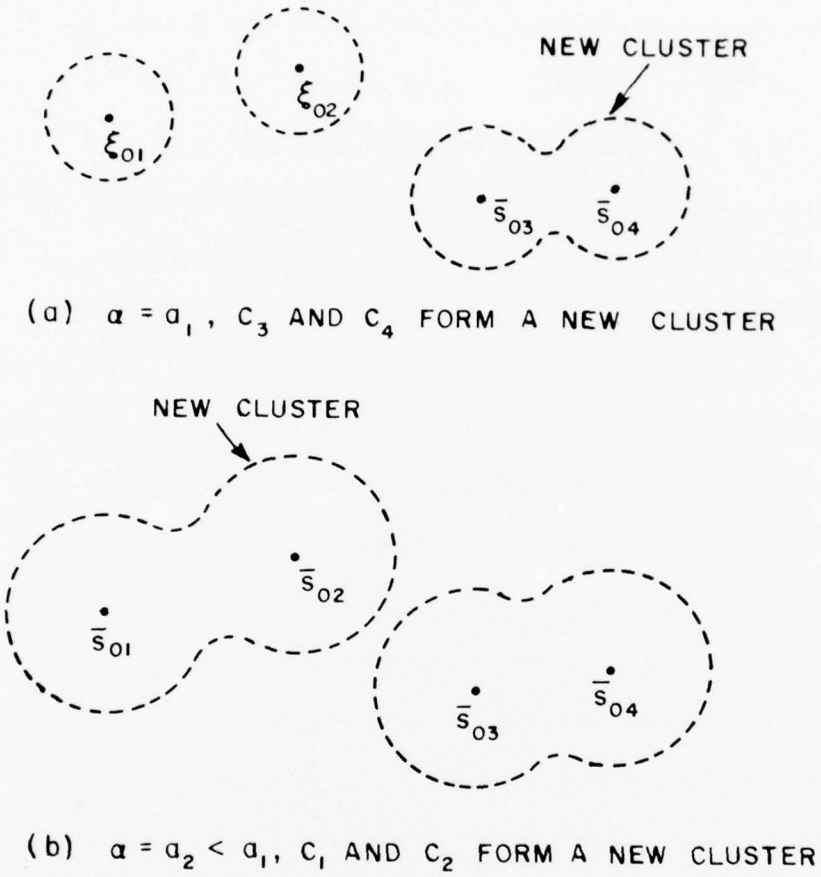


Figure 27. When α decreases, some of the subclasses merged to form new cluster(s).

E. Identifying Unlisted Aircraft - An Application

The above simulation technique is applied to an aircraft identification problem. It has been shown[6,7] that a low frequency method can effectively identify a large variety of airplanes. The features used are the electromagnetic scattering returns from the observed objects. Signal amplitude as well as phase information and the orthogonal polarizations of each scattering return are used to do the identification. An additive Gaussian noise with zero mean is assumed to be added to each noise free signal. A set of four aircraft (F104, F4, MIG19 and MIG21) is chosen to form the listed class. The a priori probabilities are assumed to be 25% for each class of aircraft. The scattering data of all the aircraft were computed by the moment method[14] at Ohio State University ElectroScience Laboratory and the features (frequencies) are chosen to optimize a nearest neighbor classification of the four aircraft. Based on the above assumptions, the probability density function when a known object is present is

$$g(\bar{x}) = \frac{1}{4} \sum_{i=1}^{4} \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{|\bar{x}-\bar{S}_1|^2}{2\sigma^2}} \qquad (107)$$

where

$\bar{x}$ is the observed signal vector,

$\bar{S}_i$ is the noise free signal vector of the ith aircraft,

σ is the noise standard deviation, and

n is the number of features used.

The aircraft is assumed to be facing the observer nose on. We first apply the hypersphere approximation to this problem. The region Z_N is constructed by the hyperspheres centered at each signal point with radius r. The object is identified as one of the listed objects if the observed vector lies in Z_N. From the previous argument, the probability contribution from Z_N should total $1-\alpha$. For α=5%, the total probability contribution over the region Z_N is therefore 0.95. The Monte Carlo simulation is used to find the radius for three cases using different numbers of features and the results are shown in Table 2.

TABLE 2
The Approximated Radii Obtained by the Monte Carlo Simulation for Four Aircraft Data, α=0.05, σ^2=1

Dimension	Radius for α=0.05	Radius for α=0.05 when there is only one object in the known class
2	2.136	2.42
4	2.864	3.08
8	3.813	3.94

Forty-thousand tested vectors were generated each time to assure the accuracy of the simulation. The first row in the table employs both components of a complex signal (the amplitude and phase of a scattering return constitute a complex signal) for a single frequency, horizontally polarized wave. The second row uses both vertical and horizontal polarizations of the same frequency signal. The third row uses both polarizations of two frequency returns simultaneously.

The second column in the table lists the data obtained by the simulation. The last column lists the radii for the cases when there is only one element in the known class. This happens to be the same as the distance for the x^2 distribution with the cumulative probability being 95%. It is obvious from the table that the deviation of the radius from that of the single element case decreases as the dimensionality goes up (13% in the two-dimensional case and only 3% in the eight-dimensional case). This confirms our argument before that the influence of the subclasses on one another decreases as the dimensionality increases. At the same time, it also demonstrates that the single element radius is an increasingly good approximation as dimensionality increases.

The implementation of the above is also very simple. A test vector $\bar{x}$ whose minimum distance to any of the four signal points is less than or equal to the obtained radius can be considered from the known class (one of F104, MIG19, F4 and MIG21) and vice versa.

The second scheme is to obtain the threshold for a prefixed α directly from the Monte-Carlo simulation. This was carried out over the similar cases and the results are listed in Table 3. Again, 10,000 test vectors were generated in each simulation. The accompanying CPU time in the table is the computer time for each simulation needed for the Datacraft 6400 at The Ohio State University ElectroScience Laboratory, which is about three times slower than an IBM 370/165. For the most complicated case here, only 604.39 ms is needed to finish the simulation. The accuracy, from Equation (105), is in the range of 4.36% of the correct value for α being 5%. This illustrates that the method is effective and efficient.

An interesting experiment was carried out by introducing four other objects -- MIG25, SR71, B1 and F14, chosen to represent a wide range of different shapes and sizes of aircraft. The probabilities of these new objects being classified to the listed class by the proposed classifier (described in Figure 24) were computed and the results were tabulated in Tables 4 and 5 for a two-dimensional case. Also computed in the tables were the probabilities of classifying the listed objects into the unlisted class when the test vector originated from the known objects. Both tables list the probability of classifying each object out of the eight into the unlisted class. In the process, the test vector is substituted into Equation (107) and the resultant scalar is compared with the threshold ξ_T, obtained by presetting α to be sequentially 0.05 and 0.10. Ten thousand test vectors were

TABLE 3
The Threshold ξ_T Obtained by the Monte-Carlo Simulation for α=0.05. σ is assumed to be 0.1 of the Average Signal of the Responses of the Listed Objects

Case	Dimension	ξ_T	CPU(ms)	Noise Standard Deviation σ	Feature(s) Used
ξ_T ①	2	0.01063026	46.93	0.436	Horizontal Pol., 1 freq.
ξ_T ②	2	0.00919883	45.46	0.463	Vertical Pol., 1 freq.
ξ_T ③	4	2.6652×10^{-3}	75.20	0.377	Vertical Pol., 2 freq.
ξ_T ④	4	0.0965×10^{-4}	76.29	0.494	Horizongal Pol., 2 freq.
ξ_T ⑤	4	1.3180×10^{-3}	80.76	0.450	Both Pol., 1 freq.
ξ_T ⑥	8	5.809×10^{-4}	149.84	0.439	Both Pol., 2 freq.

TABLE 4

Probability of Classifying the Object into the Unlisted Class by Using the Proposed Scheme when the Set of Four Aircraft -- F104, MIG19, F4, MIG21 -- is considered as the Listed Class. The Corrupting Noise is Gaussian with Zero Mean and σ=0.1 of the Average Signal of the Responses of the Listed Objects

ξ_T	F104	MIG19	F4	MIG21	MIG25	SR71	B1	F14
0.01063 (α=0.05)	0.0496	0.0479	0.0471	0.0508	1.0	1.0	1.0	1.0
0.02179 (α=0.10)	0.1038	0.1044	0.1024	0.1045	1.0	1.0	1.0	1.0

TABLE 5
Probability of Classifying the Object into the Unlisted Class by Using the Proposed Scheme when the Set of Four Aircraft -- F104, MIG19, F4, MIG21 -- is Considered as the Listed Class. The Corrupting Noise is Gaussian with Zero Mean and σ=0.2 of the Average Signal of the Responses of the Listed Objects

ξ_T	F104	MIG19	F4	MIG21	MIG25	SR71	B1	F14
0.003148 (α=0.05)	0.0429	0.0541	0.0382	0.0538	1.0	0.9531	0.9996	1.0
0.00627 (α=0.10)	0.0818	0.1195	0.0830	0.1178	1.0	0.9785	1.0	1.0

generated for each case and the noise added to the objects was again assumed to be Gaussian with zero mean and σ set to be equal to 10 per cent of the average signal of the responses of the four listed objects. For the first four (listed) aircraft, the probabilities of classifying them into the unlisted class are close to 0.05 in the first rows and to 0.1 in the second rows of the table because we prefixed α to be 0.05 and 0.1 respectively. The probability of classifying an unlisted object into the listed class is zero for each of the unlisted objects in Table 4, indicating that the classifier is an excellent discriminator in this case. Note that the thresholds ξ_T change almost linearly as α changes from 0.05 to 0.10, indicating that the instability discussed in Chapter III-C does not occur in this case, although the thresholds ξ_T are small. This is a reasonable result since the rate of change of a Gaussian distribution (the function $g(\overline{x})$ in this example) is always proportional to the value of the function at the point considered. This eliminates the flatness of $g(\overline{x})$ over any regions in the observational space.

When σ increases to twenty percent of the average signal, the same conclusion can also be drawn on the performances (Table 5) except the probabilities for SR71 and B1 being classified to the listed class are not zero. This happens because (1) the responses of SR71 and B1 are closer to those of the listed objects (Table 6), and (2) Z_N covers a larger area in the observation space when σ is larger. Also when α gets larger (i.e., the probability of misclassifying a listed object into the unlisted class gets larger),

the region Z_N associated with the listed class shrinks, making the probability of misclassifying an unlisted object into the listed class smaller. This is seen by comparing the first row and the second row in Table 5.

Table 6

The Noise Free Responses of the Eight Aircraft at the Considered Frequency (24 MHz) at Nose-on Aspect

	F104	MIG19	F4	MIG21	MIG25	SR71	B1	F14
$\overline{S}$ =	-5.130	-0.526	-2.904	-5.076	2.687	-3.723	4.426	17.864
	3.244,	-5.908,	4.258,	-5.283,	-20.057,	2.828,	-9.145,	0.363

Incidentally, the probability of classifying no object (null class) to the listed class was also computed and the results are zero for all the cases considered here.

When the dimensionality increases to four and higher, the classifier performs even better. The probability of misclassifying any of the listed objects into the unlisted class becomes 0.05 for all of the four listed objects and that of identifying an unlisted object as the listed class is zero for all of the four unlisted objects. This demonstrates that the proposed scheme is indeed a very effective one even when applying to a quite noisy environment.

CHAPTER IV
COMPLETE CLASSIFICATION

A. Introduction

As described in the first chapter, there are two steps in a "complete" classification procedure. One is to decide whether the object to be identified is in the list of the known objects. If it is one of the catalogued objects, in the next step a conventional scheme is then employed to do the classification. If not, the object is designated to be a new object and a learning process is employed to estimate its characteristics.

This chapter attempts to show how this complete classification procedure can be conducted. The influence of the preclassification on the final classification and the strategies to be used are investigated. Some related problems are also discussed.

B. The Effect of Preclassification

The technique developed in the previous two chapters involved the separation of the uncatalogued class from the catalogued class. For the convenience of the following discussion, this step will be called "preclassification" and the step of classifying the observed object as one of the listed objects after the preclassification, "final classification", or just "classification". The preclassification approach minimizes the region of

AD-A064 843 OHIO STATE UNIV COLUMBUS ELECTROSCIENCE LAB F/G 12/2
IDENTIFICATION OF CATALOGUED AND UNCATALOGUED CLASSES.(U)
DEC 78 H LIN AFOSR-74-2611
UNCLASSIFIED ESL-783815-7 AFOSR-TR-79-0055 NL

2 OF 2
AD
A064 843

END
DATE
FILMED
4-79
DDC

the catalogued class in the observation space, while keeping the probability of misclassifying the catalogued class to a fixed value. This tends to maximize the probability of correctly classifying an uncatalogued object while ensuring that the probability of misclassifying a catalogued target is below a prespecified level. We have shown that this region is constructed in such a way that no information about the uncatalogued class is needed. This enables the scheme to be useful in a practical situation where usually no information about the uncatalogued class is available.

In the final classification, only the response that falls into Z_N, the region associated with the listed class defined in the last chapter, is used. Most of the observation space will thus be excluded. This will have some influence on the final classification after the observed object is determined to be in the listed class.

In Chapter III-A, we have shown that Z_N excludes the region where the probability density of the observed vector $\bar{x}$ is small, compared with its density in Z_N. This indicates that the preclassification procedure excludes only the portion where the measured vectors are least likely to occur in the feature space. If α is small enough, the region Z_N should contain all the regions which are significant in the final classification process. In this case, the preclassification has very little effect on the final classification. As α increases, Z_N shrinks and the impact of the preclassification becomes noticeable. However, the regions

corresponding to the lowest probability density are usually far away from the uncorrupted signals of the listed objects, where the classification is more likely to make an error. Consequently, the exclusion of these regions is usually not detrimental to the overall performance of the complete classification process. Yet, the total impact of preclassification depends on the data distribution of the listed objects as well as the misclassification probability α.

To demonstrate the influence of preclassification, an example, using Bayes approach as a means of final classification, is shown here. Consider the example given in Figure 20. Let the distance between the two objects be 2d, σ be the standard deviation of the noise added to each signal, and the probability of misclassifying any of the listed objects (S_1 or S_2 in Figure 20) as unlisted be α. If we use Bayes classifier to do the classification of the two objects directly (without going through the preclassification), the average probability of misclassification is shown by the curve designated as α=0 in Figure 28. If we first apply the preclassification process and use Bayes classifier to do the classification after determining the object to be in the listed class, the results are shown by the rest of the curves in Figure 28, for various values of α. As mentioned above, at the final classification the feature space is shrunk to Z_N since only the response that falls into Z_N is used for the Bayes test. Therefore the probability of misclassification for the final classification is a conditional probability obtained from dividing the

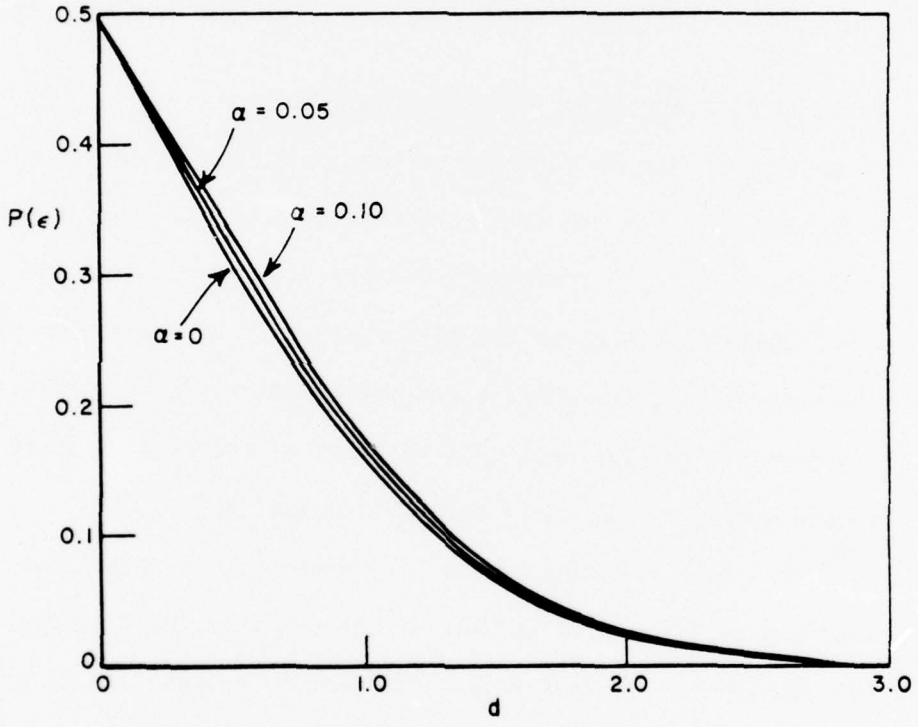


Figure 28. The error probability of Bayes classification, carried out after the preclassification, as a function of the distance between the two classes.

misclassification probability by the probability of the response falling into Z_N. When α increases the average probability of misclassification increases, but not significantly. All the curves for different α's converge to 0.5 as d approaches zero. They all go down to zero as d goes up to infinity, which would be expected since the classification error probability approaches zero when the two noise free signals are separated by an infinite distance.

C. Identification Among the Known Objects

We demonstrate here, with a practical aircraft identification problem, how the two step classification works. As shown in Chapter III-E, the electromagnetic scattering returns from the observed object are used as the test features. The selection of the features (e.g., *frequencies and polarizations*)[6,7] *should be* solely dependent on the data distributions of the catalogued objects since they are the only information available.

In the case that the probability distributions are known the optimum decision rule is that of Bayes, where the misclassification probability is minimized. The average probability of error will be

$$P_e = 1 - \int_Z \max_K (P_K P(\bar{x}|C_K))d\bar{x} \qquad (108)$$

where

C_K denotes the kth object in the catalogued class,

P_K is the a proiri probability of C_K,

$P(\bar{x}|C_K)$ is the probability distribution of the test vector $\bar{x}$ when C_K is present, and

Z is the feature space defined before.

The integration is carried over the whole feature space by picking the maximum of $p_K P(\bar{x}|C_K)$ among all K such that the error probability is minimized.

In many cases it may be difficult to evaluate P_e analytically. This is so even if the distribution $P(\bar{x}|C_K)$ is known exactly since it is difficult to analytically describe the region which gives the maximum value. Also the computation of P_e becomes extremely complicated when the number of objects is large, therefore, a nearest neighbor (N.N.) method is used below.

An N.N. rule can be described as follows. Given training samples $\{\bar{S}_i^K\}$ for each object C_K, the rule is to classify the tested point $\bar{x}$ as a member of C_r to which its nearest neighbor belongs, i.e.,

$$\bar{x} \epsilon C_r, \text{ if } ||\bar{x}-\bar{S}_j^r|| = \min_{i,k} ||\bar{x}-\bar{S}_i^K|| \tag{109}$$

where C_r is one of the known classes and we use the notation for Euclidean distance,

$$||x-S_i^K|| = [\bar{x}-\bar{S}_i^K]^T[\bar{x}-\bar{S}_i^K]^{1/2} \quad . \tag{110}$$

An analytical calculation of probability of error for the nearest neighbor classification is also very involved because it would require the integration of a multivariate density function

over extremely complicated boundaries. For this reason, the errors are obtained by Monte Carlo simulation.

The N.N. classifier has a practical advantage over Bayes classifier in that the difficulties of determining the boundaries of the integration are eliminated, and of course, it is nonparametric and consequently is perfectly applicable to classification problems where the statistics of the noise and signals are not known. Cover and Hart[15] have shown that in the large sample case, the probability of error for the N.N. rule is bounded above by twice the Bayes probability of error, and is clearly bounded below by the Bayes error. For most practical situations these bounds are sufficiently tight to indicate the merit of the N.N. rule. Therefore, in the following examples, the N.N. rule is employed as the final classification to compute the overall performance of the classifier. In the aircraft identification problem, the training samples are just the noise free responses from the aircraft to be identified and the test samples are these responses to which a Gaussian noise is added. As stated in Chapter III-E, the a priori probabilities of the aircraft subclasses in the catalogued class are assumed to be all the same. Therefore, the probability function of the test vector $\bar{x}$ when N is true is

$$g(\bar{x}) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{|\bar{x}-\bar{S}_i|^2}{2\sigma^2}} \qquad (111)$$

where M is the number of objects in the catalogued class and $\bar{x}$, $\bar{S}_i$, σ, n are all defined as those in Equation (106).

D. Two Step Classification

Let the listed class consist of five American airplanes, F104, F4, SR71, B1 and F14. The data of the electromagnetic scattering returns were, as mentioned before, numerically computed at Ohio State University ElectroScience Laboratory[14]. The optimum frequencies were selected by using the N.N. rule to do the classification. The optimum frequencies were found to be 24 MHz for single frequency features, and 20 MHz and 24 MHz for two frequency features.

A preclassification process is first employed to test whether an observed object is listed. To do this, we obtain the probability density function of an observed vector $\bar{x}$ when a catalogued object is present. *This is* $g(\bar{x})$ *in Equation (111), where* $g(\bar{x})$ *is shown to be a function of all scattering returns as well as that of the noise. The thresholds* ξ_T*'s are therefore computed in terms of the noise standard deviation* σ *and listed in Table 7 for* α=0.05.

The numbers shown in the table under the "case" column are explained in Table 3 in Chapter III-E. For instance, (1) represents the system utilizing one frequency, horizontally polarized electromagnetic returns and so on. $\bar{A}$ represents the average amplitude of all the listed objects.

When α is greater than $1/\sqrt{2\pi}$, which is what we have in all the cases here, it is seen that as σ increases, ξ decreases in the table. This is because $g(\bar{x})$ in Equation (111) goes down as

TABLE 7
The Thresholds Computed for α=0.05 at Three Different Noise Levels. The Listed Class Consists of Five American Aircraft: F4, F104, SR71, B1 and F14

Case	Dimensions	σ=0.1 $x\bar{A}$	0.2 $x\bar{A}$	0.3 $x\bar{A}$
ξ_T ①	2	6.3054×10^{-3}	2.1445×10^{-3}	1.3814×10^{-3}
ξ_T ②	2	5.0052×10^{-3}	1.3613×10^{-3}	6.3608×10^{-4}
ξ_T ⑤	4	5.0133×10^{-4}	3.133×10^{-5}	6.22×10^{-6}
ξ_T ⑥	8	5.6813×10^{-6}	2.2192×10^{-8}	8.6591×10^{-10}

σ becomes larger, reducing the value of amplitude of the probability density in the vicinity of its maximum. This is also true when the number of features n increases in Equation (111).

Once the ξ_T's are decided, the experiment described in Figure 24 in Chapter III-D can be used to test the classification of these five American made airplanes. Three other, foreign made, aircraft (MIG19, MIG21 and MIG25) are added to the preclassification test and the results are shown in Table 8 for σ=0.1 $x\bar{A}$ for case ①.

The probabilities of misclassifications are computed by Monte Carlo simulation and the result for each aircraft is listed. The average probability of preclassification error for the five listed aircraft is 0.0477, which is close to the theoretical value 0.05 (since we set α=0.05). The difference is caused by computational error and is so small that it can be considered in

TABLE 8

Probabilities of Misclassification for σ=0.1 $x\bar{A}$ When Using Two Features (Horizontally Polarized Returns). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0442	0.0418	0.0485	0.0508	0.0533	0.0	0.0	0.0
	Average = 0.0477					Average = 0.0		
Error Probability of Classification	0.0074	0.0073	0.0	0.0	0.0			

TABLE 9

Probabilities of Misclassification for σ=0.2 $x\bar{A}$ When Using Two Features (Horizontally Polarized Returns). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0256	0.0334	0.0472	0.0601	0.0732	0.0	0.0	0.0
	Average = 0.0479					Average = 0.0		
Error Probability of Classification	0.1707	0.1325	0.0518	0.007	0.0001			

agreement with the design value 0.05. The classification by using the N.N. rule, after the object is assigned to be in the listed class, is also computed and the errors are negligible when the noise is ten percent of the average amplitude return.

Table 9 shows the probabilities of misclassification when σ equals twenty percent of the average amplitude returns. The overall performance at the preclassification stage is still very good, however, the classification among the listed objects deteriorates quite significantly. The fact that the listed airplanes can be distinguished from the unlisted ones while they cannot be well identified among themselves stems from the way the two classifications are set up. The preclassification process classifies two major classes, namely, listed and unlisted. The errors committed in the detailed classification among the subclasses of the listed class are not considered as errors in the preclassification process unless the objects are identified as unlisted.

When the noise level increases, the identification of the listed objects worsens further, although it is still relatively easy to distinguish them from the unlisted ones (Table 10).

As discussed in [6,7], one way of improving the overall performance of classification among the known objects is to increase the dimensionality of the feature vector. This was done by employing both vertically and horizontally polarized radar returns. The results are shown in Tables 11-13 for σ=0.1 to 0.3 of the average return of the listed aircraft.

TABLE 10
Probabilities of Misclassification at σ=0.3 $x\bar{A}$ When Using Two Features (Horizontal Polarization). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0172	0.0156	0.0617	0.0712	0.0774	0.0006	0.0005	0.0038
	Average = 0.0486					Average = 0.0016		
Error Probability of Classification	0.362	0.2064	0.1423	0.0620	0.338			

TABLE 11
Probabilities of Misclassification at σ=0.1 $x\bar{A}$ When Using Four Features (Two Polarizations). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.049	0.0407	0.0454	0.042	0.0452	0	0	0
	Average = 0.0444					Average = 0.0		
Error Probability of Classification	0	0	0	0	0			

TABLE 12
Probabilities of Misclassification at σ=0.2 $x\bar{A}$ When Using Four Features (Two Polarizations). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0489	0.0407	0.0453	0.0420	0.0452	0	0	0
	Average = 0.0444					Average = 0		
Error Probability of Classification	0	0	0	0	0			

TABLE 13
Probabilities of Misclassification at σ=0.3 $x\bar{A}$ When Using Four Features (Two Polarizations). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0490	0.0409	0.0455	0.0424	0.0453	0.007	0.0005	0.0024
	Average = 0.0446					Average = 0.0015		
Error Probability of Classification	0	0	0	0	0			

At the preclassification stage, the overall performance is still similar to the previous results. Yet, identification among the listed aircraft improves drastically when more features are used. This is in agreement with the identification results of an N.N. classifier without going through the preclassification step, where the more features are used, the better a classifier preforms. Note that the increase of the dimensionality also improves the separability between the unlisted and listed aircraft, although not significantly in this example.

When one increases the dimensionality to eight, i.e., utilizing two frequency returns simultaneously, the overall performance is further improved. The separation between the listed and unlisted classes (for the three aircraft added here) becomes very large. The error probabilities of classifying the unlisted aircraft as the listed ones approach zero even when σ increases to thirty percent of the average return (Tables 14-16).

E. Effect of the Type One Error Probability α in the Preclassification Process

The threshold ξ_T is constructed by presetting type one error probability α to a fixed value, therefore the change of α has some influence on the overall performance. Some implications of this was discussed in Section C when a Bayes classifier was employed to do the second step classifications. In this section, we demonstrate the effect of changing α in using an N.N. classifier to do the classification of the catalogued objects.

TABLE 14
Probabilities of Misclassification at σ=0.1 $x\bar{A}$ When Using Eight Features (Two Frequencies, Two Polarizations). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.049	0.0522	0.0474	0.0474	0.0479	0	0	0
	Average = 0.0488					Average = 0		
Error Probability of Classification	0	0	0	0	0			

TABLE 15
Probabilities of Misclassification at σ=0.2 $x\bar{A}$ When Using Eight Features (Two Frequencies, Two Polarizations). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0489	0.0522	0.0474	0.0473	0.0479	0	0	0
	Average = 0.0487					Average = 0		
Error Probability of Classification	0	0	0	0	0			

TABLE 16
Probabilities of Misclassification at σ=0.3 $x\bar{A}$ When Using Eight Features (Two Frequencies, Two Polarizations). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0489	0.0522	0.0474	0.0473	0.0479	0	0	0
	Average = 0.0487					Average = 0		
Error Probability of Classification	0	0	0	0	0			

Since for α=0.05, the other misclassification probability is zero. There is an unbalance between the two types of error and the overall error is larger than need be. Reducing α will thus tend to minimize the overall error. Its value is decreased to 0.01 to test the performance.

Again, the thresholds for the preclassification are computed first. The five American airplanes are included in the listed class. Since the listed aircraft are still the same, the optimum features selected before are unchanged too. The thresholds for five different cases at three noise levels are computed and listed in Table 17. Since α is five times smaller than before, according to Equation (106), the number of tested vectors required to yield the same computational accuracy should increase almost five fold, implying that if the same number of random vectors are used in the computation of ξ's, the computation error will increase. Still, if one uses 10,000 random vector, the deviation of the classification error will be kept to less than 10% of the actual value.

Again, when only using the data of the horizontally polarized wave at the frequency of 24 MHz, the classifier performs quite satisfactorily even when α is decreased to 0.01 and σ equals ten percent of the average response of the listed aircraft. The average error probability of the preclassification becomes 0.00804, which is close to the specified value 0.01 (Table 18). The misclassification among the listed aircraft at the second step classification is still kept as low as it is at

TABLE 17
The Thresholds Computed for α=0.01 at Three Different Noise Levels. The Listed Class Consists of Five American Aircraft: F4, F104, SR71, B1 and F14

Case	Dimensions	σ=0.1 $x\bar{A}$	0.2 $x\bar{A}$	0.3 $x\bar{A}$
ε_T ①	2	1.2504×10^{-3}	5.3065×10^{-4}	3.153×10^{-4}
ε_T ②	2	9.982×10^{-4}	2.745×10^{-4}	1.510×10^{-4}
ε_T ⑤	4	8.524×10^{-5}	0.533×10^{-5}	0.108×10^{-5}
ε_T ⑥	8	4.382×10^{-7}	1.712×10^{-9}	6.679×10^{-11}

α=0.05. However, since α decreases, the region Z_N in the feature space expands, increasing the tendency of identifying an unlisted object as listed. Although this is not evident in Tables 18 and 19 where the noise levels are low, it shows up in Table 20 where the noise level increases to thirty percent of the average response of the listed aircraft. Of course, this kind of error probability depends on the response of the unlisted objects. The affect of α on the overall performance is: the smaller α is, the more probable it is that the preclassifier will identify an unlisted object as one of the listed objects. Note that the influence of α on the performance of the second step classification is negligibly small. The results in Tables 18-20 do not change very much from those in Tables 9-11. A general conclusion cannot be drawn based on these because the differences are so small. They are probably contributed by the numerical computation error.

TABLE 18
Probabilities of Misclassification at σ=0.1 $x\bar{A}$ When Using Two Features (Horizontal Polarization). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0074	0.0073	0.0087	0.0079	0.0089	0	0	0
	Average = 0.00804					Average = 0		
Error Probability of Classification	0.0088	0.0078	0.0002	00	0.0			

TABLE 19
Probabilities of Misclassification at σ=0.2 $x\bar{A}$ When Using Two Features (Horizontal Polarization). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.005	0.0042	0.0102	0.0099	0.0126	0	0	0
	Average = 0.00838					Average = 0		
Error Probability of Classification	0.1705	0.1333	0.0517	0.0085	0.0015			

TABLE 20
Probabilities of Misclassification at σ=0.3 $x\overline{A}$ When Using Two Features (Horizontal Polarization). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0019	0.0026	0.0122	0.0129	0.0110	0.0055	0.0062	0.0256
	Average = 0.00812					Average = 0.0124		
Error Probability ov Classification	0.3622	0.2975	0.1383	0.0602	0.0340			

TABLE 21
Probabilities of Misclassification at σ=0.1 $x\overline{A}$ When Using Two Features (Vertical Polarization). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.009	0.0082	0.0083	0.007	0.0083	0	0	0
	Average = 0.0082					Average = 0		
Error Probability of Classification	0	0	0	0	0			

Tables 21-23 list some of the results when using the vertically polarized returns at the same frequency. The data distribution of the noise free responses of the listed objects are not the same as that in the previous case. Hence the classification among the known objects changes quite a bit. Nevertheless, the overall performance in separating the listed class from the unlisted class is still very good.

The error probability of identifying the unlisted class as the listed class is on the average less than ten percent even when the standard noise deviation σ is thirty percent of the average response. The error probability for the opposite direction of identification is close to 0.01, in agreement with the specified value.

Finally, we show that, for the special case of using both polarized wave returns at one frequency the results are good enough for the preclassified error even under the constraint of α=.01. The identification among the known objects also performs very satisfactorily when utilizing the N.N. rule to do the second step classification. The results are shown in Tables 24-26.

F. Groupings and Strategies

The scheme developed to do the classification between the listed and unlisted classes can serve not only as a means of preclassification, but also an intermediate step in the complete classification. This is especially beneficial in a classification involving a large number of classes.

TABLE 22
Probabilities of Misclassification at σ=0.2 $x\bar{A}$ When Using Two Features (Vertical Polarization). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0012	0.0026	0.0122	0.0129	0.0110	0.0055	0.0062	0.0259
	Average = 0.008					Average = 0.0125		
Error Probability of Classification	0.3623	0.2975	0.1383	0.0602	0.0397			

TABLE 23
Probabilities of Misclassification at σ=0.3 $x\bar{A}$ When Using Two Features (Vertical Polarization). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.01	0.009	0.009	0.0054	0.0064	.2182	.0633	0.0154
	Average = 0.00796					Average = 0.0990		
Error Probability of Classification	0	0	0.0002	0.0077	0.0070			

TABLE 24

Probabilities of Misclassification at σ=0.1 $x\bar{A}$ When Using Four Features (Vertical and Horizontal Polarizations). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0089	0.0081	0.0094	0.0066	0.0076	0	0	0
	Average ≈ 0.00814					Average = 0		
Error Probability of Classifacation	0	0	0	0	0			

TABLE 25

Probabilities of Misclassification at σ=0.2 $x\bar{A}$ When Using Four Features (Vertical and Horizontal Polarizations). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Proabaility of Preclassification	0.0090	0.0081	0.0094	0.0067	0.0076	0	0	0
	Average = 0.00816					Average = 0		
Error Probability of Classification	0	0	0	0	0			

TABLE 26
Probabilities of Misclassification at σ=0.3 $x\bar{A}$ When Using Four Features (Vertical and Horizontal Polarizations). The Listed Class Consists of Five American Aircraft

AIRCRAFT	F104	F4	SR71	B1	F14	MIG19	MIG21	MIG25
Error Probability of Preclassification	0.0090	0.0083	0.0094	0.0066	0.0080	0.0113	0.0041	0.0107
	Average = 0.00826					Average = 0.0087		
Error Probability of Classification	0.0001	0.0	0.0001	0.0002	0.0			

Considering a situation where an object is to be identified as one of many possible objects in the known list, a conventional method will require the response of this object to be compared with those of known objects in some predetermined way. For instance, in an N.N. classifier, the Euclidean distances of this response to the noise free points of all the listed objects are computed and the minimum one of these computed distances is then used to classify the observed object. The computation and the comparisons of these distances usually consume a lot of time if the number of the listed objects is large. In general, the time for picking a minimum (or maximum) value is proportional to n, the number of items from which the choice is made. A conventional scheme becomes very slow and inefficient due to this. Moreover, the classifier becomes inevitably complicated when the number of the objects is large.

One way of tackling the above problem is to group all the listed objects into several subgroups and use the preclassification scheme developed in this work as an intermediate step of classification. One of the subgroups can be considered as the listed class and the scheme is employed to do the preclassification. Once it is decided that the object is in this subgroup, a conventional scheme can be used to identify it as one of the objects in this subgroup. If not, the next subgroup can be considered as the listed class and the same process is applied until the observed object is identified. Of course, in the whole procedure the objects with higher a priori probabilities should be

subgrouped and used as the first listed class. By doing so, the number of tests would be reduced since we test the most likely occuring objects first.

The procedure would at least yield the following advantages:

(1) The basic classifiers can be greatly simplified since the number of the objects in each identification is small.

(2) The objects with similar responses can be grouped into the same subgroup in the first classification such that the classification error probabilities can be made small. Incidentially, the clustering process described in Chapter III-D can serve as a grouping method.

(3) The classifier identifies each subgroup in sequence, hence the objects considered most important (or with highest a priori probabilities) can be subgrouped and used as the first listed class. In this way, the objects can be grouped according to the importance of the identification of each one, and the classifier carries out the classification in a specified order of priorities.

Note that the preclassifier developed by the proposed criterion only compares the value $g(\bar{x})$ with the designated threshold ξ_T, the process is much simpler than that needed for an ordinary classification of several objects. Therefore, the preclassification is much less time consuming as compared to a conventional classifier.

However, the preclassification process yields some classification error in each step of identification. For successive identifications of several subgroups this process will definitely worsen the overall performance, the extent to which this will happen depends on the prefixed error probability α as well as the data distribution. Since this is also related to the order of the subgroups chosen, an optimum strategy for this system has to be defined. Further study is necessary to determing the stragegy.

CHAPTER V
SUMMARY AND RECOMMENDATIONS

A technique has been developed to discriminate listed objects from unlisted ones. It is based on the principle of minimizing the probability of error in an identification process. Since no information regarding the unlisted objects is available, instead of minimizing the overall probability of misclassification the method prefixes the probability of misclassifying a listed object as an unlisted one and minimizes the region associated with the listed class in the feature space. This minimizes the likelihood of misclassifying unlisted objects as listed ones.

It was proved that the devised classifier could be implemented as a threshold test. The employment of the latter greatly simplifies the design of the classifier. The classifier was applied to an aircraft identification problem. It was shown that the error probability of misclassifying catalogued targets as uncatalogued and vice versa can be made very small, while keeping a high probability of correct identification when the presence of a listed object is detected. The misclassification probability for a specific case of three unlisted objects and five listed ones was computed and was found to approach zero when the number of the features used was as low as four. The additional step of discriminating listed objects from unlisted ones produced very

little degradation of the overall classification performance. The overall misclassification probability for all cases considered was changed less than five percent. The implementation of the developed scheme was shown to be simple and efficient.

The technique does not need to utilize any information about the unlisted class to carry out the classification. However, some a priori knowledge of the unlisted class is sometimes available. For instance, the responses of the unlisted objects might well be confined to a restricted region of the observation space. We can use the technique developed in Chapter II-D to tackle this kind of a problem.

Some information on unlisted objects can be obtained from the observed response. For example, once the observed object is determined to be a new one, a learning process is employed to estimate the characteristic of this new object, which can be used as the a priori information for the classification process. This would lead to a modification of the minimum volume criterion. Since the response of any unlisted object was assumed to be unknown in this study, no further investigation was conducted along this line.

Another problem that occurs frequently is that the parameters of the listed class are not known. This turns out to be a nonparametric classification problem of identifying the unlisted class as distinct from the listed class. No attempt has been made to deal with this problem, but it should be investigated.

A procedure for identifying a subclass of a large number of objects was discussed. The classification utilizing the devised classifier is essentially an elimination process. The subgroup bearing the highest a priori probabilities is tested first and eliminated from the list if it is decided that the observed object does not belong to this subgroup. The detailed classification procedure depends on the cost function assigned to the number of tests needed, the probability of misclassification and the complexity of the classifier[16]. This approach, which was outlined in Chapter IV, requires more thorough study.

BIBLIOGRAPHY

[1] L. Kanal, "Patterns in Pattern Recognition: 1968-1974," IEEE Trans. Inform. Theory, Vol. IT-20, November 1974, pp. 697-722.

[2] Y.C. Ho and A.K. Agrawala, "On Pattern Classification Algorithms - Introduction and Survey," Proc. IEEE, Vol. 56, December 1968, pp. 2101-2114.

[3] G. Nagy, "State of the Art in Pattern Recognition," Proc. IEEE, Vol. 56, May 1968, pp. 836-862.

[4] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, New York, 1973.

[5] J.D. Gibson and J.L. Melsa, *Introduction to Nonparametric* Detection With Applications, Academic Press, New York, 1975.

[6] A.A. Ksienski, Y.T. Lin and L.J. White, "Low-Frequency Approach to Target Identification," Proc. of IEEE, Vol. 63-12, December 1975.

[7] H.C. Lin and A.A. Ksienski, "Optimum Frequency Selection for Aircraft Classification," Report 3815-6, to be published. Ohio State University ElectroScience Laboratory.

[8] A.N. Mucciardi and E.E. Gose, "A Comparison of Seven Techniques for Choosing Subjects of Pattern Recognition," IEEE Trans. Computers, Vol. C-20, September 1971.

[9] A. Papoulis, Probability, Random Variables, and Stocastic Processes, McGraw-Hill, New York, 1965, pp. 125-137.

[10] H.L. Van Trees, Detection, Estimation, and Modulation, Pt. I, John Wiley and Sons, New York, 1968.

[11] A.M. Mood, Introduction to the Theory of Statistics, Chap. 9, McGraw-Hill, New York, 1950.

[12] D. Middleton, Introduction to Statistical Communication Theory, Chap. 19, McGraw-Hill, New York, 1960.

[13] R.V. Hogg and A.T. Craig, Introduction to Mathematical Statistics, Chap. 4, Macmillan, London, 1970.

[14] Y.T. Lin and J.H. Richmond, "E.M. Modeling of Aircraft at Low Frequencies," IEEE Trans. Antennas and Propagat., Vol. 23-1, March 1975.

[15] T.M. Cover and P.E. Hart, "Nearest Neighbor Pattern Classification," IEEE Trans. Informa. Theory, Vol. 13-1, January 1967.

[16] H.J. Payne and W.S. Meisel, "An Algorithm for Constructing Optimal Binary Decision Trees," IEEE Trans. on Computers, Vol. 26-9, September 1977.