UNCLASSIFIED

AD 273 235

Reproduced
by the

ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA

UNCLASSIFIED

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. 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.

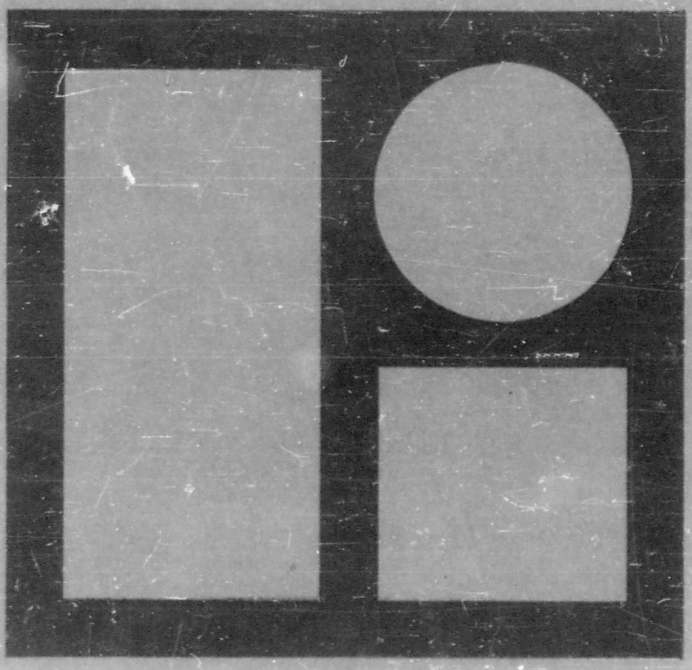
LITTON PROPRIETARY

62-2-5

COPY NO. 38

273 235

CATALOGED BY ASTIA
273235



Data Systems Division, Litton Systems, Inc.

ASTIA
RETRIEVED
MAR 20 1962
TISIA

LITTON PROPRIETARY

AD 273 235

20 March, 1962

ASTIA (TIPAA)
Arlington Hall Station
Arlington 12, Virginia

Subject: Final Report
Contract AF19(604)-8024

Reference: Letter of 16 March, 1962 from W. C. Donelan,
Subject same as above

1. In accordance with the referenced letter 10 copy (copies) of the final report on the subject contract were sent to your facility. Covers stamped "LITTON PROPRIETARY" were erroneously used on this report. Since this report contains no Litton Proprietary information, please disregard this stamped notation.

2. We are enclosing 10 copy (copies) of this letter, one copy to be enclosed or filed with each report submitted to you.

LITTON SYSTEMS, INC.

W. C. Donelan
Business Manager
Communication Sciences Laboratory

WCD:deb

ASTIA
RECEIVED
MAR 22 1962
TISIA A

AFCRL 62-65

DATA SYSTEMS DIVISION
LITTON SYSTEMS, INC.
221 Crescent Street
Waltham 54, Massachusetts

Contract No. AF19(604)-8024

Project No. 5632

Task No. 563205

Prepared for:

ELECTRONICS RESEARCH DIRECTORATE
Air Force Cambridge Research Laboratories
Office of Aerospace Research
UNITED STATES AIR FORCE
Bedford, Massachusetts

Period Covered
January 1, 1961 through December 31, 1961

Final Report

# Study Program of PATTERN RECOGNITION RESEARCH

31 December 1961

by

George S. Sebestyen/Alice K. Hartley

"Requests for additional copies by Agencies of the Department of Defense, their contractors, and other Government agencies should be directed to the:

ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA

Department of Defense contractors must be established for ASTIA services or have their 'need-to-know' certified by the cognizant military agency of their project or contract."

"All other persons and organizations should apply to the:

U. S. DEPARTMENT OF COMMERCE
OFFICE OF TECHNICAL SERVICES
WASHINGTON 25, D. C."

Pattern Recognition Research

Contract AF19(604)-8024

31 December 1962

Prepared by

George Sebestyen
Alice Hartley

Approved by

George Sebestyen
Technical Director

John Gerdes
Assistant Manager

LITTON SYSTEMS, INC.
DATA SYSTEMS DIVISION

## ABSTRACT

The basic element in the solution of pattern recognition problems is a requirement for the capability to recognize membership in classes. This report considers the automatic establishment of decision criteria for measuring membership in classes that are known only from a finite set of samples. Each sample is represented by a point in a suitably chosen, finite-dimensional vector space in which a class corresponds to a domain that contains its samples. Boundaries of the domain in the vector space can be expressed analytically with the aid of transformations that cluster samples of a class and separate classes from one another. From these geometrical notions a generalized discriminant analysis is developed which, as the sample size goes to infinity, leads to decision-making that is consistent with the results of statistical decision theory.

The approximations of the generalized non-linear discriminant analysis are developed in the form of generalized multivariate power series and in the form of polynomial approximations. Applications of the linear and non-linear methods of partitioning the vector space are programmed on general purpose digital computers and are verified on problems drawn from the area of speech signal analysis. Automatic recognition of spoken words, sounds and speakers are described. In addition, classes with specially chosen properties are generated on a computer to test the performance of the analysis technique.

Certain practical problems concerned with recognition in the presence of noise and recognition with only partially operating equipment are discussed. The latter problem concerns decision making on a lower dimensional manifold of the vector space. Approximation methods and

adaptive techniques for classification are discussed as a means of doing recognition less perfectly but in real time. The adaptive techniques discussed are algorithmic approximations of likelihood ratio computers for arbitrary distributions of the class probability densities.

## ACKNOWLEDGMENT

The study reported herein could not have been completed without the cooperative efforts of many individuals. The following are deserving of special note. Mr. Stanley Petrick, Applied Mathematics Branch, AFCRL for his guidance and constant encouragement throughout the technical phases of this work; Mr. C. P. Smith of the AFCRL Communications Laboratory for providing speech data, in digitalized vocoder formant, and for his suggestions and comments provided during informal discussions of speech data processing, which were helpful in assuring the successful performance of the program; Miss Willett of AFCRL for her assistance in helping to gather speech data, and to members of the Applied Mathematics Branch, AFCRL for samples of their speech in tape format.

TABLE OF CONTENTS


ABSTRACT

ACKNOWLEDGMENT

I. INTRODUCTION 1

II. LINEAR TECHNIQUES 5

2.1 Summary of Pertinent Earlier Work 5

2.1.1 Linear Clustering Techniques 5
2.1.2 Other Linear Techniques 11

2.2 Extensions of the Theory 11

2.2.1 Linear Separating Transformations 11
2.2.2 Other Linear Discrimination Techniques 19

2.3 Experimental Studies 32

2.3.1 Linear Clustering Transformations 32
2.3.2 Linear Separating Transformations 34

III NON-LINEAR TECHNIQUES 37

3.1 Summary of Pertinent Earlier Work 37
3.2 Extensions of the Theory 38

3.2.1 Polynomial Approximation of the Optimum Non-Linear Transformations 38
3.2.2 Generalized Non-Linear Discriminant Functions 44
3.2.2 Computational Methods for Non-Linear Discrimination 46

3.3 Experimental Studies 53

3.3.1 Univariate Examples 53
3.3.2 A Two-Dimensional Example 66

TABLE OF CONTENTS (Contd.)


IV. INCOMPLETE AND NOISY DATA 73

4.1 Classification Decisions Based on Incomplete Sets of Observations 73

4.1.1 Introduction 73
4.1.2 Decision Rules 75
4.1.4 Conclusions 79

4.2 Implications of Results on Missing Parameters 84

4.3 Classificatory Decisions on Noisy Observations 88

V. SPECIAL TECHNIQUES 98

5.1 Automatic Subclass Generation 98

5.2 A Proximity Algorithm and Selective Sampling 103

5.2.1 Selective Sampling 104
5.2.2 Experimental Results Using the Proximity Algorithm 107

5.3 Optimum Small-Sample Discrimination 109

VI RECOMMENDATIONS FOR FUTURE WORK 120

LIST OF REFERENCES 122

Appendix 2A. Program for Eigenvalues and Eigenvectors of a Real Symmetric Matrix 123

Appendix 2B. Program for Making Classificatory Decisions Through Use of Linear Class-Separating Transformation 131

Appendix 3A. Program for Non-Linear Discrimination Problem 136

Appendix 5A. Generation of Univariate Densities 140

# I. INTRODUCTION

Pattern Recognition, as dealt with here, is concerned with methods of automatically establishing decision criteria for classifying events as members of one or another class when the only information available about any class is that which is contained in a given finite set of samples known to belong to that class.

We envision recognition systems with two distinct modes of operation: a learning mode in which the system is exposed to a sequence of sample events each labeled according to the class or category to which it belongs, and a recognition mode in which new unlabeled events are classified as members of one or another of these classes. During the learning mode the system develops class-characterizing criteria from the labeled samples submitted to it and during the recognition mode it uses these criteria for classifying unlabeled samples.

A sample event is represented by an N-dimensional vector or point whose components are the values of N measurable attributes or parameters describing the event. Sample events belonging to the same category are represented by points scattered throughout some region of N-dimensional sample space in accordance with an unknown N-dimensional probability distribution function.

The classification criteria we shall be concerned with are based on the assumption that there exist transformations of the sample space such that sample points belonging to the same category lie close to each other after transformation when the Euclidean measure of distance is used. We may list the principal types of transformation studied as follows.

(a) Linear Clustering Transformations. These are transformations of sample space which minimize the mean squared Euclidean distance between pairs of points in a single sample set after transformation. In the learning mode a recognition system based on this type of transformation finds as many (clustering) transformations as there are categories, each minimizing mean squared distance for the sample set belonging to a single category. In the recognition mode an unlabeled sample is subjected to each clustering transformation in turn, and its mean squared Euclidean distance to the corresponding transformed cluster is found. The nearest cluster determines the classification.

(b) Linear Separating Transformations. These transformations minimize some measure of the intraset diameter, such as the sum of the sample set variances, subject to a constraint that maintains the distance between sample sets. A system for discriminating between two classes according to this principle determines a single transformation, in the learning mode, and in the recognition mode it uses this transformation on unlabeled samples, classifying them according to the closest sample set after transformation. A special case of this method is the well-known linear discriminant.

(c) Nonlinear Transformation. A wider class of clustering and separating transformations is made available by removing the linearity restriction. Only certain types of non-linear transformations have been studied so far. The simplest type of non-linear clustering transformation is a scalar set function which assigns a single number to each vector in a sample set in such a way that the variance of the resulting one dimensional sample set is minimized. Similarly, a non-linear separating transformation of the simplest type finds a scalar set function which minimizes the sum of sample set variances subject to a constraint on the separation of transformed means.

Chapters II and III of this report are devoted to theoretical and experimental results obtained in the two general areas of linear and non-linear techniques. In each case a summary of pertinent earlier work is given.

In addition to the study of basic recognition techniques it is important to deal with certain peripheral problems that arise immediately when applications are attempted. One of these occurs when some of the N-dimensional sample vectors on which recognition is based are incomplete, i. e., one or more component values are missing due to sensor failure or unavailability of data. Another problem is the effect of noise on the performance of these recognition systems. Some work in these two areas is reported in Chapter IV.

The basic techniques described above apply when the number of categories or classes is known in advance and a few sample events from each are available for learning. In many cases, however, the number of classes is unknown, so that methods are needed for automatically constructing sample classes from unlabeled data. Here adaptive techniques for handling samples arriving in sequence are particularly important for recognition systems working in real time. Closely related to this is the need for discrimination techniques that can be used when the number of samples is extremely small, as it would be for example during the early stages of the sequential adaptive operation mentioned above. These topics are treated in Chapter V.

Pattern recognition is a relatively new, rapidly developing field whose early successes justify the confident expectation that notable contributions to the automatization of many recognition tasks long regarded as

intractable will be forthcoming in the near future. Many of the studies reported here must be regarded as preliminary steps towards the solution of outstanding problems which must be dealt with before the full potentialities of the field can be realized. Chapter VI contains a brief discussion of some of the steps that must be taken next.

# II. LINEAR TECHNIQUES

## 2.1 SUMMARY OF PERTINENT EARLIER WORK

### 2.1.1 Linear Clustering Techniques

The optimum linear clustering transformation mentioned in the Introduction has been derived in work already reported[1]. The problem and its solution are reviewed here to provide background for the experimental work described below in Section 2.3.1, and to introduce the reader to some of the distance and transformation concepts that underlie much of our work in this field.

Let $f_m = (f_{m1} \cdots f_{mN})$ be a row vector denoting one of M samples, each with N dimensions, available from a certain class. After a linear transformation A this vector becomes $f'_m = (f'_{m1} \cdots f'_{mN})$:

$$f'_m = f_m A. \tag{2.1}$$

The mean squared Euclidean distance between pairs of vectors after transformation may be written in either of the following two forms:

$$\overline{D^2} = \overline{(f'_p - f'_m)(f'_p - f'_m)^T} \tag{2.2}$$

$$\overline{D^2} = \overline{(f_p - f_m) AA^T (f_p - f_m)^T} \tag{2.3}$$

---

[1] George Sebestyen, Classification Decisions in Pattern Recognition, MIT RLE Report 381, Parts II and II; also Litton Systems, Inc. Technical Memorandum No. 13 (Preprint of RLE Report).

where the superscript T denotes the transpose of a vector or matrix, and the overscore denotes the sample expectation, here taken over all $\underline{p}$ and $\underline{m}$. We seek the choice of A that minimizes $D^2$.

It may be shown that the transformation A which minimizes (2.1) and (2.2) subject to the constraint that the volume of the space remains unchanged is given by

$$A = \Lambda^{-1/2} C, \tag{2.3}$$

where $\Lambda$ is a diagonal matrix with elements

$$\lambda_{ii} = \sigma_i^2 \left( \prod_{k=1}^{N} \sigma_k^2 \right)^{-1/N}, \quad i=1, \ldots, N \tag{2.4}$$

and C is an orthogonal matrix whose columns are the eigenvectors of the N x N sample covariance matrix $U = (u_{ij})$:

$$U = \overline{(f_m - \bar{f}_m)^T (f_m - \bar{f}_m)} \tag{2.5}$$

$$\sigma_i = \overline{(f_{mi} - \bar{f}_{mi})^2} = \overline{f_{mi}^2} - \overline{f_{mi}}^2 = u_{ii} \tag{2.6}$$

$$C^T U C = \Lambda \tag{2.7}$$

The transformation (2.3) is termed an optimum linear clustering transformation.

A recognition system using this type of transformation makes classification decisions by subjecting the input vector to as many clustering transformations as there are sample sets corresponding to categories and then measuring the mean squared distance from the transformed input vector to each of the transformed sample sets. The decision is in favor of the nearest sample set.

It may be shown that the decision rule thus obtained becomes equivalent in form to the one known to be optimum in the limit as sample size increases, when the sample distributions are Gaussian. Suppose there are two classes, for example, labeled F and G, with samples denoted $\underline{f}$ and $\underline{g}$ respectively. The mean squared distance between an input vector $\underline{v}$ and the sample set $\{f_m\}$ after the optimum clustering transformation $A_F$ (previously determined from the sample set) has been applied, may be written in either of the equivalent forms:

$$S(v, \{f_m\}) = \overline{(v' - f'_m)(v' - f'_m)^T} \qquad (2.8)$$

$$S(v, \{f_m\}) = \overline{(v - f_m) A_F A_F^T (v - f_m)^T} \; . \qquad (2.9)$$

From (2.3) and (2.7) we have

$$A_F A_F^T = C \Lambda^{-1} C^T = U_F^{-1} \; , \qquad (2.10)$$

so that (2.9) becomes

$$S(v, \{f_m\}) = (v - \bar{f}_m) U_F^{-1} (v - \bar{f}_m)^T + B_1 \qquad (2.11)$$

$$B_1 = \overline{(f'_m - \bar{f}'_m)(f'_m - \bar{f}'_m)^T} \; . \qquad (2.12)$$

The last term is a bias equal to the sum of the coordinate variances of the transformed sample set $\{f'_m\}$ and independent of the input vector $\underline{v}$. Equation (2.8) may be written equivalently:

$$S(v, \{f_m\}) = (v' - \overline{f}'_m)(v' - \overline{f}'_m)^T + B_1 \qquad (2.13)$$

showing that the mean squared Euclidean distance between the transformed input vector and the transformed sample set is the same as the squared Euclidean distance between the transformed input vector and the mean of the transformed sample set, except for a bias constant.

The distance of $\underline{v}$ from the sample set $\{g_m\}$ may be expressed similarly in any of the equivalent forms (2.8), (2.9), (2.11) or (2.13) with of course $U_G$ and $A_G$ replacing $U_F$ and $A_F$. It should be noted that C and $\Lambda$ in (2.10) depend on $A_F$ and $U_F$, although subscripts to indicate this have been omitted to simplify writing, so that when $A_F$ and $U_F$ change to $A_G$ and $U_G$, C and $\Lambda$ change also.

The decision rule for classification is then as follows:

$$\begin{aligned} &\text{Decide } v \in F \text{ if } S(v, \{f_m\}) < S(v, \{g_m\}) \\ &\text{Decide } v \in G \text{ if } S(v, \{f_m\}) > S(v, \{g_m\}) \end{aligned} \qquad (2.14)$$

In order to compare this rule with the optimum decision theory solution, we must bear in mind that the latter is derived only for known sample distribution functions. In the above example, if the sample distribution functions were known to be f(v) and g(v) the optimum decision rule would be as follows:

$$\text{Decide } v \in F \text{ if } \log f(v) > \log g(v) + B_2$$

$$\text{Decide } v \in G \text{ if } \log f(v) < \log g(v) + B_2 \qquad (2.15)$$

where the bias $B_2$ depends on prior probabilities and the relative costs of the two possible ways of misclassifying $\underline{v}$.

If f(v) and g(v) are N-dimensional Gaussian densities with covariance matrices $K_F$ and $K_G$ respectively, (2.15) becomes

$$\text{Decide } v \in F \text{ if } (v-\bar{v}_F)K_F^{-1}(v-\bar{v}_F)^T < (v-\bar{v}_G)K_G^{-1}(v-\bar{v}_G)^T + B_3$$

$$\text{Decide } v \in G \text{ if } (v-\bar{v}_F)K_F^{-1}(v-\bar{v}_F)^T > (v-\bar{v}_G)K_G^{-1}(v-\bar{v}_G)^T + B_3 \qquad (2.16)$$

where $\bar{v}_F$ and $\bar{v}_G$ are the means of the two distributions f(v) and g(v) respectively. We need only observe that if the populations from which the finite samples $\{f_m\}$ and $\{g_m\}$ are drawn were Gaussian, then as the sample size increases without limit,

$$U_F \longrightarrow K_F, \qquad U_G \longrightarrow K_G$$

$$\bar{f}_m \longrightarrow \bar{v}_F, \qquad \bar{g}_m \longrightarrow \bar{v}_G \ , \qquad (2.17)$$

so that the decision rules (2.16) and (2.14) take on the same form, differing only in the bias constant.

These considerations also suggest a useful interpretation of the clustering transformation in statistical terms. We note from (2.11) and (2.13) that the loci of input vectors equidistant from a sample set are ellipsoidal surfaces in N dimensions before transformation, and spherical

surfaces in N-dimensions after transformation. From (2.16) and (2.17) we see that if the sample distribution were actually Gaussian, the loci of equiprobable values of $\underline{v}$ would also be ellipsoidal surfaces before and spherical surfaces after transformation. Thus the distance measure S may be regarded as a measure of probability in the limit of large samples.

For finite samples we see that the requirement that $\overline{D^2}$ of (2.3) be minimized by proper choice of transformation results in the definition of ellipsoidal loci in the sample space which may be regarded as the constant probability density loci of some fictitious N-dimensional Gaussian population distribution. If the population distribution were actually Gaussian, these loci would change as sample size increased unitl in the limit they coincided with the Gaussian contours.

There is some justification therefore for interpreting the linear clustering transformation as an operation which fits an N-dimensional Gaussian distribution function to the finite sample set. It follows that the classification technique based on linear clustering may be regarded as an application of standard decision theory in which hypothetical Gaussian distributions are fitted to the finite samples in a special way, viz., the means and covariance matrices of the hypothetical distributions are set equal to the corresponding means and covariance matrices of the sample sets.

Experimental results based on this method are given in Section 2.3.1. During the course of preparing and refining the linear clustering method, a program for the solution of eigenvalues and eigenvectors of a real symmetric matrix was written. The resulting computer program was published by Autonetics Indistrial Products, as Recomp II Users' Program No. 1068. The technical description of this program is given in Appendix 2A of this report.

### 2.1.2 Other Linear Techniques

In work already reported[2] an alternate way of rendering classificatory decisions consisted of subjecting the vector space, in which known samples of the classes are imbedded, to a transformation that clusters samples within the same class while maintaining the separation between different sets by the imposition of any one of a number of different constraints. The solutions of the resulting minimization problems were obtained under the condition that the transformation which accomplishes the clustering and class separating operations is from the class of linear transformations. Details of several variants of this method, which we have called the linear separating method, are given in Section 2.2.1.

## 2.2 EXTENSIONS OF THE THEORY

### 2.2.1 Linear Separating Transformations

Let us consider the separation of two classes F and G, containing F and G sample points respectively. Denote the points belonging to F by $x_i (i = 1, \ldots, F)$, and those belonging to G by $x_i (i = F+1, \ldots, F+G)$. Consider an arbitrary row vector $a = (a_1, \ldots, a_N)$ in the N-dimensional sample space whose direction is to be chosen so that when the clusters F and G are projected upon it they overlap as little as possible. It is possible to conceive of several maximization and minimization problems that might satisfy this general requirement, and our purpose here is to consider some of these and to compare the separations achieved by each.

A recognition system based upon such a linear separating technique determines the direction $\underline{a}$ in the learning mode, and then classifies a new point in the recognition mode by projecting it onto $\underline{a}$ and measuring its mean squared Euclidean distance along $\underline{a}$ to the projections of the two sets of sample points F and G. The nearest projected set determines the classification.

---

[2] Ibid, Section 3

A "best" direction $\underline{a}$ may be defined in a number of ways. Given some measures of set diameter and separation, for example, we might define $\underline{a}$ as the direction along which the sum of projected set diameters is minimized with set separation held fixed. Other possibilities are to minimize the sum of diameters while holding the separation of the means constant, or while holding the diameter of the combined sets F and G fixed.

A suitable measure of distance is the mean-squared Euclidean distance between pairs of points after projection:

$$\overline{D^2} = a\overline{(x_i - x_j)^T (x_i - x_j)} a^T, \tag{2.18}$$

in which the average is taken over pairs of points belonging to the same set to obtain set diameter or intraset distance, and over pairs of points in different sets to obtain set separation or interset distance. The following notation is convenient:

Intraset Distance: $\overline{D_W^2} = a\ W\ a^T$ (2.19)

Interset Distance: $\overline{D_B^2} = a\ B\ a^T$ (2.20)

Total Distance: $D_T^2 = a\ T\ a^T$ (2.21)

Distance Between Means: $D_M^2 = a\ M\ a^T$ (2.22)

in which W, B and T are N x N matrices of the form $\overline{(x_i - x_j)^T (x_i - x_j)}$, with the average taken over regions W, B and T defined as

Region W: $1 \leq i, j \leq F$ and $F + 1 \leq i, j \leq F + G$ (2.23)

$$\text{Region B:} \quad \begin{cases} 1 \le i \le F \text{ with } F+1 \le j \le F+G, \text{ and} \\ 1 \le j \le F \text{ with } F+1 \le i \le F+G . \end{cases} \tag{2.24}$$

$$\text{Region T:} \quad 1 \le i,\ j \le F+G. \tag{2.25}$$

The N x N matrix M is given by

$$M = (\bar{x}_F - \bar{x}_G)^T (\bar{x}_F - \bar{x}_G) , \tag{2.26}$$

where $\bar{x}_F$ and $\bar{x}_G$ are the means of the sample sets F and G.

It is readily verified that W, B, T and M are related to the covariance matrices $U_F$ and $U_G$ as follows:

$$W = \frac{2F^2 U_F + 2G^2 U_G}{F(F-1) + G(G-1)} \tag{2.27}$$

$$B = U_F + U_G + M \tag{2.28}$$

$$T = \frac{2F(F+G)\,U_F + 2G(F+G)\,U_G + 2FGM}{(F+G)\,(F+G-1)} \tag{2.29}$$

$$U_F = \frac{1}{F} \sum_{i=1}^{F} (x_i - \bar{x}_F)^T (x_i - \bar{x}_F) \tag{2.30}$$

$$U_G = \frac{1}{G} \sum_{i=F+1}^{F+G} (x_i - \bar{x}_G)^T (x_i - \bar{x}_G) \tag{2.31}$$

When the numbers of samples in the two classes are equal and large, we have

$$W \longrightarrow U_F + U_G \tag{2.32}$$

$$B \longrightarrow U_F + U_G + M \tag{2.33}$$

$$T \longrightarrow U_F + U_G + \frac{1}{2} M = \frac{1}{2}(W + B) \tag{2.34}$$

$$M \longrightarrow B - W. \tag{2.35}$$

In order to find "best" directions $\underline{a}$ for linear separating as discussed above, we may now make use of the fact that a ratio of quadratic forms takes on its smallest value when its argument is the eigenvector corresponding to the smallest eigenvalue of the corresponding characteristic equation. Thus we have:

1. Minimize ratio of intraset distance to interset distance:

$$\text{Min}\left[\overline{D_W^2} \Big/ \overline{D_B^2}\right] = \lambda_{\min}, \tag{2.36}$$

where $\lambda_{\min}$ is the smallest eigenvalue of

$$a(W - \lambda B)a^T = 0. \tag{2.37}$$

The corresponding eigenvector is the direction $\underline{a}$ which minimizes the ratio of intraset distance to interset distance (or maximizes the inverse ratio).

2. Minimize ratio of interset distance to distance between the means:

$$\text{Min}\left[\overline{D_W^2} \Big/ \overline{D_M^2}\right] = \beta_{\min}, \tag{2.38}$$

where $\beta$ is the smallest eigenvalue of

$$a(W - \beta M)a^T = 0. \tag{2.39}$$

Expressing M in terms of B and W with the help of (2.35) we find that this is equivalent to

$$a(W - \frac{\beta}{1+\beta} B)a^T = 0, \tag{2.40}$$

which is the same as (2.37) with $\lambda = \beta/(1+\beta)$. Thus the direction $\underline{a}$ that minimizes the ratio of intraset distance to interset distance is the same as the direction that minimizes the ratio of intraset distance to distance between the means.

3. Minimize ratio of intraset distance to total distance:

$$\text{Min} \left[ \overline{D_W^2} \Big/ \overline{D_T^2} \right] = \gamma_{\min} , \tag{2.41}$$

where $\gamma_{\min}$ is the smallest eigenvalue of

$$a(W - \gamma T)a^T = 0. \tag{2.42}$$

Again with the help of (2.34) this may be rewritten as

$$a(W - \frac{\gamma}{2-\gamma} B)a^T = 0 , \tag{2.43}$$

which is equivalent to (2.37) with $\lambda = \gamma/(2-\gamma)$.

We have thus shown that the directions $\underline{a}$ which minimize the ratios of intraset distance to interset distance, to distance between the means and to total distance all coincide.

To separate several classes it is possible to find a single direction in which, for example, the sum of the distances between the means is a maximum, subject to holding the sum of intraset distances constant. Since maximizing the sum does not necessarily maximize the addends, it is clear that there will be situations in which this method will give poor results. A better procedure is to develop a discriminant function for each pair of classes. Thus, for K classes, we may find a set of $\frac{K(K-1)}{2}$ directions, $a_{ij}$, in which the $i^{th}$ and $j^{th}$ classes are best separated. We may then determine whether a sample is more likely to belong to class i or class j. If it is more likely a member of the class i than of class j for all $j \neq i$, we classify it as a member of class i. In this way, although we must determine $\frac{K(K-1)}{2}$ directions, only K discriminant functions must be evaluated to classify a sample vector. This procedure is identical to that used in "maximum likelihood ratio" decisions. The determination of the directions, $a_{ij}$, proceeds as for two classes.

Another approach to the problem of several classes is to find a direction for each class in which that class is best separated from all other classes by minimization of the intraset distance of one class keeping the total distance constant. In this way we obtain K "two class" problems, separating class "i" from class "not i". It is,however, not difficult to imagine cases for which this approach will be inferior to the one previously described.

Computational Methods

Problems of the form

$$a(A - \lambda B)a^T = 0 \qquad (2.44)$$

may be solved in several different ways. The most straightforward method is to solve $a(AB^{-1} - \lambda I)a^T = 0$. The disadvantage of this method is that

$AB^{-1}$ is not always symmetric. Since we prefer to deal with symmetric matrices, a different method of solution has been used, as described step-by-step below.

1. Factor B into two triangular matrices

$$B = S^T S \tag{2.45}$$

$$(S^{-1})^T B S^{-1} = I \tag{2.46}$$

2. Compute $S^{-1}$

3. Compute $(S^{-1})^T A S^{-1} \equiv K$ (2.47)

4. Find the largest (or smallest) eigenvalue $\underline{\lambda}$ and corresponding eigenvector $\underline{b}$ of K:

$$b(K - \lambda I)b^T = 0 \tag{2.48}$$

$$bb^T = 1 \tag{2.49}$$

5. Then the desired solution for $\underline{a}$ is given by

$$a = b(S^{-1})^T \tag{2.50}$$

It is readily verified that this satisfies the original equation (2.44).

In this method B must be non-singular. If B is singular and A is not, A and B may be interchanged. This involves changing from maximization to minimization or vice-versa. A program for this method is described in Appendix 2B.

The definitions of A and B may be chosen from any of the previously formulated statements of the separation problem. We note that the problem of minimizing the ratio of intraset distance to distance between the means (Eq. 2.39) is not amenable to this treatment, because the matrix

M is singular. By interchanging W and D and solving for the maximum eigenvalue and corresponding eigenvector however we maximize the ratio of distance between the means to intraset distance, and the present method applies. A further advantage of this choice is that less computation is necessary to find the largest eigenvalue than to find the smallest. In addition the spectral resolution of the eigenvalues is greater in this form than in any of the others.

A still simpler technique may be used for the latter problem, which takes advantage of the fact that M is of rank 1 and W is symmetric and positive definite. The problem may be stated as follows:

$$\text{Maximize} \quad \frac{a d^T d a^T}{a W a^T} \quad \text{by choice of } \underline{a}. \tag{2.51}$$

Here M has been written as $d^T d$ in accordance with (2.26). For the matrix W we may write a Schwarz inequality of the form

$$\frac{(x W y^T)^2}{x W x^T} \leq y W y^T. \tag{2.52}$$

Taking $x \equiv a$ and $y \equiv dW^{-1}$, we have

$$\frac{a d^T d a^T}{a W a^T} \leq d W^{-1} d^T \tag{2.53}$$

showing that the choice of $\underline{a}$ which maximizes (2.51) is

$$a = d W^{-1}. \tag{2.54}$$

Thus the direction $\underline{a}$ may be found as the solution of a set of linear equations.

Experimental results obtained with the linear separating transformations are given in Section 2.3.2.

### 2.2.2 Other Linear Discrimination Techniques

In this section we consider some techniques for linear discrimination which apply when certain assumptions about the N-dimensional population densities can be made. Our starting point is the well-known fact that if f(v) and g(v) are the density functions of the two populations F and G to be discriminated, then the optimum decision rule for classifying a sample $\underline{v}$ is as follows:

$$\text{Decide } v \in F \text{ when } f(v)/g(v) > 1 \tag{2.55}$$

$$\text{Decide } v \in G \text{ when } f(v)/g(v) < 1 \tag{2.56}$$

This rule is optimum in the sense that it maximizes the probability of correct decision when the prior probabilities of $\underline{v}$'s membership in the two classes are equal, and when the costs of misclassifying $\underline{v}$ are the same for the two possible types of misclassification. In the following we consider f(v) and g(v) to be Gaussian densities, so that from (2.16) the rule (2.55) becomes equivalent to

$$\text{Decide } v \in F \text{ when } (v-\bar{v}_F)\, K_F^{-1}\, (v-\bar{v}_F)^T < (v-\bar{v}_G)\, K_G^{-1}\, (v-\bar{v}_G)^T, \tag{2.57}$$

where $K_F$ and $K_G$ are the covariance matrices, and $\bar{v}_F$ and $\bar{v}_G$ the means of f(v) and g(v) respectively.

We may show that when $K_F = K_G$ the optimum rule (2.57) is equivalent to a distance comparison after projection of the sample vector $\underline{v}$ onto the same line found to be optimum in Section 2.2.1. Expansion and rearrangement of (2.57) yields:

$$\text{Decide } v \in F \text{ when } (\bar{v}_F - \bar{v}_G) K^{-1} [\, v - \tfrac{1}{2} (\bar{v}_F + \bar{v}_G) \,]^T > 0 \qquad (2.58)$$

we see that this rule projects the vector difference between the sample $\underline{v}$ and the average of the sample means onto the line

$$a = dK^{-1}, \quad d \equiv (\bar{v}_F - \bar{v}_G) \qquad (2.59)$$

and bases its decision on the polarity of the result. The direction (2.59) has the same form as the one found for finite samples above which maximized the projected distance between the means holding projected intraset distance constant. As sample size increases the sample covariance matrix U approaches K and the two results become identical.

In the discussion of the finite sample case in 2.2.1 the problem was to find a direction $\underline{a}$ such that when the sample sets were projected onto $\underline{a}$ they were optimally separated in the sense that intraset distance was minimized subject to various constraints. Let us now consider the (large sample) case of two known Gaussian distributions f(v) and g(v) which are projected onto a line $\underline{a}$, and ask what the direction $\underline{a}$ should be to minimize the probability of error when a likelihood decision rule is used on the projected one-dimensional distributions.

If the new one-dimensional variable is $y = av^T$, its distributions f(y) and g(y) will be one-dimensional Gaussian distributions with means and variances given by

$$\bar{y}_F = a\bar{v}_F^T, \qquad \sigma_F^2 = aK_F a^T$$
$$\bar{y}_G = a\bar{v}_G^T, \qquad \sigma_G^2 = aK_G a^T \,. \qquad (2.60)$$

The likelihood ratio decision rule is

$$\begin{aligned} &\text{Decide } v \in F \text{ when } f(y)/g(y) > 1 \\ &\text{Decide } v \in G \text{ when } f(y)/g(y) < 1 \ . \end{aligned} \tag{2.61}$$

We assume for the remainder of this section that $\bar{y}_F = \bar{y}_G = 0$, reserving consideration of the general case for the next section.

For equal means but different variances, two univariate normal densities intersect at two points:

$$y^{+,-} = \pm \sigma_F \sigma_G \sqrt{\frac{\log \sigma_G^2 / \sigma_F^2}{\sigma_G^2 - \sigma_F^2}} \tag{2.62}$$

The probability of error is thus

$$\begin{aligned} P(\epsilon) &= 1 - 2\Phi\left(\frac{y^+}{\sigma_G}\right) + 2\Phi\left(\frac{y^+}{\sigma_F}\right), \quad \sigma_F > \sigma_G \\ P(\epsilon) &= 1 - 2\Phi\left(\frac{y^+}{\sigma_F}\right) + 2\Phi\left(\frac{y^+}{\sigma_G}\right) \quad \sigma_G > \sigma_F \end{aligned} \tag{2.63}$$

where $\Phi(t)$ is the probability integral

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} \, dx. \tag{2.64}$$

For $\sigma_G > \sigma_F$ and y scaled so that $\sigma_1 = 1$ for convenience, the problem is then to minimize

$$\frac{P(\epsilon)-1}{2} = \Phi\left[\sqrt{\frac{\log \sigma_G^2}{\sigma_G^2-1}}\right] - \Phi\left[\sqrt{\frac{\sigma_G^2 \log \sigma_G^2}{\sigma_G^2-1}}\right] . \quad (2.65)$$

Since the first and second terms of (2.65) are decreasing and increasing functions of $\sigma_2$, respectively, we wish to maximize $\sigma_2^2$ subject to the constraint that $\sigma_1^2$ remains fixed. In view of (2.60) this problem and its solution may be stated as follows:

$$\text{Max} \frac{a\, K_G\, a^T}{a\, K_F\, a^T} = \lambda_{max} \quad (2.66)$$

where $\lambda_{max}$ is the largest eigenvalue of

$$a(K_G - \lambda K_F) a^T = 0 \quad (2.67)$$

and the solution $\underline{a}$ is the eigenvector corresponding to $\lambda_{max}$. For $\sigma_F > \sigma_G$ (2.66) and (2.67) become

$$\text{Max} \frac{a\, K_F\, a^T}{a\, K_G\, a^T} = \psi_{max} \quad (2.68)$$

$$a(K_F - \psi K_G) a^T = 0 \quad (2.69)$$

Since the inequality involving the $\sigma$'s depends on the direction $\underline{a}$, both (2.67) and (2.69) must be solved and the solution corresponding to the larger of ( $\lambda_{max}$, $\psi_{max}$) chosen as the one producing the smallest error. We note that $\psi_{max} = (\lambda_{min})^{-1}$ so that (2.66) may be solved for the larger of $[\lambda_{max}, (\lambda_{min})^{-1}]$.

We now remove the condition $\bar{v}_F = \bar{v}_G$ imposed above and consider the general case of normal distributions with different means and different covariances. As seen in Figure 2-1, which shows the two distributions of $y = av^T$, two types of error can occur when a threshold $y_o$ is used for decision making according to the rule:

$$\begin{aligned} &\text{Decide } v \in F \text{ when } y < y_o \\ &\text{Decide } v \in G \text{ when } y > y_o \end{aligned} \qquad (2.70)$$

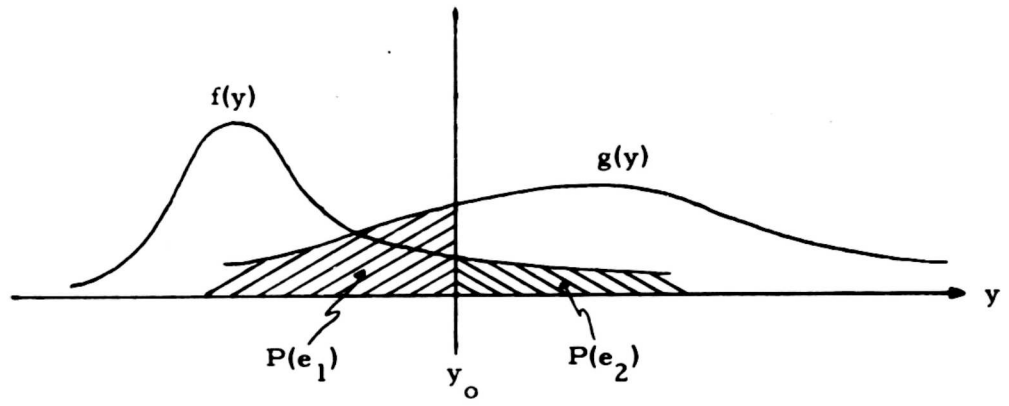



Figure 2-1.

The probability of deciding $v \in F$ when actually $v \in G$ is denoted $P(e_1)$ and is given by

$$P(e_1) = \int_{-\infty}^{y_o} g(y)\,dy \tag{2.71}$$

or

$$P(e_1) = 1 - \Phi\left(\frac{\bar{y}_G - y_o}{\sigma_G}\right) , \tag{2.72}$$

in which $\Phi(t)$ is defined by (2.64). Similarly, $P(e_2)$ is the probability of deciding $v \in G$ when actually $v \in F$:

$$P(e_2) = \int_{y_o}^{\infty} f(y)\,dy \tag{2.73}$$

or

$$P(e_2) = 1 - \Phi\left(\frac{y_o - \bar{y}_F}{\sigma_F}\right) . \tag{2.74}$$

A problem of considerable interest is that of finding the direction $\underline{a}$ which minimizes $P(e_2)$ with $P(e_1)$ held fixed. This may be accomplished as follows. We introduce two quantities $x_1$ and $x_2$ defined by:

$$P(e_1) = 1 - \Phi(x_1) \tag{2.75}$$

$$P(e_2) = 1 - \Phi(x_2) . \tag{2.76}$$

From (2.72) and (2.74) $x_2$ may be expressed in terms of $x_1$ according to

$$x_2 = \frac{(\bar{y}_G - \bar{y}_F) - \sigma_G x_1}{\sigma_F} \tag{2.77}$$

We now choose $\underline{a}$ to maximize $x_2$ with $x_1$ held fixed, which from (2.75) and (2.76) maximizes $P(e_2)$ with $P(e_1)$ held fixed. With (2.59) and (2.60), (2.77) becomes

$$x_2 = \frac{ad^T - (aK_G a^T)^{1/2} x_1}{(aK_F a^T)^{1/2}} \tag{2.78}$$

The usual calculus procedures show that $x_2$ is a maximum for $\underline{a}$ satisfying

$$a(U_1 + \lambda U_2) = d, \tag{2.79}$$

where

$$\lambda = \frac{ad^T - (aK_G a^T)^{1/2} x_1}{aK_F a^T} \tag{2.80}$$

Equation (2.79) is non-linear in $\underline{a}$, since $\lambda$ is a function of $\underline{a}$. It may, however, be solved by an iterative procedure to be described in a later Section of 2.2.2.

In the above paragraphs we have been concerned with minimizing the probability of error in recognizing two classes. Another approach maximizes the so-called "average information for discrimination" contained in a sample to be classified. The two approaches are closely related, as we now proceed to demonstrate.

According to information theory a sample $\underline{y}$ which comes from a class F with probability $P(F \mid y)$ and from G with probability $P(G \mid y)$ conveys

$$- \log \frac{P(F \mid y)}{P(F)} \tag{2.81}$$

bits of information about F, and

$$- \log \frac{P(G \mid y)}{P(G)} \tag{2.82}$$

bits of information about G, where P(F) and P(G) are the prior probabilities that the sample comes from each of the classes (assumed to be equal in what follows). The difference between (2.81) and (2.82) is the information conveyed by $\underline{y}$ for discrimination between F and G:

$$\log \frac{P(F \mid y)}{P(G \mid y)} = \log \frac{f(y)}{g(y)} \quad , \tag{2.83}$$

where f(y) and g(y) are the density functions of the two classes. When (2.83) is averaged with respect to f(y) we obtain the "average information conveyed by a sample for discrimination in favor of F", denoted I(F:G):[3]

$$I(F:G) = \int f(y) \log \frac{f(y)}{g(y)} \, dy \, . \tag{2.84}$$

Similarly,

$$I(G:F) = \int g(y) \log \frac{f(y)}{g(y)} \, dy \, . \tag{2.85}$$

The difference between (2.84) and (2.85) is the "average information for discrimination per sample", and is thus a measure of the distinguishability of the two classes.[3]

---

[3] S. Kullback, Information Theory and Statistics, Wiley, 1959

$$J(F:G) = I(F:G) - I(G:F) \tag{2.86}$$

Now consider as before that $\underline{y}$ is the one-dimensional projection of an N-dimensional sample point $\underline{v}$ on the line $\underline{a}$: $y = av^T$, and that f(y) and g(y) are Gaussian with statistics given in (2.60). We find that (2.84), (2.85) and (2.86) then become:

$$I(F:G) = \frac{1}{2} \log \frac{\sigma_G^2}{\sigma_F^2} - \frac{1}{2} + \frac{1}{2} \left[ \frac{\sigma_F^2 + (\bar{y}_F - \bar{y}_G)^2}{\sigma_G^2} \right] \tag{2.87}$$

$$I(G:F) = -\frac{1}{2} \log \frac{\sigma_G^2}{\sigma_F^2} - \frac{1}{2} + \frac{1}{2} \left[ \frac{\sigma_G^2 + (\bar{y}_F - \bar{y}_G)^2}{\sigma_F^2} \right] \tag{2.88}$$

$$J(F:G) = -1 + \frac{1}{2} \left[ \frac{\sigma_F^2 + (\bar{y}_F - \bar{y}_G)^2}{\sigma_G^2} \right] + \frac{1}{2} \left[ \frac{\sigma_G^2 + (\bar{y}_F - \bar{y}_G)^2}{\sigma_F^2} \right] \tag{2.89}$$

The usual calculus procedures show that I(F:G) is maximized by an $\underline{a}$ which satisfies

$$a(K_F + \lambda K_G) = d \tag{2.90}$$

with

$$\lambda = \frac{(ad^T)^2 - (aK_G a^T - aK_F a^T)}{(ad^T)(aK_G a^T)} \tag{2.91}$$

Similarly, J(F:G) is maximized by an $\underline{a}$ satisfying

$$a(K_F + \lambda K_G) = d \tag{2.92}$$

in which

$$\lambda = \frac{aK_F a^T - aK_G a^T}{(aK_G a^T)(ad^T)} + \frac{a(K_F a^T)(ad^T)}{(aK_G a^T)(aK_F a^T + aK_G a^T)} . \quad (2.93)$$

In the case of equal covariance matrices, the information measures I(F:G) and I(G:F) are equal for any direction $\underline{a}$, and maximization of either (or of J(F:G)) yields the same solution. Moreover, this solution is the same as that obtained by minimizing the probability of misclassification.

If the mean vectors of the two populations are equal, maximization of any of the information measures yields an equation of the same form as that satisfying the minimum error, namely

$$a(K_F - \lambda K_G) a^T = 0. \quad (2.94)$$

That is, the largest and smallest eigenvalues of (2.94) are equal to the largest and smallest values of the ratio

$$r = \frac{aK_F a^T}{aK_G a^T} = \frac{\sigma_F^2}{\sigma_G^2} \quad (2.95)$$

Since the information measures in this case may be expressed in terms of $\underline{r}$ as follows:

$$I(F:G) = r - \log r \quad (2.96)$$

$$I(G:F) = \frac{1}{r} + \log r \quad (2.97)$$

$$J(F:G) = \frac{1}{r} + r, \quad (2.98)$$

the eigenvalue of (2.94) for which each of the information measures is a maximum is either the largest, $\lambda_{max}$, or the smallest $\lambda_{min}$. It is possible, however, that I(F:G) is maximized for $\lambda_{max}$, and I(G:F) for $\lambda_{min}$ (e.g. $\lambda_{max} = 1.9$, $\lambda_{min} = 1/2$).

It was shown in Section 2.2.2 that for equal means, the probability of error is minimized in the direction corresponding to $\lambda_{max}$ if $\lambda_{max} > \frac{1}{\lambda_{min}}$ and $\lambda_{min}$ if $\lambda_{max} < \frac{1}{\lambda_{min}}$.

This is the same solution as that obtained by maximizing J(F, G), since if $\lambda_{max} > \frac{1}{\lambda_{min}}$ and $\lambda_{max} < 1$, then $\lambda_{max} + \frac{1}{\lambda_{max}} > \lambda_{min} + \frac{1}{\lambda_{min}}$.

In the general case maximization of each information measure results in a different solution. It has not been shown that any one will give a direction in which the probability of error is minimum. However, all the approaches give good results as does that of assuming equal covariance matrices.

An Iterative Technique for Solving the General Case

Equations of the form $a(K_F + \lambda K_G) = d$, where $\underline{\lambda}$ is a function of $\underline{a}$ such that $|K_F + \lambda K_G| \neq 0$, may be solved by an iterative method which uses the recursion relation

$$a_n = d(K_F + \lambda_{n-1} K_G)^{-1} \tag{2.99}$$

$$\lambda_n = f(a_n) . \tag{2.100}$$

This involves inversion of a matrix of the form $(A + \lambda B)$ for each iteration. The inversion may be carried out step-by-step as follows:

1) Introduce a matrix H with the following property

$$H(A + \lambda B)H^T = I$$
$$(A + \lambda B)^{-1} = H^T H, \tag{2.101}$$

and consider the decomposition of H:

$$H = DCL \tag{2.102}$$

in which D, C and L are respectively diagonal, orthonormal and triangular matrices. Equation (2.101) then may be written

$$DCLAL^T C^T D + \lambda DCLBL^T C^T D = I \tag{2.103}$$

2) Determine L such that

$$LBL^T = I \tag{2.104}$$

Then $CLBL^T C^T = I$ (B non-singular), and (2.103) becomes

$$DCKC^T D + \lambda D^2 = I \tag{2.105}$$

$$K \equiv LAL^T \tag{2.106}$$

3) Determine C such that

$$CKC^T = \Lambda \tag{2.107}$$

where $\Lambda$ is a diagonal matrix of eigenvalues. Equation (2.103) is thereby further reduced to

$$D \Lambda D + \lambda D^2 = I \ , \tag{2.108}$$

so that D is given by

$$D^{-2} = ( \Lambda + \lambda I) \tag{2.109}$$

if the right-hand side is non-singular.

4) From (2.101) we now have

$$(A + \lambda B)^{-1} = L^T C^T ( \Lambda + \lambda I)^{-1} CL \tag{2.110}$$

with C and L independent of $\lambda$. The recursion relation corresponding to (2.99) becomes

$$a_n = DL^T C^T ( \Lambda + \lambda_{n-1} I)^{-1} CL \tag{2.111}$$

5) With the following abbreviations

$$b \equiv dCL \tag{2.112}$$

$$M \equiv L^T C^T \equiv (m_{ij}) \tag{2.113}$$

$$\Lambda \equiv (\mu_{ii}), \tag{2.114}$$

we may write the desired solution for $\underline{a}$ in the form

$$a_{nk} = \sum_{j=1}^{N} \frac{m_{kj} b_j}{\mu_{jj} + \lambda_{n-1}} \ . \tag{2.115}$$

Each iteration requires the multiplication of a matrix by a vector for the determination of $a_n$.

In the applications discussed in previous paragraphs, $\lambda_n$ is a function of the quadratic forms $a_n K_F a_n^T$ and $a_n K_G a_n^T$, which thus must be evaluated at each stage to continue the iteration. The relation

$$a_n K_F a_n^T + \lambda_{n-1} a_n K_G a_n^T = a_n d^T$$

may be helpful in this regard.

The iterative procedure is not guaranteed to converge, but it has been used successfully by other investigators.[4]

## 2.3 EXPERIMENTAL STUDIES

### 2.3.1 Linear Clustering Transformations

In work already reported,[5,6] classificatory decisions on N-dimensional vector representations of the physical world were rendered by a decision rule that evaluates and compares a set of non-euclidean distances. These distances compare the proximity of the vector to be classified with the means of each class to which the vector may belong. It has been shown that if the non-euclidean distance is defined as the euclidean distance after linear transformation of the vector space, the above described procedure is equivalent to fitting the set of known sample vectors of each class with the best fitting Gaussian probability density, and basing subsequent classificatory decisions on the maximum "likelihood" ratio principle, using the probability densities thus obtained. (See Section 2.1 above.)

---

[4] Ibid

[5] op cit, Sebestyen, Sections 2 and 3

[6] G. Sebestyen, "Recognition of Membership in Classes", IRE Transactions PGIT, January 1961.

To test the applicability of fitting "difficult looking" unknown data with Gaussian densities, a small sample problem was prepared and run on the computer. Recognition of two consonant sounds was investigated. These were the sound of "T" in the spoken word "two" and the middle of the sound "X" in the word "six". Sounds were represented as 25-dimensional vectors. The first 18 of these represent the magnitude of the spectrum of the sound resolved into 18 different bands of frequencies by the AFCRL Vocoder.* Additional dimensions of the space were the quantities computed from the spectrum listed below:

19 Centroid of the sound spectrum (first moment arm)

20 Variance of spectrum about Centroid

21 The number of spectrum peaks

22 Frequency of first spectrum peak

23 Frequency of second spectrum peak

24 Frequency of third spectrum peak

25 Binary variable stating whether or not there were exactly three spectrum peaks in the sound

Recognition of the T sound was learned on 26 samples while X was learned on 30 samples.

Recognition testing of these sounds, using the learning samples as test data, results in the confusion matrix:

---

*The vocoder outputs used in this experiment were provided through the courtesy of C. P. Smith, AFCRL Communication Laboratory.

| sound spoken as | sound recognized as: T | sound recognized as: X |
|---|---|---|
| T | 26 | 0 |
| X | 0 | 30 |

A brief recognition test was also performed on several sample sounds not employed in the learning program, resulting in the following confusion matrix:

| sound spoken as | sound recognized as: T | sound recognized as: X |
|---|---|---|
| T | 7 | 3 |
| X | 0 | 6 |

Other experiments with this computational technique are described in references 1 and 6.

2.3.2 Experiments with Linear Separating Transformations

To test the linear class separating techniques, the "T" and "X" sounds described in the previous section were subjected to 2 variants of the class-separating technique. In variant I the direction "a" was found in which the distance between the means is maximized keeping the intraset distance constant. The decision procedure was to decide that the instantaneous sound spectrum, v, is more likely a T sound than an X sound if the inequality below holds.

$$\left( v - \frac{M_T + M_X}{2} \right) \cdot a > 0$$

Here $M_T$ and $M_X$ are the means of the T and X sounds, while "a" is the direction found through solving the minimization problems given above.

In variant II, two directions were found: "$a_T$", the direction in which the intraset distance of T sounds is minimized, keeping the total of intra and interset distances constant, and "$a_X$", the direction in which intraset distances of the sound X are minimized, subject also to the constraint stated above.

The decision procedure was to decide that

$$v \in T \text{ if } |(v - M_T) \cdot a_T| < |(v - M_X) \cdot a_X|$$

The results are summarized by the two confusion matrices shown below. Variant I of the method, achieved more separation and fewer errors. As the number of classes (sounds) increases, however, results of Variant II are not expected to be degraded while the opposite is expected of results obtained by Variant I.

sound recognized as

| sound spoken as | T | X |
|---|---|---|
| T | 35 | 1 |
| X | 0 | 36 |

Variant I

sound recognized as

| sound spoken | T | X |
|---|---|---|
| T | 34 | 2 |
| X | 5 | 31 |

Variant II

Linear separating techniques [$a(D - \lambda (U_1 + U_2)) = 0$], were also applied to the problem of voiced-unvoiced discrimination. In this example a sample group of voiced and unvoiced sounds (in 18 channel vocoder representation) was selected for learning. The number of samples of each sound in the group is proportional to the frequency of occurrence of that sound in the English language. The results are summarized in the following confusion matrix.

| | | Recognized as | |
|---|---|---|---|
| | | voiced | unvoiced |
| spoken as | voiced | 175 | 14 |
| | unvoiced | 2 | 49 |

Total error rate 6.6 percent

The direction of best separation is given by the following list of component values:

| Channel | Component of "a" | Channel | Component |
|---|---|---|---|
| 1 | .42458 | 10 | .25909 |
| 2 | .23173 | 11 | .21405 |
| 3 | .15665 | 12 | .18047 |
| 4 | .18447 | 13 | .05617 |
| 5 | -.00081 | 14 | .21853 |
| 6 | .37618 | 15 | -.16202 |
| 7 | .12593 | 16 | -.06934 |
| 8 | .30841 | 17 | .34683 |
| 9 | .01263 | 18 | -.32712 |

Two of the key programs realizing the computational techniques of Section II of this report are given in Appendices 2A and 2B.

# III NON-LINEAR TECHNIQUES

## 3.1 SUMMARY OF PERTINENT EARLIER WORK

It has previously been shown[7] that the mean-square distance between vectors of a set, after transformation, is equivalent to the sample variance of that set about the mean vector. Minimization of the mean-square distance between members of a set, subject to the constraint of holding the distance between the sets fixed, may thus be formulated as the problem of finding a transformation u(v) which accomplishes the objectives stated in Equation 3.1, where. $\sigma_F^2(u)$ and $\sigma_G^2(u)$ are sample variances of class F and G, while $\overline{u_F}$ and $\overline{u_G}$ are sample means after transformation u(v).

$$\text{minimize } Q = \sigma_F^2(u) + \sigma_G^2(u) \quad \text{while} \tag{3.1a}$$

$$K = \overline{u_F} - \overline{u_G} \quad . \tag{3.1b}$$

The generalization of the decision rule of Equation 2.14 of Section 2.1 may be expressed in Equation 3.2.

$$\text{decide } v \in F \text{ if } \quad |u(v) - \overline{u_F}| < |u(v) - \overline{u_G}| \tag{3.2}$$

The function u(v) in the above decision rule expresses the transformation which clusters members of the sets while separating sets, an objective expressed as a readily solved mathematical problem in Equation 3.1. The more general the class of transformations u is, the more perfectly the classes may be distinguished from each other.

---

[7] op cit., Sebestyen, Section 5.

The minimization problem of Equation 3.1 is solved in Section 3.2.1 for the general function u(v) in terms of the probability densities of the two classes. The function u(v) is then approximated by a polynomial (of degree p) such that the expected value of the error of approximation (squared) is minimized. The details for the univariate case are given in Section 3.2.1 and for the multivariate case in Section 3.2.2.

## 3.2 EXTENSIONS OF THE THEORY

### 3.2.1 Polynomial Approximation of the Optimum Nonlinear Transformation

In this section the minimization problem of Equation 3.1 will be solved for the optimum nonlinear transformation, u(v).

This transformation, when substituted into the decision rule of Equation 3.2 will be shown to equate this rule to a Bayes Decision Rule, which minimizes the probability of error in making decisions.

The optimum polynomial approximation of degree R which approximates the function u(v) will be derived by minimizing the expected error in making the approximation.

The coefficients, $a_n$, of this polynomial are obtained in terms of the higher moments of the sample vectors.

First the rule of Equation 3.2 will be reduced to an equivalent but simpler form. Note that if the inequalities of Equation 3.2 hold, so do those of 3.3.

$$\text{decide } v \in F \text{ if } [u(v) - \overline{u_F}]^2 < [u(v) - \overline{u_G}]^2$$

$$\text{decide } v \in G \text{ if } [u(v) - \overline{u_F}]^2 > [u(v) - \overline{u_G}]^2 \tag{3.3}$$

Squaring both sides of the inequalities, cancelling like terms and rearranging the inequality yields Equation 3.4

$$u^2(v) - 2u\,\overline{u_F} + \overline{u}_F^2 < u^2(v) - 2u\,\overline{u_G} + \overline{u}_G^2 \tag{3.4a}$$

$$\overline{u}_F^2 - \overline{u}_G^2 < 2u\,(\overline{u_F} - \overline{u_G}) \tag{3.4b}$$

$$\frac{\overline{u_F} + \overline{u_G}}{2} < u(v) \tag{3.4c}$$

Now we solve the minimization problem of Equation 3.1a subject to the constraint stated in Equation 3.1b. To facilitate the solution, we minimize 3.5a subject to 3.5b instead, and show later that the two problems have identical solutions.

$$\text{minimize } Q' = \overline{u_F^2} + \overline{u_G^2} \tag{3.5a}$$

$$K = \overline{u_F} - \overline{u_G} > 0 \tag{3.5b}$$

Using the method of Lagrange multipliers and writing out the expression in integral form, we obtain Equation 3.6

$$Q' + \lambda K = \int \left\{ u^2(v)[p_F(v) + p_G(v)] + \lambda\, u(v)\,[p_F(v) - p_G(v)] \right\} dv \qquad (3.6)$$

Setting the variation of Equation 3.6 equal to zero, Equation 3.7 is obtained.

$$\delta(Q' + \lambda K) = 0 = \int \left\{ 2u(v)[p_F(v) + p_G(v)] + \lambda\,[p_F(v) - p_G(v)] \right\} \delta u\, dv \qquad (3.7)$$

Since $\delta u$ is arbitrary, however, the expression $\{\ \}$ must be identically zero for all $\delta u$. Solving for u(v), we obtain 3.8.

$$u(v) = -\frac{\lambda}{2}\,\frac{p_F(v) - p_G(v)}{p_F(v) + p_G(v)} \qquad (3.8)$$

The Lagrange multiplier may now be evaluated by substituting 3.8 into the constraint at Equation 3.5b.

$$K = \overline{u_F} - \overline{u_G} = -\frac{\lambda}{2} \int \frac{[p_F(v) - p_G(v)]^2}{[p_F(v) + p_G(v)]}\, dv > 0 \qquad (3.9)$$

We have to note only that the integral is a positive quantity, establishing the fact that the constant, $\lambda$, is negative. This fact permits us to rewrite 3.8 as shown below.

$$u(v) = \left|\frac{\lambda}{2}\right| \left( \frac{p_F(v) - p_G(v)}{p_F(v) + p_G(v)} \right) \qquad (3.10)$$

Substitution into the simplified decision rule, 3.4c yields the further simplified rule of 3.11

$$u(v) > \frac{\overline{u_F} + \overline{u_G}}{2} = \left|\frac{\lambda}{4}\right| \int [p_F(v) - p_G(v)]dv = 0 \qquad (3.11a)$$

$$\text{decide } v \in F \text{ if } u(v) > 0 \qquad (3.11b)$$

$$\text{decide } v \in G \text{ if } u(v) < 0 \qquad (3.11c)$$

When Equation 3.10 is substituted in the decision rule of Equation 3.11, we obtain Equation 3.12, which is satisfied only if $p_F(v) > p_G(v)$, the criterion required by Bayes Decision Rule.

$$\text{decide } v \in F \text{ if } u(v) = \left|\frac{\lambda}{2}\right| \frac{p_F(v) - p_G(v)}{p_F(v) + p_G(v)} > 0 \qquad (3.12)$$

Thus the two rules are proved to be equivalent if u(v) is a solution of the minimization problem stated in 3.5a and 3.5b. It now remains to show only that the problem stated in 3.3 and 3.1 is equivalent to that stated in 3.5a and 3.5b.

It is seen from 3.11a and 3.1b that $\overline{u_F(v)} = -\overline{u_G(v)} = K/2$. Substituting this value in Equation 3.1, we obtain 3.13.

$$Q = \sigma_F^2(u) + \sigma_G^2(u) = \overline{u_F^2} + \overline{u_G^2} - \overline{u_F}^2 - \overline{u_G}^2 + \overline{u_F}^2 + \overline{u_G}^2 - \frac{K^2}{2} \qquad (3.13)$$

Since $K^2/2$ is a constant, the minimization of Q leads to the same solution as the minimization of Q'.

Thus the nonlinear transformation is given by the function

$$u(v) = \left|\frac{\lambda}{2}\right| \frac{p_F(v) - p_G(v)}{p_F(v) + p_G(v)} \tag{3.14}$$

The expected error in approximating u(v) by a polynomial of degree R is given in 3.15, where the squared expression is the squared error due to the approximation when the observation v is made, and the other bracketed expression is the probability that v will occur (neglecting differences in the a priori probabilities of the two classes).

$$E = \int \left[ u(v) - \sum_{n=0}^{R} a_n v^n \right]^2 [p_F(v) + p_G(v)]\, dv \tag{3.15}$$

Using variational methods, we may solve for the coefficients of the polynomial approximations by setting $\partial E / \partial a_s = 0$ for all $s = 0, 1, \ldots, R$.

$$\frac{\partial E}{\partial a_s} = 0 = -2 \int \left[ u(v) - \sum_{n=0}^{R} a_n v^n \right] v^s [p_F(v) + p_G(v)]\, dv \tag{3.16}$$

By interchanging the order of summation and integration and by substituting 3.14 into 3.16, we obtain 3.17. The constant $\frac{\lambda}{2}$ has been absorbed into the polynomial coefficients, $a_n$,

$$\int u(v)\, v^s [p_F(v) + p_G(v)]\, dv = \sum_{n=0}^{R} a_n \int v^{n+s} [p_F(v) + p_G(v)]\, dv \tag{3.17a}$$

$$\int v^s [p_F(v) - p_G(v)]dv = \sum_{n=0}^{R} a_n \int v^{n+s} [p_F(v) + p_G(v)]\, dv \text{ for } s = 0, 1, \ldots, R \tag{3.17b}$$

By letting the integral on the left side of the equation be $\beta_s$ and that on the right be $\gamma_{ns} = \gamma_{sn}$, the general element of the matrix $\gamma$, we can write 3.17 in matrix notation as shown in 3.18, where b is the row vector of $\beta_s$ and a is that of $a_n$.

$$\beta_s = \sum_{n=0}^{R} a_n \gamma_{ns} \quad \text{for } s = 0, 1, \ldots, R \tag{3.18}$$

$$b = a\gamma$$

The coefficients, $a_n$, may now be solved as in 3.19 which is the solution of the optimum polynomial approximation of u(v).

$$a = b\gamma^{-1} \tag{3.19}$$

The solution involves only moments of the probability densities. In case only a finite number of points are available, moments of the samples can be computed, although it has not been shown whether or not this is optimum. It is known, however, that as the sample size increases, this procedure becomes optimum.

### 3.2.2 Generalized Non-Linear Discriminant Functions

In the multivariate case the optimum non-linear discriminant function u(v) is determined exactly as in the univariate case and the solution $u(v) = \frac{p_F(v) - p_G(v)}{p_F(v) + p_G(v)}$ is of the same form but v is now an n dimensional vector. Thus we wish to approximate the function $\frac{p_F(v)-p_G(v)}{p_F(v)+p_G(v)}$ $= u(v)$ by a polynomial of the form $\sum_{k=1}^{K} a_k \alpha_k$, where $\alpha_k$ is a function of the n variables, and has the form $v_1^q v_2^r \ldots . v_n^t$, $q + r \ldots + t = p$; there are $\binom{n+p}{p} = K$ such terms. We will choose the polynomial which minimizes the expected value of the mean-square deviation from the function u(v).

Thus, we wish to minimize

$$I = \int_{v_1} \int_{v_2} \cdots \int_{v_n} u(v) - \sum_{k=1} a_k \alpha_k \Big)^2 \quad \frac{[p_F(v) + p_G(v)]}{2} \, dv_1 dv_2 \ldots dv_n \qquad (3.20)$$

$$\frac{\delta I}{\delta a_j} = \int_{v_1} \int_{v_2} \cdots \int_{v_n} u(v) - \sum_{k=1}^{K} a_k \alpha_k \Big) \alpha_j [p_F(v) + p_G(v)] \, dv_1 \ldots dv_n = 0 \qquad (3.21)$$

$$\int_{v_1} \int_{v_2} \cdots \int_{v_n} [p_F(v) - p_G(v)] \alpha_j dv_1 \ldots dv_n = \int_{v_1} \int_{v_2} \cdots \int_{v_n} \sum_{k=1}^{K} a_k \alpha_k \alpha_j [p_F(v) + p_G(v)] dv \qquad (3.22)$$

Call

$$\int_{v_1}\int_{v_2}\cdots\int_{v_n}[p_F(v) - p_G(v)]\,\alpha_j dv_1\cdots dv_n = b_j \tag{3.23}$$

and

$$\int_{v_1}\int_{v_2}\cdots\int_{v_n}\alpha_k\alpha_j[p_F(v) + p_G(v)]\,v_1\cdots dv_n = \gamma_{jk}\ (=\gamma_{kj}) \tag{3.24}$$

Then $b_j = \sum_{k=1}^{K} a_k\gamma_{jk}$ and $b = a\gamma,\ a = b\gamma^{-1}$ (3.25)

To form the matrix $\gamma$ and the vector b, we note that

$$b_j = E_f(\alpha_j) - E_g(\alpha_j) \tag{3.26a}$$

and

$$\gamma_{jK} = E_f(\alpha_j\alpha_K) + E_g(\alpha_j\alpha_K). \tag{3.26b}$$

Thus, if we have $M_f$ samples from class F and $M_g$ samples from class G, and if we call the value of $\alpha_k$ in the $m^{th}$ sample of class F, $(\alpha_{mk})_f$, then

$$\gamma_{jK} = \frac{1}{M_f}\sum_{m=1}^{M_f}(\alpha_{mk})_f(\alpha_{mj})_f + \frac{1}{M_g}\sum_{m=1}^{M_g}(\alpha_{mk})_g(\alpha_{mj})_g. \tag{3.27}$$

Let $A_f$ be an $M_f$ x K matrix $[(\alpha_{mj})_f]$.

Then $A_f^T A_f = [p_{jk}]$

$$\text{where } p_{jk} = \sum_{m=1}^{M_f} (\alpha_{mj})_f (\alpha_{mk})_f = p_{kj}. \tag{3.28}$$

$$\text{Hence } \gamma_{jK} = \frac{A_f^T A_f}{M_f} + \frac{A_g^T A_g}{M_g} . \tag{3.29}$$

The $\alpha_k$ are arranged according to increasing values of p, $\alpha_1 = 1$, and $E_f(\alpha_j) = E_f(\alpha_j \alpha_1)$. Hence $\beta$ may be formed by taking the difference between the first row of $\frac{A_f^T A_f}{M_f}$ and the first row of $\frac{A_g^T A_g}{M_g}$ .

This method was programmed for the Recomp II and tested on samples of one and two-dimensional random processes. The program is described in Appendix 3A. Experimental studies follow in Section 3.3.

### 3.2.3 Computational Methods for Non-Linear Discrimination

In the previous section it has been shown that the determination of the polynomial approximation to the optimum discriminant requires the solution of the set of linear equations

$$b = a\gamma \quad \text{where } \gamma \text{ is the symmetric matrix of moments} \tag{3.30}$$

$$\gamma_{jk} = \int \alpha_j \alpha_k (f + g)\, dv \quad \text{where f and g are the probability densities of F and G.}$$

and b is the vector, $b_j = \int \alpha_j (f-g)\, dv$

Many of the exact (non-iterative) methods for the solution of a set of linear equations with a symmetric matrix are based on the factorization of the matrix into two triangular matrices

$$\gamma = SS^T \text{ where S is a lower triangle.} \qquad (3.31)$$

Such a factorization is possible if the sequence of matrices

$$\left|\gamma_{11}\right| \quad , \quad \begin{vmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{vmatrix} \quad , \ldots \gamma \quad \text{are all non-singular.}$$

One such method is the square-root method in which S is computed by columns and two sets of linear equations are solved. $kS^T = b$, $aS = k$.

The procedure is as follows

$$S_{ii} = \sqrt{\gamma_{ii} - \sum_{k=1}^{i-1} S_{ik}^2} \qquad S_{ji} = \frac{\gamma_{ji} - \sum_{k=1}^{i-1} S_{ik} S_{jk}}{S_{ii}}$$

$$j = i+1,\ i+2 \ldots n$$

Then $k_i = \frac{1}{S_{ii}} \left( b_i - \sum_{k=1}^{i-1} S_{ik} k_k \right) \quad i = 1, 2 \ldots n$

$$\text{and } a_i = \frac{k_i - \sum_{j=i+1}^{n} S_{ji} k_j}{S_{ii}} \qquad i = n,\ n-1 \ldots 1$$

The number of multiplications required is

$$\frac{n(n+1)(2n+1)}{6} \quad \text{for S}$$

$$\frac{n(n+1)}{2} \quad \text{for K}$$

$$\frac{n(n+1)}{2} \quad \text{for a}$$

Another method in which the loss of significant figures is somewhat less troublesome is the so-called escalator method. Here S and $S^{-1} = \phi$ are computed simultaneously from the requirements:

1) $\phi S = I$

2) $\gamma \phi^T = S$ (since $\gamma = SS^T = S(\phi^T)^{-1} \Rightarrow \gamma \phi^T = S$)

Thus given i-1 rows of $\phi$ and i-1 columns of S, the $i^{th}$ row of $\phi$ is computed as follows:

$$\phi S = I \Rightarrow \sum_{k=j}^{i} \phi_{ik} S_{kj} = 0 \qquad (3.32)$$

and since $\phi_{ii} \neq 0$ $\dfrac{\phi_{ij}}{S_{ii}} = -\dfrac{1}{S_{jj}} \sum_{k=j+1}^{i} \left(\dfrac{\phi_{ik}}{\phi_{ii}}\right) S_{kj}$ $j = i-1, i-2 \ldots 1$ (3.33)

Compute $\phi_{ii} = \dfrac{1}{S_{ii}}$ from $\gamma \phi^T = S$

$$\phi_{ii} \sum_{k=1}^{j} \gamma_{ik} \left(\frac{\phi_{ik}}{\phi_{ii}}\right) = S_{ii} = \frac{1}{\phi_{ii}} \Rightarrow$$

$$\sum_{k=1}^{i} \gamma_{ik} \left(\frac{\phi_{ik}}{\phi_{ii}}\right) = \left(\frac{1}{\phi_{ii}}\right)^2 \qquad (3.34)$$

and compute $i^{th}$ column of S (Also using $\gamma\phi^T = S$)

$$\sum_{k=1}^{i} \gamma_{jk} \phi_{ik} = S_{ji} \quad j = i+1, \ldots n \qquad (3.35)$$

Now to find $a = b\gamma^{-1}$, use either $a = b\phi^T\phi$, or a $S = b\phi^T$.

The number of multiplications required is the same as for the square root method

$\dfrac{n(n+1)(n+2)}{6}$ for $\phi$

or $\dfrac{n(n+1)(2n+1)}{6}$ for $\phi$ and S

$\dfrac{n(n+1)(n-1)}{6}$ for S

and 2 $\left(\frac{n(n+1)}{2}\right)$ for a by either alternative.

## Orthonormal Polynomials

An interesting fact is exposed in the above method. The $i^{th}$ row of the matrix $\phi$ supplies the coefficients of the $i^{th}$ orthonormal polynomial with respect to the weighting function f + g. This is shown as follows:

Let the $i^{th}$ polynomial

$$\phi_i = \sum_{j=1}^{i} \phi_{ij} \alpha_j$$

and in order that the polynomials be orthonormal

$$\int \phi_i \phi_j (f+g)\, dv = \delta_{ij} \tag{3.36}$$

Thus the following equation must be satisfied

$$\int \sum_{n=1}^{j} \sum_{m=1}^{i} \phi_{im} \alpha_m \phi_{jn} \alpha_n (f+g)\, dv = \delta_{ij} \tag{3.37}$$

Interchanging summation and integration

$$\sum_{n=1}^{j} \sum_{m=1}^{i} \phi_{im} \phi_{jn} \int \alpha_m \alpha_n (f+g)\, dv = \delta_{ij} \tag{3.38}$$

and from the definition of $\gamma$

$$\sum_{n=1}^{j} \sum_{m=1}^{i} \phi_{im} \phi_{jn} \gamma_{mn} = \delta_{ij} \qquad (3.39)$$

Or in matrix notation

$$\phi \gamma \phi^T = I \text{ which is equivalent to } \gamma\phi^T = S = \phi^{-1}.$$

Thus the coefficients of the orthonormal polynomials of f + g are given by the rows of the matrix $\phi$.

The function $u = \frac{f-g}{f+g}$ may be approximated by a sum of the polynomials $\phi_i$; $u = \sum_{i=0}^{n} c_i \phi_i$, where $c_i = \int \phi_i u(f+g)\, dv = \int \phi_i (f-g)\, dv = (\phi b)_i$

That this form is equivalent to $u = \sum_{j=1}^{n} a_j \alpha_j$ is apparent from

$$u = \sum c_i \phi_i = \sum_i (\phi b)_i \phi_i = \sum_i (\phi b)_i \sum_j \phi_{ij} \alpha_j$$ where the coefficient of $\alpha_j$ is the $j^{th}$ element of the vector $b\phi^T\phi = a$.

Thus if we are interested in one dimensional problems a considerable simplification may be introduced since the orthonormal polynomials may be obtained recursively.

Let $\phi_{i+1} = v\, d\phi_i - \alpha\phi_i - \beta\,\phi_{i-1}$ ($\phi_{i+1}$ unnormalized)

then $\int \phi_i\phi_{i+1}\,(f+g)\,dv = 0 \Rightarrow \alpha_i = \int v\phi_i^2\,(f+g)\,dv$ (3.40)

$$\int \phi_{i-1}\,\phi_{i+1}(f+g)\,dv = 0 \Rightarrow \beta_i = \int v\,\phi_i\phi_{i-1}(f+g)\,dv \qquad (3.41)$$

$\phi_{i+1}$ is now automatically orthogonal to all $\phi_j$ ($j < i-1$) since

$\int \phi_{i+1}\,\phi_j\,dv = \int v\phi_j\phi_i\,dv - \alpha 0 - \beta 0$ and $v\,\phi_j$ is of degree less than i, and may be expressed as a sum of polynomials of degree $\leq$ i-1, hence $\int v\,\phi_j\phi_i\,dv = 0$. The expressions for $\alpha$ and $\beta$ may be still further simplified

$$\alpha_i = \int v\,\phi_i^2\,(f+g)\,dv$$

$$= \sum_{m=0}^{i}\sum_{n=0}^{i} \phi_{im}\,\phi_{in} \int v^{m+n+1}\,(f+g)\,dv$$

$$= \phi_{i,i-1} \sum_{n=0}^{i} \phi_{in} \int v^{i+n} (f+g)\, dv + \phi_{ii} \sum_{n=0}^{i} \phi_{in} \int v^{i+1+n} (f+g)\, dv$$

$$= \frac{\phi_{i,i-1}}{\phi_{ii}} + \phi_{ii} \int v^{i+1} \phi_i (f+g)\, dv \qquad (3.42)$$

$$\beta_i = \sum_{m=0}^{i-1} \sum_{n=0}^{i} \phi_{i-1,m} \phi_{in} \int v^{m+n+1} (f+g)\, dv$$

$$= \phi_{i-1,i-1} \sum_{n=0}^{i} \phi_{in} \int v^{i+n} (f+g)\, dv$$

$$= \frac{\phi_{i-1,i-1}}{\phi_{ii}} \qquad (3.43)$$

It is possible that simplification may also be effected in the multivariate case. The method is not, however, immediately apparent.

## 3.3 EXPERIMENTAL STUDIES

### 3.3.1 Univariate Examples

Random samples from known densities were used to test the rate of convergence of the polynomial approximation of the discriminant $u(x) = \frac{f(x) - g(x)}{f(x) + g(x)}$ as a function of both the degree of the polynomial and the number of samples used in "learning" (estimating the moments).

a. Normal distributions

Polynomials were computed using random samples from normal distributions with a variety of means and variances. To test the functions, error rates were computed using a different group of random samples. As a control we also computed the error rates using the likelihood ratio computed from the sample means and variances. The results are tabulated in Table 3.1.

From these results we observe that for small sample sizes (10-20 samples in this case) the lower degree polynomials (second and third) give error rates closer to those of the estimated likelihood ratio. As the sample size increases there is little difference in error rates as the degree of the polynomial is increased. It is, of course, expected that higher degree polynomials will have little meaning for small sample sizes since the higher moments are not well estimated. In addition, a higher degree polynomial is not actually necessary in this case since the function $\frac{f-g}{f+g}$ has only two roots. The only advantage to be gained from a high degree polynomial is a small improvement in the location of the real roots of the polynomial. Although these results are only qualitative they are encouraging.

Table 3.1a Normal Distribution Results $\mu_F = 0$, $\sigma_F = 1$ $\mu_G = 1.5$ $\sigma_G = 0.7$

(Expected Error Rates, $e_F = .24$, $e_G = .14$

| No. of Samples to form functions | Degree of Polynomial | Error for Samples from F | Error for Samples from G |
|---|---|---|---|
| 10 | 1 | .150000 | .150000 |
| | 2 | .120000 | .220000 |
| | 3 | .140000 | .180000 |
| | 4 | .120000 | .220000 |
| | 5 | .130000 | .330000 |
| | Est. likelihood ratio | .180000 | .110000 |
| 20 | 1 | .160000 | .240000 |
| | 2 | .160000 | .260000 |
| | 3 | .170000 | .250000 |
| | 4 | .150000 | .280000 |
| | 5 | .160000 | .220000 |
| | Est. likelihood ratio | .160000 | .220000 |
| 30 | 1 | .160000 | .270000 |
| | 2 | .150000 | .290000 |
| | 3 | .180000 | .250000 |
| | 4 | .170000 | .260000 |
| | 5 | .160000 | .240000 |
| | Est. likelihood ratio | .170000 | .230000 |
| 40 | 1 | .160000 | .200000 |
| | 2 | .140000 | .220000 |
| | 3 | .170000 | .220000 |
| | 4 | .160000 | .230000 |
| | 5 | .170000 | .200000 |
| | Est. likelihood ratio | .170000 | .200000 |

able 3. 1a (Contd.)

| No. of Samples to form functions | Degree of Polynomial | Error for Samples from F | Error for Samples from G |
|---|---|---|---|
| 50 | 1 | .280000 | .230000 |
| | 2 | .260000 | .250000 |
| | 3 | .310000 | .220000 |
| | 4 | .300000 | .230000 |
| | 5 | .300000 | .220000 |
| | Est. likelihood ratio | .300000 | .170000 |
| 60 | 1 | .250000 | .190000 |
| | 2 | .250000 | .190000 |
| | 3 | .260000 | .190000 |
| | 4 | .260000 | .190000 |
| | 5 | .250000 | .190000 |
| | Est. likelihood ratio | .280000 | .150000 |
| 70 | 1 | .180000 | .180000 |
| | 2 | .190000 | .140000 |
| | 3 | .210000 | .180000 |
| | 4 | .200000 | .190000 |
| | 5 | .190000 | .190000 |
| | Est. likelihood ratio | .190000 | .170000 |
| 80 | 1 | .150000 | .110000 |
| | 2 | .170000 | .090000 |
| | 3 | .160000 | .090000 |
| | 4 | .160000 | .100000 |
| | 5 | .160000 | .100000 |
| | Est. likelihood ratio | .160000 | .090000 |
| 90 | 1 | .230000 | .120000 |
| | 2 | .240000 | .100000 |
| | 3 | .240000 | .110000 |
| | 4 | .240000 | .120000 |
| | 5 | .240000 | .120000 |
| | Est. likelihood ratio | .240000 | .100000 |

Table 3. 1a (Contd.)

| No. of Samples to form functions | Degree of Polynomial | Error for Samples from F | Error for Samples from G |
|---|---|---|---|
| 100 | 1 | .210000 | .170000 |
| | 2 | .250000 | .130000 |
| | 3 | .250000 | .140000 |
| | 4 | .240000 | .140000 |
| | 5 | .250000 | .140000 |
| | Est. likelihood ratio | .250000 | .130000 |

Table 3.1b Normal Distribution ( $\mu_F = 0$, $\sigma_F = 1$, $\mu_G = 1$, $\sigma_G = .7$)

Expected Error Rates $e_F = .36$, $e_G = .18$

| No. of Samples to form functions | Degree of Polynomial | Error for Samples from F | Error for Samples from G |
|---|---|---|---|
| 10 | 1 | .200000 | .400000 |
| | 2 | .260000 | .280000 |
| | 3 | .230000 | .340000 |
| | 4 | .180000 | .380000 |
| | 5 | .240000 | .410000 |
| | Est. likelihood ratio | .260000 | .230000 |
| 20 | 1 | .310000 | .290000 |
| | 2 | .340000 | .240000 |
| | 3 | .340000 | .240000 |
| | 4 | .260000 | .360000 |
| | 5 | .250000 | .440000 |
| | Est. likelihood ratio | .390000 | .160000 |
| 30 | 1 | .180000 | .200000 |
| | 2 | .240000 | .160000 |
| | 3 | .230000 | .160000 |
| | 4 | .210000 | .200000 |
| | 5 | .220000 | .210000 |
| | Est. likelihood ratio | .270000 | .140000 |
| 40 | 1 | .340000 | .240000 |
| | 2 | .390000 | .140000 |
| | 3 | .390000 | .140000 |
| | 4 | .390000 | .130000 |
| | 5 | .390000 | .130000 |
| | Est. likelihood ratio | .410000 | .100000 |

Table 3.1 b (Contd.)

| No. of Samples to form functions | Degree of Polynomial | Error for Samples from F | Error for Samples from G |
|---|---|---|---|
| 50 | 1 | .200000 | .260000 |
| | 2 | .250000 | .240000 |
| | 3 | .250000 | .240000 |
| | 4 | .190000 | .260000 |
| | 5 | .180000 | .310000 |
| | Est. likelihood ratio | .340000 | .150000 |

b. Other than Normal Distributions

$$f(x) = \frac{1}{\pi}(1-x^2)^{-1/2} \qquad -1 < x < 1$$

$$g(x) = 0 \qquad -1 < x < 0$$

are the assumed probability densities

$$g(x) = \frac{4}{\pi}(1-x^2)^{1/2} \qquad 0 \leq x < 1$$

We computed the polynomials using sample moments and tested them with a different set of 250 random samples. As a control we used a function having the same roots as $\frac{f-g}{f+g}$. These are 0 and .866. The results are tabulated in Table 3.2. Here we note that when 50 samples were used in forming the functions the error rates obtained are quite large. Thus it would appear that larger sample sizes are necessary when attempting to approximate a discontinuous function. As in the previous case, there does not appear to be a great advantage in increasing the degree of the polynomial beyond three or four. Again the reason is that there are only two roots of $\frac{f-g}{f+g}$.

In order to determine more closely the effect of increasing the degree of the polynomial we also computed the polynomial discriminants using the true moments of f and g. This was done for the discontinuous functions obtained from

$f(x) = \frac{1}{\pi}(1-x^2)^{-1/2}$, $-1 < x < 1$ and $g(x) = \frac{4}{\pi}(1-x^2)^{1/2}$, $0 \leq x < 1$. (See Figure 3-1) and for the continuous function obtained from the same f and g, with g now defined over the entire range (-1, 1) (see Figure 3-2). An unfortunate characteristic of this method of discrimination is apparent in these graphs.

In approximating a discontinuous function, as in the example, a high degree polynomial is required to fit the discontinuities. For this application all we really desire is a good approximation to the zero's of the polynomial, a problem we will discuss later.

Table 3.3 is a comparison of the coefficients of the polynomials obtained from true moments and from estimated moments (200 samples). We note that the coefficients are quite close, as are the moments with one exception. The odd moments of f are weighted toward the positive side. Since these are the only moments which are noticably in error, it is reasonable to attribute the discrepancies to some peculiarity of the method of generation.

Table 3.2 Non-Linear Technique Using

$$f(x) = \frac{1}{\pi} (1-x^2)^{-1/2} \qquad -1 < x < +1$$

$$g(x) = \frac{4}{\pi} (1-x^2)^{1/2} \qquad 0 \leq x < 1 \quad \text{(0 elsewhere)}$$

250 samples to test

| No. of Samples to estimate moments | Degree of Polynomial | Error Rate Samples of f | Error Rate Samples of g |
|---|---|---|---|
| | 1 | .408000 | .440000 |
| | 2 | .360000 | .080000 |
| 50 | 3 | .308000 | .128000 |
| | 4 | .296000 | .204000 |
| | 5 | .292000 | .204000 |
| | x(x-.866) | .336000 | .040000 |
| | 1 | .368000 | .368000 |
| | 2 | .424000 | .120000 |
| 100 | 3 | .316000 | .100000 |
| | 4 | .316000 | .104000 |
| | 5 | .300000 | .128000 |
| | x(x-.866) | .332000 | .060000 |
| | 1 | .416000 | .348000 |
| | 2 | .444000 | .092000 |
| 150 | 3 | .328000 | .104000 |
| | 4 | .320000 | .108000 |
| | 5 | .308000 | .132000 |
| | x(x-.866) | .364000 | .068000 |
| | 1 | .336000 | .292000 |
| | 2 | .360000 | .060000 |
| 200 | 3 | .252000 | .060000 |
| (used in comparison) | 4 | .256000 | .064000 |
| | 5 | .232000 | .088000 |
| | x(x-.866) | .276000 | .040000 |
| | 1 | .420000 | .344000 |
| | 2 | .352000 | .088000 |
| 250 | 3 | .328000 | .076000 |
| | 4 | .324000 | .080000 |
| | 5 | .316000 | .080000 |
| | x(x-.866) | .376000 | .044000 |

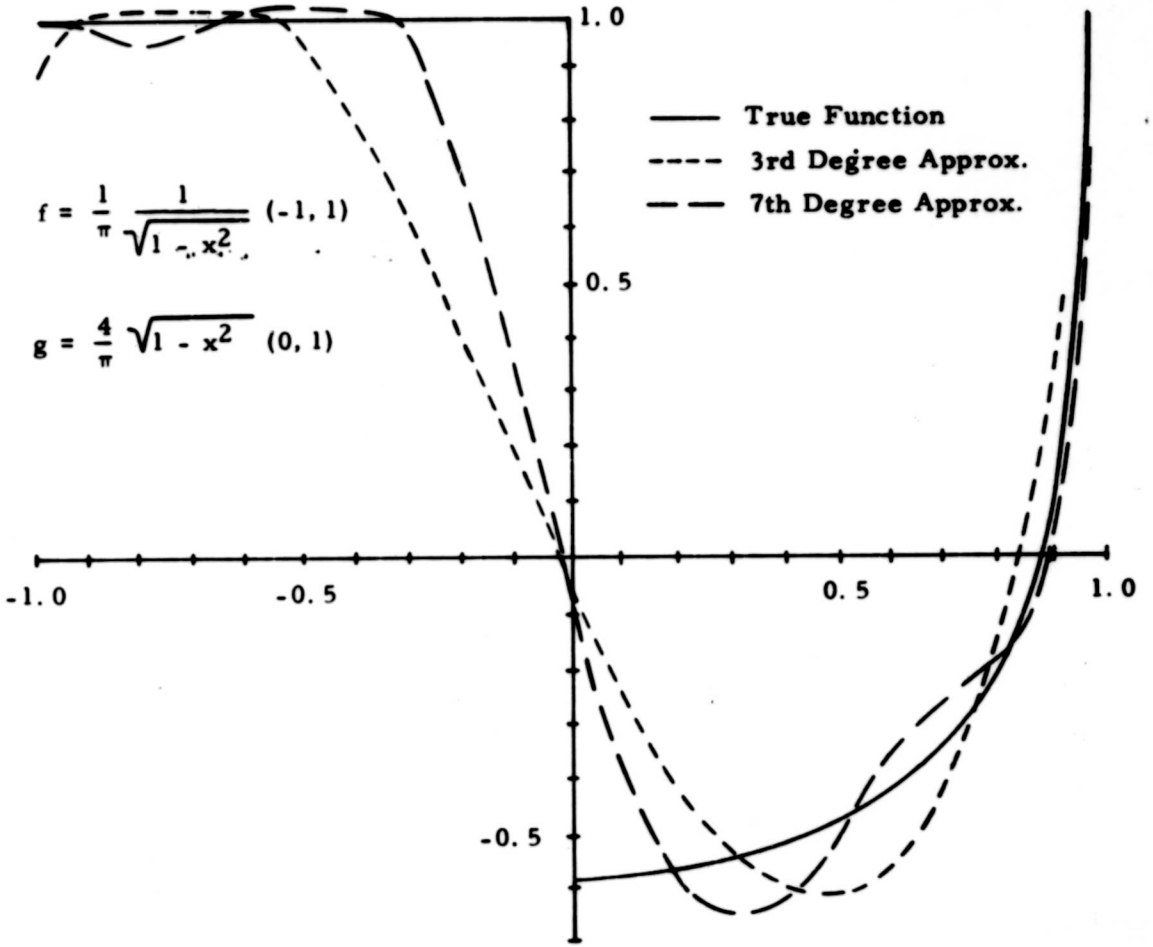



Figure 3-1. Comparison of $\frac{f-g}{f+g}$ and Polynomial Approximations Using True Moments

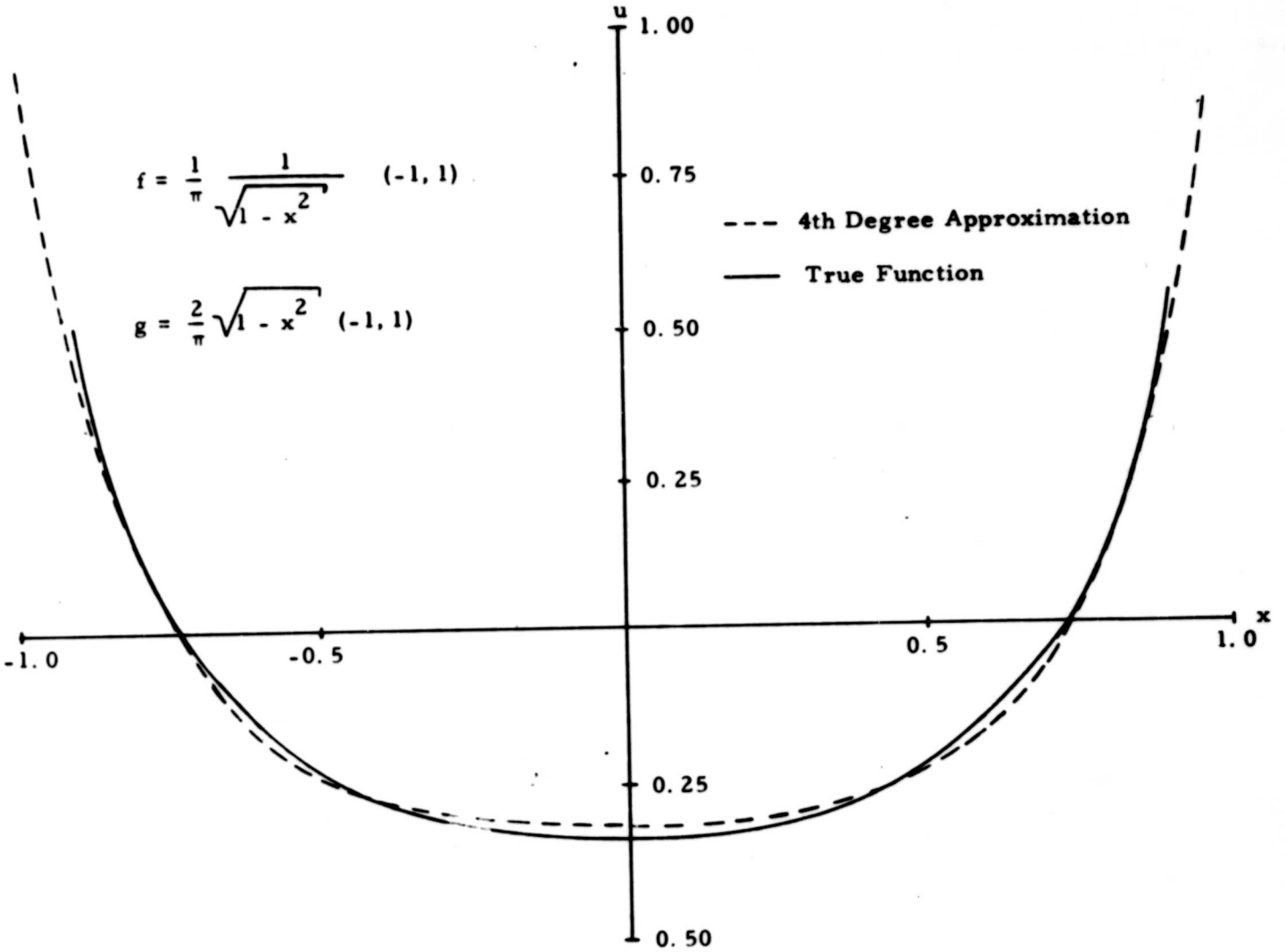



Figure 3-2. Comparison of $\frac{f-g}{f+g}$ and Polynomial Approximation Using True Moments

Table 3. 3

Comparison of Results Using True Moments and Using

Estimated Moments for $f = \frac{1}{\pi}(1-x^2)^{-1/2}(-1, 1)$ $g = \frac{4}{\pi}(1-x^2)^{1/2}(0, 1)$

| Moments | 1 | $\bar{x}$ | $\overline{x^2}$ | $\overline{x^3}$ | $\overline{x^4}$ | $\overline{x^5}$ | $\overline{x^6}$ | $\overline{x^7}$ |
|---|---|---|---|---|---|---|---|---|
| True Moments of f | 1 | 0 | .5000 | 0 | .3750 | .0 | .3125 | 0 |
| Estimated Moments of f | 1 | .0976 | .4871 | .0668 | .3795 | .0517 | | |
| True Moments of g | 1 | .4244 | .2500 | .1698 | .1250 | .0970 | .0781 | .0647 |
| Estimated Moments of g | 1 | .4234 | .2576 | .1681 | .1241 | .0966 | | |

| | Polynomials | | | | | | |
|---|---|---|---|---|---|---|---|
| Coefficient of | 1 | x | $x^2$ | $x^3$ | $x^4$ | $x^5$ | Zeros |
| From True Moments, p = 3 | -.0368 | -1.0644 | +.4636 | +1.0459 | | | -.07, +.83 |
| From Est Moments, p = 3 | +.0063 | -1.1283 | +.4256 | +1.1177 | | | +.006, +.83 |
| From True Moments, p = 5 | -.0268 | -1.2749 | +.5037 | +1.8957 | -.0752 | -.6929 | -.02, +.811 |
| From Est Moments p = 5 | +.0246 | -1.2830 | +.3234 | +1.8213 | +.0817 | -.5878 | +.019, +.812 |
| Zeros of $\frac{f-g}{f+g}$ | | | | | | | 0, +.866 |

### 3.3.2 A Two Dimensional Example

To test the non-linear method in more than one dimension, we used an artificial example in two dimensions with bimodal densities. Plots of the decision regions for various degree polynomials are shown in Figures 3-3 to 3-7. It can be seen that all the polynomials of degrees two to six will separate the two classes with a very small percentage of error (one point out of 111). Increasing the degree of the polynomial simply brings the decision regions closer to the actual clusters of points. From the graphs and from Table 3.4 listing coefficients of the polynomials, it is apparent that the function has not begun to converge. Thus, accurate decisions may be made with lower degree polynomials than are required to closely approximate the decision function $u = \frac{f-g}{f+g}$.

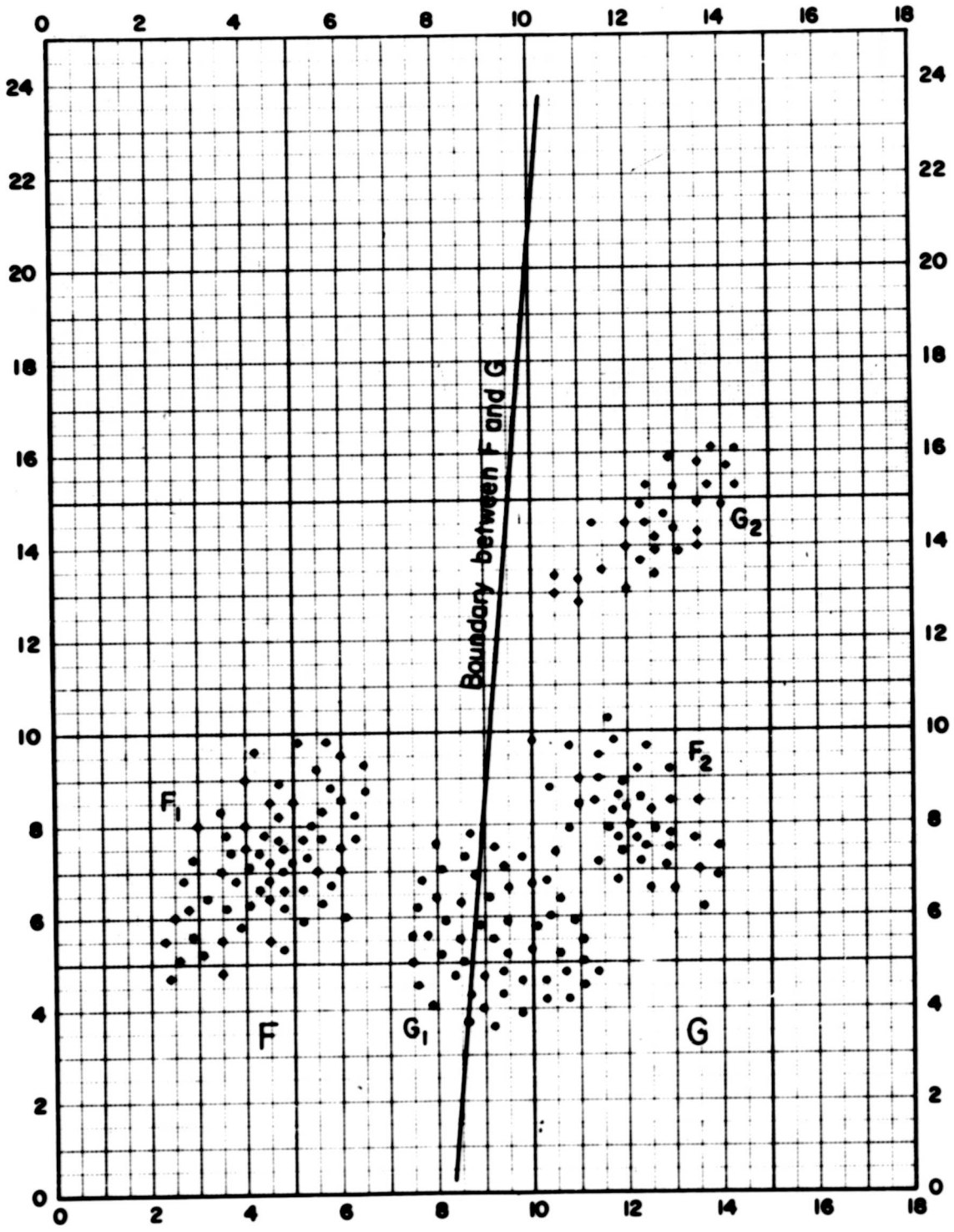



Figure 3-3. Polynomial Discriminant, R = 1

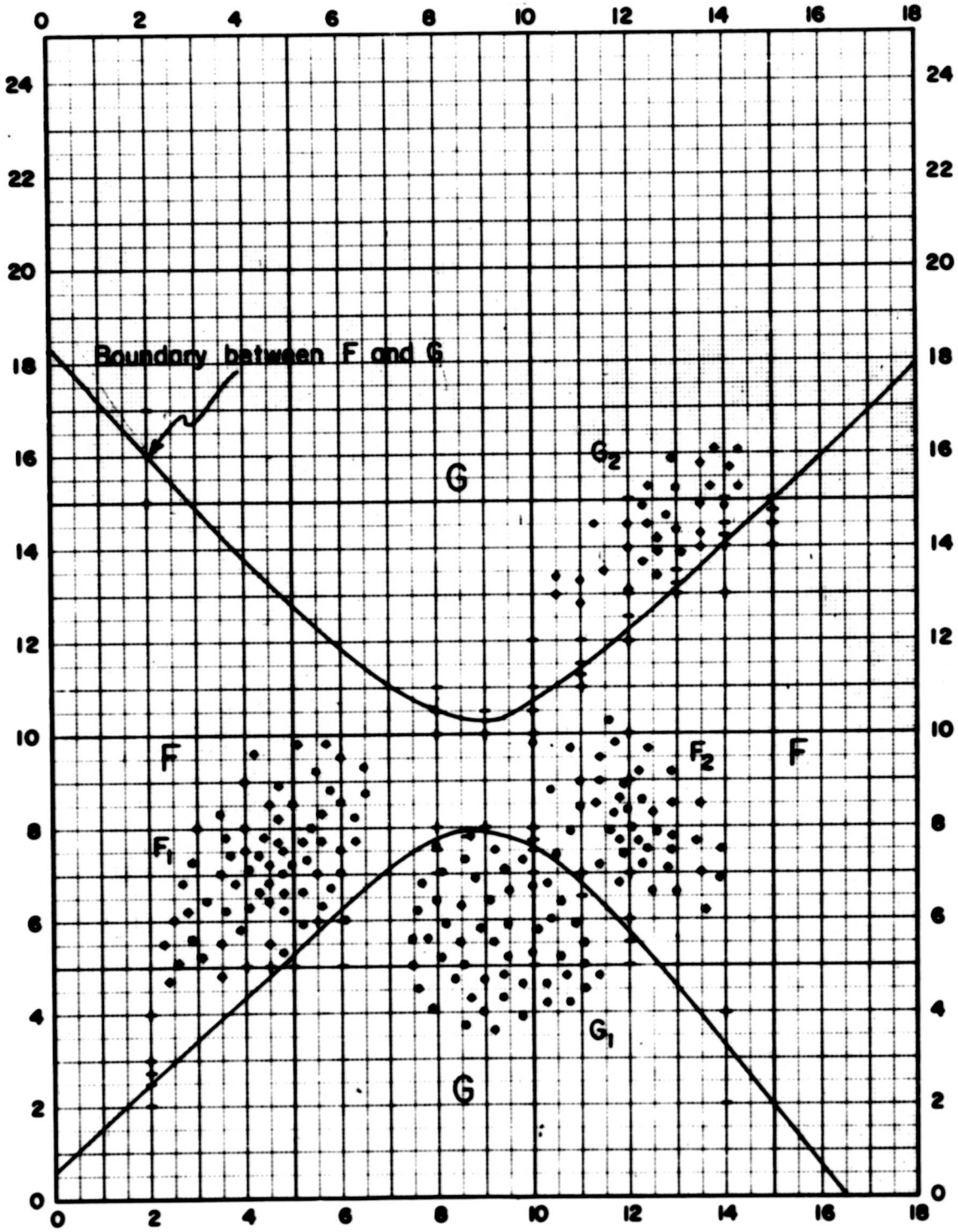



Figure 3-4. Polynomial Discriminant, R = 2

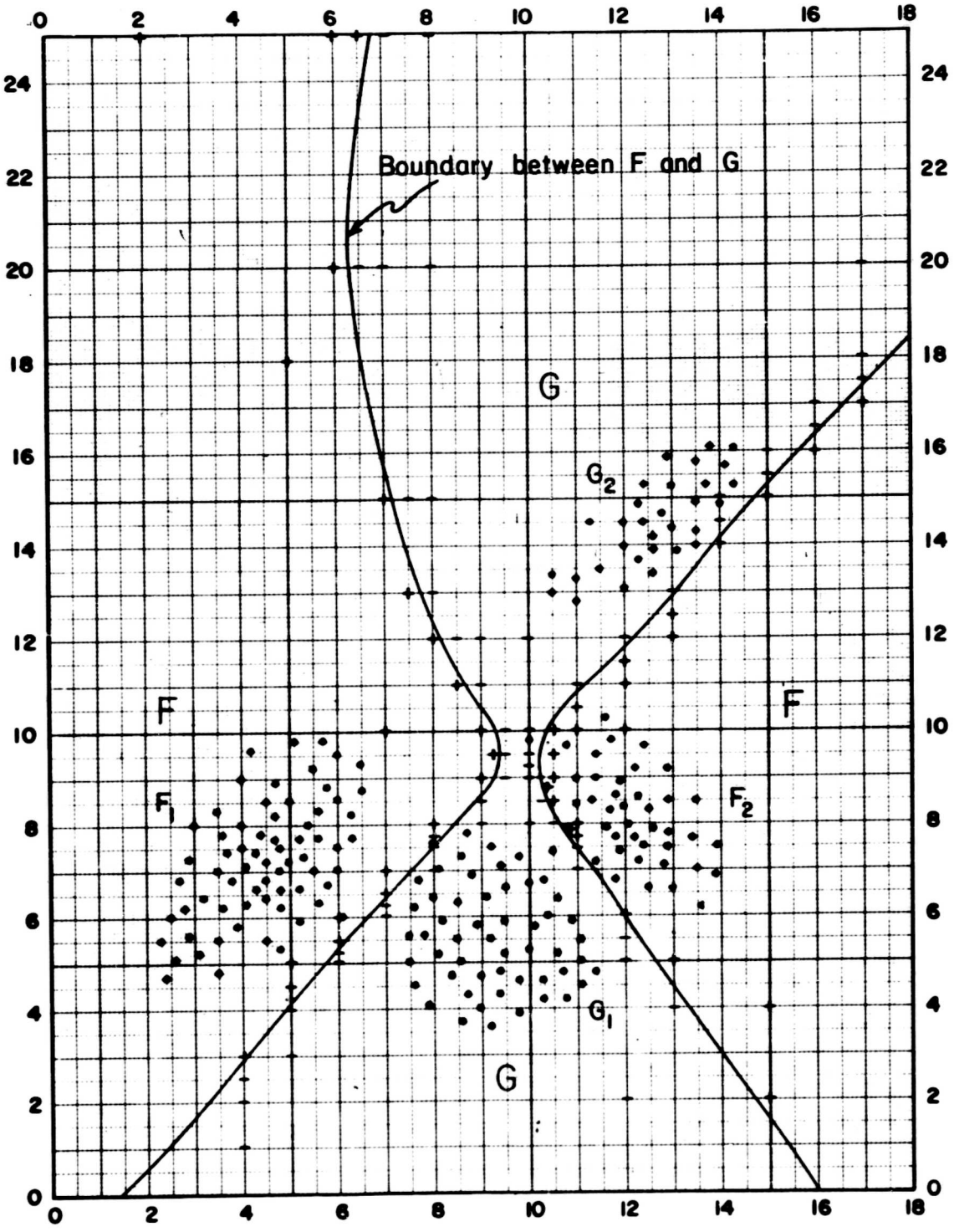



Figure 3-5. Polynomial Discriminant, R = 3

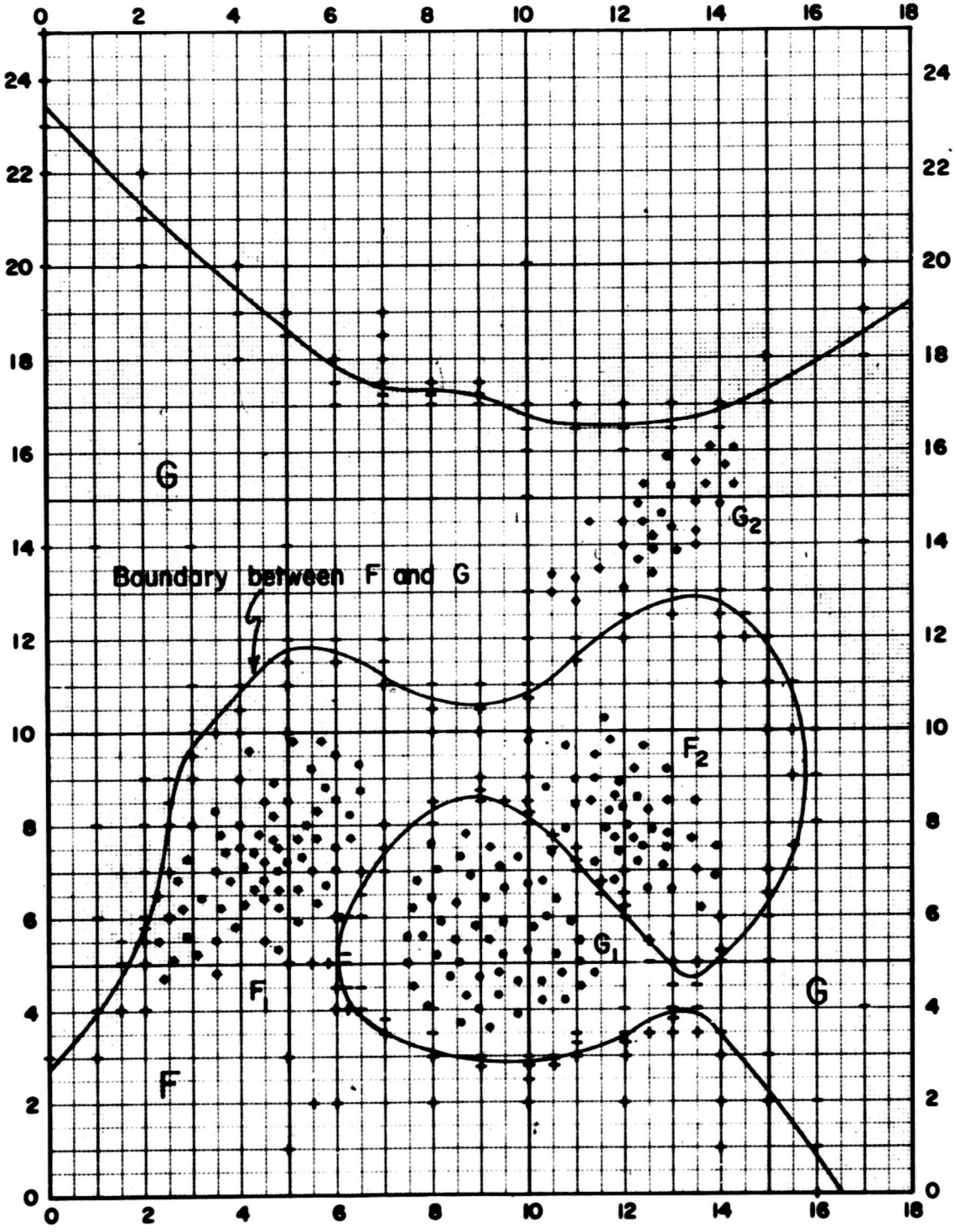



Figure 3-6. Polynomial Discriminant, R = 4

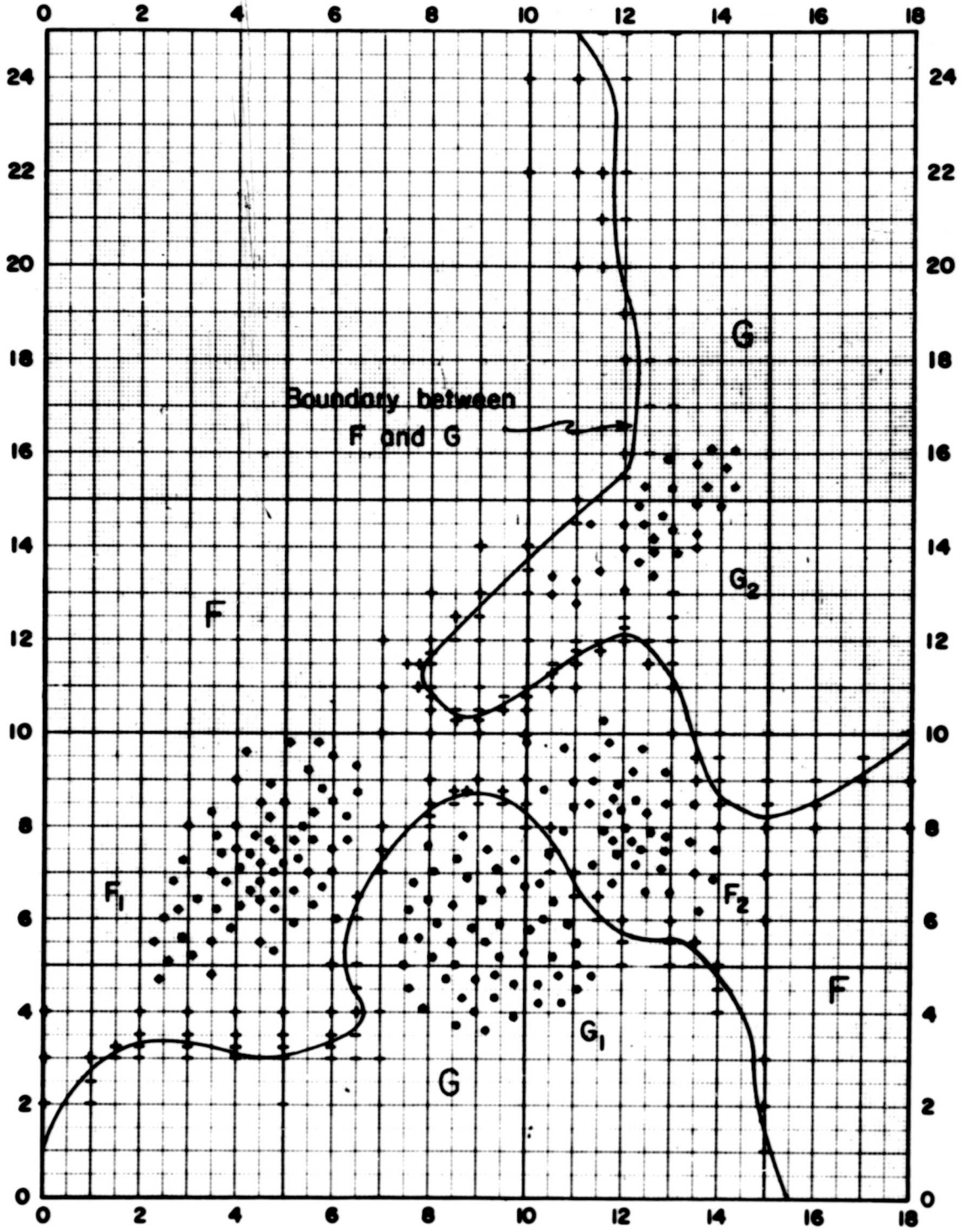



Figure 3-7. Polynomial Discriminant, R = 6

## Table 3.4

### Coefficients of Polynomials

### 2 Dimensional Example

| Coefficient of | Degree of Polynomial 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1.13708 | -.75964 | 1.10401 | 16.55251 | -7.14585 | -46.99235 |
| x | -.36626 | -.94135 | -.35972 | 3.03229 | 6.87856 | - 8.71724 |
| y | .010997 | 1.09983 | .152413 | -9.42789 | - .264435 | 42.99605 |
| $x^2$ | | .054570 | -.10027 | -1.03109 | .249397 | 4.20525 |
| xy | | -.0029499 | .14440 | .39371 | -3.37622 | .230318 |
| $y^2$ | | -.058878 | -.034275 | 1.48893 | 1.273182 | -13.70016 |
| $x^3$ | | | .0075139 | .098045 | -.179408 | -.944689 |
| $x^2y$ | | | -.0041719 | -.209358 | .227650 | .224987 |
| $xy^2$ | | | -.0047572 | -.0036441 | .409823 | -.455595 |
| $y^3$ | | | .0013776 | .109865 | -.235561 | 2.37556 |
| $x^4$ | | | | -.0030357 | .014483 | .119397 |
| $x^3y$ | | | | .0013612 | .0068251 | -.105636 |
| $x^2y^2$ | | | | - .00099659 | -.043274 | .118663 |
| $xy^3$ | | | | .00076929 | .00040861 | -.00028709 |
| $y^4$ | | | | .0026330 | .0097179 | .210432 |
| $x^5$ | | | | | -.00045213 | -.0079295 |
| $x^4y$ | | | | | .000082047 | .0116804 |
| $x^3y^2$ | | | | | -.00068595 | -.0092108 |
| $x^2y^3$ | | | | | .00266577 | -.00317675 |
| $xy^4$ | | | | | -.0016180 | .00230716 |
| $y^5$ | | | | | .00023290 | .00871901 |
| $x^6$ | | | | | | .000208697 |
| $x^5y$ | | | | | | -.00041510 |
| $x^4y^2$ | | | | | | .000405810 |
| $x^3y^3$ | | | | | | -.00029510 |
| $x^2y^4$ | | | | | | .00049388 |
| $xy^5$ | | | | | | -.00035191 |
| $y^6$ | | | | | | -.000071513 |

# IV INCOMPLETE AND NOISY DATA

## 4.1 CLASSIFICATORY DECISIONS BASED ON INCOMPLETE SETS OF OBSERVATIONS

### 4.1.1 Introduction

The problem of classificatory analysis (that of making multiple, binary-choice decisions) has been treated in the literature[8] as a problem of optimally deciding, on the basis of N observed measurements $(x_1, x_2, \ldots, x_N)$, which of several classes produces the particular set of N measurements. If the joint probability densities of the N measurements are known under all assumptions of class membership for the set of N measurements, classificatory decisions can be rendered by computing the N-dimensional likelihood ratios and then comparing these ratios with each other or with a threshold. With the joint probability densities not known, but a finite number of N-tuple measurements of each class available, decision rules have been devised and tested which approach the likelihood ratio computation in the limit, as the number of N-tuple measurements approaches infinity.

The problem which will be considered in this section concerns the method of making the optimum, multiple, binary-choice decision when not all of the N-measurable parameters of an N-dimensional process are available. Two practical applications of this problem are listed below.

Application to Classification Problems

If there are N observables of an object to be classified as a member of one of several classes, then the ratio of a posteriori probabilities for all possible sets of N values of the observables is known. If there are only

---

[8] Sebestyen, op. cit.

N-K measurements on an object because of malfunction of K instruments which normally measure K different observables, how are the N-K given measurements to be used, in an optimum decision?

Application to Detection Problems

If the presence or absence of a specific waveform of 2TW product equal to N is to be detected in additive noise, optimum detection consists of comparing the values of the joint probability density of 2TW samples of the received signal under two assumed conditions: first, that the value belongs to the "desired waveform plus noise" population, second, that the value belongs to "noise alone". The resulting computation or likelihood ratio is usually instrumented by correlating the 2TW samples of the input signal with a replica of the sequence of 2TW samples of the desired signal. However, the problem under present consideration is how to detect optimally the presence of the desired waveform if it is known that only 2TW - K samples were transmitted; that is, a correlator for the receipt of a specific T-long signal is constructed, with the fact known that not all of the signal was transmitted; part of it was interrupted, jammed, or is missing.

The question raised by this problem is: whether knowledge of the higher order densities can or cannot help to make multiple, binary-choice decisions if the measurements needed for insertion into these densities are not all actually available.

If the following discussion, several ways of utilizing knowledge of the higher order densities are explored, and the optimum method of making the classificatory decisions on N-K measurements is derived. The different methods of making classificatory decisions will be stated in the form of Decision Rules.

### 4.1.2 Decision Rules

Without any loss of generality, it will be assumed in the following that the set of N-K measurements may be a member of only one of two classes A or B, that the cost of deciding that the set belongs to A when indeed it is a member of B (The cost of false alarm) is $c_1$, and the cost of deciding in favor of B when actually the set belongs to A is $c_2$ (the cost of false dismissal).

If all the N measurements on an input were available, and $c_1$ were equal to $c_2$, an imminently reasonable way of making classificatory decisions would be decided by the following equation:

$$\text{Decide } (x_1, x_2, \ldots, x_N) \text{ belongs to A iff } \frac{p(A/x_1, x_2, \ldots, x_N)}{p(B/x_1, x_2, \ldots, x_N)} > 1 \qquad (4.1)$$

This decision states that if, given $x_1, x_2, \ldots, x_N$, A is more likely than B, then one should decide that the N-dimensional vector x belongs to A. The opposite decision is made of the inequality does not hold. By applying Bayes' Rule, Equation 4.1 may be written as shown in Equations 4.2a and 4.2b, where P(A) and P(B) are the a priori probabilities of occurrence of class A and B, respectively,

$$\text{Decide } x \in A \text{ if } \frac{p(A/x_1, x_2, \ldots, x_N)}{p(B/x_1, x_2, \ldots, x_N)} = \frac{p_A(x_1, x_2, \ldots, x_N)P(A)}{p_B(x_1, x_2, \ldots, x_N)\,P(B)} > 1 \qquad (4.2a)$$

or

$$\frac{p_A(x_1, x_2, \ldots, x_N)}{p_B(x_1, x_2, \ldots, x_N)} = \ell(x) > \frac{P(B)}{P(A)} = T = \text{constant threshold} \qquad (4.2b)$$

The function $\ell(x)$ is the N-dimensional likelihood ratio. Several other likelihood ratios will be introduced later; they will be distinguished from each other by subscripts which will indicate the Decision Rule with which the likelihood ratios are associated.

Suppose that of the N-measurable parameters on which decisions should be based, only N-K are available. In the following, four of several possible decision rules are discussed for deciding, from available measurements, which of the two classes, A or B, is most likely.

Decision Rule 1 (Decision Based on the Marginal Densities)

This rule states that, given the measurements $(x_1, x_2, \ldots, x_{N-K})$, decide that they belong to A if A is more likely than B as follows:

$$\text{Decide } (x_1, x_2, \ldots, x_{N-K}) \in \text{A iff } \frac{p(A/x_1, x_2, \ldots, x_{N-K})}{p(B/x_1, x_2, \ldots, x_{N-K})} > 1 \qquad (4.3)$$

By employing Bayes' Rule, this equation may be rewritten as:

$$\frac{p(A/x_1, x_2, \ldots, x_{N-K})}{p(B/x_1, x_2, \ldots, x_{N-K})} = \frac{P(A)}{P(B)} \frac{p_A(x_1, x_2, \ldots, x_{N-K})}{p_B(x_1, x_2, \ldots, x_{N-K})} > 1 \qquad (4.4a)$$

or

$$\text{Decide } (x_1, x_2, \ldots, x_{N-K}) \in \text{A iff } \frac{p_A(x_1, x_2, \ldots, x_{N-K})}{p_B(x_1, x_2, \ldots, x_{N-K})} > \frac{P(B)}{P(A)} = T \qquad (4.4b)$$

Let $\frac{p_A(x_1, x_2, \ldots, x_{N-K})}{p_B(x_1, x_2, \ldots, x_{N-K})} = \ell_1(x_1, x_2, \ldots, x_{N-K})$ be the N-K dimensional likelihood ratio, abbreviated as $\ell_1(x)$ to simplify the notation. This likelihood ratio is the ratio of marginal densities of the actually observed variables. Decision Rule 1 states: to determine if x should be classed in A or B, compare $\ell_1(x)$ with a threshold; which in effect, means that since $x_1, x_2, \ldots, x_{N-K}$ are the only measurements made, these measurements alone should be the basis for decision.

Decision Rule 2 (Decision Uses the Most Probable Values of the Missing Measurements $x_{N-K+1}$ through $x_N$)

After the N-K measurements $(x_1, x_2, \ldots, x_{N-K})$ are made, the probability density of the missing K measurements ($x_{N-K+1}$ through $x_N$) can be calculated and most probable values of the K measurements chosen for use in the N-dimensional likelihood ratio $\ell(x)$. The value of the ratio $\ell(x)$ when the most probable values of the K missing measurements are used is $\ell_2(x)$. The most probable values are those which maximize the probability density given in Equation 4.5 and are denoted by $\ell_2(\hat{x})$.

$$p(x_{N-K+1}, x_{N-K+2} \cdots x_N / x_1, x_2, \ldots, x_{N-K}) \tag{4.5a}$$

$$p(\hat{x}_{N-K+1}, \ldots, \hat{x}_N / x_1, \ldots, x_{N-K}) \geq p(x_{N-K+1}, \ldots, x_N / x_1, \ldots, x_{N-K}) \tag{4.5b}$$

The decision rule states to

$$\text{Decide } (x_1, \ldots, x_{N-K}) \in A \text{ iff } \ell_2(x) = \frac{p_A(x_1, \ldots, x_{N-K}, \hat{x}_{N-K+1}, \ldots, \hat{x}_N)}{p_B(x_1, \ldots, x_{N-K}, \hat{x}_{N-K+1}, \ldots, \hat{x}_N)} \frac{P(B)}{P(A)} = T \tag{4.6}$$

This rule predicts the most likely values of the missing measurements and uses them as if they had actually been measured.

Decision Rule 3 (Decision uses Most Probable Value of Likelihood Ratio)

When only N-K measurements $(x_1, x_2, \ldots, x_{N-K})$ are made, the likelihood ratio $\ell(x)$ is a random variable, a function of the unmeasured random variables $x_{N-K+1}$ through $x_N$. To indicate this fact, the random likelihood ratio is denoted by $\ell_3(x)$ which is defined in Equation 7.

$$\ell_3(x) = \frac{p_A(x_{N-K+1}, \ldots, x_N / x_1, \ldots, x_{N-K})}{p_B(x_{N-K+1}, \ldots, x_N / x_1, \ldots, x_{N-K})} \tag{4.7}$$

There is a probability density $p(\ell_3(x))$ associated with $\ell_3(x)$. The most probable value of the likelihood ratio, given the observed $x_1, \ldots, x_{N-K}$ measurements, can now be chosen; that is, a decision can be based on the most probable value of the likelihood ratio $\ell_3(x)$, defined as follows:

$$p(\hat{\ell}_3(x)) \geq p(\ell_3(x)) \tag{4.8}$$

The decision rule is to

$$\text{Decide } (x_1, \ldots, x_{N-K}) \in A \text{ iff } \hat{\ell}_3(x) > T \tag{4.9}$$

**Decision Rule 4 (Decision Uses Average Value of Likelihood Ratio)**

In this decision rule, the likelihood ratio is again treated as a random variable $\ell_3(x)$, a function of the missing K measurements. Of course, the observed N-K measurements are known parameters in $\ell_3(x)$. Instead of using in this rule the most probable value, the average value of the likelihood ratio is used with $x_1, \ldots, x_{N-K}$, given as a basis for deciding between A and B.

$$\text{Decide } (x_1, \ldots, x_{N-K}) \in A \text{ iff } E[\ell_3(x)] > T \tag{4.10a}$$

where

$$E[\ell_3(x)] = \int_{-\infty}^{\infty} \ell_3(x)\, p(\ell_3(x))\, d\ell_3 \tag{4.10b}$$

In certain very seldom occurring situations, where $\ell_3(x)$ is a monotonic function of $x_{N-K+1}$ through $x_N$, Equation (4.10b) may be written as follows:

$$E[\ell_3(x)] = \int_{-\infty}^{\infty} \cdots \int \int \ell_3(x_{N-K+1}, \ldots, x_N \; p(x_{N-K+1}, \ldots, x_N / x_1, \ldots, x_{N-K}) dx_{N-K+1}, \ldots, dx_N \tag{4.11}$$

4.1.3 Comparison of Decision Rules

To compare the different decision rules, the probabilities of error (both kinds) are computed so as to determine which decision rule results in the smallest probability of error. The two error rates, probability of false alarms and probability of false dismissals, P(FA) and P(FD), are given in Equations (4.12) and (4.13), where Y is the region of the N-dimensional space within which the decision rule in vogue decides that

$$P(FA) = \iint_{Y} \cdots \int P_B(x_1, x_2, \ldots, x_N)\, dx_1, dx_2, \ldots, dx_N \qquad (4.12)$$

$$P(FD) = \iint_{Y'} \cdots \int P_A(x_1, x_2, \ldots, x_N) dx_1, dx_2, \ldots, dx_N \qquad (4.13)$$

the set of measurements $(x_1, x_2, \ldots, x_{N-K})$ belong to A.. Similarly, Y' is the region in which the decision favors class B. If there are only two classes, Y' is the complement of Y.

The differences between the decision rules described in the preceding section arise only in the manner in which the region Y is selected. Hence, a subscript is used to denote the decision rule which gave rise to the choice of the region Y. For instance, $Y_2$ is the region in the N-dimensional space within which Rule 2 decides in favor of class A.

According to each of the decision rules, the comparison of a likelihood ratio with a threshold determines the region Y. If the ratio exceeds the threshold T, a decision is made that $(x_1, \ldots, x_{N-K}) \epsilon Y$; if the ratio is less than T, the decision is that the set of measurements belong to Y'. It should be noted that, given the values $(x_1, x_2, \ldots, x_{N-K})$, the likelihood ratios $\ell_1(x)$ through $\ell_3(x)$ are all functions of the N-K given measurements alone. No matter how complicated a function the likelihood ratios may be, they are, for a specified choice of $P_A(x_1, \ldots, x_N)$ and $P_B(x_1, \ldots, x_N)$, deterministic functions of $x_1, x_2, \ldots, x_{N-K}$. The region Y, whichever rule is used, has an infinite extent in the K-dimensional manifold of the missing measurements. The integrals of Equations 4.12 and 4.13 may be written as shown in Equations 4.14 and 4.15, where the region y denotes the region of the N-K dimensional space in which the measurement values are assigned to class A by the rule in vogue. Similarly, y' is the complement of y.

$$P(FA) = \iint_{y} \cdots \int p_B(x_1, x_2, \ldots, x_{N-K})\, dx_1, dx_2, \ldots, dx_{N-K} \qquad (4.14)$$

$$P(FD) = \iint_{y'} \cdots \int p_A(x_1, x_2, \ldots, x_{N-K})\, dx_1, dx_2, \ldots, dx_{N-K} \qquad (4.15)$$

If the positive constants $c_1$ and $c_2$ are the costs of false alarms and false dismissals, $p_A(x_1, \ldots, x_{N-K})$ and $p_B(x_1, \ldots, x_{N-K})$ are the marginal densities of the random processes A and B, and $y_s$ and $y_g$ are the regions in the N-K dimensional subspace of measured values in which decision rules s and g, respectively, decide that the observations should belong to A, then rule s is better than g if the inequality of Equation 4.16 holds in the specified direction. Each side of the inequality expresses the probability of error according to the corresponding decision rule.

$$P(B)c_1 \iint_{y_s} \cdots \int p_B(x_1, \ldots, x_{N-K})\, dx_1, \ldots, dx_{N-K} + P(A)c_2 \iint_{y'_s} \cdots \int p_A(x_1, \ldots, x_{N-K})\, dx_1, \ldots, dx_{N-K} <$$

$$P(B)c_1 \iint_{y_g} \cdots \int p_B(x_1, \ldots, x_{N-K})\, dx_1, \ldots, dx_{N-K} + P(A)c_2 \iint_{y'_g} \cdots \int p_A(x_1, \ldots, x_{N-K})\, dx_1, \ldots, dx_{N-K} \qquad (4.16)$$

Given the densities $p_A(x_1, \ldots, x_{N-K})$ and $p_B(x_1, \ldots, x_{N-K})$, obviously the decision rule which minimizes Q is best (minimizes the probability of error).

$$Q = P(B)c_1 \iint\cdots\int_{y_s} p_B(x_1, \ldots, x_{N-K})\, dx_1, \ldots, dx_{N-K} + P(A)c_2 \iint\cdots\int_{y_s'}$$

$$p_A(x_1, \ldots, x_{N-K})\, dx_1, \ldots, dx_{N-K} \qquad (4.17)$$

But, if there are only two classes, Equation 4.18 is true; this allows simplification of 4.17 as shown in Equation 4.19.

$$\iint\cdots\int_{y_s'} p_A(x_1, \ldots, x_{N-K})dx_1, \ldots, dx_{N-K} = 1 - \iint\cdots\int_{y_s} p_A(x_1, \ldots, x_{N-K})$$

$$dx_1, \ldots, dx_{N-K} \qquad (4.18)$$

$$Q = c_2P(A) + \iint\cdots\int_{y_s} c_1P(B)p_B(x_1, \ldots, x_{N-K}) - c_2P(A)p_A(x_1, \ldots, x_{N-K})$$

$$dx_1, \ldots, dx_{N-K} \qquad (4.19)$$

$Q_{min}$ occurs if $y_s$ is the region in which the integrand is always negative, where Equation 4.20 holds:

$$y_{optimum} \text{ where } c_2P(A)\, p_A(x_1, \ldots, x_{N-K}) > c_1P(B)p_B(x_1, \ldots, x_{N-K}) \qquad (4.20)$$

For the special case where $c_1 = c_2$, the region $y_{opt}$ is determined and hence class A is selected when

$$\frac{p_A(x_1, x_2, \ldots, x_{N-K})}{p_B(x_1, x_2, \ldots, x_{N-K})} > \frac{P(B)}{P(A)} \qquad (4.21)$$

We recognize that this is just the decision made by the marginal densities of Decision Rule 1.

### 4.1.4 Conclusions

Since results of this study apply to the two types of problems given at the beginning of this paper, it may now be concluded that:

1. The optimum classificatory decision on the N-K observed measurements consists of comparing the ratio of a posteriori probabilities of the actually observed N-K measurements with a threshold. No useful purpose is served by knowing to what higher dimensional process the N-K measurements belong.
2. In the optimum detection of a known signal in noise, where it is known that only part of the signal was transmitted, it is best to correspondingly truncate the stored reference signal and correlate the input with the truncated reference. No prediction of the missing portions of the waveform should be employed.

## 4.2 IMPLICATIONS OF RESULTS ON MISSING PARAMETERS

In Section 4.1 it was shown that, given an incomplete set of observations, the optimum decision procedure is that based on the marginal densities of the observed parameters. We will show, under the assumption of one missing parameter, the additional computation which must be performed for both the linear clustering and linear separating techniques.

In the linear clustering method we evaluate $\log f_1(x) - \log f_2(x)$ where

$$2 \log f_1(x) = - \log |U_1| - (x-M_1) U_1^{-1} (x - M_1)^T$$

If one parameter, say $x_n$ is missing, the desired function becomes (ignoring the factor of $\sqrt{2\pi}$ which appears in both $f_1$ and $f_2$)

$$2 \log f_1^*(x) = - \log |U_1| + \log \frac{|U_1|}{|U_1^*|} - (x^*-M_1^*) U_1^{*-1} (x^*-M_1^*)^T \quad (4.22)$$

where $^*$ refers to quantities omitting terms concerning $x_n$.

If the matrix $U_1^{-1}$ is partitioned as follows, where $A_{11}$ is an n-1 x n-1 matrix and a is an n-1 dimensional row vector

$$U_1^{-1} = \begin{bmatrix} A_{11} & a^T \\ a & a_{nn} \end{bmatrix}, \text{ then } \frac{|U_1|}{|U_1^*|} = \frac{1}{a_{nn}}$$

and $(U_1^*)^{-1} = A_{11} - \frac{a^T a}{a_{nn}}$ (see proof below).

Thus $2 \log f_1^*(x) = -\log |U_1| - \log a_{nn} - (x^* - M_1^*)(A_{11} - \frac{a^T a}{a_{nn}})(x^* - M_1^*)^T$

$$= -\log |U_1| - \log a_{nn} - (x^* - M_1^*) A_{11} (x^* - M_1^*)^T$$

$$+ \frac{1}{a_{nn}} [(x^* - M_1^*) a^T]^2 \qquad (4.23)$$

Proof that $(U_1^*)^{-1} = A_{11} - \frac{a^T a}{a_{nn}}$ .

If $U_1^{-1}$ and $U_1$ are partitioned as previously described

$$U_1^{-1} U_1 = I \Rightarrow \begin{bmatrix} A_{11} & a^T \\ a & a_{nn} \end{bmatrix} \begin{bmatrix} U_1^* & u^T \\ u & u_{nn} \end{bmatrix} = I \qquad (4.24)$$

$$\Rightarrow A_{11} U_1^* + a^T u = I$$

and $a U_1^* + a_{nn} u = 0 \Rightarrow u = -\frac{a^T U_1^*}{a_{nn}}$ $(a_{nn} \neq 0)$

Thus $A_{11}\, U_1^* - \frac{a^T a}{a_{nn}} \quad U_1^* = I$ (4.25)

or equivalently

$$\left(A_{11} - \frac{a^T a}{a_{nn}}\right) \quad U_1^* = I \tag{4.26}$$

$$\Longrightarrow U_1^{*-1} = A_{11} - \frac{a^T a}{a_{nn}} \tag{4.27}$$

In the linear separating transformations the optimum direction "a" is a function of the densities $f_1(x)$ and $f_2(x)$. Thus if a parameter is missing the direction must be re-evaluated.

For example if $a = d(U_1 + U_2)^{-1} = d\,\gamma^{-1}$

Then $a^* = d^*(U_1^* + U_2^*)^{-1} = d^*\gamma^{*-1}$ (4.28)

and as before $\left( \gamma^{-1} = \begin{bmatrix} V_{11} & v^T \\ v & v_{nn} \end{bmatrix} \right) \; ; \; \gamma^{*-1} = V_{11} - \frac{v^T v}{v_{nn}}$ (4.29)

$$a^* = d^*\left(V_{11} - \frac{v^T v}{v_{nn}}\right), \text{ but } d^* V_{11} = d\gamma^{-1} - d_n v = a - d_n v \tag{4.30}$$

Hence

$$a^* = a - d_n v - \frac{d^* v^T v}{v_{nn}}$$

$$a^* = a - \left[ d_n - \frac{d^* v^T}{v_{nn}} \right] v \qquad (4.31)$$

For more than one missing parameter the process may be continued - proceeding from $(U_1^*)^{-1}$ to $(U_1^{**})^{-1}$, etc.

Although the occurrence of incomplete sets of measurements does not necessitate re-learning (inversion of new matrices), it does require additional computation. For this reason decision procedures which do not involve transformations of the variables (some of which are discussed in Section 5) have the advantage of being easily adaptable to incomplete sets of observations.

## 4.3 CLASSIFICATORY DECISIONS ON NOISY OBSERVATIONS

Suppose that a collection of objects are to be classified into two categories from which the objects are known to be drawn. Further suppose that a set of examples from each class are available. The classification process must proceed in two stages: a learning stage in which the invariant characteristics of each category are determined from the known examples and a recognition stage in which unknown objects are classified on the basis of results obtained from the learning stage.

If measurements on the unknown objects can be considered as random variables, then "learning" consists of estimating the probability density functions of the two categories over the observation space. This would require an estimation of as many parameters as are required to specify the probability distribution.

On the basis of the results of observations made on known objects during the learning stage, a decision rule is formulated which partitions the observation space into regions $D_F$ and $D_G$ corresponding to classes F and G. Then, when an observation on an unknown object falls in $D_F$ it is classified as belonging to F. The reverse is true if it falls in $D_G$. A common decision rule is based on comparing the likelihood ratio (defined as the ratio of the two probability density functions) with some threshold $\theta$. $\theta$ is computed from relative costs of the two types of errors and the a priori probabilities so as to minimize the expected loss due to incorrect decisions. Thus,

$$\text{decide } v \in F \text{ if } \ell(v) = \frac{p_F(v)}{p_G(v)} > \theta \; ; \; v \in F \text{ if } \ell(v) = \frac{p_F(v)}{p_G(v)} \leq \theta, \tag{4.32}$$

where $v = (v_1, v_2, \ldots, v_N)$ is a point in the sample space, and $p_i(v)$ is the probability density associated with the $i^{th}$ class. Choosing $\theta = 1$ is equivalent to assuming equal a priori probabilities and costs and results in minimization of the probability of error.

The two types of errors occur as follows: Type 1 (false dismissal) occurs when the sample object to be classified belongs to F but $v \in D_G$. Type 2 (false alarm) occurs when the sample object belongs to G but $v \in D_F$. The probabilities of the two types of errors are

$$P_{e_1} = \int_{D_G} p_F(v)\, dv_1, dv_2, \ldots, dv_N \tag{4.33}$$

and

$$P_{e_2} = \int_{D_F} p_G(v)\, dv_1, dv_2, \ldots, dv_N \tag{4.34}$$

The total probability of error is $P_e = qP_{e_1} + (1-q)P_{e_2}$ where q is the a priori probability of class F.

Now, suppose that $p_F(v)$ and $p_G(v)$ are not known correctly. The decision regions $D'_F$ and $D'_G$ determined from the erroneous density functions $p'_F$ and $p'_G$ will not, in general, be equal to the regions $D_F$ and $D_G$ corresponding to the true distributions. Therefore, a greater cost (or probability of error) will result than if $p_F$ and $p_G$ were known correctly.

A particularly interesting case in which the density functions may not be known correctly is the instance in which additive, independent noise is present in the observed variables*. That is, suppose that the observed

*Henceforth, to isolate effects, the estimation or learning process will be presumed accurate.

variables $v_i = y_i + n_i$ where $y_i$ is due strictly to the object to be classified and $n_i$ is another random variable called noise due to some independent cause.

If the noise is present during "learning", i.e., while the densities $p_F$ and $p_G$ are being estimated, and during the recognition stage, then the resulting probability of error will be minimized. This probability will, of course, be greater than the probability of error that would result had there been no noise.

If, however, the conditions change between the learning and recognition, the decision regions will not be optimum and the probability of error will not be minimized. Specifically, if learning goes on in the absence of noise but noise is present during recognition the probability of error will be greater than if noise had been present during learning. It is not clear, however, just how the probability of error varies with the amount of noise added or just what deviation from optimum occurs.

Suppose that the density functions of two classes have been estimated under certain noise conditions and the decision regions have been determined. There will be, in general (or we would have no real problem), a non-zero probability of error, $P_L \neq 0$. Now, suppose that the amount of noise is increased. This will increase the spread of the class distributions and tails of the distribution will be increased. But, the probability of error is given by integrals over the tails of the distributions so that the probability of error will increase with increasing noise (or correspondingly, decrease with decreasing noise.) Figures 4-1 and 4-2, for one-dimensional probability densities, show this effect. The sketches do not aid much in answering the questions: What is the deviation from the minimum probability of error for decisions based on fixed decision regions? How strongly does the probability of error depend on additive

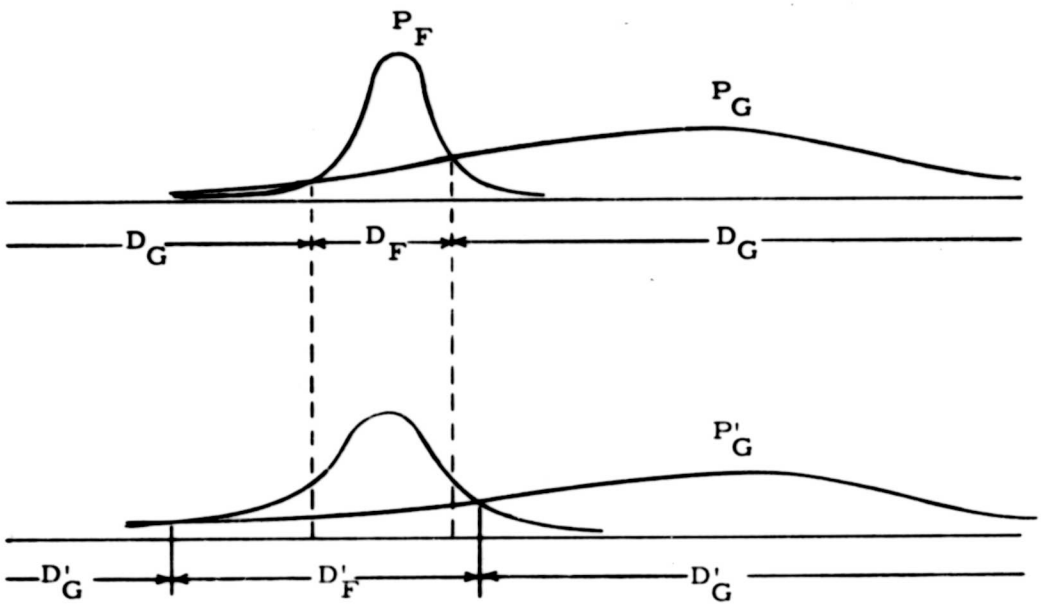



Figure 4-1. Variation in Decision Regions with Additive Noise

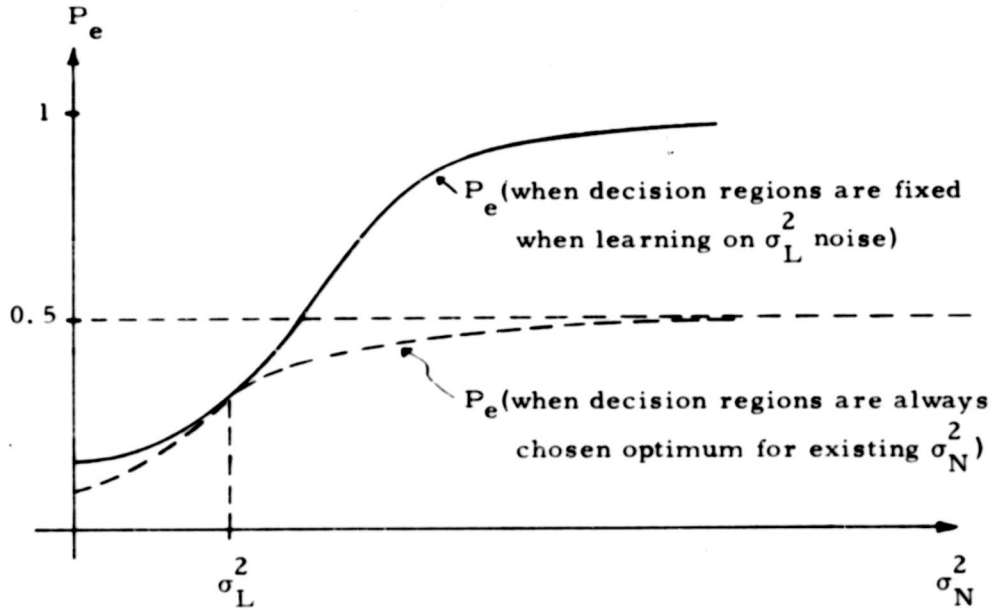



Figure 4-2. Probability of Error as a Function of Noise Power

noise? Certainly the answers to these questions depend on the distributions involved, but, in general, the curves of probability of error for fixed and optimum regions will appear much as in Figure 4-2. The curves will be continuous unless the distributions are discrete, but will not be smooth unless the probability density functions are continuous. The curves are plotted here as $P_e$ versus noise variance which corresponds to the one-dimensional case. In higher dimensions, it is not always clear what to use as a measure of noise intensity. The noise variance during learning is $\sigma_L^2$. It is assumed that when the decision regions are fixed, they are fixed at the time of termination of the learning stage. Therefore the regions are identical to the optimum regions at a noise power of $\sigma_L^2$. The probability of error when using these decision regions is, of course, always higher than when the optimum regions for each prevailing noise condition is used.

To illustrate the above remarks, we shall derive the properties of the probability of error curves for a simple case, viz., both classes a priori equiprobable and having univariate distributions.

Let the density functions of the classes be

$$p_F(v) = \frac{1}{\sqrt{2\pi}\ (\sigma_F^2 + \sigma_N^2)^{1/2}} \exp\left[ -\frac{1}{2} \frac{v}{(\sigma_F^2 + \sigma_N^2)} \right] \qquad (4.35)$$

$$p_G(v) = \frac{1}{\sqrt{2\pi}\ (\sigma_G^2 + \sigma_N^2)^{1/2}} \exp\left[ -\frac{1}{2} \frac{(v-m)^2}{(\sigma_G^2 + \sigma_N^2)} \right] \qquad (4.36)$$

where $\sigma_N^2$ is the additive noise variance and is the independent parameter. Note that due to symmetry of the Gaussian distribution, only the case of $\sigma_F^2 \leq \sigma_G^2$ and $m \geq 0$ need be considered. In fact, here we shall suppose that $\sigma_F^2 < \sigma_G^2$.

Setting the likelihood ratio equal to one is equivalent to setting $\ell n[\ell(v)] = 0$. This results in a quadratic expression in v which has solutions $v'_o$ and $v''_o$, where for convention $v'_o < v''_o$,

$$v''_o = \frac{1}{\sigma_G^2 - \sigma_F^2}\left[\left\{(\sigma_F^2 + \sigma_N^2)(\sigma_G^2 + \sigma_N^2)\left[m^2 + 2(\sigma_G^2 - \sigma_F^2)\,\ell n\,\frac{\sigma_G^2 + \sigma_N^2}{\sigma_F^2 + \sigma_N^2}\right]\right\}^{1/2} - (\sigma_F^2 + \sigma_N^2)m\right] \quad (4.37)$$

$$v'_o = \frac{1}{\sigma_G^2 - \sigma_F^2}\left[\left\{(\sigma_F^2 + \sigma_N^2)(\sigma_G^2 + \sigma_N^2)\left[m^2 + 2(\sigma_G^2 - \sigma_F^2)\,\ell n\,\frac{\sigma_G^2 + \sigma_N^2}{\sigma_F^2 + \sigma_N^2}\right]\right\}^{1/2} + (\sigma_G^2 + \sigma_N^2)m\right] \quad (4.38)$$

Now it may be observed that $v''_o$ is a strictly increasing monotonic function of $\sigma_N^2$ (and m) and $v'_o$ is strictly monotonic decreasing with $\sigma_N^2$ (and m). The interval $v'_o < v < v''_o$ is that in which $p_F > p_G$.

The expression for the minimum or optimum probability of error is

$$P_e(\text{opt}) = \frac{1}{2}\left[\int_{v'_o}^{v''_o} p_G + 1 - \int_{v'_o}^{v''_o} p_G\right] = \frac{1}{2}\left[1 + \int_{v'_o}^{v''_o}(p_G - p_F)\right] \quad (4.39)$$

where the notation $\int_a^b f = \int_a^b f(v)\,dv$ has been used.

Now suppose learning goes on in the presence of noise with variance $\sigma_L^2$, i.e., the decision region is fixed at the optimum for $\sigma_N^2 = \sigma_L^2$. Let $v' = v_o'(\sigma_L^2)$ and $v'' = v_o''(\sigma_L^2)$. Then the probability of error for this fixed decision region is

$$P_e(\text{fixed}) = \frac{1}{2}\left[1 + \int_{v'}^{v''} (p_G - p_F)\right]. \qquad (4.40)$$

Hence, $P_e(\text{opt}) - P_e(\text{fixed}) = \int_{v_o'}^{v_o''} (p_G - p_F) - \int_{v'}^{v''} (p_G - p_F) \geq 0$, where equality holds only at $\sigma_N^2 = \sigma_L^2$ since $(p_G - p_F)$ cannot be zero over any interval.

The slopes of the probability of error curves are given by

$$\frac{d\,P_e(\text{opt})}{d\,\sigma_N^2} = (p_G - p_F)\Big|_{v=v_o''} \frac{d\,v_o''}{d\,\sigma_N^2} - (p_G - p_F)\Big|_{v=v_o'} \frac{d\,v_o'}{d\,\sigma_N^2} + \int_{v_o'}^{v_o''} \frac{d}{d\,\sigma_N^2}\,(p_G - p_F)$$

$$= \int_{v'_o}^{v''_o} \frac{d}{d\sigma_N^2} (p_G - p_F) \qquad (4.41)$$

and $$\frac{d P_e(\text{fixed})}{d\sigma_N^2} = \int_{v'}^{v''} \frac{d}{d\sigma_N^2} (p_G - p_F). \qquad (4.42)$$

Therefore, the probability of error curves are tangent at $\sigma_N^2 = \sigma_L^2$.

To evaluate the slopes of the curves, note that

$$\frac{d}{d\sigma_N^2} p_\nu = \frac{1}{2\sqrt{2\pi}\,(\sigma_\nu^2 + \sigma_N^2)^{3/2}} \exp\left\{-\frac{1}{2}\,\frac{(x - m_\nu)^2}{\sigma_\nu^2 + \sigma_N^2}\right\}\left[\frac{(x - m_\nu)^2}{\sigma_\nu^2 + \sigma_N^2} - 1\right] \qquad (4.43)$$

The behavior of this derivative is roughly as is shown in Figure 4-3.

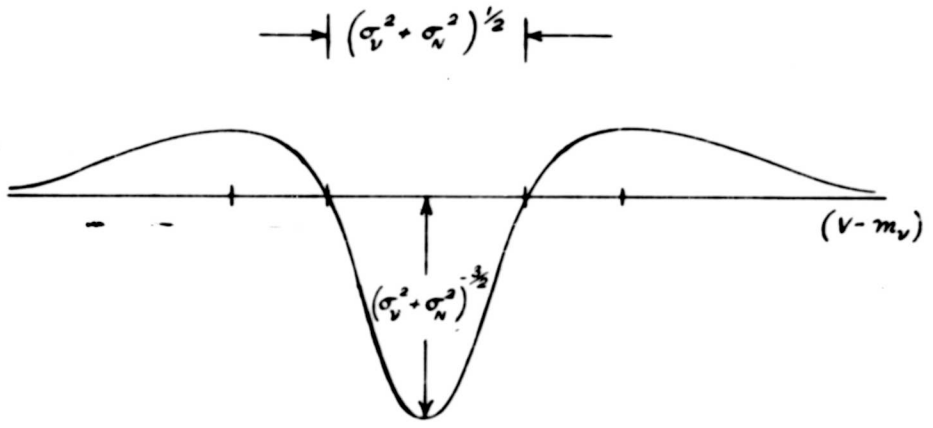



Figure 4-3.

Therefore,

$$\frac{d}{d\sigma_N^2}(p_G - p_F) = \frac{1}{2}\,\frac{1}{2\pi}\,\frac{1}{(\sigma_G^2 + \sigma_N^2)^{3/2}} \exp\left\{-\frac{1}{2}\,\frac{(v-m)^2}{(\sigma_G^2 + \sigma_N^2)}\right\}$$

$$\left[\frac{(v-m)^2}{(\sigma_G^2 + \sigma_N^2)} - 1\right]$$

$$-\frac{1}{2}\,\frac{1}{2\pi}\,\frac{1}{(\sigma_G^2 + \sigma_N^2)^{3/2}} \exp\left\{-\frac{1}{2}\,\frac{v^2}{(\sigma_F^2 + \sigma_N^2)}\right\}\left[\frac{v^2}{(\sigma_F^2 + \sigma_N^2)} - 1\right]$$

$$= \frac{1}{2}\left\{\frac{p_G}{(\sigma_G^2 + \sigma_N^2)}\left[\frac{(v-m)^2}{(\sigma_G^2 + \sigma_N^2)} - 1\right] - \frac{p_F}{(\sigma_F^2 + \sigma_N^2)}\left[\frac{v^2}{(\sigma_F^2 + \sigma_N^2)} - 1\right]\right\}$$

(4.44)

is positive in an interval around the origin. Further, (heuristically) since

$\int_{-\infty}^{\infty} \frac{d}{d\sigma_N^2}(p_G - p_F) = 0$, an integral over a finite interval around the origin must be strictly positive. Therefore, both $P_e$(opt) and $P_e$(fixed) are strictly monotonic increasing with $\sigma_N^2$.

Furthermore, the curvature of the two curves at $\sigma_N^2 = \sigma_L^2$, as indicated by the second derivatives, are of the same sign; $\frac{d^2 P_e}{d\sigma_N^4} = \int \frac{d^2}{d\sigma_N^4}(p_G - p_F)$

$+ \frac{d}{d\sigma_N^2}(p_G - p_F)\left[\frac{d^2 v_o''}{d\sigma_N^4} - \frac{d^2 v_o'}{d\sigma_N^4}\right]$. Since the curves are smooth, the second

derivatives must have the same sign in an interval around $\sigma_N^2 = \sigma_L^2$. The same type of argument can be used on the higher derivatives.

We are led to the following conclusions: The $P_e$ curves are very snug fitting in the vicinity of $\sigma_N^2 = \sigma_L^2$ and hence decision performance does not seriously deviate from optimum for moderate changes in the amount of additive noise. A decrease in the noise level decreases the probability of error. But, the resulting probability of error is greater than the optimum for that noise condition, see Figure 4.2.

For higher dimensional observation spaces, the effects of noise can be represented in a similar manner with $P_e = P_e(\sigma_{N_1}^2, \sigma_{N_2}^2, \ldots \sigma_{N_n}^2)$, where $\sigma_{N_k}^2$ is the noise variance in the $k^{th}$ dimension. The curve in Figure 4.2 is replaced by a surface and the learning conditions are specified by the point $\tilde{\sigma}_L^2 = < \sigma_{L_1}^2, \ldots, \sigma_{Ln}^2 >$. The actual $P_e$ surface and the optimum $P_e$ surface coincide at $\sigma_L^2$ and are very close in a neighborhood around that point.

Certain important differences exist, however, between univariate and multivariate cases. These arise because the distributions for each class can have different orientations (roughly, correlation coefficients), in the observation space. In particular, the probabilities of error are not monotonic with the separation between the means of the distributions. In addition, small amounts of noise in certain of the observation dimensions can change the relative orientation of the distributions and hence result in rather large increases in the probabilities of error.

# V SPECIAL TECHNIQUES

## 5.1 AUTOMATIC SUBCLASS GENERATION

In considering arbitrarily complex regions of parameter space, it is always possible, using the non-linear method, to derive a function which effects optimum discrimination between classes. Since in many instances the required computations may be excessive, using the non-linear method, it is worthwhile to consider alternate simpler techniques.

The complexity of a region containing members of a class results in many cases from the existence of subclasses within the class. If the probability density has several modes, the samples in a region surrounding a mode can be considered to constitute a subclass. For example a multi-modal density would certainly occur for any representation of printed characters. The letters "A" and "a" are subclasses in the general class of a's and are clustered in different regions of the space. A similar situation arises in the case of speaker recognition where the subclass of utterances of a particular speaker correspond to the particular sounds spoken. In Figure 5.1 two such multi-modal densities are depicted. In this case the non-linear method would require at least a third-degree polynomial (10 x 10 matrix inversion) to effectively distinguish the two classes. If, however, the subclasses are isolated, discrimination can be accomplished by fitting Gaussian densities to the subclass densities (four 2 x 2 matrix inversions), a much simpler problem. It is also possible that two unimodal densities such as those of Figure 5.2 cannot be distinguished by linear methods. In this case, although there may be no intuitive argument for a division into subclasses, the division of B shown would permit the separation of A and B by linear methods.

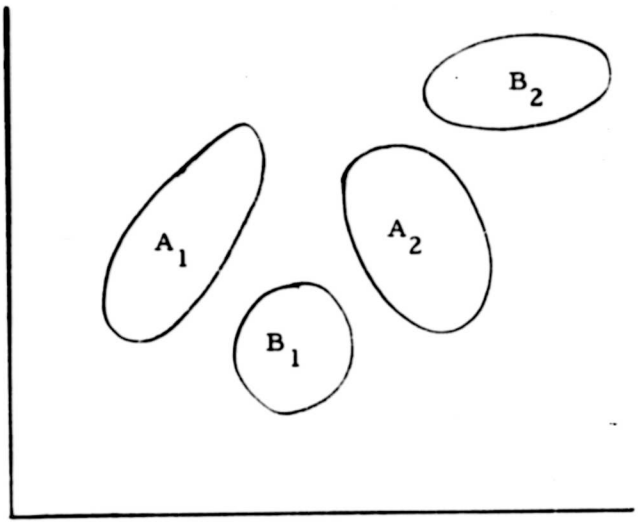



Figure 5.1 Subclasses of Bimodal Densities

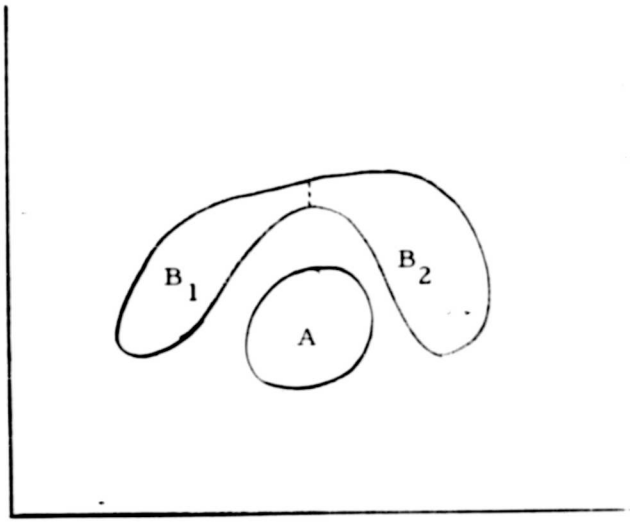



Figure 5.2 Subclasses of Unimodal Density

In the Automatic Subclass Generation program we have attempted to locate subclasses which are separated in space and have made no attempt to restrict the shape of the regions. Once the subclasses are obtained we use linear methods for discrimination. In the example tried, the results by this method were superior to those obtained using linear methods on the entire class. There is of course no guarantee that subclasses so obtained will be separable by linear methods, but it is expected that they will be more easily separable than the original classes.

Method of Subclass Generation

We define a subclass to be a group of points which are "pseudo-connected" in that each point in the subclass is within a distance T of at least one other point of the set. (distance may be either Euclidean or $\sum|X_i-Y_i|$ ). Thus if we have two points which are within a distance T of each other they form a subclass and if a third point is within T of one of the first two, it also is a member of the subclass. By a sequential processing of the samples, new members may be added to the subclass and new subclasses may be formed when a point is a member of no previous subclass. Two subclasses may be merged into one if a point is obtained which is within T of one sample of each subclass; the new point "connects" the two subclasses. This is shown geometrically in Figure 5-3.

For a given value of the threshold there is no way of knowing, a priori, how many subclasses will result. For this reason we have adopted a method of subclass generation in which the distances between classes and samples are retained. In this way the threshold may be increased to obtain a reduction in the number of subclasses without recomputing the distances.

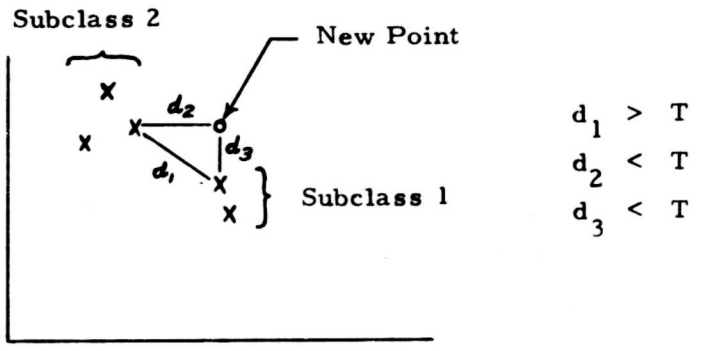



Figure 5-3. A New Point Connecting Two Subclasses

The method is outlined below:

Definitions: $C_i$ is a subclass, $Y_{ik}$ are members of $C_i$

X is a sample point

$$D(X, C_i) = \min_k D(X, Y_{ik})$$

1) Select a number of points $(X_1, \ldots, X_I)$, calling each point a subclass.

2) Compute the (I x I) matrix M where $m_{ij} = D(X_j, C_i)$ [$= D(X_j, X_i)$ at this stage].

Note that $m_{ij} = 0$ means $X_j \in C_i$.

3) Subclasses $C_k$ and $C_j$ may be merged if $m_{ki} = 0$ and $m_{ji} < T$. (There is a point $X_i$ in $C_k$ which is within T of $C_j$). When two subclasses are merged the elements of M must be changed as follows:

$$m^1_{kn} = \text{MIN}\ (m_{kn},\ m_{jn})$$

$m^1_{jn}$ becomes empty

4) All possible mergers are effected in the original set of points.

5) A new point is now called for and its distance is computed to all the subclasses. [A new column is added to the matrix M.] If this column makes possible one or more mergers, the distances of the point to all previous points are computed, and merging takes place. If there are no mergers but there is space for another subclass, the new point is included as an additional subclass. This also requires computation of the distances of the new point to all previous points. If neither of the above possibilities exists, the point is labeled "isolated", for the present.

6) Step 5 is repeated for the rest of the library of points.

7) If in Step 6 any of the subclasses were unmerged, Step 5 is repeated for the "isolated" points. [One of the points which was initially "isolated" may become "connected" to a class due to the addition of a later point to the class.]

8) The process is repeated until no further merging is possible.

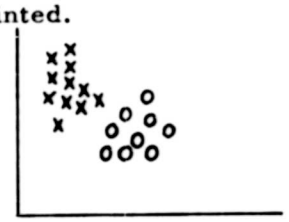
9) If at this point the number of sublcasses is too large, the threshold may be increased and the process repeated.

10) The threshold is increased until the number of classes falls within some preset limit. (Note that there may still be "isolated" points remaining.)

If some method has been devised for selecting a threshold ahead of time, the procedure is much simpler. One need keep track only of which points are members of which subclasses. As a new point is added its distance is computed to each subclass. If it is within T of one or more subclasses, the point and all the classes to which it is "connected" become one class. Otherwise the new point becomes a new subclass.

## 5.2 A PROXIMITY ALGORITHM AND SELECTIVE SAMPLING

An extension of the concept of subclass generation would be the labeling of each sample point as a different sublcass. An unknown sample might then be identified according to the point to which it is closest (by Euclidean distance). This type of identification procedure would certianly work if the regions of the space occupied by members of various classes were disjointed.



If the regions overlap, the decisions made in this way might be termed probabilistic decisions. Suppose for example there is a volume of space in which samples of one class are uniformly distributed with density $\frac{N_1}{V}$ and samples of another class with density $\frac{N_2}{V}$. Then a new sample, randomly placed in this region would be closest to a member of class 1 with probability $\frac{N_1}{N_1 + N_2}$, or to class 2 with probability $\frac{N_2}{N_1 + N_2}$.

This method requires the availability of all sample points for decision making. For this reason it is not practical. Hence we are led to the sample selection technique described in the next section.

### 5.2.1 Selective Sampling

As stated above, the determination of decision regions using the proximity algorithm requires the retention of all the sample points. It is, however, possible to select two subsets of the original sample points (one subset for each class) in such a way that the original sample points will be classified correctly (using the proximity algorithm) considering only distances to the points of the subsets. The procedure for selecting the subsets is based on the principle that a sample need not be included in the subset if there is a point already in the subset which served to correctly classify the point. The disadvantage of this method of sample selection is that the addition of a new point to the subset representing one class may result in the misclassification of a point in the other class (not included in the representative subset) which was previously classified correctly. Thus the representative sample set is not up-date-able.

From the above discussion we will adopt the principle of selecting a subset of sample points which in some manner "represents" the original set (leads to correct decisions) in view of the decision procedure to be employed. In addition, we will attempt to devise a "sample selection-decision procedure" combination in which the sample set is easily updateable. In order to achieve these objectives it appears necessary to include more than positional information about the selected samples. That is, one might want some idea of the number of original points which have been found within a small neighborhood of a selected point. If this information is included with a suitably selected set of points the result will be that the density of the original set of points is roughly approximated by the union of

a set of uniform spherical densities with the selected samples as their means and the number of original samples in the spherical neighborhoods of the means as their relative heights (a priori probabilities.)

In keeping with the above reasoning the representative samples are selected as follows: if a new point (known classification) is within the neighborhood (of pre-determined size and shape) of a previous representative sample, the a priori probability of that representative sample is suitably increased; and if the new point is not near a previous representative sample it is included as a new representative sample.

There is, however, no guarantee that the samples which have been associated with a particular representative sample point will be distributed in any special way or that the mean of the associated samples will correspond to the first point chosen in the group. To overcome the second objection we will compute the representative sample as the mean of the points associated with it; when a new point is within the neighborhood of a representative sample, that point is used to up-date the mean.

Control over or determination of the distribution of the points surrounding a representative sample is less easily solved; however, it may be argued that if the size of the neighborhood is small enough the actual distribution within it will not be particularly important.

In the description that follows it is assumed that the points associated with a representative sample are normally distributed about the representative sample with equal variance in all directions (covariance matrix $\sigma^2 I$) and that each normal density [ $\phi_i^F(x)$ and $\phi_j^G(x)$ ] has an a priori probability proportional to the number of samples included in it ($N_i^F$ and $N_j^G$). The density of class F, a union of normal densities is then:

$$f(x) = \sum_i N_i^F \phi_i^F (x)$$

and decisions according to Bayes Rule are made as follows:

$$x \in f \text{ if } \sum_i N_i^F e^{-\frac{1}{2}\frac{(x-M_i^F)(x-M_i^F)^T}{\sigma^2}} > \sum_j N_j^G e^{-\frac{1}{2}\frac{(x-M_j^G)(x-M_j^G)^T}{\sigma^2}}$$

where $N_i^F$ is the a priori probability (in F) of the class represented by $M_i$ (the representative sample or mean of the class).

The selection of the samples or means may be performed by the following method. ($V_i^F$ are samples known to be members of f).

Let $V_1^F = M^F$ and $N_1^F = 1$ representing $C_1^F$

if $d(V_2^F, M_1^F)^T < T$ (a suitable threshold),

then $V_2^F \in C_1^F$

and $M_1^F$ becomes $\dfrac{N_1^F M_1^F + V}{N_1^F + 1}$

and $N_1^F$ increases to $N_1^F + 1$.

Otherwise $V_2^F = M_2^F$ .

In general,

if $d(V, M_i^F) < T$, $V \in C_i^F$

$M_i^F$ and $N_i^F$ are updated as above,

if $d(V, M_i^F) > T$ for all i

V becomes the mean of a new class $C_x^F$

$M_x^F = V \qquad N_x^F = 1$

5.2.2 Experimental Results Using the Proximity Algorithm

In this example the proximity algorithm was used as a means for detection rather than discrimination; that is, to decide on the basis of samples of one class whether or not a new sample is a member of that class. The event to be detected in this case is the speech of a particular or "desired" speaker. The speech of any other "undesired" speakers should be dismissed. The detection is to be based on measurements of the four formant frequencies (the first four maxima of the spectrum) in 20 ms samples of voiced sounds. Thus for each (4 dimensional) 20 ms sample of speech the machine must make a decision as to whether or not the desired speaker has spoken. The decision is based on the minimum distance, $D_m$, of the sample point to the set of samples of speech by the desired speaker stored in the machine. If $D_m < T$ the machine notes that the desired speaker has spoken; if $D_m > T$ the desired speaker has not spoken and the sample is dismissed as undesired.

To obtain the results summarized below, 100 (20 ms) samples of speech by the desired speaker were stored in the machine. 70 samples of each group, desired and undesired were used for the test. The threshold was determined afterwards to minimize the error rates. The value of the threshold was $T^2 = 2.1 \times 10^4$.

| | | Speaker | |
|---|---|---|---|
| | | Detected as Desired | Dismissed as Undesired |
| Sounds Spoken by | Desired Speaker | 57 | 13 |
| | Undesired Speaker | 9 | 61 |

False alarm rate 12.4 o/o

False dismissal rate 18.6 o/o

Of course the results would improve if decisions were based on a sequence of 20 ms samples (by the same speaker) instead of a single sample. In other experiments reported on Contract AF30(602)-2499, this subject is discussed in more detail.

## 5.3 OPTIMUM SMALL SAMPLE DISCRIMINATION

The classical problem in the testing of statistical hypotheses involves the making of multiple choice decisions when the statistical properties of the classes about which the observations are made are known. The statistical properties are assumed known in the form of the probability densities associated with the distributions of members of the various classes.

The application of decision theory to pattern recognition may take either of two forms. In one, where sufficient knowledge exists about the nature of the distribution of the set of observable parameters in the observation space, the probability densities are actually known, thus satisfying the fundamental requirements necessary in the making of multiple choice decisions. The second method of meeting this requirement, where the probability densities are not known from physical or other reasons is that this knowledge must be gained by taking a very large number of samples of the classes. Although it is possible to estimate - at least in certain cases - the number of samples sufficient to characterize the classes adequately, in many cases a large number of samples are not available for the construction of the probability densities required for the statistical treatment of the classification problem. Hence a way must be found for making optimum classificatory decisions based on a small number of samples.

The purpose of this section is to show how a solution that minimizes the cost of incorrect decisions on a finite number of samples can be obtained. The problem is stated and solved only in its simplest form. A number of problems, stated at the end, must still be considered. These involve the extensions of the method to multivariate cases and the investigation of the properties of the solution as the sample size goes to infinity and the solution must approach that offered by decision theory.

If we observe a random variable v, which may belong to either of two random processes A or B with probability densities $p_A(v)$ and $p_B(v)$, the optimum decision rule (that minimizes the cost of misclassification) requires that we test the variable v to see if it is contained in an interval of v where $p_A(v)$ is larger than $p_B(v)$ or in one where the reverse is true. Since $p_A(v)$ and $p_B(v)$ are between 0 and 1, $p_A(v) - p_B(v) = 0$ should mark the boundary that separates decisions favoring class A from those favoring class B.

If we had a function u(v) which was positive over the intervals of v where $p_A(v) > p_B(v)$ and negative elsewhere, the optimum decision would be given in the following equation.

decide $v \in A$ if $u(v) > 0$

decide $v \in B$ if $u(v) < 0$

In earlier reports the problem of solving for the optimum function u(v) was solved. The essential aspects of u(v) are, however, only its zero crossings. The problem considered in this section deals with the optimum placement of zero crossings.

The problem of making optimum classificatory decisions, based on a finite number of samples of each class, is stated below in its most elementary form. The required decision rule should be limited to a maximum of Z degrees of freedom (Z is the storage limitation or the degree of complexity of the machine).

Problem

Assume that M members of two classes A and B are given. Of these, $M_A$ belong to class A and $M_B$ belong to class B. Let us assume that a member of a class is a variable which may assume any

value between $-\infty$ and $+\infty$. The two classes may be two different one-dimensional random processes with arbitrary probability densities; and the $M_A$ and $M_B$ samples of the two processes may be finite but independently chosen samples.

Suppose that a decision procedure which can distinguish between members of A and those of B is constructed as a binary valued function of v, (f(v) has values of $\pm$ 1) and the decision is made to

decide $v \in$ A if $f(v) = +1$

decide $v \in$ B if $f(v) = -1$

Also, suppose that the function f(v) can have only Z or less zeros in order to limit the complexity of the resulting equipment. The problem to be solved can be stated as follows.

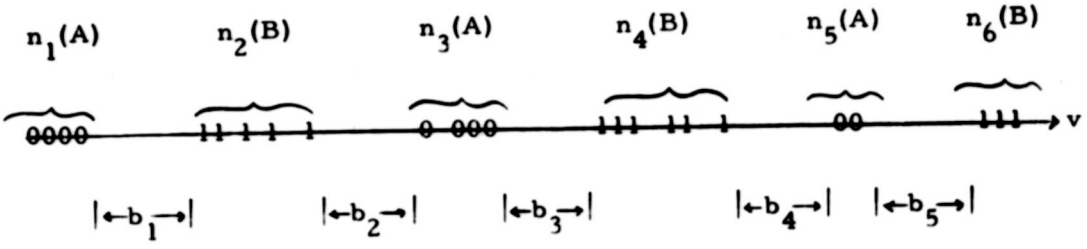
If $C_A$ is the cost of false alarms (that of deciding A when B occurred) and $C_D$ is the cost of false dismissal (that of deciding B when A occurred), where should the Z or less zeros of f(v) be placed to minimize the total cost incurred by using the decision rule?

Total cost $C = C_A N_A + C_D N_D$, where $N_A$ is the number of samples of B misclassified as A and $N_D$ is the number of A misclassified as B.

Solution

Let R be the number of intervals on v. An interval is defined as a region of v in which only members of one class, and in which at least one member of some class is contained. In the Figure below, for instance, R = 6, $M_A$ = 10, and $M_B$ = 14. To facilitate the description of the solution, additional definitions must be made.

Define $K_A$ as the number of A intervals, intervals that contain only A's. Similarly, let $K_B$ be the number of B intervals. In the figure below $K_A = 3 = K_B$. Note that $|K_A - K_B| \leq 1$ always. Let $n_p(A)$ be the number of given samples of the $p^{th}$ interval. The argument A denotes that the samples of the $p^{th}$ interval are members of class A.

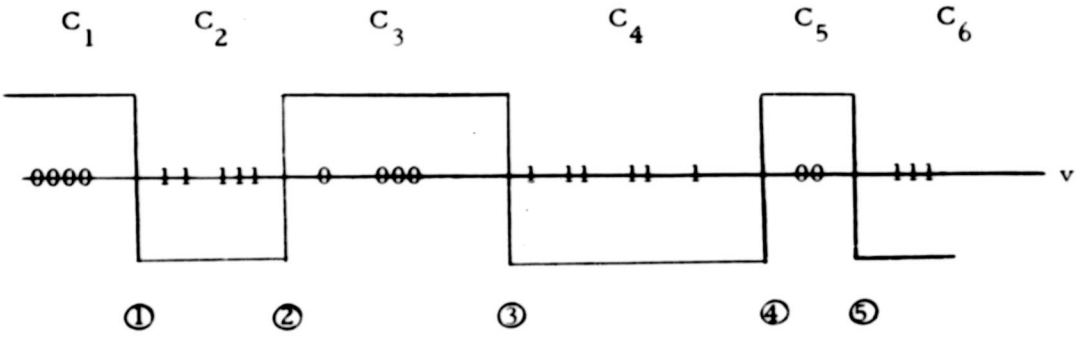



a) If the number of zeros Z of the binary function f(v) satisfies the inequality $Z \geq R-1$, the zeros can always be located on the line v so that the cost C = 0; i.e. so that no errors in classification decisions are made. R - 1 of the zeros should be located in the R-1 boundary regions. The remaining Z - R + 1 zeros should be located at very high or very low values of v, where they do not influence the decisions made on the given samples. A boundary region, denoted by b, is a line segment that terminates on a sample of class A at one of its ends and on a sample of class B at the other end. A boundary region, furthermore, contains no samples of either class.

b) If $Z < R - 1$, no error free decisions can be made on all given samples. The best that can be achieved, therefore, is to place the zero crossings of f(v) so that the cost is minimized. An algorithm is presented in the following for achieving the objective described. Steps of the algorithm are as follows.

**Step 1**

Temporarily locate a zero of f(v) in every boundary region of the R intervals. Thus a total of R-1 zeros are temporarily introduced. If decisions were made using the function f(v) thus obtained, no errors in decision would take place and the total cost C = 0. There are, however, R - 1 - Z excess zeros in the resulting solution. Thus the number of zeros must be reduced by R - 1 - Z. It is clear that the zeros remaining in the final solution must also be placed in the boundary regions. If, for example, a region in the final solution where f(v) is positive (classified as A) contains members of B adjacent to the zeros bounding the region (zeros not in boundary regions), the cost of misclassification could be reduced by moving the zeros bounding the region to the boundary regions so that these samples of B are excluded from the A region and are correctly classified. Thus the optimum solution may be obtained by a successive elimination (in optimum fashion) of the temporary zeros. The result of locating R - 1 = 5 temporary zeros in the example of the above figure is illustrated below by the encircled numbers.



Step 2

Assign a cost to the misclassification of each interval that would result if the sign of the function f(v) in the interval were changed. Such a change would come about if the temporary zeros bounding the interval were eliminated during the attempt to reduce the number of excess zeros of the temporary function f(v). Let the cost of misclassification of the $p^{th}$ interval be $C_p$. If the $p^{th}$ interval contains elements of class A, $C_p = C_D n_p(A)$, if it contains elements of B, $C_p = C_A n_p(B)$.

2.1

If R-Z-1 = 1, remove the temporary zeros (or zero) bounding the interval corresponding to the minimum $C_p$. Let $C_m$ be the minimum $C_p$ and the corresponding interval be m.

2.1.1

If m is an end interval, one zero is removed, yielding a function f(v) with exactly Z zeros.

2.1.2

If m is not an end interval, as is the case in the figure if $C_A = C_D$, two zeros (4 and 5 in the example) must be removed, yielding an optimum solution of Z - 1 zeros of the function f(v). In either case, we conclude that the solution of minimum cost is obtained.

2.2

If R - Z - 1 = 2, and $c_m$ is not an end interval, the two zeros bounding interval m are removed to yield a solution with exactly Z zeros.

2.2.1

If $R - Z - 1 = 2$, and $C_m$ is an end interval, the two cases described in a and b below may exist.

a) The costs of the two extreme intervals $C_1$ and $C_R$ are such that $C_1 + C_R < C_T$ where $C_T$ is the smallest cost of all the internal intervals. If this is true, remove the first and last zeros, yielding a solution with exactly Z zeros and minimum cost.

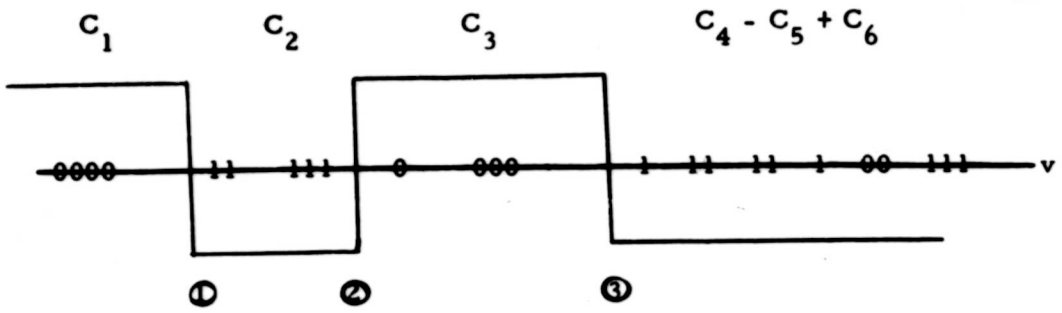
b) If alternative a above is false, then $C_1 + C_R > C_T$ and the removal of the two zeros bounding $C_T$ yields a minimum cost solution of exactly Z zeros.

2.3

If $R - Z - 1 \geq 3$, at least two intervals must be eliminated. First the interval corresponding to the minimum cost $C_m$ is eliminated by removal of the zero (or zeros) that bound it. The number of excess zeros, after this step is completed, is still equal to or greater than one. We proceed to Step 3.

Step 3

If 2.3 holds and one interval, m, is eliminated, the required solution is not yet obtained. The modified tentative function f(v) still has a number of zeros in excess of Z. The elimination of interval, m, however, merges the intervals adjacent to m. In the above example, if interval 5 is eliminated, 4, 5 and 6 are merged into one. The cost of misclassifying this new interval, illustrated below, is $C_4 - C_5 + C_6$. In general, the cost of misclassifying the newly created interval is $C_{m-1} - C_m + C_{m+1}$. Note that if the new interval is misclassified, i.e., sign of f(v) is changed, interval 5 would again be classified correctly.




This accounts for the minus sign of $C_5$. Therefore, in step 3 we reassign costs of misclassification to the intervals as they now exist. We thus find the state of affairs identical to that which existed before the removal of any of the zeros. That is, we have $R' < R$ regions with the associated costs of misclassification, and $Z < R' - 1$. Hence the algorithm continues the iterative removal of excess zeros by executing the applicable steps of steps 1 through 3 on the new arrangement of regions until the required number of zeros are removed.

A special case arises when the cost associated with more than one interval is the same and equal to the minimum cost, $C_m$. In this case, if the three steps of the algorithm would still permit a non unique method of eliminating intervals, the interval that, after its merger with its neighbors, results in a new interval of smallest cost is to be eliminated. (If Step 2. 3 is applicable, the interval that, when eliminated, reduced the number of zeros most, is to be absorbed into its neighbors.)

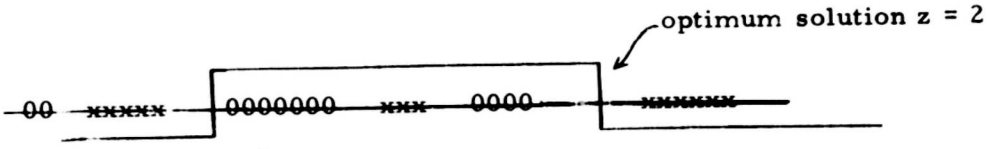
The above discussion of the algorithm is not intended to prove that a solution of minimum cost will, in fact, result. The algorithm has been applied successfully to several examples (some of which are

given below). Because of the limited applicability of one-dimensional discrimination, it was not deemed necessary to attempt the lengthy arguments involved in a rigorous proof of optimality.

A set of illustrative examples of the algorithm will now be given.

Example 1

Consider 13 samples of class 0 and 14 samples of class x given and distributed along v as shown, and we wish the optimum solution with 2 or less zeros if $C_D = C_A$.



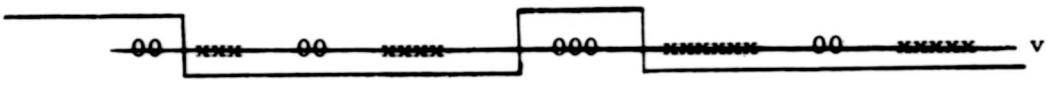

The sequence of costs are 2 5 7 3 4 6 and 5 zeros are necessary for a solution with no error. The excess zeros R - 1 - Z = 3 and we apply Step 2. 3 and reassign costs according to Step 3 to obtain the new sequence of costs 3 7 3 4 6, a solution that has 4 zeros or 2 excess zeros. The total cost of decision thus far is 2. We now follow the special procedure above, and eliminate the interval cost 3 with the most zeros. This is the middle 3, since elimination of the end 3 would still leave an excess zero. The final solution of 2 zeros is thus shown below. The cost of decisions is 5, since an additional 3 was accumulated in this last step.

The solution can be shown in a short hand notation given below. Here the cost is incurred by the last step, the cumulative cost of all the steps, and the instantaneous interval cost structure of every step in the algorithm is shown together with the number of zeros of the tentative function and the number of surplus zeros we must still eliminate.

| Cost of this step | Cumulative Cost | Interval cost structure | No. of Present Zeros | Excess Zeros |
|---|---|---|---|---|
| 0 | 0 | 2 5 7 3 4 6 | 5 | 3 |
| 2 | 2 | 3 7 3 4 6 | 4 | 2 |
| 3 | 5 | 3 8 6 | 2 | 0 |

Example 2

Three or less zeros of the function f(v) are desired.




| Cost of this step | Cumulative Cost | Interval cost structure | No. of Present Zeros | Excess Zeros |
|---|---|---|---|---|
| 0 | 0 | 2 3 2 4 3 6 2 5 | 7 | 4 |
| 2 | 2 | 2 5 3 6 2 5 | 5 | 2 |
| 2 | 4 | 2 5 3 9 | 3 | 0 |

There are a number of problems that must yet be investigated. Some of these are listed below.

a) Where in the boundary regions should the zeros of f(v) be placed?

b) How does overspecification of the number of zeros affect their locations? Suppose $p_A(v) - p_B(v) = 0$ in less than Z places.

Experiment with Optimum Small Sample Discrimination

In this experiment samples from two normal densities were used, class one with mean zero and variance one, class two with mean zero and variance nine. There were approximately 100 samples of each class, all greater than zero. That is, we used only the positive half of each distribution. Application of the algorithm to achieve a solution with one zero, resulted in a zero at 1.57. This is exactly the point at which the decision according to Bayes rule would change from class 1 to class 2 (moving left to right).

The program used to generate the two normal densities is described in Appendix 5A.

# VI RECOMMENDATIONS FOR FUTURE WORK

The material discussed in this report is concerned with decision making and classificatory problems in cases where a parametric representation of the physical world is available. Machine learning in decision making is considered as the task of automatically partitioning the parameter space into regions where in each region only members of one class are contained. Partitioning the parameter space is accomplished by subjecting the space to linear or non-linear transformations which cluster members of the same class and separate from each other members of different classes. The resulting partitioning of the parameter space assured the minimum probability of error in the decision making schemes discussed in this report. For each specification of the allowable complexities of the decision procedure, the methods described in this report provided the optimum decisions. The price paid for this performance was the computational complexity of the methods which, although not very complex, are not conveniently performed in real time.

The methods discussed in Section 5 of this report are intended to overcome this problem by providing methods that are inherently simple, from a computational point of view, and yet result in only slightly degraded performance from the best achievable results obtainable theoretically. These methods are motivated by decision theoretic considerations that attempt to minimize the probability of decision errors. Decision theoretic considerations, however, are compromised to the extent necessary to guarantee real-time machine-learning and recognition. Results achieved with these methods on a number of practical problems requiring extremely rapid data processing (speaker recognition) serve to demonstrate the potential of these methods.

Future work should be directed toward the continued development of adaptive, algorithmic techniques which can simulate and approximate likelihood ratio computation, a procedure known as best in the decision theoretic sense. In striving to achieve real time data processing techniques, however, one must not compromise performance to such an extent that undue limitations are placed on the ways in which parameter space can be partitioned.

## LIST OF REFERENCES

George Sebestyen, Classification Decisions in Pattern Recognition, MIT RLE Report 381, Parts II, III and V, also Litton Systems, Inc. Technical Memorandum No. 13 (Preprint of RLE Report)

S. Kullback, Information Theory and Statistics, Wiley, 1959

G. Sebestyen, "Recognition of Membership in Classes", IRE Transactions PGIT, January 1961.

Von Neumann, Various Techniques Used in Connection with Random Digits, Monte Carlo Method, National Bureau of Standards, 1951, pp. 36-38.

## APPENDIX 2A

## PROGRAM FOR EIGENVALUES AND EIGENVECTORS OF A REAL SYMMETRIC MATRIX

(Recomp II Users' Program No. 1068)

METHOD - JACOBI METHOD ref. Goldstein, Murray, and von Neumann. The Jacobi Method for Real Symmetric Matrices: Journal of the ACM, Vol. 6, (Jan., 1959), 59-96.

A sequence of orthonormal transformations, $U^i$, are found such that the matrix $A^i = U^i A^{i-1} U^{i*}$ is similar to $A^{i-1}$ and the sum of the squares of the off-diagonal elements, $\tau^i$, of $A^i$ is less than $\tau^{i-1}$ of $A^{i-1}$. In this way the sequence of matrices $A^i$ converges to a diagonal matrix with the eigenvalues on the diagonal. That is $U^i U^{i-1} \ldots . U^o A^o A^{o*} U^{1*} \ldots U^{i*}$ and $U^i U^{i-1} \ldots . U^o$ is the approximation to the matrix of eigenvectors.

For the $i^{th}$ iteration we pick an off-diagonal element of $A^{i-1}$, say $a^{i-1}_{jk}$, whose magnitude is greater than or equal to the average magnitude of the off-diagonal elements, i.e. $a^{i-1}_{jk}{}^2 \geq \frac{\tau^{i-1}}{n(n-1)}$ ($\tau^{i-1}$ is the sum of the squares of the off-diagonal elements of $A^{i-1}$). We then determine a matrix $U^i$, such that in the matrix $A^i = U^i A^{i-1} U^{i*}$, $a^i_{jk}$ is zero and $\tau^i = \tau^{i-1} - 2\left(a^{i-1}_{jk}\right)^2$.

Let $U^i$ be defined as follows:

$u^i_{mm} = \delta mn$ for $m \neq j$ or $k$ $\quad n \neq j$ or $k$

$u^i_{jj} = u^i_{kk} = \cos \phi$

$$u^{i}_{jk} = u^{i}_{kj} = \sin\phi$$

Then
$$a^{i}_{jj} = \frac{1}{2}(a^{i-1}_{jj} + a^{i-1}_{kk}) + \frac{1}{2}(a^{i-1}_{jj} - a^{i-1}_{kk})\cos 2\phi - a^{i-1}_{jk}\sin 2\phi$$

$$a^{i}_{kk} = \frac{1}{2}(a^{i-1}_{jj} + a^{i-1}_{kk}) - \frac{1}{2}(a^{i-1}_{jj} - a^{i-1}_{kk})\cos 2\phi + a^{i-1}_{jk}\sin 2\phi$$

$$a^{i}_{mj} = a^{i}_{jm} = a^{i-1}_{mj}\cos\phi - a^{i-1}_{mk}\sin\phi \quad m \neq j,k$$

$$a^{i}_{km} = a^{i}_{mk} = a^{i-1}_{mj}\sin\phi + a^{i-1}_{mk}\cos\phi \quad m \neq j,k$$

and
$$a^{i}_{jk} = a^{i}_{kj} = \frac{1}{2}(a^{i-1}_{jj} - a^{i-1}_{kk})\sin 2\phi + a^{i-1}_{jk}\cos 2\phi$$

All other elements remain unchanged

In order that $a^{i}_{jk} = 0$, $\phi$ must be defined as follows:

$$\tan 2\phi = \frac{2a^{i-1}_{jk}}{a^{i-1}_{kk} - a^{i-1}_{jj}}$$

Then $\cos\phi$ and $\sin\phi$ may be computed directly from:

$$R = \left[ (a^{i-1}_{kk} - a^{i-1}_{jj})^2 + 4(a^{i-1}_{jk})^2 \right]^{1/2}$$

$$\cos 2\phi = \frac{\left| a^{i-1}_{kk} - a^{i-1}_{jj} \right|}{R}$$

$$\sin 2\phi = \frac{2a_{jk}^{i-1}}{R} \qquad \left[\text{Sign } (a_{kk}^{i-1} - a_{jj}^{i-1})\right]$$

$$\cos \phi = \left[\frac{1}{2}(1 + \cos 2\phi)\right]^{1/2}$$

$$\sin \phi = \frac{1}{2} \quad \frac{\sin 2\phi}{\cos \phi}$$

The $i^{th}$ iterant of the matrix of eigenvectors $X^i = U^i X^{i-1}$ is also computed

$$x_{jm}^{i} = x_{jm}^{i-1} \cos \phi - x_{km}^{i-1} \sin \phi \qquad m = 1, 2, \ldots, n$$

$$x_{km}^{i} = x_{km}^{i-1} \cos \phi + x_{jm}^{i-1} \sin \phi$$

All other elements remain unchanged.

To show that the process converges:

$$\text{Clearly } \tau^i = \tau^{i-1} - 2(a_{jk}^{i-1})^2$$

and since $(a_{jk}^{i-1})^2 \geq \frac{\tau^{i-1}}{n(n-1)}$

$$\tau^i \leq \tau^{i-1}\left(1 - \frac{2}{n(n-1)}\right) \leq \tau^i \left[e^{-\frac{2}{n(n-1)}}\right] \leq \tau_o \, e^{-\frac{2i}{n(n-1)}}$$

## STORAGE

The program occupies relative locations 0000-0367 (OCTAL). For an N x N matrix $4N^2$ additional consecutive locations are required.

## INSTRUCTIONS FOR USE

a) INPUT

In rel. loc 0000 store $A_o$, the location at which the matrix begins in the left half address.

In rel. loc. 0001 store N, the size of the matrix, in binary at b18

In rel. loc. 0062.1 store exit; the location to which you wish to return on completion of computation (do not destroy the rest of 0062)

Beginning in location $A_o$ store the matrix row-wise in floating point. Although a square matrix is required as input, it is not actually necessary to store those elements which are below the principal diagonal. Only the upper triangle is used in computation. Start by transferring to relative location 0002.0.

b) OUTPUT

At the completion of the computation the eigenvalues are stored on the principal diagonal of the original input matrix. (The upper half of the input matrix is destroyed.) The eigenvectors are stored row-wise immediately following the diagonal matrix of eigenvalues. (in locations $A_o + 2N^2$ to $A_o + 4N^2 - 1$.) Control is transferred to the address stored in 0062.1.

## RELOCATION

The program is relocatable. The tape includes the relocation matrix and AN-004. The program may only be relocated to locations of the form XX00: XX10: XX50: XX60.

## LIMITATIONS

The minimum matrix size is 2 x 2.

The maximum size is 30 x 30, assuming optimum placement of program and matrix. There is no check to determine whether the capacity of the machine has been exceeded (when $A_o + 4N^2 > 7760$).

## TIME

The computation takes approximately $.7N^3$ seconds for an N x N matrix.

## ACCURACY

Theoretically, if $\tau^i \leq \epsilon$, the maximum error in each of the eigenvalues is $\epsilon^{1/2}$, $\left[\left|a^i_{jj} - \lambda_k\right| < \epsilon^{1/2}\right]$. The error in the components of the eigenvectors is of the same order of magnitude.

Exit from the program takes place when $\tau^i < \epsilon$. $1/2\,\epsilon$ is stored in floating point in relative locations 0076 and 0077 and is presently $1/2 \times 2^{-38}$. Thus the error in the eigenvalues is $\pm .7 \times 2^{-19} = \pm 1.3 \times 10^{-6}$.

In practice, for large matrices the error may be greater than $\epsilon^{1/2}$ because of cumulative computational errors. To increase the accuracy it may be desirable to decrease $\epsilon$. An estimate for the optimum value of $\epsilon$ is $\epsilon^{1/2} = 2(N^{1/2} + 7)N^2 2^{-s}$ where s is the number of significant binary digits in the input matrix. The error in the eigenvalues for this value of $\epsilon$ is

$$E < \epsilon^{1/2}\left[1 + .69\left|\log_2 \epsilon^{1/2}\right|\right] + 13N^{5/2}2^{-s}$$

(Experimentally it has been found that this overestimates the error by about a factor of 10).

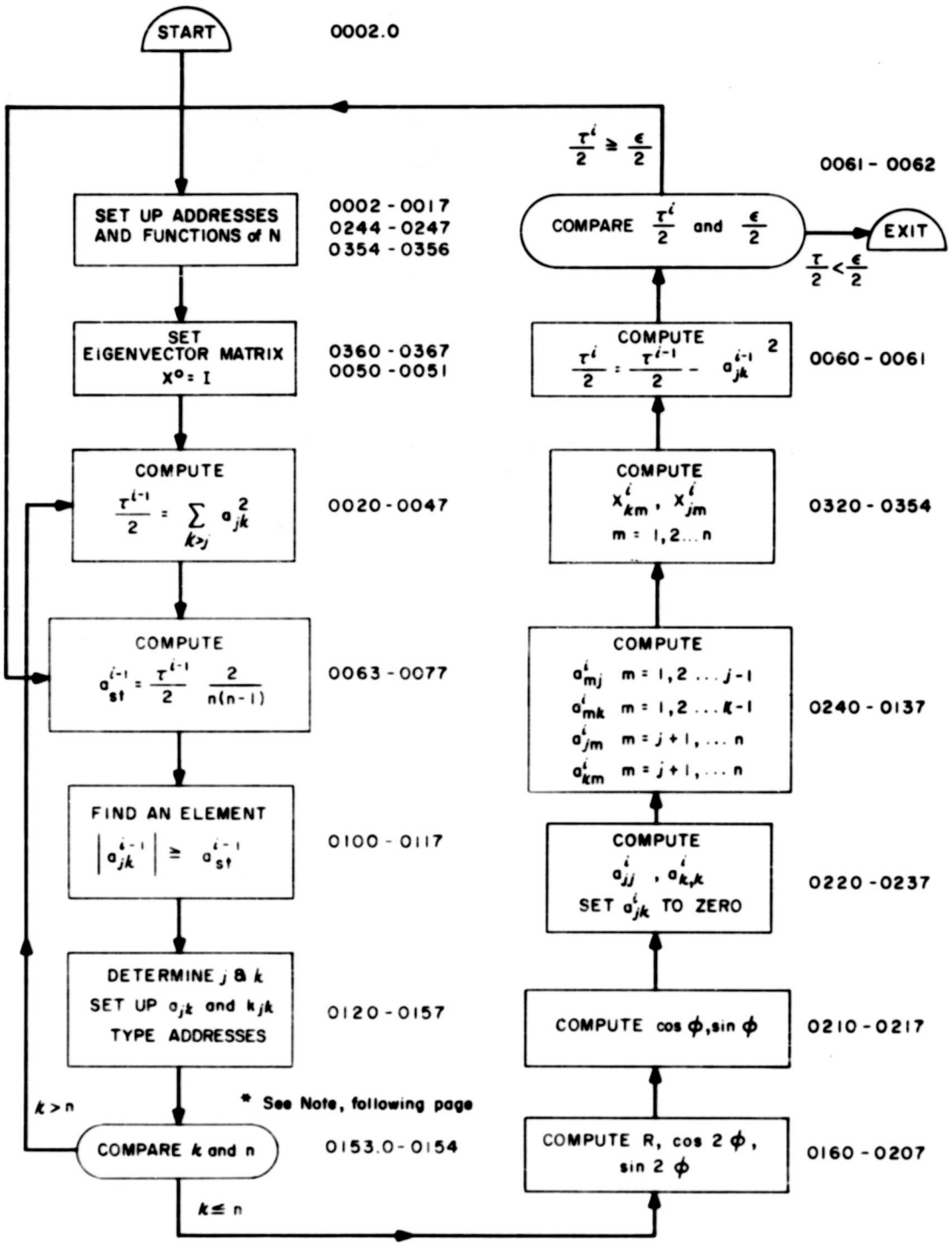
## COMMENTS

1. The matrix may be singular
2. There is no error halt.
3. The program resets itself. N, $A_o$ and the exit are not destroyed.
4. If only the eigenvalues are desired, the eigenvectors may be avoided by inserting the following changes.

   in rel. loc. 0356 put +57 0054.0 + 40 0000.0

   in rel. loc. 0315 put +57 0345.0 + 40 0000.0

   Store some large floating point number in location $A_o + (2N^2)_8$ (large means greater than $1/2$.)
5. A matrix consisting of all zeros will not be tolerated.




FLOW CHART FOR EIGENVALUE PROGRAM

*NOTE: Since $\frac{\tau^i}{2}$ is computed from $\frac{\tau^{i-1}}{2} - \left(a_{jk}^{i-1}\right)^2$ it is possible that the errors that accumulate because of the subtractions will be such that $\frac{\tau^i}{2}\left(\frac{2}{n(n-1)}\right)$ is greater than any off diagonal element. When this happens $\tau^i$ is recomputed from the actual off-diagonal elements. Notice that the estimate of $\frac{\tau^i}{2}$ will always be greater than it should be, which is actually an advantage until the above phenomenon occurs.

## APPENDIX 2B.

## DESCRIPTION OF PROGRAM FOR MAKING CLASSIFICATORY DECISIONS THROUGH USE OF LINEAR CLASS-SEPARATING TRANSFORMATIONS

Computation of a transformation that maximizes the distance between different classes while clustering members of the same class was programmed for the Recomp II computer in the following manner.

Given: a data matrix $[F]_k$ containing M samples of N dimensional vectors for class k.

Restrictions: $2 \leq N \leq 30$

$1 \leq k \leq 20$

no restriction on M

Block One

The given data is converted to floating point form and is punched, with proper identification, on tape. Due to storage limitations, if the number of points (M x N) exceeds 900, output is separated in blocks of $[900/N^2] \times N^2$ values. ([ ] indicates the integer part of the division.)

Block Two

Column mean vectors, $[\bar{f}_{kn}]$, where $\bar{f}_{kn} = \frac{1}{M} \sum_{m=1}^{M} f_{kmn}$, for $n = 1, 2, \ldots, N$ are computed, stored and punched.

Difference matrices,

$$[D]_{kl} = [\bar{f}_{kn} - \bar{f}_{ln}]^T [\bar{f}_{kn} - \bar{f}_{ln}],$$

where k and l denote different classes are formed and punched.

Block Three

The covariance matrix of a class is defined as

$$[U]_k = \frac{1}{M} [Q^T Q]$$

where $[Q] = [F]_k - [\mu]_k$

$$[\mu]_k = \begin{bmatrix} \bar{f}_{k1} & \bar{f}_{k2} & \cdots\cdots\cdots & \bar{f}_{kn} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \bar{f}_{k1} & \bar{f}_{k2} & \cdots\cdots\cdots & \bar{f}_{kn} \end{bmatrix}$$

To eliminate any restriction on the number of samples, portions of $Q^T Q$ are formed, stored and later added together to form $[U]_k$ which is then punched.

Block Four

To afford flexibility in the method of forming $[\alpha]$ and $[\beta]$ matrices for input to $\alpha - \lambda\beta$ routine, the following options are available.

Option No. 1 -

$$[\beta] = \sum_{I=1}^{C_3} [D]_{k1} + \sum_{I=1}^{C_2} [U]_k$$

$$[\alpha] = \sum_{I=1}^{C_1} [U]_k$$

Option No. 2 - $$[\beta] = \sum_{I=1}^{C_1} [U]_k$$

$$[\alpha] = \sum_{I=1}^{C_2} [U]_k + \sum_{I=1}^{C_3} [D]_{kl}$$

where $C_1$, $C_2$ and $C_3$ are any desired number of matrices to be summed and $C_2$ may be zero.

If $N \leq 20$, the $[\alpha]$ and $[\beta]$ matrices are punched on one tape sector. If $N > 20$, $[\beta]$ is punched followed by a start 0000.0 code and then $[\alpha]$.

Block Five

Two versions of $\alpha - \lambda\beta$ routine are available. One, for $N \leq 20$, handles all computation and options within the machine; the other, for $20 < N \leq 30$, requires two instruction tapes and punches some intermediate results.

The computation performed is as follows:

Input - $[\alpha]$ and $[\beta]$ to $\alpha - \lambda\beta$ routine

Let $[\lambda_i]$ = eigenvalues of $\beta$

$[C_{ij}]$ = eigenvectors of $\beta$

$$[F^*] = [\lambda_i^{-\frac{1}{2}}] [C_{ij}]$$

$$[A^*] = F^* A F^{*T}$$

At this point there are two choices of output:

No. 1

Let [ V] = eigenvector associated with the principal eigenvalue of $A^*$. Then output is:

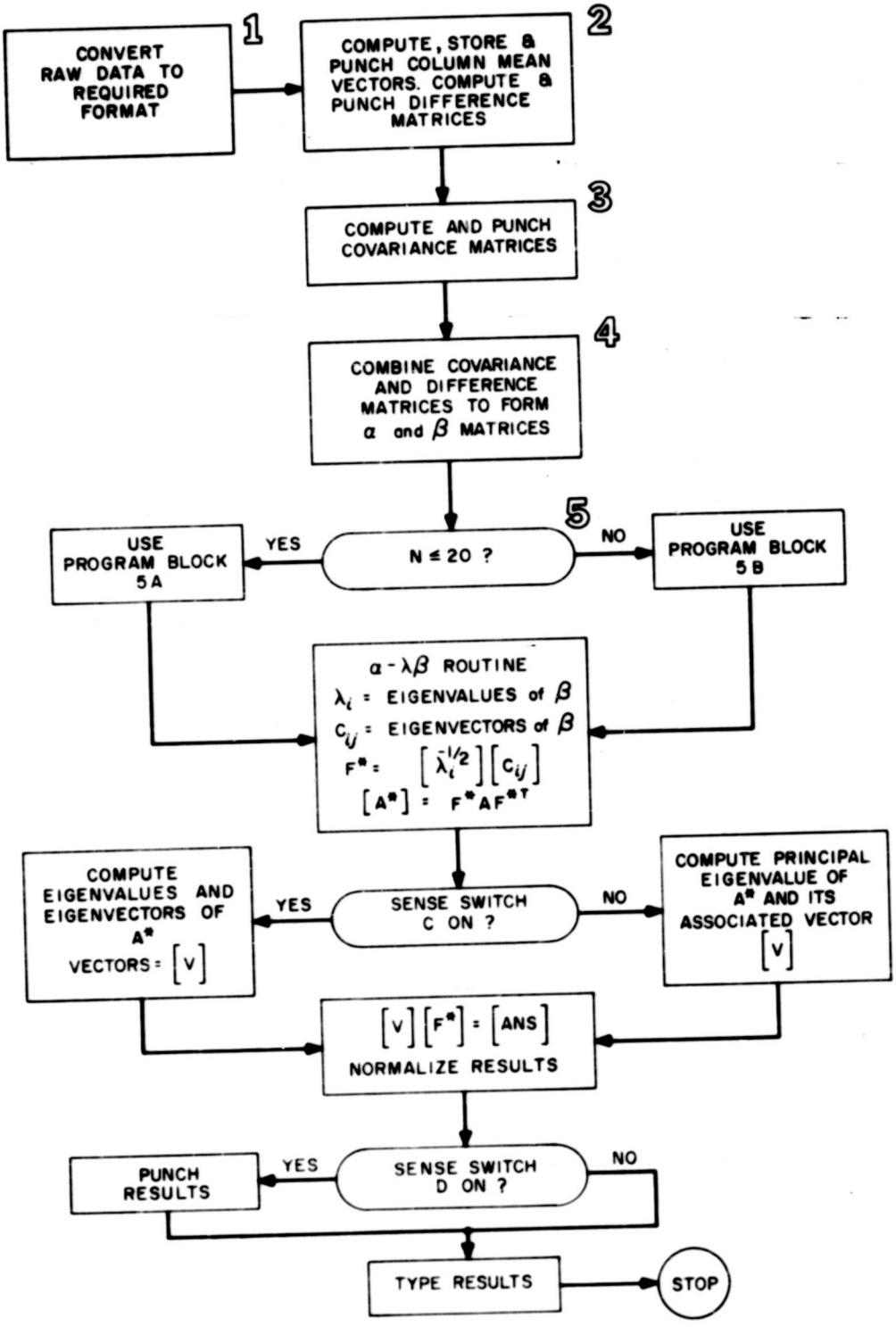
1. Principal eigenvalue of $A^*$
2. [ V] [ $F^*$ ] (normalized)
3. Normalizing factor

No. 2

Let [ V] = eigenvectors of $A^*$. Then output is:

1. Eigenvalues of $A^*$
2. [ V] [ $F^*$ ] (normalized if $N \leq 20$)
3. Normalizing factor, if any.

Normal output is typed. Punched output is available on a sense switch option.




FLOW CHART OF PROGRAM FOR LINEAR-SEPARATING METHOD

## APPENDIX 3A

## PROGRAM FOR NON-LINEAR DISCRIMINATION PROBLEM

METHOD: It has been shown that the function which minimizes the sum of the variances of two classes while keeping the distance between the means constant is

$$v = \frac{f(x_1, x_2, \ldots, x_n) - Kg(x_1, x_2, \ldots, x_n)}{f(x_1, x_2, \ldots, x_n) + Kg(x_1, x_2, \ldots, x_n)}$$

where $f(x_1)$ and $g(x_1)$ are the n-variable probability densities of the two classes. In addition, the decision rule is

$$v > 0 \text{ (implies } \frac{f(x_i)}{g(x_i)} > K) \text{ decide } x_i \in f$$

$$v < 0 \text{ (implies } \frac{g(x_i)}{f(x_i)} > \frac{1}{K}) \text{ decide } x_i \in g$$

We obtain a polynomial approximation to the function v by first finding the orthonormal polynomials, $\phi_i$, of $h(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n) + g(x_1, x_2, \ldots, x_n)$.

The approximation, $v_N$, is then expressed as a sum of the orthonormal polynomials as follows.

$$v_N = \sum_{i=0}^{N} v_i \phi_i \quad \text{where } v_i = \int \phi_i(x_i)\, dh$$

$v_N$ is then the N term polynomial which minimizes the expected mean-square error $\left[\int (v_N - v)^2 \, h \, d\, x_i\right]$. (Where $N = \binom{n+p}{p}$ and p is the maximum degree of any term.)

THE PROGRAM: The program is divided into four parts.

Part 1. Formation of the vectors $\alpha_K$. This program, given a set of n-variable data, will form the vectors $\alpha_K = x_1^d x_2^m, \ldots, x_n^r$ for $d + m + \ldots + r \leq p$. That is, it forms a vector consisting of all products of all n variables of degree zero to p. There are $N = \binom{n+p}{p}$ such products.

Part 2. Formation of the matrix

Given $N_1$ samples from class f and $N_2$ samples from class g in the form of vectors $\alpha_K$, this program forms the matrix $[\gamma_{ij}] = [\frac{1}{2}(\overline{\alpha_i \alpha_j}^f + \overline{\alpha_i \alpha_j}^g).]$

This is accomplished by first forming

$$U^F = A^{F^T} A^F$$

$$U^G = A^{G^T} A^G$$

where $A^F$ and $A^G$ are $N_1$ x N and $N_2$ x N matrices whose rows are vectors $\alpha$ from class f and g respectively.

The matrix $\gamma$ is then given by

$$\gamma = \frac{1}{2}\left(\frac{1}{N_1} U^F + \frac{1}{N_2} U^G\right)$$

and is an N x N matrix.

Part 3. Formation of the functions, $v_n$.

Given the matrix $\gamma$ from part 2, this program computes the N polynomial approximations to the function, v. There are N of these polynomials, the $n^{th}$ one having n terms.

The coefficients, $a_{ik}$, of the polynomial $\phi_i = \sum_{k=0}^{i} a_{ik} \alpha_k$ (where $\alpha_k$ is an element of the form $\alpha_k = x_1^d x_2^m, \ldots, x_n^r$) are given by the solutions to the set of equations

$$\sum_{k=0}^{i} \left(\frac{a_{ik}}{a_{ii}}\right) \gamma_{kj} = -\gamma_{ji} \quad \text{for} \quad j = 0, 1, \ldots, i-1$$

and $$a_{ii}^2 = \frac{D_i}{D_{i+1}}$$

where $$\gamma_{ij} = \frac{1}{2} (\overline{\alpha_i \alpha_j}^f + \overline{\alpha_i \alpha_j}^g)$$

and $D_i$ is the determinant of the (i x i) $\gamma$ matrix and $D_{i+1}$ of the [(i+1) x (i+1)] $\gamma$ matrix.

To express the function $v_N$ in terms of the polynomial, we determine the coefficients

$$v_i = (1+K)\, E_f(\phi_i) \qquad [E \text{ is expectation}]$$

$$v_0 = 1 - K$$

Hence, writing $v_n$ as $\sum_{k=0}^{n} b_{nk} \alpha_k$

the coefficients are

$$b_n = \sum_{i=1}^{n} (1+K)\, E_f(\phi_i)\, a_{ik}$$

$$b_{n0} = 1 - K$$

Part 4. Testing the functions, $v_n$.

In order to test the decision rule ($v(x_i > 0 \quad x_i \in f$ etc.), this program accepts vectors from part 1, substitutes them into the function $v_n$ and determines error rates for all the functions $v_n$. That is, if a vector $\alpha$ is actually a member of class f, but the decision rule identifies it as a member of g, an error (type 1) is counted: (The converse for member of g (type 2)). Thus, the ratio of errors to total samples tested is computed for classes f and g for all functions $v_n$.

## APPENDIX 5A

## GENERATION OF UNIVARIATE DENSITIES

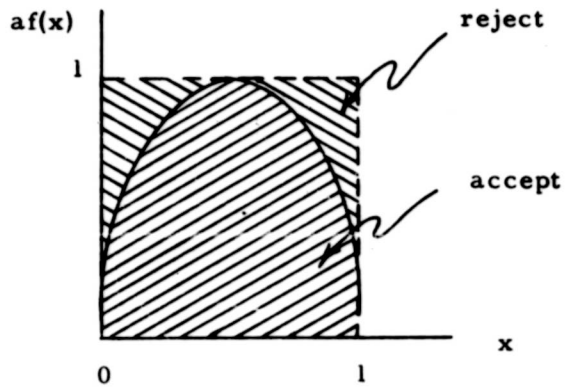
It is easily shown that the cumulative distribution function of a random variable is uniformly distributed on (0, 1).

$$t = F(x) = \int_0^x f(x)\,dx \quad \text{and} \quad x = \phi(t)$$

Thus if a random variable T is uniformly distributed on (0, 1), $X = \phi(T)$ is a random variable with density f(x).

To generate samples from a normal distribution the above method was used with a table look-up substituted for the explicit determination of $\phi(T)$. F(x) was tabulated for $0 \leq x < 5$ in increments of .01. Then, for a given value of T = F(x), the corresponding value of x may be found from the table. (if $T < 1/2$, $\phi(1-T) = -X$). The standardized variable, x, is then converted to a variable with the desired mean and variance, $y = \sigma x + \mu$.

Often the inverse function $x = \phi(t)$, is awkward to determine. If both f(x) and x are bounded, a different approach is sometimes feasible. Scale f(x) such that $a\,f(x) \leq 1$. Then produce two random variables, X and Y, each uniformly distributed on (0, 1). If $Y > af(x)$, the pair is rejected. If $Y \leq a\,f(x)$ is accepted as the next sample. The accepted x's are then distributed according to f(x).




This method was used to generate samples with the density

$$af(x) = (1-x^2)^{1/2} \quad (0 \leq x \leq 1)$$

The test $Y \leq (1-X^2)^{1/2} \Longrightarrow X^2 + Y^2 \leq 1$.

We also considered the distribution

$$f(x) = \frac{1}{\pi} \quad \frac{1}{(1-x^2)^{1/2}} \quad (-1 \leq x \leq 1).$$

Since f(x) is unbounded, the previous trick does not apply. However

$$t = F(x) = \frac{1}{\pi} (\sin^{-1} x + \pi/2)$$

and

$$x = \sin(\pi t + \pi/2 = \sin(\pi t - \pi/2)$$

To avoid explicit computation of the sine, the following method is used. We again generate two variables, Z and Y, uniformly distributed on (0, 1)

to obtain the angle $\phi = \tan^{-1} Z/Y$. To insure that $\phi$ is unifo
on $(0, \pi/2)$ we accept only those Z, Y pairs lying within the u
$(Z^2 + Y^2 \leq 1)$. Then $x = \sin(2\phi - \pi/2) = -\cos 2\phi = \sin^2\phi -$

$$x = \frac{Z^2 - Y^2}{Z^2 + Y^2}$$

---

[9] Von Neumann, Various Techniques Used in Connectio
Random Digits, Monte Carlo Method, National Bureau of Sta
1951, pp. 36-38.

UNCLASSIFIED

UNCLASSIFIED