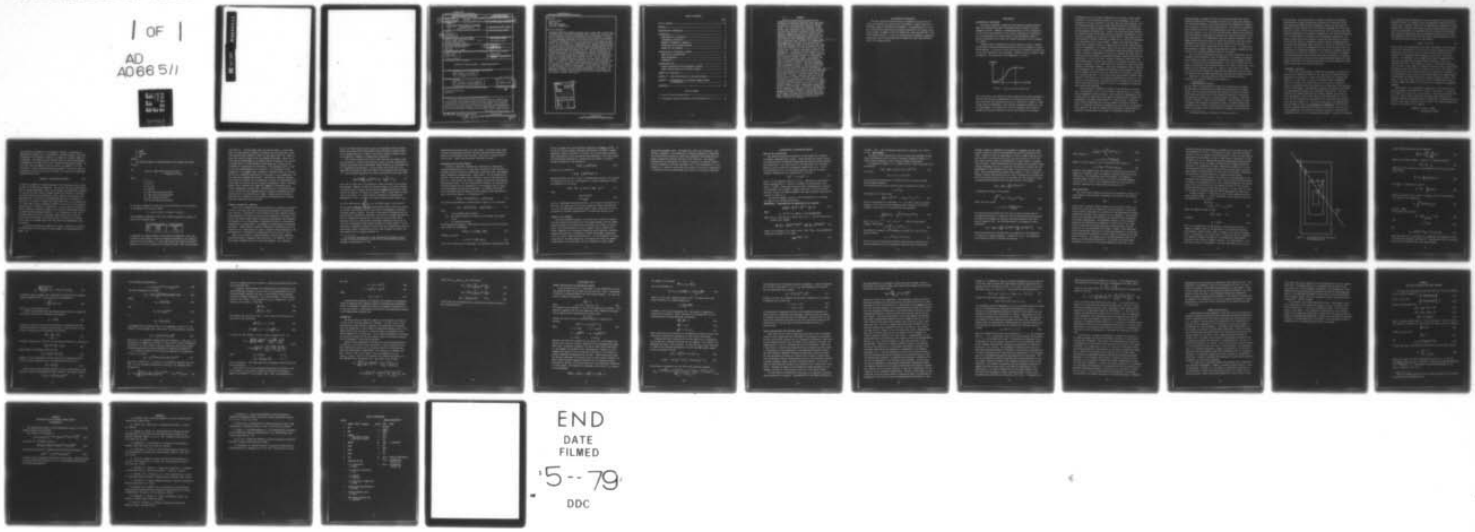
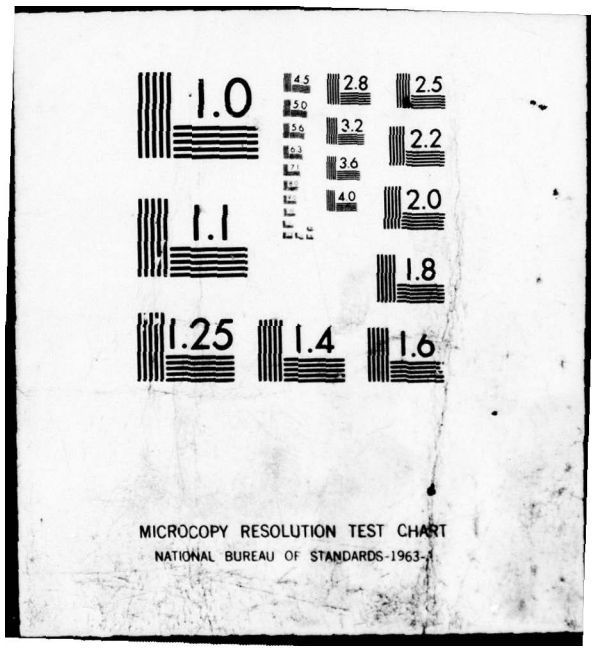
AD-A066 511 DAVID W TAYLOR NAVAL SHIP RESEARCH AND DEVELOPMENT CE--ETC F/G 12/1
ACCURACY AND EFFICIENCY IN PATTERN CLASSIFICATION.(U)
MAR 79 S BERKOWITZ

UNCLASSIFIED DTNSRDC-79/029 NL

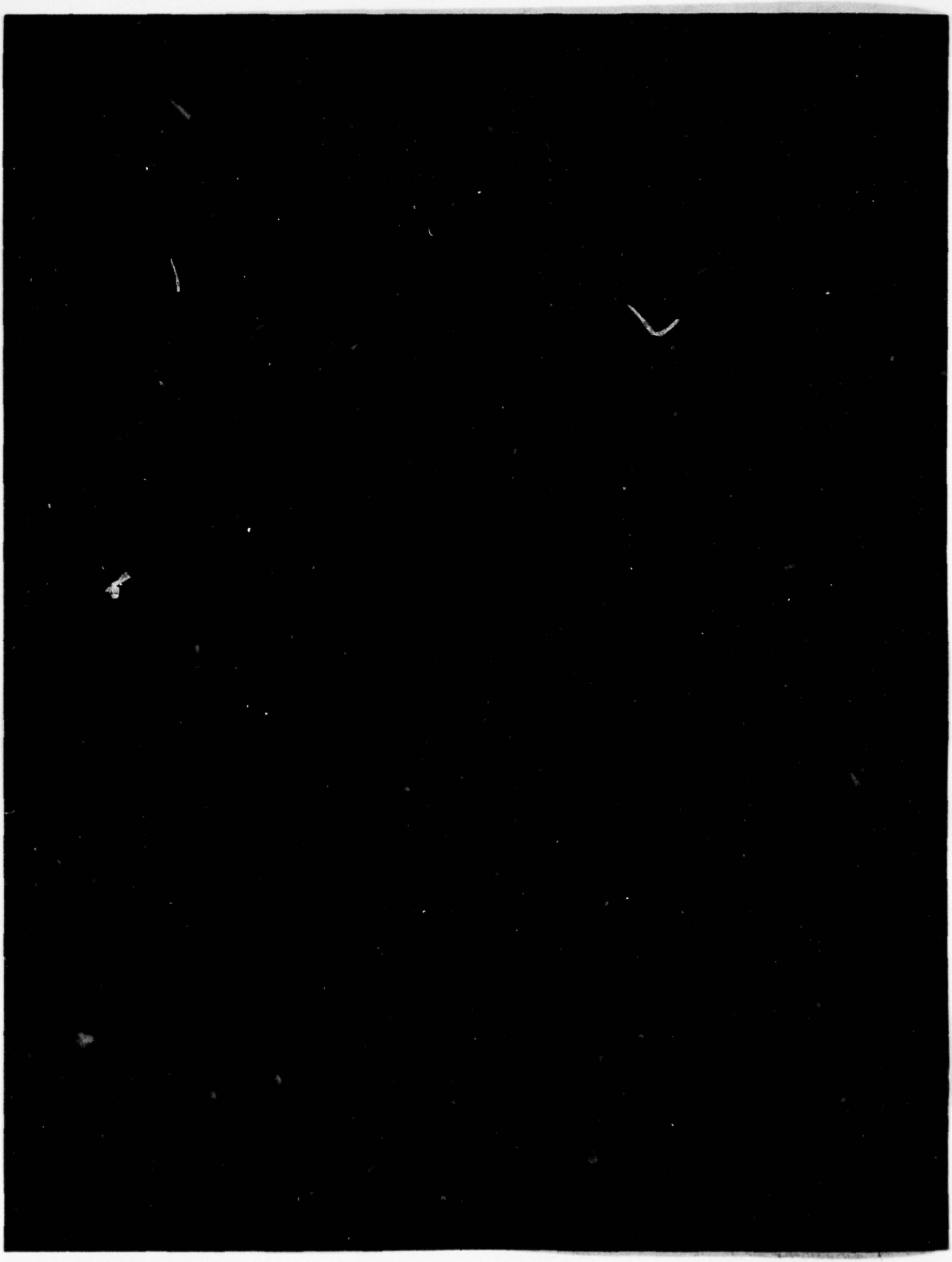
| OF |
AD
A066 511



END
DATE
FILMED
5-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

DDC FILE COPY AD A066511



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

(14) 1. REPORT NUMBER: DTNSRDC-79/029

2. GOVT ACCESSION NO.

(9) Research and development rept.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

(6) ACCURACY AND EFFICIENCY IN PATTERN CLASSIFICATION

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

(10) S. Berkowitz

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
David W. Taylor Naval Ship Research
and Development Center
Bethesda, Maryland 20084

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
(See reverse side)

11. CONTROLLING OFFICE NAME AND ADDRESS
Naval Sea Systems Command
Code 03F
Washington, D.C. 20360

(11)

12. REPORT DATE
March 1979

13. NUMBER OF PAGES
44

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

(12) 45 p.

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

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

(16) SR01403

18. SUPPLEMENTARY NOTES

(17) SR014031

DDC
RECEIVED
MAR 28 1979
A

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

Pattern Classification
Statistics

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

A pattern classification scheme which is grounded in classical probability theory may be associated with confidence intervals that represent an estimate of the predictive capability of the scheme. As a practical matter, realistic allocations of data acquisition and processing resources may severely constrain acceptable levels of predictability. Motivated by pattern classifications techniques which represent radical departures from

(Continued on reverse side)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

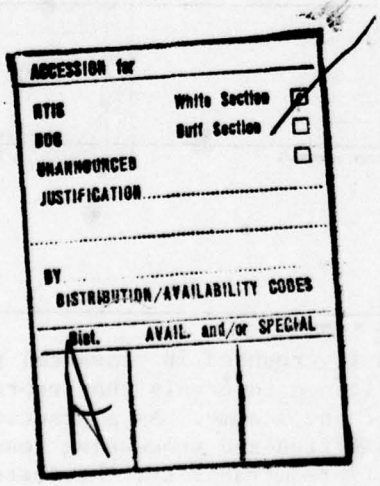
387 682
79 03 27 001

(Block 10)

Element 61153N
Task 15321
Task Area SR0140301
Work Unit 1-1808-010-14

(Block 20 continued)

standard statistical hypothesis testing, such as those based on fuzzy logic, we examine some of the basic assumptions which underlie the standard techniques and briefly discuss their justification or usefulness. In particular, we show that fuzzy logic effectively produces conservative estimates for the conditional probability of the union of sets since, in that case, it neglects information related to the intersection. We propose that such neglect can be remedied, at a computational cost, without resorting explicitly to the usual procedure of integrating over irregularly shaped volumes. To this end, but only by way of extended example, we introduce a class of probability density distributions which, under conditions developed in the paper, possesses (hyper) rectangular contour. Explicit formulas for the normalization constant and the probability of error are then derived for typical distributions. By the introduction of suitable formal approximations, the binary classification problem is solved for two sample distributions of truncated domain and differing roll-off. Finally, liberties taken with the additive property of density functions permit the realization of a piecewise linear discrimination logic for non-rectangular contours. The result is a sample classification scheme that avoids some of the computational expense or unduly constraining assumptions implied in the derivation of classical discriminants. At the same time it takes some rational advantage of all the information available to the analyst. The suggestion is implicit that the methods demonstrated in the example are applicable to diverse situations which provide widely varying information as to domain, roll-off, and contour.



## TABLE OF CONTENTS



## LIST OF FIGURES

## ABSTRACT

A pattern classification scheme which is grounded in classical probability theory may be associated with confidence intervals that represent an estimate of the predictive capability of the scheme. As a practical matter, realistic allocations of data acquisition and processing resources may severely constrain acceptable levels of predictability. Motivated by pattern classifications techniques which represent radical departures from standard statistical hypothesis testing, such as those based on fuzzy logic, we examine some of the basic assumptions which underlie the standard, statistical techniques, and briefly discuss their justification or usefulness. In particular, we show that fuzzy logic effectively produces conservative estimates for the conditional probability of the union of sets since, in that case, it neglects information related to the intersection. We propose that such neglect can be remedied, at a computational cost, without resorting explicitly to the usual procedure of integrating over irregularly shaped volumes. To this end, but only by way of extended example, we introduce a class of probability density distributions which, under conditions developed in the paper, possesses (hyper) rectangular contour. Explicit formulas for the normalization constant and the probability of error are then derived for typical distributions. By the introduction of suitable formal approximations, the binary classification problem is solved for two sample distributions of truncated domain and differing rolloff. Finally, liberties taken with the additive property of density functions permit the realization of a piecewise linear discrimination logic for non-rectangular contours. The result is a sample classification scheme that avoids some of the computational expense or unduly constraining assumptions implied in the derivation of classical discriminants. At the same time it takes some rational advantage of all the information available to the analyst. The suggestion is implicit that the methods demonstrated in the example are applicable to diverse situations which provide widely varying information as to domain, roll-off, and contour.

## ADMINISTRATIVE INFORMATION

The basic ideas for distributions of rectangular contour were conceived in 1970 when the author was at the Technion - Israel Institute of Technology under a post-doctoral fellowship. An implementation of one of the classification algorithms was realized successfully in a speech recognition project at NSRDC in 1973 for NAVSEA 0311, Task Area SR0140301, Task 16565, Element 61153N. In 1977, the subject was re-considered, sharpened, generalized, and is here documented for NAVSEA 03F, Task Area SR0140301, Task 15321, Element 61153N.

# INTRODUCTION

## PREDICTABILITY, EFFICIENCY

This report describes a scheme for discriminating between patterns drawn from distributions whose isoprobable density contours are hyper-polygonal surfaces. However, an underlying purpose of this Introduction is not so much the motivation of yet another classification algorithm as it is the discussion and clarification of some fundamental issues in pattern classification for which the proposed algorithm is but an example of a remedy.

Current trends in hypothesis testing as applied to pattern classification exhibit certain characteristics which this work attempts to identify and exploit. If one were to plot the accuracy of prediction against the cost of data acquisition and processing, a graph like that in Figure 1 might be expected. The graph shows two breakpoints: the first

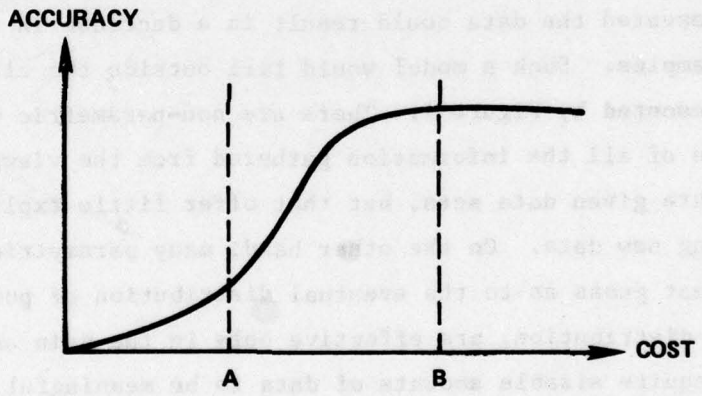



Figure 1 - Form of Accuracy versus Cost

(A) at the start of an interval in which enough data have been assembled to initiate the structural approximation of the density distributions under test; the second (B) indicating the onset of saturation, a region of diminishing return corresponding to the limiting convergence property of a law of large numbers. The lower end of the curve suggests that either not enough data have been collected to reveal the structure of the sample

distributions or that the model chosen does not represent a large enough investment in data processing capability to take advantage of the fine information structure available in the data that have been collected. The gain region indicates an area in which sufficient data are available and one has room to trade off models of increasing complexity and (hopefully) accuracy against increasing cost. Finally, the saturation region reflects our axiomatic belief in the limit nature of frequency distributions, as reflected, for example, in a weak law of large numbers. This region is the least sensitive to model trade-offs, although it can certainly be depressed by a sufficiently poor choice of model. Attempts have been made both to measure and to improve accuracy at reduced cost in the gain region between A and B. Measurement, for example, might take the form of calculating confidence intervals, albeit for an arbitrarily chosen distribution, often the Normal. Similarly, improvement of accuracy at reduced cost might involve assuming distributions whose resulting discriminant requires relatively few computations. On the other hand, choosing a parametric model which seriously misrepresented the data could result in a decrease in accuracy for an increase in samples. Such a model would fall outside the class of decision surfaces represented by Figure 1. There are non-parametric models that will take advantage of all the information gathered from the viewpoint of being able to separate given data sets, but that offer little explicit rationale for forecasting new data. On the other hand, many parametric models, which represent a best guess as to the eventual distribution of points or a statistic of the distribution, are effective only in the gain and saturation regions and require sizable amounts of data to be meaningful, discounting prior engineering knowledge.

Even the axioms of probability theory are not impervious to distortion - intentionally or not - for the sake of reducing cost. In a sense, some non-parametric models represent a tacit option for computational convenience and cost reduction to the neglect of the axiom that requires the convergence of frequency distribution to a limit function. Some parametric models acknowledge the axiom by assumption but make no serious effort to empirically justify the assumption on the basis of the available data, again for the sake of computational convenience. Models based on potential functions which are related to clustering procedures and

nearest-neighbor techniques[1*] present an interesting compromise between classical parametric and non-parametric models in the sense that densities are effectively assumed at each sample point and are independently summed to form an effective, overall density distribution. The axioms of convergence and of additivity are both thus manipulated for the sake of efficiency. In this case, however, it is not clear that arbitrarily assumed local densities can be effectively extrapolated to determine the asymptotic behavior of the whole distribution. The independence and usual spherical symmetry of the assumed densities may especially be called into question, but even if these assumptions are accepted, the local density should not be expected to do more than predict where future samples are likely to fall. In other words, distributions of bounded support are perhaps more realistic assumptions and they will be accordingly addressed later in this report.

As in the case of potential functions, an axiom of probability that is sometimes tacitly modified is that of additivity. For example, the application of fuzzy set theory to pattern classification can be regarded in practice - theoretical protests notwithstanding - as a theory derived from the axioms of probability with the modification of the axiom of additivity. This subject will be reviewed further on. Indeed, a basic aim of this report will be to show how fundamental assertions about probability can be reasonably modified and yet yield classification algorithms with predictive power.

Basically, it is argued here that the two principal functional objectives of a pattern classification scheme are

- o predictability
- o computational efficiency

and that, while the theoretical foundations for prediction are fragile, the knowledge and predilection for computational efficiency are robust. Indeed, even if one knew of methods that realized a strong relationship between past and future events, the financial limitations of any particular application would force a trade-off between the amount of data that could be collected - hence, the predictability - and the amount of human or computer time available to process the data. As matters stand, many experiments are by nature data-limited, and no amount of processing justifies the confidence that is often placed in forecasts that are derived

---

*A complete listing of references is given on page 39.

from the data. On the other hand, one should not jump to the extreme conclusion that the gentle skepticism with which we review some of the basic assumptions underlying both hypothesis testing and the concept of probability itself implies that classification techniques are of no value. Quite the contrary, the notion that the future lies in the past, even when sketchily described, is the essential ingredient in the human understanding, of or at least belief in, causality and the meaning of experience. The ideas in this report do not advocate a break with the empirical tradition of letting the results of experiment determine whether assumptions about the statistical mechanisms used to describe a process are justified or need revision. All that will be suggested is that, on the one hand, since the reliability of forecasting techniques is ill-founded in principle and sufficient data are often not available in fact, one is entitled to indulge a bent toward assumptions about nature that permit convenience in computing tests of prediction; on the other hand, these assumptions need not be made rigidly and blindly, but rather should take into account engineering knowledge or biases about the process in question.

CONVERGENCE, ADDITIVITY

At the root of any attempt to predict the occurrence of events by an application of probability theory is the idea that the bounded relative frequency of occurrence is a point set function that will converge with sample size in the limit to a point set measure called a probability. That is, for sample size n drawn say from the real line, a sequence $F_n(x \leq X)$ is hypothesized which monotonically increases with X and converges with n to some $P(x \leq X)$. Alternatively, frequency distributions f(x) might be considered which converge to a probability density p(x). The probability of the random event $A \subseteq R^n$ is, roughly speaking, the value of a real, non-negative, additive set function P(A) whose domain is a Borel field of sets B and which is normalized in the sense that $P(U_B)=1$, where $U_B$ is the unit element of B. The two jarring notes in the applied probability scheme of things are convergence and additivity. While we have intuitive explanations available to bolster our belief in the efficacy of those assumptions, our desire for theoretical elegance or computational convenience may be in fact the dominating prejudice. Moreover, the evidence

as to whether limited experiments of the coin-tossing or even statistical mechanics variety are consistent with the assumptions noted does not provide us with a logically tight case for employing the assumptions in other problem areas, especially when there are severe restrictions on the amount of data to be gathered. The possibilities that frequency distributions might radically change their form or parameters their value as more data are accumulated, the weak law of large numbers notwithstanding, or that, for mutually exclusive events A,B, the equation

$$P(A \cup B) = P(A) + P(B) \tag{1}$$

may not hold, are considerations that temper our dogmatic attitude toward the peculiar representation of the future offered by probability theory. Indeed, in recent years at least one theory, namely Zadeh's fuzzy set theory[2,3], has challenged these assumptions and offered - with reservation[3,4] - a weakening of the assumptions and hence an apparently flexible means of describing the persistence of trends or at least our intuition about them. For example, in lieu of a conditional probability limit function, we may speak of a membership function $\mu(A|x)$ that describes the degree to which an element is a member x of a set A. The empirical means by which we capture the function is left unspecified, although it is not unreasonable that a frequency distribution might be an adequate means of expressing the normative quality of concepts regarded as fuzzy - like young. On the other hand, we are free to cast the quantitative expression of a membership function in the world of our subjective, engineering judgment.

With respect to the assumption of additivity, first note how inclusion and union are calculated in the theory of fuzzy sets. A is said to be included in B, $A \subseteq B$ if $\forall x$, $\mu(A|x) \leq \mu(B|x)$. The union of two fuzzy subsets A,B - each of which contains by definition every element of the universe with some non-negative degree of membership - is defined to be the smallest fuzzy subset contained in both A and B. The resulting membership function is calculated by the equation

$$\begin{aligned} \mu(A \cup B|x) &\triangleq \mu(A|x) \nabla \mu(B|x) \\ &\equiv \max\left[\mu(A|x), \mu(B|x)\right] \end{aligned} \tag{2}$$

The similarity to Equation (1) is apparent. However, the estimate of membership in Equation (2) is conservative - as will be shown - since the contribution of the subset in which x is a member to a smaller degree is disregarded. This neglect arises properly only for those x whose degree of membership is neither zero nor one in either of the joined classes. To the extent that the degree is zero for some x, the subsets - in a non-fuzzy sense - may be considered mutually exclusive. In order to take the overlapping of subsets into account, Equation (1) viewed as a conditional probability would read

$$P(A \cup B|x) = P(A|x)+P(B|x)-P(A \cap B|x) \qquad (3)$$

in order to eliminate the double-counting in the intersection which occurs because of the additivity assumption. One can now see that the fuzzy estimate is indeed conservative since, if for example $\mu(A \cup B|x)=P(A|x)$, Equation (3) and the fact that $P(B|x) \geq P(A \cup B|x)$ imply that $\mu(A \cup B|x) \leq P(A \cup B|x)$. For those disciplined in probability theory, Equation (3) may seem a more reasonable way to consider the knowledge gathered from all classes, but Equation (2) is easier to compute and lends itself well to the construction of a fuzzy logic devoid of Lebesque integration over the real line which is characteristic of probability calculations. A recent paper by Stallings[5] offers discussion, empirical examples, and references contrasting the fuzzy set approach with Bayesian statistics. It is interesting to note that a published comment by Jain[6] on Stallings' paper asserts that fuzzy set theories may be developed for different axioms of additivity, although, as the reply of Stallings notes, the axiom used here predominates in the literature.

A non-trivial analytical example is in order. Consider the situation in which people of various ages are being judged as to youthfulness and vigor. Let

y = young

v = vigorous

a = age .

Define:

PWL $[X; \{(X_i, Y_i)\}]$ as a piecewise-linear curve joining the points $(X_i, Y_i)$.

Let

$$p(v,y,a) = K[(1-a/50)v(\delta(y)S_0 = \delta(y-1)S_1) = 10S_2(\delta(y) = \delta(y-1))] \quad (4)$$

where

$0 \leq V \leq 10$

$y = 0$ or $1$

$0 \leq a \leq 100$

$K = 1/11000$

$S_0$ = PWL [a;0,0;30,0;50,1;100,1]

$S_1$ = PWL [a;0,1;30,1;50,0;100,0]

$S_2$ = PWL [a;0,0;40,.5;100,1]

$\delta(Y)$ is the Dirac delta function.

If we define $\mu(v \leq V | a) \triangleq P(v < V | a)$ and $\mu(y < Y | a) \triangleq P(y < Y | a)$, the calculation of the degree of the union is simply

$$\mu_1 = \mu(v \leq V \cup y \leq Y | a) = P(v \leq V | a) \nabla P(y \leq Y | a)$$

Some elementary integration of the joint density expressed by Equation (4) elicits the following table:

| a | $P(v \leq 5 \cup y \leq 0 \mid a)$ | $\mu(v \leq 5 \cup y \leq 0 \mid a)$ |
|---|---|---|
| 20 | 9/16 | 13/32 |
| 40 | 65/88 | 1/2 |
| 60 | 27/37 | 77/148 |

As expected, the fuzzy estimate is less than the probability. The interpretation of the results - which are themselves dependent on the arbitrary choice of the joint density - is that young people are either not youthful appearing or not very vigorous with probability greater than 1/2 and degree

less than 1/2. Intuition might favor the fuzzy estimate. On the other hand, the interpretation continues, declaring that old people are of the sort mentioned with probability about 2/3 and degree about 1/2. Here, it seems that intuition might favor the probability estimate. While it is true that the issue hinges on the dubious choice of Equation (4), the use of probabilistic or fuzzy estimates nevertheless has in principle the potential for counter-intuitive decisions. This phenomenon may be a comment on the reasonableness of our intuition - not necessarily of the assumptions - and suggests that conformity of derived results to intuition may not be an especially good test of the validity of the assumptions. On the other hand, Kahneman and Tversky[7] - motivated by empirical evidence that people's evaluation of the probability of events is significantly different than that derived from the classical assumptions - offer a heuristic measure which represents the subjective judgment of likelihood and is largely independent of sample size. Nevertheless, it is debatable whether or not counter-intuitive findings suffice to justify discarding basic assumptions about the persistence of trends. Perhaps we should rather be satisfied that confidence in the structure of distributions can be determined only empirically after the fact.

## SYMMETRY, UNIMODALITY, NORMALITY

Even if one attributes predictive potential to a collection of pattern vectors in the sense that their distribution approximates an unspecified limit function, calculation of the distribution is often rendered difficult or meaningless by the choice of grid and the sparseness of the data. As a result, the analyst takes refuge in positing a distribution and estimating its parameters or in constructing a distancelike function to characterize membership in a pattern class - for example, the fuzzy functions mentioned previously. These efforts to achieve computational convenience may be influenced by the traditional notion that a pattern class is generated by the perturbation of a hypothetical standard pattern due to noise or error.

With no prior reason to judge otherwise, the error model suggests a distribution which is symmetric about one point - the "pattern" - or at worst along orthogonal axes but at any rate unimodal. As a practical

matter, the ideal standard pattern cannot be determined, and the circumstance of the mean vector being symmetrically perturbed or being located at the maximum density is a special case with no logical necessity other than formalistic elegance or computational convenience.

Still another inference that may be made from the error model, often with recourse to a central limit theorem, is that the error constitutes a normally distributed sum of independent random variables $\xi_k$, each generated by a hypothetical microprocess. In the peculiar event that the $\xi_k$ are identically distributed, with distribution F(X) possessing finite variance $\sigma^2$, the sum indeed approaches the normal distribution in the sense that[8]

$$\lim_{n\to\infty} P\left[\left(\frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}\xi_k - n\int x\,dF(x)\right)<x\right] = \frac{1}{2\pi}\int_{-\infty}^{x} e^{-z^2/2}\,dz \qquad (5)$$

Here, F(X) is said to be in the domain of normal attraction of the normal distribution. However, if F(X) has infinite variance, the limit distribution may be any of a large class of so-called <u>stable</u> distributions - which include the unitary, the Cauchy, the Pearson (type v) among others in addition to the normal - in a sense similar to Equation (5) [Gnedenko and Kolmogorov,[8] p. 181]. In fact, for a doubly indexed random variable $\xi_{nk}$, the limit distribution of $\sum_{k=1}^{mn}\xi_{nk}$ - for arbitrary $F_{nk}(X)$ - may be any of the so-called <u>infinitely divisible</u> distributions* which include the stable, the Poisson, and distributions of finite sums of independent, infinitely divisible random variables. With such a wide range of assumptions and associated limit distributions available, the choice of a normal distribution must be influenced by ignorance of the nature of the fictional microprocesses and by the computational convenience afforded by facts such as "the sum of independent normal random variables is itself normal." On the other hand, computing discriminant surfaces and error probabilities even for multivariate normal distributions is not as easy as might be desired, and some remedies will be suggested further on.

---

*A canonical representation of the characteristic function of an infinitely divisible distribution is given by the Levy-Khintchine formula [Gnedenko and Kolmogorov,[8] p. 76].

Computational efficiency aside, the error model, if insisted upon, might be better interpreted by seeking a function that empirically matched the shape of the distribution near the mean - roll-off at a distance being non-critical - rather than arbitrarily choosing the normal.

## HYPERPLANES AND THE HINGE PROBLEM

In order to avoid the computational difficulties of empirically estimating distributions or of using them, analysts have had recourse to distance, potential function, correlation, clustering, fuzzy, and nonparametric sequential interpretations - among others - of the classification problem. Since each scheme tacitly reflects a class of distributions for which it is best suited, the scheme in general must perform sub-optimally, and sometimes poorly. For some classification schemes, the computational efficiency gained does not compensate for the consequent lack of flexibility in adapting to practical situations. For example, Sebestyan,[9] p.44, effectively seeks to maximize the discriminant which separates the sets F and G with respect to a clustering of F:

$$\delta(\underline{f},\underline{g}) = E\left[ (\underline{f}-\underline{g})^T W(\underline{f}-\underline{g} - (\underline{f}-\overline{\underline{f}})^T W(\underline{f}-\overline{\underline{f}})\right] \quad (6)$$

over positive semi-definite transformations W subject to the constraint

$$\eta(\underline{f},\underline{g}) = E\left[(\underline{f}-\underline{g})^T W(\underline{f}-\underline{g}) - (\underline{g}-\overline{\underline{g}})^T W(\underline{g}-\overline{\underline{g}})\right] = K \quad (7)$$

where

- E is the ensemble mean operator;
- f,g are random variables in classes F,G with means f,g, respectively;
- k is a positive constant.

By setting $\delta(\underline{x},\underline{f})=\delta(\underline{x},\underline{g})$ for a test point $\underline{x}$, one can see that the discriminant surface is a hyperplane

$$\underline{x}^T W(\underline{g}-\underline{f}) = 1/2\ (\overline{\underline{g}}^T W\overline{\underline{g}} - \overline{\underline{f}}^T W\overline{\underline{f}}) \quad (8)$$

Consider the points

$$\underline{x} = d\overline{\underline{f}} + (1-d)\overline{\underline{g}},\ (0\leq d\leq 1) \quad (9)$$

on the line connecting the class means. The solution of Equations (8) and

(9) for $\underline{d}$ reveals that the discriminant hyperplane is hinged at d=1/2. It is easy to discover linearly separable distributions for which a hinged hyperplane, however efficient to calculate, has no orientation which provides near-optimal or even satisfactory discrimination. One could unhinge the discriminant hyperplane by reformulating the problem to include a variable threshold $\underline{s}$, thus maximizing

$$\delta(\underline{f},\underline{g}) - \frac{\underline{s}}{2}(\bar{\underline{g}}-\bar{\underline{f}})^T W(\underline{g}-\underline{f}) \tag{10}$$

subject to the constraint

$$\eta(\underline{f},\underline{g}) - \frac{\underline{s}}{2}(\bar{\underline{g}}-\bar{\underline{f}})^T W(\underline{g}-\underline{f}) = K$$

Whereas Equations (6) and (7) have a straightforward solution, the solution to Equations (10) and (11) (see Appendix A) requires maximizing the maximum eigenvalue of the matrix

$$(R(\underline{g}) + R(\bar{\underline{f}}) - \frac{\underline{s}}{2} Y)(R(\underline{f}) + R(\bar{\underline{g}}) - \frac{\underline{s}}{2} Y)^{-1} \tag{11}$$

where

$$\begin{aligned} R(\underline{h}) &\triangleq E(\underline{h},\underline{h}^T) \\ Y &\triangleq R(\bar{\underline{g}}-\bar{\underline{f}}) \end{aligned} \tag{12}$$

over $\underline{s}$, a challenging computational task. In the light of difficulties of this sort, one might just as well have assumed two normal distributions, for which the hyperplane discriminant has no known closed form but may be calculated by recursive approximation,[10] as indeed may Equation (12).

## SUMMARY OF THE ARGUMENT

From the preceding discussion, one should have received the impression that statistical decisions are often bigger gambles than the numbers indicate because of the arbitrariness with which one chooses the assumptions. Since predictability is such a slippery notion, analysts have locked themselves into simplifying assumptions about empirical processes, ranging from additivity of measures to hinged discriminant hyperplanes, all in the name of computational efficiency. In the discussion to follow, an approach to choosing assumptions is introduced for which the computational penalty is not excessive, and all information is used in some

intuitively agreeable sense. The method will offer the flexibility - and arbitrary complexity - of determining or ignoring dissymetry, polymodality, and roll-off while relaxing the additivity assumption and dealing with unhinged, arbitrarily oriented, piecewise linear surfaces. The proposed technique has its limitations and cannot be unique in an arena of choices dictated primarily by intuitive agreement, but it may nevertheless be a contribution to a means for exploratory data analysis and offers an opportunity to escape from the confining assumptions inherent in statistical handbooks.

## DISTRIBUTIONS OF RECTANGULAR CONTOUR

### FORMS OF THE DISTRIBUTIONS

The density distributions discussed in this chapter are of hyperrectangular isoprobabilistic contour. These contours are closed, concentric surfaces which circumscribe hyperellipsoids whose principal semi-axes are proportional to the square roots of the eigenvalues of a covariance matrix $Q_i$ and lie along the corresponding eigenvectors of $Q_i$. If the covariance matrices $Q_1$, $Q_2$ are positive definite, they can be simultaneously diagonalized by a transformation A such that

$$AQ_1A^T = I \text{ and } AQ_2A^T = \Lambda \tag{13}$$

where $\Lambda$ is the eigenvalue matrix of $Q_1^{-1}Q_2$. Therefore, in the case of binary discrimination, it suffices to discuss two distributions in $R^n$, namely: $P_1(\underline{X}^T\underline{X}^x)$, whose mean is zero and whose isoprobabilistic contours are hypercubes; and $P_2((\underline{X}-\underline{\bar{X}})^T \Lambda (\underline{\bar{X}}-\underline{\bar{X}}))$, whose mean is $\underline{\bar{x}}$, whose isoprobabilistic contours are hyperrectangular, and whose principal semi-axes are parallel to those of $P_1$.

The hyperrectangular contours of the distributions to be analyzed are based on the following proposition:

PROPOSITION. A hypercube may be represented by the equation

$$\lim_{m\to\infty} \sum_j \Theta_j \left( \sum_i x_i^{2m} + \beta_j^m \right)^{h/m} = \sum_j \Theta_j \rho_j^h \tag{14}$$

where

$$\rho_j = \max \{X^2, \beta_j\}, \text{ and } \theta_j, X, \beta_j \text{ are arbitrary.}$$

Proof For $x_k = \phi_k X$, $0 \leq |\phi_k| \leq 1$ except that $|\phi_1| = 1$, the left-hand side of Equation (14) becomes

$$\lim_{m\to\infty} \sum_j \Theta_j \left( \left( \sum_j \phi_k^{2m} \right) X^{2m} + \beta_j^m \right)^{h/m} = \lim_{m\to\infty} \sum_j \Theta_j \left( NX^{2m} + \beta_j^m \right)^{h/m} \tag{15}$$

Where N is the number of $\phi_k$ equal to one. When $X^2 \geq \beta_j$, the parenthesized expression in Equation (15) yields

$$\lim_{m\to\infty} N^{h/m} X^{2h} = X^{2h} \tag{16}$$

and when $X^2 \leq \beta_i$, the parenthesized expression in Equation (15) reduces to $\beta_j^h$. **END OF PROOF.**

The power of the Proposition is that it permits the synthesis of symmetric distributions with roll-offs of extensive complexity subject to a finite volume constraint, as the following examples demonstrate.

An exponential squared - i.e., normal roll-off - may be achieved by the distribution

$$p_a(\underline{x}) = \lim_{m\to\infty} \kappa_a \exp\left[-1/2\left(\sum_i (\lambda_i x_i^2)^m\right)^{1/m}\right] \tag{17}$$

By letting

$$B_1=1,\ \Theta_1=1,\ \beta_j=\Theta_j=0\ (j\neq 1)$$

one sees that the Proposition is satisfied and the contours are consequently hyperrectangular.

The distribution has finite volume and may be normalized as follows. Let

$$\underline{z} \triangleq \Lambda^{1/2}\underline{x}.$$

Then Equation (17) implies that

$$\lim_{m\to\infty} \left(\sum_i (z_i^2)^m\right)^{1/m} = -2\ln(p_a\, \kappa_a^{-1} \det \Lambda^{1/2}) \tag{18}$$

which is the equation of a hypercube. If $z_j=0\ (j\neq i)$, then the squared semi-axis $z_i^{*2}$ is given by the right-hand side of Equation (18) and the integral of Equation (17) is calculated by the expression

$$\int_0^A \prod_{i=1}^n (2z_i^*)\, dp_a = \int_0^A 2^n (2/\ln(p_a^{-1}A))^{n/2} dp_a \tag{19}$$

where $A= \kappa_a \det \Lambda^{-1/2}$. Let $q \triangleq p_a A^{-1}$. Then the normalization integral (19) becomes

$$A\, 2^{3n/2} \int_0^1 (\ln q^{-1})^{n/2} dq = 1 \quad . \tag{20}$$

The solution to Equation (20) (Gradshteyn, Ryzlink,[11] p. 525, 4.215.1) gives the result

$$\kappa_a = 2^{-3n/2} \det \Lambda^{1/2}/\Gamma(1+n/2) \tag{21}$$

The fact that the elliptically-contoured normal density distribution is simply approximated by the rectangularly-contoured normal distribution

through a change of eigenvalues (see Appendix B) suggests that any computational difficulty with one distribution will not be alleviated for the other. Moreover, since the requirement for finite volume is met by a normal roll-off on the order of exp $[-nx^2]$, but not as slow as $x^{-n}$ for non-zero tails, we will resort to distributions of finite support to achieve other roll-offs. These distributions with truncated roll-off exhibit computational advantages and are in accord with the physical intuition that all measurements of real phenomena fall within a finite distance of the class mean. Before discussing the truncated case, however, it will be helpful to display a density distribution which not only illustrates the necessity for fast roll-off when non-zero tails are included, but also serves as a means of generating distributions with truncated roll-off. The density

$$p_b(\underline{x}) = \lim_{m\to\infty} \kappa_b \left(\sum_i^n (\lambda_i x_i^2)^m + \beta^m\right)^{-h/m} \tag{22}$$

is normalized by means of the integral

$$\int_0^{\kappa_b \beta^{-h}} 2^n \det \Lambda^{1/2} (P_b/\kappa_b)^{-n/2h} dp_b = 1 \tag{23}$$

which, for $2h>n$, yields

$$\kappa_b = 2^{-n} \det \Lambda^{1/2} \left(\frac{2h-n}{2h}\right) \beta^{h-n/2} \tag{24}$$

The restriction $2h>n$ implies that the roll-off must be faster than $x^{-n}$ as mentioned previously. Slower roll-off for distributions with truncated roll-off can be achieved by carefully combining densities of the form given by Equation (22). The following density is an example of such a combination.

$$p_c(\underline{x}) = \lim_{m\to\infty} \kappa_c \left[\left(\sum_{i=1}^n (\lambda_i x_i^2)^m + \beta_2^m\right)^{h/m} - \left(\sum_{i=1}^n (\lambda_i x_i^2)^m + \beta_1^m\right)^{h/m}\right] \tag{25}$$

Since the distribution form satisfies the conditions of the Proposition, the contours are hyperrectangular. Suppose that $\beta_2>\beta_1$. The normalization proceeds as before, producing the equation

$$\kappa_c h 2^n \det \Lambda^{-1/2} \int_{\beta_1}^{\beta_2} q^{n/2-h-1} dq = 1 \tag{26}$$

whose solution is

$$\kappa_c = 2^{-n} \det \Lambda^{1/2} d / h(\beta_2^d - \beta_1^d) \tag{27}$$

where $d = (n/2)-h$ unless $h = n/2$, in which case the constant is

$$\kappa_c = 2^{-n} \det \Lambda^{1/2} / (n/2) \ln(\beta_2/\beta_1) \tag{28}$$

Except for the singular point h=0, useful solutions exist for all non-negative h. Moreover, when h is negative, solutions exist for the condition $\beta_1 > \beta_2$. The formulation (25) will generate, for example, a truncated linear roll-off for h=1/2, $\beta_1 > \beta_2$, and a truncated quadratic roll-off for h=-1, $\beta_1 > \beta_2$. More complex profiles can be achieved by a judicious combination of forms derived from Equation (25).

## ERROR EXPRESSIONS

Due to the simple geometry associated with rectangular contours, the error induced by partitioning a measurement space with the discriminant hyperplane

$$\underline{s}^T \underline{x} + t \tag{29}$$

can be easily calculated without reference to the analytic forms of the densities by reducing the n-dimensional volume integral to a one-dimensional integral along density, as was done for the normalization constant. Although the resultant error expression is not amenable to linear optimization, the expression itself serves as an intuitive aid in determining sub-optimal discriminants. The deliberate attempt to find an optimal hyperplane by minimizing the probability of error leads to more computational difficulties than some approaches involving other criterion functions, like the mean-square error (MSE). However, those approaches produce discriminants that may not be good approximations to the discriminant minimizing the probability of error, which is regarded as the fundamental criterion. For example, since the MSE hyperplane, which is an asymptotic minimum MSE approximation to the optimal (nonlinear) Bayes discriminant, is related to the density magnitude rather than to the

distance from the decision surface, the MSE does not necesarily vary with the probability of error.[12] The decision to deal, then, with the probability of error is noteworthy and made possible by the fact that explicit expressions are available, at least in some cases of importance, for the probability of error. Without such explicit expressions, placing bounds on error would be the only means of formulating an optimum hyperplane. Haralick,[13] in considering the Bayes classification problem for a class of distributions which include those discussed in this report, develops lower bounds on the probability of error due to the difficult integration associated with a domain bounded by a hyperplane discontinuity.

In the following discussion, the density p will be presumed to vary inversely with increasing values of $|\underline{x}|$. Since a density distribution may be viewed as the superposition of a set of unimodal densities, the assumption is not as restrictive as it may appear, at least not for the well-behaved densities one believes exist in practice. In any event, the overall strategy of derivation would not differ appreciably for other p(x).

As seen in Figure 2 - shown for the case of two dimensions and for a given hyperplane $\underline{s}^T\underline{x} = t$ - as the isoprobable contours expand, they meet the hyperplane at $2^{n-1}$ vertices

$$\begin{aligned} \underset{\sim}{\underline{x}}_i &\triangleq \alpha_i \underline{\eta}_i \\ &\triangleq \alpha_i (\pm \lambda_1^{-1/2}, \ldots, \pm \lambda_n^{-1/2})^T \end{aligned} \tag{30}$$

where the signs are chosen so that

$$\underline{s}^T\underline{\eta}_i < \underline{s}^T\underline{\eta}_j \quad , \quad i<j \tag{31}$$

and where

$$\underline{\alpha}_i = t/\underline{s}^T\underline{\eta}_i \tag{32}$$

Here it is presumed that the hyperplane is situated above the mean. A symmetric argument exists if the hyperplane is located below the mean. The degeneracy caused by $\underline{s}^T\underline{\eta}_i = \underline{s}^T\underline{\eta}_j$ in Equation (31) will be discussed later. From Figure 2, one can see that, at a given $\underline{x}$, the error consists of a weighted triangular area ABC or its n-dimensional right pyramid equivalent, less a similar weighted area CDE - for large enough x - for each vertex $\underset{\sim}{\underline{x}}$ above the hyperplane. The error expression for any of the

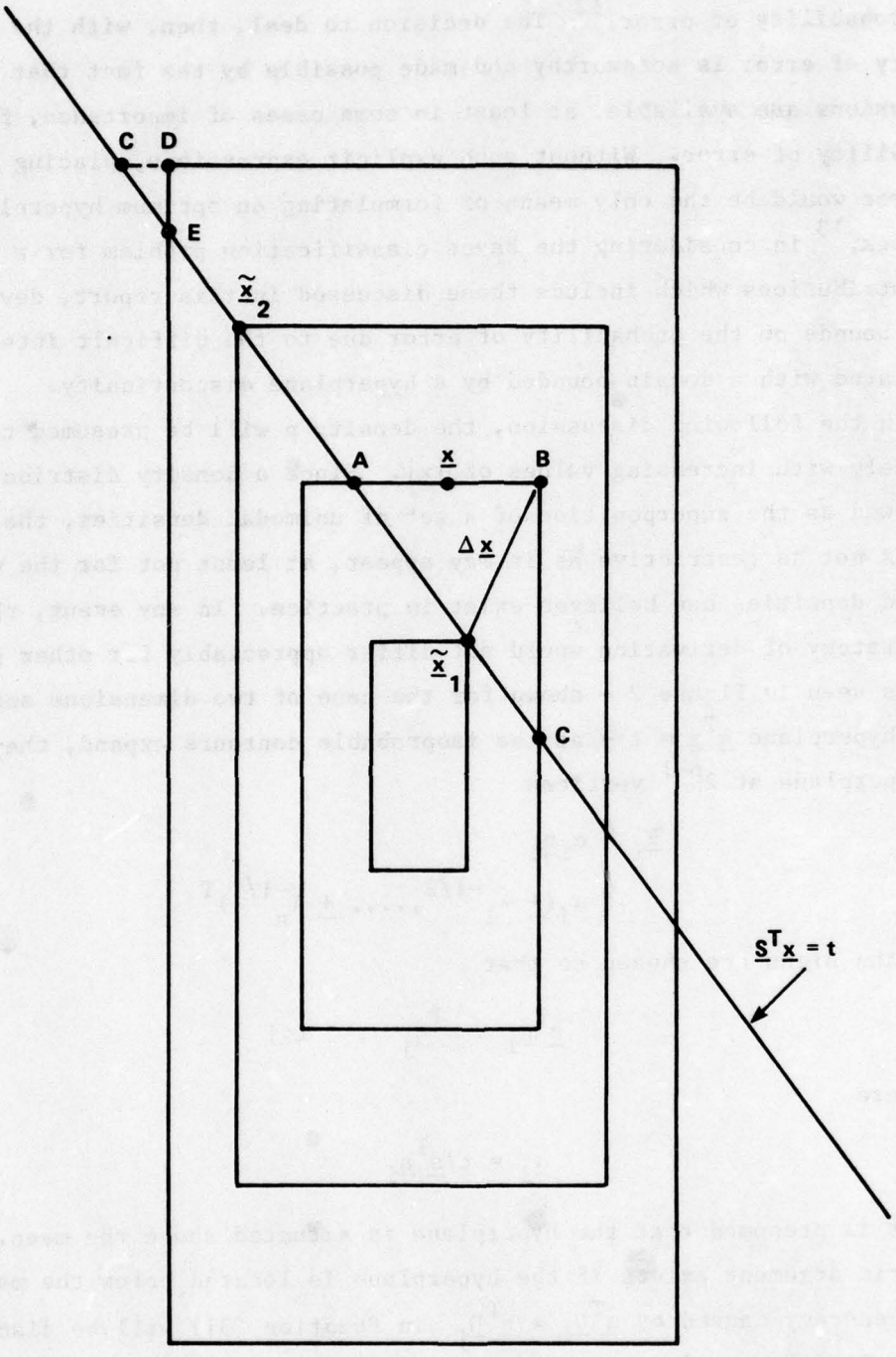



Figure 2 - Rectangular Contours Intersected by Discriminant Line

areas (volumes) mentioned has the expansion

$$\int_{\tilde{p}} d\underline{x} = pV\Big|_{v=0}^{p=0} - \int_{p(\tilde{\underline{x}})}^{0} V\, dp \tag{33}$$

where V is a volume variable. A side of the indicated volume is

$$\hat{x}_{ik} + \Delta x_k - x_k \tag{34}$$

where $\hat{x}_k$ is the intersection of the hyperplane and the $x_k$-axis, given implicitly by

$$\hat{x}_k = s_k + \sum_{j=k} (\tilde{x}_{ij} + \Delta x_j) s_j = t \tag{35}$$

Since $\tilde{\underline{x}}_i \underline{s} = t$, Equation (35) yields

$$\hat{x}_k = \hat{x}_{ik} - \sum_{j \neq k} \Delta x_j s_j / s_k \tag{36}$$

Therefore, Equation (34) is equal to $(\Delta \underline{x}^T \underline{s})/s_k$ and the error Equation (33) results in the expression

$$E_i = \int_0^{\tilde{p}_i} (\Delta \underline{x}^T \underline{s})^n / n! \prod_j s_j \, dp \tag{37}$$

where $\tilde{p}_i = p(\tilde{\underline{x}}_i)$.
If the profile is defined as

$$p_0 = p(\underline{x})\Big|_{x_k=0, k\neq 1} \equiv f(x_1) \tag{38}$$

then

$$x = f^{-1}(p_0) \tag{39}$$

and

$$x_k = \pm\sqrt{\lambda_1/\lambda_k}\, f^{-1}(p_0) = \sqrt{\lambda_1}\, \eta_{ik} f^{-1}(p_0) \tag{40}$$

where the signs for $\underline{\eta}_i$ are chosen to conform with the appropriate vertex as indicated by Equations (30) and (31). When Equation (40) is introduced into the error expression (37), the latter becomes

$$E_i = \frac{\left(\sqrt{\lambda_i \underline{s}^T \underline{\eta}_i}\right)^n}{n! \prod_j s_j} \int_0^{\tilde{p}} (f^{-1}(p_0) - f^{-1}(\tilde{p}_i))^n dp_0 \qquad (41)$$

In general, then, the total error expression for densities not truncated before extending the $2^{m-1}$ th vertex beyond the hyperplane is

$$E = \sum_{i=1}^{2^{m-1}} (2\delta_{i1} - 1) E_i \qquad (42)$$

where $\delta_{jk}$ is the Kronecker delta.

Discrimination between two distributions whose means are separated by the vector $\underline{\delta}$ requires the coordinate transformation

$$\underline{x}' = -(\underline{x} + \underline{\delta}) \qquad (43)$$

in order to retain the form of the preceding error expressions for the distribution which is displaced from the origin. From the point of view of such a distribution, the hyperplane has the form

$$\begin{aligned} \underline{s}^T \underline{x}' &= \underline{s}^T \delta - \underline{s}^T \underline{x} \\ &= \underline{s}^T \delta - t \end{aligned} \qquad (44)$$

Therefore, Equation (41) holds for the displaced distribution if and only if

$$\underline{s}^T \underline{\eta}_i (\sqrt{\lambda_i} f^{-1}(\tilde{p})) = \underline{s}^T \underline{\delta} - t \qquad (45)$$

which yields the following solution for $\tilde{p}_i$

$$\tilde{p}_i = f((\underline{s}^T \underline{\delta} - t)/\sqrt{\lambda_i}\, \underline{s}^T \underline{\eta}_i) \qquad (46)$$

Similarly, if the distribution at the origin has the normalized vertex vector $e_i$, whose components are $\pm 1$, the expression for $\tilde{p}_i$ is then

$$\tilde{p}_i = f(t/\underline{s}^T \underline{e}_i) \qquad (47)$$

With the aid of the preceding formulas, the error expressions for the previously defined distributions may now be derived. For normal roll-off,

$$f^{-1}(p_{ao}) = (\lambda_1^{-1} \ln(\kappa_a / p_{a0})^2)^{1/2} \qquad (48)$$

If one makes the transformation

$$q = (\ln(\kappa_a/p_{ao})^2)^{1/2} - (\ln(\kappa_a/\tilde{p}_i)^2)^{1/2} \tag{49}$$

the error expression (41) becomes

$$E_{ia} = \kappa_a (A_i/n!) \int_0^\infty q^n (q+\tilde{q}_i) \exp(q+\tilde{q}_i)^2/2) dq \tag{50}$$

where

$$A_i = (\underline{s}^T \underline{\eta}_i)^n / \prod_j s_j \tag{51}$$

and

$$\tilde{q}_i = (\ln(\kappa_a/\tilde{p}_i)^2)^{1/2} \tag{52}$$

or

$$\tilde{q}_i = (\underline{s}^T \underline{\delta} - t)/\underline{s}^T \underline{\eta}_i \tag{53}$$

The expanded form of Equation (50) is (cf. Gradshteyn, Ryzhik,[11] p. 337, 3.462.1 for the integral and p. 1066, 9.247.1 for the recurrence relation)

$$E_{ia} = \kappa_a A_i \exp(\tilde{q}_i^2/4) D_{-n}(\tilde{q}_i^2) \tag{54}$$

where $D_{-n}(\cdot)$ is a parabolic cylinder function of degree n. Since $D_{-n}(\cdot)$ can be directly, if laboriously, related to a normal error function (Gradshteyn and Ryzhik,[11] p. 1067, 9.254.1, 2 and the recurrence relation just cited), further analysis does not seem profitable (cf. Anderson and Bahadur[10]).

By a similar derivation, one can display component error expressions for $p_b$, $p_c$ as follows:

$$E_{ib} = \kappa_b 2^{n+2} h((4h-2n-2)!!/(4h)!!) \tilde{q}_i^{n-2h} \tag{55}$$

where $k!! \equiv k(k-2)(k-4) \ldots$, and $A_i$, $\tilde{q}_i$ are defined in Equations (51) and (53). For the truncated distribution $p_c$ with $h \geq 1$, the component error expression is

$$E_{jc} = \kappa_c A_i \sum_{j=0}^{n} \binom{n}{j} \left(\frac{2h}{n!(n+2h-j)}\right) \left[\left(\sqrt{\beta_2}\right)^{n+2h-j} - (\tilde{q}_i)^{n+2h-j}\right] (-\tilde{q}_i)^j \tag{56}$$

For h<1, a similar derivation produces a complicated expression which will not be detailed here.

In general, the preceding error expressions are not amenable to the usual minimizing techniques, due not only to the nonlinearity of the resulting extremal equations, but also to the variety of forms produced by various boundary constraints. As a trivial example, if the distributions for two classes - with eigenvectors $\underline{e}_i$, $\underline{\eta}_i$, respectively - do not overlap, one can find $\underline{s}$, t such that

$$\sqrt{\beta_2'}\ \underline{s}^T\underline{e}_i < t \tag{57}$$

$$\sqrt{\beta_2''}\ \underline{s}^T\underline{\eta}_i < \underline{s}^T\underline{\delta}-t \tag{58}$$

which imply that the error is zero. A more useful, non-trivial case is specified by the constraints

$$\sqrt{\beta_2'}\ \underline{s}^T\underline{e}_i > t > \beta_1'\ \underline{s}^T\underline{e}_i \tag{59}$$

$$\underline{s}^T(\underline{\delta}-\sqrt{\beta_1''}\ \underline{\eta}_i) > t > \underline{s}^T(\underline{\delta}-\sqrt{\beta_2''}\ \underline{\eta}_i) \tag{60}$$

for which the $i^{\underline{th}}$ component of error - based on Equation (56) - is

$$E_{ic} = \sum_{j=0}^{n}\binom{n}{j}\left[B_{i1}\left(\left(\sqrt{\beta_2'}\right)^{w_1} - \left(\frac{t}{\underline{s}^T\underline{e}_i}\right)^{w_1}\right)\left(-\frac{t}{\underline{s}^T\underline{e}_i}\right)^j + (\mu B_{i2})\left(\left(\sqrt{\beta_2''}\right)^{w_2} - \left(\frac{\underline{s}^T\underline{\delta}-t}{\underline{s}^T\underline{\eta}_i}\right)^{w_2}\right)\left(-\frac{\underline{s}^T\underline{\delta}-t}{\underline{s}^T\underline{\eta}_i}\right)^j\right] \tag{61}$$

where

$$w_k = n+2h_k-j, \qquad k = 1,2$$

$$B_{ik} = 2h_kA_{ik}\kappa_{kc}/n!W_k, \qquad k = 1,2$$

and $\mu$ is arbitrary. This constraint will be presumed to hold in the rest of the discussion.

Since the roll-off of $p_c$ decreases monotonically with density, the error component can be further refined by extending the analysis of Anderson and Bahadur[10] to the present case. Thus, the expressions for

$\underline{s}$,t are:

$$\underline{s} = (t_1 I + t_2 \Lambda)^{-1} \underline{\delta} \tag{62}$$

$$t = t_1 \underline{\delta}^T (t_1 I + t_2 \Lambda)^{-2} \underline{\delta} \tag{63}$$

where

$$|t_1| + |t_2| = 1 \tag{64}$$

Inserting these formulas into Equation (61) produces a one-parameter family of equations which appear no easier to minimize than the generalization of Sebestyan's formulation which was considered in Equation (12). In the next section, a sub-optimal solution will be discussed which depends on the degeneracies presented next.

## DEGENERACIES

The usual notion of degeneracy applies to the situation in which one or each of the covariance matrices is singular. In such a case, if the distributions occupy different subspaces, test points may be classified with zero probability of error simply by determining whether or not a component differs from the mean in a direction which is orthogonal to the reduced subspace of a distribution. If the reduced subspaces are identical, the problem is handled in the usual manner, since the covariance matrices will be nonsingular in the reduced subspace.

Another form of degeneracy occurs when the discriminating hyperplane is parallel to a surface of each distribution. This occurs when one or more components of $\underline{s}$ are zero in the space in which the covariance matrices are simultaneously diagonalized. This implies, for example, that $\underline{s}^T \underline{\eta}_i - \underline{s}^T \underline{\eta}_j$ for some i,j not equal. With the assumption that $\underline{s} = (s_1, s_2, \ldots, s_j, 0, \ldots, 0)$ it is not difficult to re-derive the formula for the component error of Equation (61). The result is as follows:

$$E_{ic} = \sum_{j=0}^{j} \binom{J}{j} \left[ B_{i1} \left( \left( \sqrt{\beta_2'} \right)^{w_1} - \left( \frac{t}{\underline{s}^T \underline{e}_i} \right)^{w_1} \right) \left( \frac{t}{\underline{s}^T \underline{e}_i} \right)^{j} \right.$$

$$\left. + (\mu B_{i2}) \left( \left( \sqrt{\beta_2''} \right)^{w_2} - \left( \frac{\underline{s}^T \delta - t}{\underline{s}^T \underline{\eta}_i} \right)^{w_2} \right) \left( - \frac{\underline{s}^T \delta - t}{\underline{s}^T \underline{\eta}_i} \right)^{j} \right] \tag{65}$$

where the $A_{i1,2}$ and $B_{i1,2}$ are re-defined as

$$A_{i1} = (\underline{s}^T \underline{e}_i)^J \prod_{j=J+1}^{n} e_{ij} / 2^{n-J} \prod_{j=1}^{J} s_j \tag{66}$$

$$A_{i2} = (\underline{s}^T \underline{n}_i)^J \prod_{j=J+1}^{n} \eta_{ij} / 2^{n-J} \prod_{j=1}^{J} s_j \tag{67}$$

$$B_{ik} = 2h_k A_{ik} \kappa_{kc} / J! w_k \ , \qquad k=1,2 \tag{68}$$

We will make use of degeneracies to determine sub-optimal solutions in the following section.

## DISCRIMINANT LOGIC

### BINARY CLASSIFICATION FOR RECTANGULAR CONTOUR

In the previous section, a formulation for the minimization of error was presented which did not permit an easy solution. Here, a sub-optimal formulation is presented which eases the problem somewhat. For each dimension, consider the hyperplane

$$\underline{s}_j^T \underline{x} = t_j$$

where

$$\underline{s}_j \equiv (0,\ldots,0,s_j,0,\ldots,0) \tag{69}$$

The problem is thus reduced to a set of n one-dimensional problems. If we substitute Equation (69) into the degenerate error component form specified by Equation (65), differentiate by $d/d\tilde{t}_j$, where

$$\tilde{t}_j \equiv t_j/s_j$$

and set the result to zero, a non-linear equation is generated as follows:

$$a_1 \tilde{t}_j^{\,w} + b_1 = a_2(\delta_j - \tilde{t}_j)^v + b_2 \tag{70}$$

where

$$a_1 = \kappa_{1c}/pe_j^w \quad ; \quad a_2 = \mu\kappa_{2c}/q\eta_j^v$$

$$b_1 = \kappa_{1c}(\sqrt{\beta_2'})^w/w \quad ; \quad b_2 = \mu\kappa_{2c}(\sqrt{\beta_2''})^v/v$$

$$w = n+2h_1-1 \quad ; \quad v = n+2h_2-1$$

Equation (70) need not be summed over the vertex set in accordance with Equation (42), since no vertex is preferred in the one-dimensional degenerate case. Solutions to Equation (70) may be found either numerically or, for example, by piecewise linear approximation. However, one must remember that Equation (70) was derived on the basis of the overlap constraints expressed by Equations (59) and (60). These constraints permit the decision hyperplane to be placed in any of seven distinct regions, for each of which the error component Equation (65) is slightly different. Nevertheless, Equation (70) has the same form - with different coefficients - in each region. The regions are determined by the value of $t_j$ relative to the breakpoints

$$\{-\sqrt{\beta_2''}\,e_i,\ \pm\sqrt{\beta_1'}\,e_i,\ \delta_j \pm \sqrt{\beta_1'}\ \eta_i,\ \delta_j\sqrt{\beta_2'}\ \eta_i\}$$

For example, in the region

$$-\sqrt{\beta_1'}\, e_i \le t_i \le \sqrt{\beta_1'}\, e_i$$

the error expression is

$$E_i' = E_i + p_{1c}(0)(\sqrt{\beta_i'}\, e_i - \tilde{t}_i) \prod_{j \neq i} (e_j \sqrt{\beta_1'}) \tag{71}$$

where $E_i$ is the error component Equation (65). The added term on the right of Equation (71) simply implies that

$$p_{1c}(0) \prod_{j \neq 1} (e_j \sqrt{\beta_i'})$$

is added to the $b_1$ term in Equation (70). The number of regions can effectively be reduced to three, indicated by the breakpoints $[0, \delta]$, by introducing the following additional constraints:

$$\sqrt{\beta_1'},\ \sqrt{\beta_1''} \to 0 \tag{72}$$

$$\sqrt{\beta_2'} >> t_j/e_j \tag{73}$$

$$\sqrt{\beta_2''} >> (\delta - t_j)/\eta_j \tag{74}$$

These constraints focus attention on the competition between roll-offs and effectively disregard the truncations in both roll-off and density.

Because each side of Equation (70) in the region $[0, \delta]$ is monotonic, one can develop a recurrence relation for the solution - if it exists - by examining the slope of the functions. At any point $\tilde{t}_{0j}$ in $[0, \delta]$ the lines of slope of the functions described by the left and right sides, respectively, of Equation (70) are

$$f_1(\tilde{t}_j) = w\tilde{t}_{0j}^{w-1}\tilde{t}_j + (a_1 - w)t_0^w + b_1 \tag{75}$$

$$f_2(\tilde{t}_j) = -v(\delta - \tilde{t}_{0j})^{v-1}(\delta - \tilde{t}_j) + (v + a_2)(\delta - \tilde{t}_{0j})^v + b_2 \tag{76}$$

The solution to Equations (75) and (76) is the recurrence relation

$$\tilde{t}_{kj} = \frac{(w - a_1)\tilde{t}_{k-1,j}^w + (v + a_2)(\delta - \tilde{t}_{k-1,j})^v + (b_2 - b_1) - v\delta(\delta - \tilde{t}_{k-1,j})^{v-1}}{w\tilde{t}_{k-1,j}^{w-1} - v(\delta - \tilde{t}_{k-1,j})^{v-1}} \tag{77}$$

The monotonicity of the functions ensures convergence. A check on Equation (77) is that in the event $w = v$, $b_1 = b_2$, both Equations (77) and (70) have the same fixed point, namely:

$$\tilde{t}_{kj} = a_2^{1/w}\delta_j/(a_1^{1/w} + a_2^{1/w}) \tag{78}$$

Finally, note that the necessary and sufficient conditions for a solution to Equation (70) to exist are

$$0 < b_1 - b_2 < a_2\delta_j^{\,v} \tag{79}$$

$$0 < b_2 - b_1 < a_1\delta_j^{\,v} \tag{80}$$

The decoupling of dimensions effected by the degenerate hyperplanes of Equation (69) in formulating the sub-optimal solution may now be removed by averaging the hyperplanes over the dimensions. Alternatively, one may choose that hyperplane intersected by the line joining the centers of the distributions, or one may keep the piecewise linear structure in a decision tree.

## BINARY CLASSIFICATION FOR POLYGONAL CONTOUR

When faced with the task of analyzing distributions whose isoprobable contour is more complicated than ellipsoidal or rectangular, one realizes how well suited those contours are for rectangular co-ordinates; their minimum and maximum distances from the origin in any subspace lie exactly and symmetrically on a set of orthogonal axes and are determined - aside from a scale factor - solely by the distance of the rms density from the origin, or equivalently, by the eigenvalues of the associated covariance matrix. One could construct dependent coordinate systems which would be suitable for the description of particular contours, but in order to handle densities of arbitrary contour, recourse must be had to the traditional selection of grids, frequency counts, and their attendant problems. On the other hand, by restricting the contours to be concentric and - aside from a scale factor - invariant in shape, one can extend the theory developed for rectangular contours in a natural way as outlined in that which follows.

Let each sample be represented by $\underline{v}_i(r_i,\underline{\theta}_i)$ in a spherical coordinate system centered at the mean $\underline{m}$. A decomposition of the sample distribution

into distributions of rectangular contour will now be sought. Select a grid based on an infinite volume solid angle $\Delta\underline{\theta}$ and compute the rms density within each grid volume element

$$\bar{v}_{rms} = \left[ \sum_{v_i \in \Delta\underline{\theta}}^{n} (v_i - m_i)^2 / n \right]^{1/2}$$

During the course of the computation, the opportunity clearly exists for detecting polymodes and wild points and for estimating fall-off, but these procedures will not be detailed here. Peak-picking on the inverse rms density over the grid elements - by a steepest descent technique, for example - will then determine the magnitude of one maximum radial vector $\tilde{\underline{v}}_j$ for each component distribution, located by definition along the central ray of a grid element. The inverse rms density in the grid elements surrounding the element containing the maximal ray may be compared with $|\tilde{\underline{v}}_j|/2$, and those elements which exceed that threshold are collected to form a half-power-point set. The set of samples thus effectively identified is now projected symmetrically through the origin at the mean $\underline{m}$. The original and image sets are joined to form a set whose mean $\underline{m}$ (by a symmetry argument) and covariance matrix $\underline{\Sigma}_j$ may be computed in some orthogonal co-ordinate system for which $\tilde{\underline{v}}_j$ is a basis vector. A Gram-Schmidt $(\Sigma_j, h_j)$ orthonormalization procedure will, for example, provide the required basis. When $\underline{m}_j$, $\Sigma_j$, and a roll-off parameter $h_j$ have been computed for each density peak - referred to a single co-ordinate system - each pair may be used as the statistic for a rectangular distribution confined to the half-space indicated by the eigenvector corresponding to the maximum eigenvalue of $\Sigma_j$. The only difference in form of the effective density from the probability densities formulated in the previous section is that the normalization constant $\kappa$ is doubled and the density is defined to be zero in the complement to the half-space just mentioned. At this point, it would be desirable to form pairwise discriminant hyperplanes between the component densities of two given distributions with arbitrary isoprobable contour, but it is first necessary to take into consideration the intersection of the component distributions. Earlier in this report, it was noted that fuzzy set theory offers a conservative means of estimating the probability of the union of sets by disregarding some information. Here,

however, it is possible to take all available information into account and at the same time not assume the computational burden of explicitly integrating over the intersection contour. To do this, consider the set probabilities $P_1(A_1;\ \Sigma_1)$, $P_2(A_2;\ \Sigma_2)$. Define the set union probability as

$$P(A_1 \cup A_2) \equiv P_1(A_1;\ \Sigma_1) + P_2(A_2;\ \Sigma_2) - P(A_1 \cap A_2) \tag{81}$$

and the set intersection probability as

$$P(A_1 \cap A_2) \approx P(A_1 \cap A_2;\ \Sigma_1 + \Sigma_2) \tag{82}$$

The rectangular solid corresponding to $\Sigma_1+\Sigma_2$ is contained within the convex hull of the intersection of the solids corresponding to $\Sigma_1$ and $\Sigma_2$. Therefore, the approximation designated by Equation (82) implies that the estimate of the set union probability is liberal. Nevertheless, all the available information has been taken into account. Indeed, one might go a step further at some computational expense and specify the intersection volume exactly by simultaneously diagonalizing the covariance matrices and then deriving the eigenvalues for the intersection covariance matrix as follows:

$$\eta_i = \min\ (1,\ \lambda_i - \delta_i) \tag{83}$$

where the diagonalized matrices have eigenvalues $\{1,\ \ldots,\ 1\}$ (the identity) and $\{\lambda_1,\ \ldots,\ \lambda_n\}$, respectively, and $\underline{\delta}$ is the vector between the set means. If $\lambda_i < \delta_i$, the intersection is null. The distribution supported by the intersection in fact has neither the roll-off nor the rectangular contour of the component distributions, although the approximation is explicitly made here to the contrary. The eigenvalues in Equation (83) have been specified along non-orthogonal co-ordinates, and an inverse transformation to normal coordinates must be made before the eigenvalues of the covariance matrix of the presumed intersection distribution can be calculated. The set union probability implied by Equation (83) has been specified for distinct means, whereas the means of the component distributions are in fact coincident. This element of generality has been purposely introduced to suggest that what has been done is not merely an approximation, but a modification of the axiom of additivity for the closed system of distributions with rectangular contour, and thus defines a logic of such distributions as does the set of Equations (81), (82). It remains now to

explicate the form of the probability of error. If the eigenvectors of the intersection covariance matrix are represented as the modal matrix

$$D^T = \left[\underline{d}_1, \ldots, \underline{d}_n\right]$$

then the error contribution due to the component matrices is reduced by an amount of the form

$$E_{Ii} = |\det D| \sum_{j=0}^{n} \binom{n}{j} B_{il}\left(\left(\beta_2''\right)^{w_2} - \left(\frac{\underline{s}^T D\underline{\delta} - t}{\underline{s}^T D\underline{\eta}_i}\right)^{w_2}\right)\left(-\frac{\underline{s}^T D\underline{\delta} - t}{\underline{s}^T D\underline{\eta}_i}\right)^{j} \qquad (84)$$

due to an intersection distribution with vertex $\eta_i$ centered (with its image) at $\underline{\delta}$, and under the constraint Equation (60). The form of Equation (84) is based on the fact that D represents a rigid notation which implies that the hyperplane takes the form $\underline{s}^T D(\underline{\delta} - \underline{s}) = t$. The matrix D is unitary, and therefore the Jacobian det D is unity. When all components, vertex components, and intersections have been analyzed, one could formulate a total error similar to Equation (81) or deal with the distributions on a component pairwise basis so as to structure a piecewise linear discriminant surface and its accompanying ordering format, a decision tree. Application of the degenerate approximation discussed at the beginning of this section would be appropriate to determine the hyperplanes in either case.

At this point, it might be appropriate to consider the various $\underline{\delta}$, $\Sigma$ as random variables and calculate confidence intervals for the decision surfaces just determined. More exactly, the expectation and variance of the probability of error should be calculated as a function of the linear classifier. The classifier, in turn, is a function of the sample means and covariance matrices, and therefore of the number of samples and dimensions. Foley[14] has performed such an anaylsis for the Fisher linear discriminant and underlying identical multivariate normally distributed densities. In the present case, however, an explicit form for the discriminant is not available, although some further approximation to Equation (70) may be possible. In lieu of an analytic solution, it is possible to acquire some understanding of the expected error and its variance by a numerical method which exploits a variation of the so-called U or "hold-one-out" rule.[15] The rule states that the hyperplane $\underline{s}$,t is designed on the basis of N-1 samples and tested on the remaining sample. Different

hyperplanes are then successively designed for each sample held out. Each hyperplane now represents a value of the probability of error, as calculated on the basis of techniques described earlier in this report. Thus, one has a means not only of calculating the mean and variance of the probability of error but also of the hyperplane coefficients themselves. Hence, confidence intervals may be computed if one chooses to presume that the samples of the probability of error are normally distributed or indeed distributed in a way that could be calculated by the functions described in this report.

## SUMMARY AND CONCLUSION

A methodology has been presented in this paper for the description of probability density distributions with concentric rectangular isoprobable contour, associated first and second order statistics, and arbitrary roll-off. Included in the description are derivations of the normalization constant and the expression for the probability of error with respect to an arbitrary but fixed hyperplane. A piecewise linear approximation to the solution of the binary classification problem for a class of rectangular distributions with bounded support and monotonically decreasing roll-off was derived on the basis of error computation along decoupled coordinate axes. The paper then discussed the approximate decomposition of overlapping rectangular distributions. An approximation to the error increment contributed by the overlap - which would constitute double-counting if not subtracted from the total error - was constructed. The approximation, in its general formulation, suggests a modification to the axiom of additivity which nevertheless takes into consideration all the information available in a reasonable fashion while still striving for computational efficiency. Finally, a procedure was discussed for estimating the expected probability of error and its variance.

The results just noted make possible the generation of a computationally efficient discriminant: namely, a piecewise linear surface - a tree of hyperplanes - together with the explicit, closed-form calculation of the probability of error and its mean and variance for a wide class of densities without unduly restricting assumptions on their forms. On the

other hand, the actual computation of the classifier may still be a prodigious task. Although generally it is presumed that this task will be done only once, a slowly changing environment or the accumulation of more samples may require updating the statistics and hence the discriminating surface. The techniques discussed in this report do not lend themselves well to such an update, and other techniques - such as sequential classification[16,17] - must be pursued.

In conclusion, the concept postulated is that a discriminant surface can be expected to have extrapolative power only if it is stable in the face of new samples. In the case of parametric models - like those presented in this report - such stability is realized only at the computational cost of gathering enough data to analyze the fine density structure. However, in designing approximations to reduce costs, one need not sacrifice accuracy if the approximations do not adhere to the traditional axioms of probability but are reasonable exploitations of the data.

## APPENDIX A

## OPEN FORM SOLUTION TO THE HINGE PROBLEM*

If f and g are uncorrelated, Equations (10) and (11) may be rewritten so as to seek the maximum of

$$\text{tr}\left[W\ (R(\underline{g})+R(\underline{f})-\frac{e}{2}\ Y)\right] \tag{A1}$$

over W, s such that

$$\text{tr}\left[W\ (R(\underline{f})+R(\underline{g})-\frac{s}{2}\ Y\right] = k \tag{A2}$$

Define the diagonalizations

$$EUE^T = R(\underline{g}) + R(\underline{f}) - \frac{s}{2}\ Y \tag{A3}$$

$$CVC^T = R(\underline{f}) + R(\underline{g}) - \frac{s}{2}\ Y \tag{A4}$$

$$B+\Omega B = V^{1/2}C^T WCV^{1/2} \tag{A5}$$

where U, V, $\Omega$ are diagonal and E, C, B are orthonormal. Then if V is positive definite, Equations (A1) and (A2) may be rewritten so as to maximize

$$\sum_i \Omega_{ii}\ Q_{II} \tag{A6}$$

over $\Omega_{ii}$, $Q_{ii}$ such that

$$\sum \Omega_{ii} = k \tag{A7}$$

and

$$Q = BV^{-1/2}C^T EUE^T CV^{-1/2}B^T \tag{A8}$$

If $Q_{ii} = \max_i [Q_{ii}]$, then Equations (A6) and (A7) have the open solution

$$\Omega_{ji} = \begin{cases} k, & i = j \\ 0, & \text{otherwise} \end{cases} \tag{A9}$$

which still leaves the task of maximizing $Q_{ii}$ over B, U, V. Equation (A8) implies that the task is equivalent to seeking the maximum eigenvalue of $V^{-1/2}C^T EUE^T CV^{-1/2}$, or equivalently the maximum eigenvalue of the matrix in Equation (12) over $\underline{s}$.

---

*This is a generalization and a more elegant statement of the problem and proof given by Sebestyen,[9] p.44.

## APPENDIX B

## APPROXIMATION OF ELLIPSOIDAL NORMAL DENSITY BY RECTANGULAR

The approximation chosen is that equiprobable contours for the same density contain the same volume.

The volume of an ellipsoid is

$$(\Pi^{n/2}/\Gamma(1+n/2))2^{n/2}\ \Lambda^{-1/2}\ \left[\ln((2\Pi)^{-n/2}\ \det\ \Lambda^{1/2})\right]^{n/2} \qquad (B1)$$

The volume of a rectangular solid is

$$2^{3n/2}\det M^{-1/2}\left[\ln(2^{-3n/2}\det M^{1/2}/\Gamma(1+n/2))\right]^{n/2} \qquad (B2)$$

The bracketed quantities in Equations (B-1) and (B-2) are equal if

$$\det M^{1/2} = (2/\Pi)^{n/2}\Gamma(1+n/2)\det\Lambda^{1/2} \qquad (B3)$$

in which case the unbracketed quantities are also equal. Therefore, Equation (B-3) effectively defines a class of transformations which produce the desired approximation.

## REFERENCES


1. Batchelor, B.G., "Practical Approach to Pattern Classification," Plenum Press, London (1974).

2. Zadeh, L.A., "Fuzzy Sets," Information and Control, 8 (1965), pp. 338-353.

3. Terano, T., Suzeno, M., "Conditional Fuzzy Measures and Their Applications," in Fuzzy Sets and their Applications to Cognitive and Decision Processes, Zadeh, L.A. et al., eds., Academic Press, New York, N.Y. (1975) pp. 151-170.

4. Kaufmann, A., "Introduction to the Theory of Fuzzy Subsets," Academic Press, New York, N.Y. (1975) pp. 250-253.

5. Stallings, W., "Fuzzy Set Theory Versus Bayesian Statistics," IEEE Transactions on Systems, Man, and Cybernetics, SMC-7, 2 (Mar 1977) pp. 216-219.

6. Jain, R., "Comments on Fuzzy Set Theory Versus Bayesian Statistics," IEEE Transactions on Systems, Man, and Cybernetics, SMC-8, 4 (Apr 1978) pp. 332-333.

7. Kahneman, D., Tversky, A., "Subjective Probability: A Judgement of Representativeness," Cognitive Psychology, 3 (1972) pp. 430-454.

8. Gnedenko, B.V., Kolmogorov, A.N., "Limit Distributions for Sums of Independent Random Variables," Addison-Wesley, Cambridge, Mass. (1954).

9. Sebestyen, G., "Decision-making Processes in Pattern Recognition," Macmillan, New York, N.Y. (1962).

10. Anderson, T.W., Bahadur, R.R., "Classification into Two Multivariate Normal Distributions with Different Covariance Matrices," Annals of Mathematical Statistics, 33 (Jun 1962) pp. 420-431.

11. Gradshteyn, I., Ryzhik, I., "Table of Integrals, Series, and Products," Academic Press, New York (1965).

12. Duda, R., and Hart, P., "Pattern Classification and Scene Analysis," Wiley, New York (1973).

13. Haralick, R., "Pattern Discrimination Using Ellipsoidally Symmetric Multivariate Density Functions," Pattern Recognition Journal, Vol. 9, No. 3 (1977) pp. 89-94.

14. Foley, D.H., "Considerations of Sample and Feature Size," IEEE Transactions on Information Theory, IT-18, 5 (Sep 1972) pp. 618-626.

15. Kanal, L., and Chandrasekaran, B., "On Dimensionality and Sample Size in Statistical Pattern Classification," Proc. 1968 National Electronics Conference, pp. 2-7.

16. Fu, K.S., "Sequential Methods in Pattern Recognition and Machine Learning," Academic Press, New York (1968).

17. Ben-Bassat, M., "Myopic Policies in Sequential Classification," IEEE Transactions on Computers, Vol. C-27, No. 2 (Feb 1978) pp. 170-174.

## INITIAL DISTRIBUTION

| Copies | |
|---|---|
| 1 | CHONR, 430/R. Lundegard |
| 1 | DNL |
| 1 | NRL |
| 2 | NAVSEA<br>1 SEA 03F/B. Orleans<br>1 SEA 0351/T. Pierce |
| 1 | NAVMAT |
| 1 | USNA |
| 1 | NOSC |
| 1 | NSWC |
| 12 | DDC |
| 1 | ROME AIR DEV CEN |
| 1 | U. of Connecticut<br>U. Chien |
| 1 | U. of Florida, Gainesville<br>J. Tou |
| 1 | U. of Kansas<br>R. Haralick |
| 1 | U. of Maryland, College Park<br>L. Kanal |
| 1 | Southeastern Massachusetts U.<br>C.H. Chen |
| 1 | Stanford Research Inst.<br>P. Hart |
| 1 | IBM, Watson Research Lab<br>T. Anderson |

### CENTER DISTRIBUTION

| Copies | Code | Name |
|---|---|---|
| 1 | 18/1809 | |
| 1 | 1802.1 | |
| 1 | 1802.1 | |
| 1 | 1803 | |
| 1 | 1805 | |
| 1 | 182 | |
| 20 | 1824 | S. Berkowitz |
| 1 | 184 | |
| 1 | 184.1 | |
| 1 | 185 | |
| 1 | 187 | |
| 10 | 5211.1 | Reports Distribution |
| 1 | 522.1 | Unclassified Library (C) |
| 1 | 522.2 | Unclassified Library (A) |