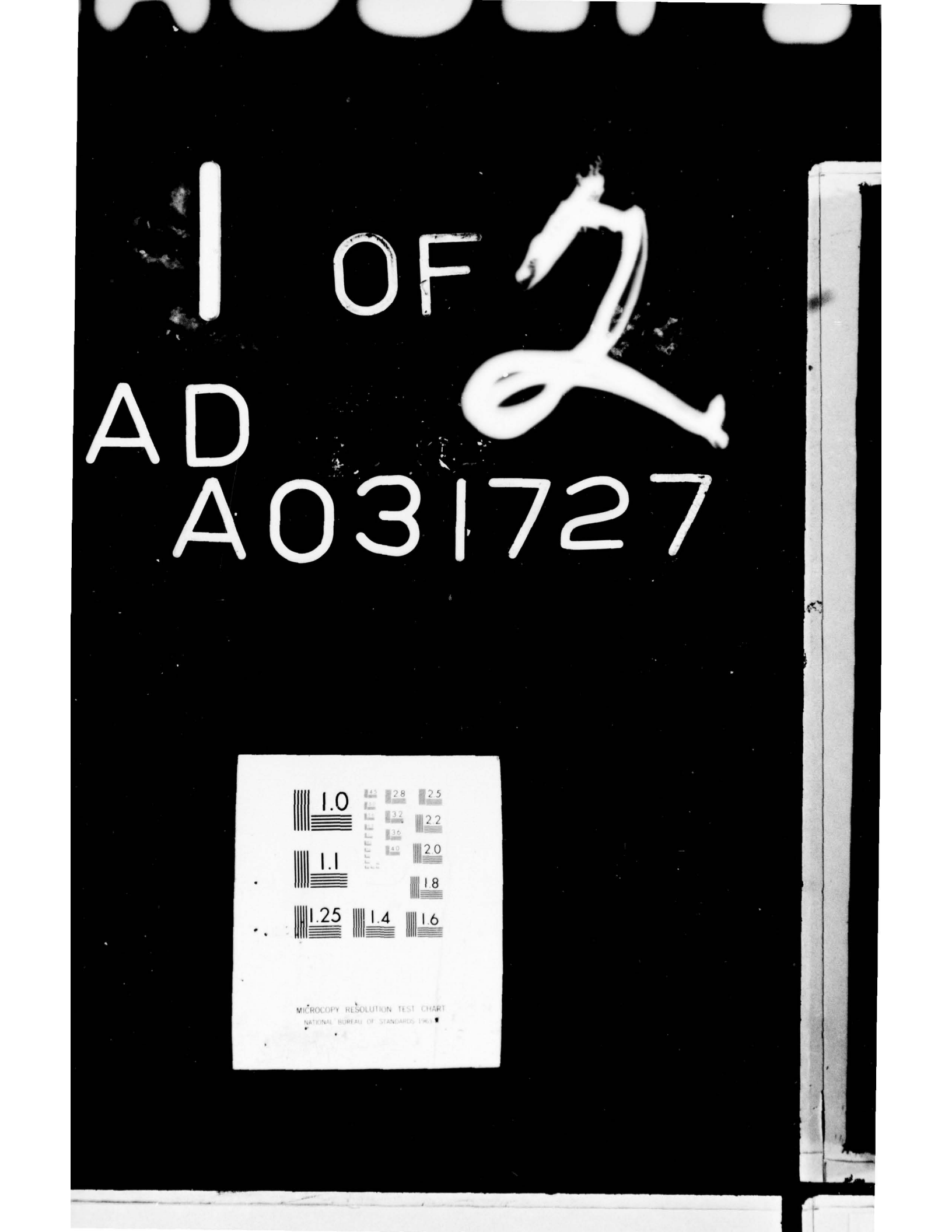
AD-A031 727 PURDUE UNIV LAFAYETTE IND SCHOOL OF ELECTRICAL ENGI--ETC F/G 9/3
A NONPARAMETRIC RECOGNITION PROCEDURE WITH STORAGE CONSTRAINT,(U)
AUG 69 E A PATRICK, F K BECHTEL F30602-68-C-0186
UNCLASSIFIED TR-EE69-24 NL

1 OF 2
AD A031727


1 OF 2
AD
A031727
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
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963

TR-EE 69-24

PURDUE UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

A NONPARAMETRIC RECOGNITION PROCEDURE WITH STORAGE CONSTRAINT

by

E. A. Patrick

and

F. K. Bechtel

August 1969

Lafayette, Indiana

Sponsored by
Rome Air Development Center, Griffiss Air Force Base, New York
under contract F 30 602-68-C-0186

Naval Ship Systems Command, Washington, D. C.
under contract N00024-69-C-1265

(14) TR-EE 69-24

Purdue University
School of Electrical Engineering

(6) A NONPARAMETRIC RECOGNITION PROCEDURE WITH STORAGE CONSTRAINT

by

(10) E. A. Patrick F. K. Bechtel

(11) August 1969
(12) 187 p.
Lafayette, Indiana

Sponsored By
Rome Air Development Center, Griffiss Air Force Base, New York
under contract F 30 602-68-C-0186
(15)
Naval Ship Systems Command, Washington, D. C.
under contract N00024-69-C-1265

292 000

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbol	Description
V^{ℓ}	ℓ-dimensional observation space
$\mathbf{x}$	A vector in V^{ℓ}
$\mathcal{D}$	A bounded domain in V^{ℓ}
ω_j	The j^{th} class
P_j	A priori probability of class ω_j
n	The total number of training observations
n_j	The number of training observations from class ω_j
Y_n	A set of n training observations
f_j	The density function describing an observation from class ω_j
L_j	Lipschitz constant for a Lipschitz condition assumed satisfied by f_j
d	A decision rule
d_o	An optimum decision rule
$d(x)$	Indicate the class chosen by d at x
$\bar{d}(x)$	Indicates the class not chosen by d at x
$Pr(\mathcal{E}\|d)$	Probability of error when using d
$Pr(\mathcal{E}\|d_o)$	Minimum attainable probability of error
α	Bound on acceptable excess of $Pr(\mathcal{E}\|d)$ over $Pr(\mathcal{E}\|d_o)$
β	Confidence to be attained that $Pr(\mathcal{E}\|d) - Pr(\mathcal{E}\|d_o)$ is acceptable

Symbol	Description
$\hat{f}_j$	Estimate of f_j
Ψ_j	A function used in the representation of f_j
$\{\Psi_{ji}\}_{i=1}^{R}$	A set of R functions used in the representation of f_j
$\hat{C}_{ji}$	Estimate of coefficient on Ψ_{ji} in representation of f_j
ISE	Integral square error
MSE	Mean square error
x_{ji}	i^{th} training observation from j^{th} class
I	A partition of $\mathcal{A}$
R	The number of intervals in I
$\mathcal{A}_I(i)$	The i^{th} interval in I; denoted $\mathcal{A}(i)$ when I is understood
$W_I(i)$	Width of $\mathcal{A}_I(i)$; denoted W(i) when I is understood
W_T	Total domain width
$p^*_{Ij}(i)$	Probability from j^{th} class in $\mathcal{A}_I(i)$; denoted $p^*_j(i)$ when I is understood
$P_j(i)$	Random variable describing uncertainty in $p^*_j(i)$; I is understood
$\hat{P}_j(i)$	The expected value of $P_j(i)$
$\underline{P}_j$	Vector of probabilities $(P_j(1),\ldots,P_j(R))^t$
$\underline{m}_j$	Vector of parameters $(m_j(1),\ldots,m_j(R))^t$ characterizing a Dirichlet density on $\underline{P}_j$
$\gamma_{j1}(i),\gamma_{j2}(i)$	Parameters characterizing a beta density on $P_j(i)$
$s_{j1}(i), s_{j2}(i)$	Contribution to $\gamma_{j1}(i),\gamma_{j2}(i)$ from a priori knowledge
$v_{j1}(i),v_{j2}(i)$	Contribution to $\gamma_{j1}(i)\gamma_{j2}(i)$ from training observations

Symbol	Description
$Pr(\mathcal{E}\|i,d)$	Probability of error when using d at x in $\mathcal{I}(i)$
$Pr(\mathcal{E}\|i,d_o)$	Probability of error when using d_o at x in $\mathcal{I}(i)$
$Pr(i)$	Probability that an observation is in $\mathcal{I}(i)$
$Q(i,d,d_o)$	$= [Pr(\mathcal{E}\|i,d) - Pr(\mathcal{E}\|i,d_o)]Pr(i)$
$\alpha(i)$	Portion of α alloted to $\mathcal{I}(i)$
$\tau(i)$	Portion of $1 - \beta$ alloted to $\mathcal{I}(i)$
$x \triangleq y$	Equality by definition
$U_j(i)$	Scaled version of $P_j(i)$
$\beta(P\|\gamma_1,\gamma_2)$	A beta density function on P with parameters γ_1,γ_2
$\beta^*(U\|\gamma_1,\gamma_2)$	A scaled beta density function
$\mu_j(i)$	Expected value of $U_j(i)$
$\sigma_j^2(i)$	Variance of $U_j(i)$
$a(i)$	$d(x)$ for $x\in\mathcal{I}(i)$
$b(i)$	$\bar{d}(x)$ for $x\in\mathcal{I}(i)$
$V_1(i),V_2(i)$	Regions of $(U_a(i),U_b(i))$ plane
$\bar{V}_1(i),\bar{V}_2(i)$	Regions of $(U_a(i),U_b(i))$ plane containing $V_1(i),V_2(i)$
$V(i)$	Region of $(U_a(i),U_b(i))$ plane containing points common to both $\bar{V}_1(i)$ and $\bar{V}_2(i)$
$\bar{f}$	Average of f over an interval
ρ	$\alpha(i)/W$; i suppressed on W(i)
C_j	$P_j L_j W/2$; i suppressed on W(i)
δ	C_b if $U_b \geq C_b$, $[(4C_b U_b)^{\frac{1}{2}} - U_b]$ if $U_b < C_b$
ΔV	Increment in V(i) when C_b increases from less than ρ to greater than ρ
$T(i)$	Bound on Pr(V(i)) when U's are beta distributed

Symbol	Description
$\Lambda(i)$	Bound on $Pr(V(i))$ when U's are Gaussian distributed
$Be(\gamma_1,\gamma_2)$	$\Gamma(\gamma_1)\Gamma(\gamma_2)/\Gamma(\gamma_1+\gamma_2)$
$g(U_j\|\mu_j,\sigma_j^2)$	A Gaussian density function on U_j with mean μ_j and variance σ_j^2
ξ_1,ξ_2	The U_b intercept and slope respectively of a straight line supporting V(i) in the (U_a,U_b) plane
$\Phi(x)$	Cumulative Gaussian distribution function evaluated at x
λ_a	The U_a coordinate of a point on the quadratic boundary of V(i)
$\hat{p}$	Estimate of mixture probability
$\hat{n}$	Estimate of training observations required
$\hat{r}$	Estimate of classification rate
W'	Variable interval width (primed quantities are variable quantities)
$\hat{r}_M$	Estimated classification rate maximized over W'
$\hat{W}_M$	W' value resulting in $\hat{r}_M$
S:N	Signal to noise ratio
b	Number system radix
K	Integer describing mapping complexity
ℓ	Integer defining dimensionality of $\underline{x}$
h_j	Multidimensional approximation to f_j
g_j	One-dimensional equivalent of h_j
L_j^*	A "pseudo-Lipschitz constant"
$\mathcal{R}$	The real line
$S_{e_1,\ldots,e_\ell}(b,K,\ell)$	A set in V^ℓ indexed by $e_1,\ldots,e_\ell$

Symbol	Description
$S_e(b,K\ell)$	A one-dimensional set in $\mathbb{R}$ indexed by e
α_{ji}	Base b digit in expansion of e_j
β_{ji}	Base b digit determined by $\{\alpha_{ji}\}$

ABSTRACT

A procedure is described for determining a decision rule for the one-dimensional, two class recognition problem with unknown, nonparametric, class-conditional density functions. A priori class probabilities are known, and the densities are assumed to satisfy Lipschitz conditions with known Lipschitz constant. The procedure is essentially a histogram approach where the partition for the histogram is changed as directed by a performance measure. It is desirable to minimize the difference between the probability of a recognition error when using the decision rule and the minimum attainable probability of recognition error. For a fixed partition conditions are stated that assure achievement of a specified confidence that this difference is below a specified constant. The variable partition procedure operates with limited storage and allows, but does not assure, attainment of the specified confidence. Computer simulated results are given that experimentally illustrate attainment of the desired confidence for the problems considered. A technique is suggested for extending the procedure to multidimensions. This technique converts the multidimensional problem to a one-dimensional problem. It operates by mapping sets in a multidimensional domain one-to-one onto sets in a one-dimensional domain. Computer simulated results are presented.

CHAPTER I

INTRODUCTION

1.1 The Problem

One of the problems in computerized recognition is that of assigning a vector observation to one of several classes. Applications include the recognition of properties of waveforms or pictures which are represented by vectors. The total recognition problem should include the following operations:

A) Select sensors for the problem and represent the sensor outputs for each waveform, picture, or etc. by an f-dimensional vector. This operation involves expert problem knowledge.

B) Represent the f-dimensional vector with a vector in a ℓ-dimensional space ($\ell < f$) called the observation space and denoted V^{ℓ}. This is accomplished using a data-dependent, dimensionality reducing mapping. Denote with $\underline{x}$ a ℓ-dimensional vector in V^{ℓ}.

C) Recognize the ℓ-dimensional vector by assigning it to one of several classes using a classification procedure conditioned on previously processed observation vectors called training observations.

Examples of applications include automatic sonar and radar detection and classification, medical diagnosis including electrocardiograms and electroencephalograms, aerial photography processing for earth resource studies, and quality control.

This report is concerned with c); thus, the problem begins with a set Y_n of n, ℓ-dimensional vector training observations (often called patterns [1]) in a ℓ-dimensional observation space denoted V^{ℓ}. Using the set Y_n together with available a priori knowledge, an observation $\underline{x}$ is assigned by the classification procedure to one of several classes.

The assumptions and constraints specifying the particular classification problem considered in this report are:

1) A vector observation $\underline{x}$ is to be assigned to one of two classes, denoted ω_1 and ω_2. The a priori probabilities P_1 and P_2 that $\underline{x}$ belongs respectively to ω_1 or ω_2 are known.

2) The observation space is 1-dimensional (Chapter V describes a method for extending the results to ℓ-dimensions).

3) The training observations are supervised; that is, the correct classification of each observation in Y_n is known. The number of training observations belonging to ω_j is n_j with $n_1 + n_2 = n$.

4) Training observations belonging to ω_j are each independently and identically distributed according to an unknown class-conditional density function f_j defined over the observation space. Class ω_1 observations are independent of class ω_2 observations. f_j is <u>not</u> assumed to be parametric; it cannot necessarily be characterized by a finite number of parameters. It is assumed that $f_j(x)$ is zero for x outside a known bounded domain $\mathcal{D}$. Without loss of generality

$$\mathcal{D} = \{x : 0 \leq x \leq 1\} \tag{1.1}$$

In addition, it is assumed that f_j satisfies a Lipschitz condition

$$|f_j(x) - f_j(y)| \leq L_j|x - y| \quad , \quad x,y \in \mathcal{X} \tag{1.2}$$

with Lipschitz constant L_j known a priori.

5) The amount of computer storage is limited.

The result of training (processing the training observations) is the specification of a decision rule d defined on $\mathcal{X}$ and taking on the values 1 or 2. The rule d divides $\mathcal{X}$ into two sets identified by their assigned classes. An observation at x is assigned to class $\omega_{d(x)}$. The choice of d minimizing the probability of a classification error is a minimum risk procedure. Details of such procedures are found in references [2,3,20]. The probability $Pr(\mathcal{E}|d)$ of classification error when using d is given by

$$Pr(\mathcal{E}|d) = \int_{\mathcal{X}} P_{\bar{d}(x)} f_{\bar{d}(x)}(x)\, dx \tag{1.3}$$

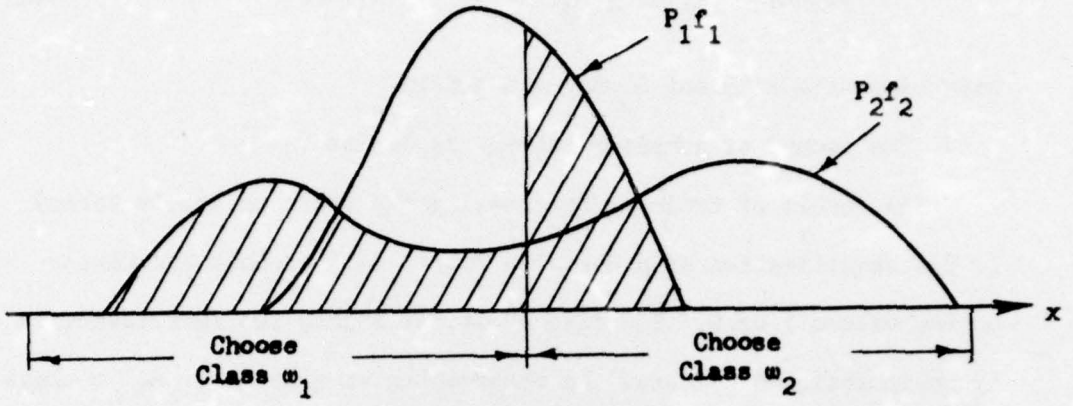
where $\bar{d}(x)$ indentifies the class <u>not</u> assigned by d to an observation at x. An optimum decision rule d_o is defined as one that minimizes $Pr(\mathcal{E}|d)$. If j is considered to be the argument of $P_j f_j(x)$, then $d_o(x)$ is given for each x in $\mathcal{X}$ by

$$d_o(x) = \text{Arg}\left[\max_{j=1,2} P_j f_j(x)\right] \tag{1.4}$$

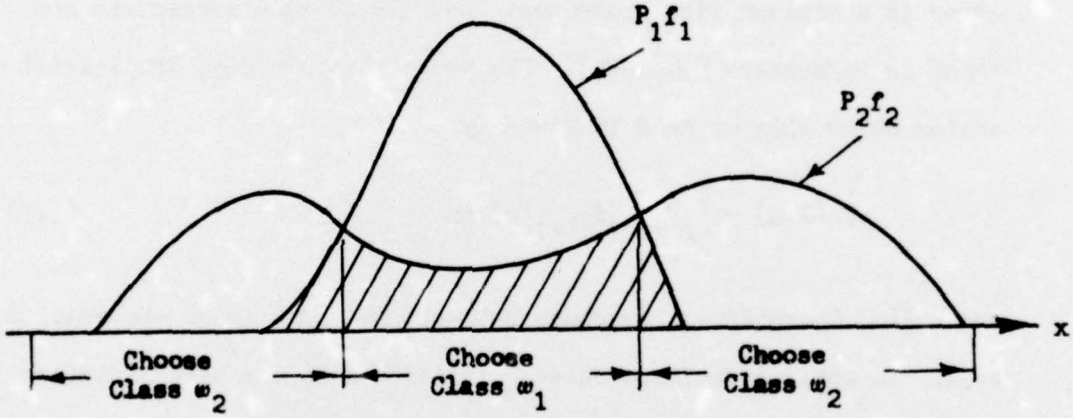
The corresponding minimum probability of error is

$$Pr(\mathcal{E}|d_o) = \int_{\mathcal{X}} \left[\min_{j=1,2} P_j f_j(x)\right] dx \tag{1.5}$$

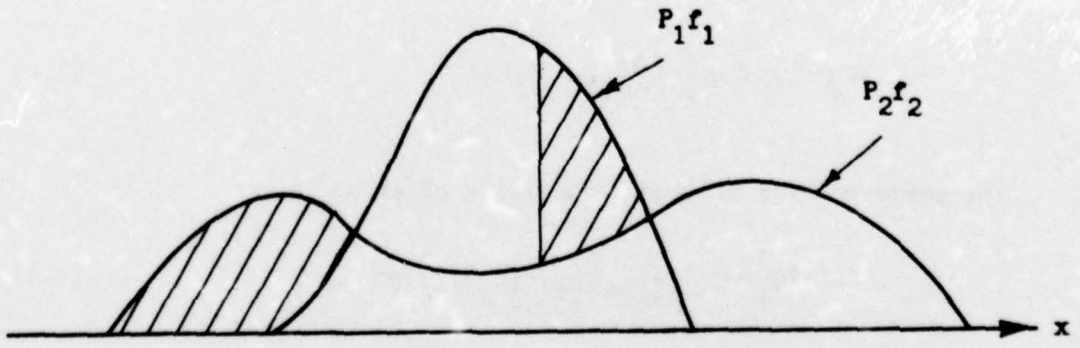
Figure 1a illustrates a decision rule d for a particular example.



(a) $Pr(e|d)$



(b) $Pr(e|d_o)$



(c) $Pr(e|d) - Pr(e|d_o)$

Figure 1. Decision Rule Illustrations

The cross-hatched area represents the corresponding probability of error, $Pr(\mathcal{E}|d)$. Similarly Figure 1b illustrates d_o and $Pr(\mathcal{E}|d_o)$. Figure 1c illustrates the excess of $Pr(\mathcal{E}|d)$ over $Pr(\mathcal{E}|d_o)$.

1.2 The Goal

An optimum decision rule d_o is defined in terms of the class-conditional density functions+ f_j. For the problem being considered these d.f.'s are unknown; however, they can be estimated and the estimates $\hat{f}_j$ substituted into (1.4) in place of f_j. The result is a decision rule d that is an estimate for the decision rule d_o. The overall objective is to satisfy

$$Pr[Pr(\mathcal{E}|d) - Pr(\mathcal{E}|d_o) \leq \alpha] \geq \beta \tag{1.6}$$

for prespecified constants α and β in the interval $[0,1]$. In words, the goal is to achieve a specified confidence that the excess of $Pr(\mathcal{E}|d)$ over $Pr(\mathcal{E}|d_o)$ is less than a specified constant.

1.3 Literature Survey

The previously described problem of obtaining a decision rule constrained by limited storage and with a goal given by (1.6) has apparently received no previous attention. The closest results are probably due to Fu and Henrichon [4] who find constants α' and β so that

$$Pr[Pr(\mathcal{E}|d) \leq \alpha'] \geq \beta \tag{1.7}$$

+Hereafter the phrase class-conditional is dropped, and f_j is referred to as a density function (abbreviated to d.f.).

is satisfied. This condition provides a statement about the size of $Pr(\mathcal{E}|d)$ whereas condition (1.6) for the current problem is concerned with the size of $Pr(\mathcal{E}|d)$ relative to $Pr(\mathcal{E}|d_o)$. For the special case when $Pr(\mathcal{E}|d_o)$ is known or is known to be negligibly small with respect to α, (1.6) and (1.7) are equivalent. Otherwise (1.6) offers the advantage of providing information concerning the amount of improvement in performance obtainable by processing additional training observations. Fu and Henrichon's procedure operates on all the training observations essentially simultaneously and requires increasing computer storage as the number of training observations increases; thus it is not applicable with the current storage constraint.

The missing link in a straight-forward application of (1.4), to obtain d, is the method of obtaining the estimate d.f.'s $\hat{f}_j$ from the n_j class ω_j training observations. The limited storage constraint complicates this estimation.

Abramson and Braverman [5], and Keehn [6] consider estimates of the form

$$\hat{f}_j = \psi_j \tag{1.8}$$

where ψ_j is a member of the family of Gaussian d.f.'s. They estimate parameters (mean vectors [5], mean vectors and covariance matrix [6]) characterizing ψ_j. With $\{\psi_{ji}\}_{i=1}^{R}$ a complete orthonormal set for the unknown f_j, Aizerman, Braverman, and Rozonoer [7] obtain estimated parameters $\hat{C}_{ji}$ in

$$f_j = \sum_{i=1}^{R} \hat{C}_{ji}\psi_{ji} \tag{1.9}$$

and show that $\hat{f}_j(\underline{x})$ converges in probability to $f_j(\underline{x})$ for each $\underline{x}$ in the domain[+] $\mathcal{D}$. Tsypkin [8] also uses an orthonormal set $\{\psi_{ji}\}_{i=1}^{R}$ to get an estimate of the form (1.9), but he does not assume it to be complete for f_j. Tsypkin obtains estimates $\hat{C}_{ji}$ with the goal to minimize the Integral Square Error (ISE),

$$\text{ISE} = \int_{\mathcal{D}} \left(f_j(\underline{x}) - \hat{f}_j(\underline{x})\right)^2 d\underline{x}$$

Kashyap and Blaydon [9] assume only that $\{\psi_{ji}\}_{i=1}^{R}$ are linearly independent functions. With an estimate $\hat{f}_j$ of the form (1.9), they consider minimizing both the ISE and the Mean Square Error (MSE),

$$\text{MSE} = \int_{\mathcal{D}} \left(f_j(\underline{x}) - \hat{f}_j(\underline{x})\right)^2 f_j(\underline{x})\, d\underline{x}$$

For the 1-dimensional case Rosenblatt [10] considers an estimate of the form (1.9) for $f_j(x)$ where $R = n_j$ and ψ_{ji} is a function obtained from the i^{th} class ω_j observation x_{ji}.

$$\hat{f}_j = \sum_{i=1}^{n_j} \frac{1}{n_j} \psi_{ji} \tag{1.10}$$

Parzen [11] shows that if

$$\psi_{ji}(x) = \frac{1}{\sigma_{n_j}} K\left(\frac{x - x_{ji}}{\sigma_{n_j}}\right) \tag{1.11}$$

where

[+]In this section on d.f. estimation, $\mathcal{D}$ can be multidimensional unless otherwise stated.

$\{\sigma_{n_j}\}$ and the function K satisfy certain conditions, then

$$\lim_{n_j \to \infty} E\left[\left(f_j(x) - \hat{f}_j(x)\right)^2\right] \to 0$$

for each x in the domain at which f_j is continuous. Because of the form (1.10), this estimate has increasing complexity as n_j increases. References [12,13,14,15] also deal with this type of d.f. estimation.

The well known histogram technique for estimating d.f.'s defines the functions $\{\psi_{ji}\}_{i=1}^{R}$ as the set of indicator functions on the regions of a R-region partition of $\mathcal{X}$. For the i^{th} region,

$$\psi_{ji}(\underline{x}) = 1, \ \underline{x} \text{ in the } i^{th} \text{ region}$$

$$= 0, \text{ otherwise}$$

$\hat{f}_j$ is given by (1.9). This is a special case of the problem considered in references [7,8,9].

The nearest neighbor decision rule or rather the more general K-nearest neighbor decision rule [16] assigns an unclassified observation to the class most heavily represented among its K nearest training observations. This rule has been shown [17,18] to have similarities with a decision rule resulting from using density function estimates in (1.4). It has been shown [16] that the nearest neighbor rule results in an asymptotic ($n_j \to \infty$) probability of error that is

less than twice the minimum attainable. Processing requires storage for all training observations; thus these results cannot be used when one operates with a storage constraint. Hart [19] has suggested an interesting storage reducing modification of the nearest neighbor rule which he calls the condensed nearest neighbor (CNN) rule. The CNN rule discards a set of training observations from the original set. The discarded set consists of training observations that, if treated as unclassified observations, are classified correctly by the nearest neighbor rule when used with the training observations retained. The storage requirement is reduced, and the criterion for discarding a training observation is based on the capability of the retained observations to make decisions. Supporting theory for the CNN rule has not yet been published.

The Gaussian assumption in Abramson and Braverman's work is too restrictive for the problem outlined in Section 1.1. The work of Aizerman, Braverman, and Rozonoer, Tsypkin, and Kashyap and Blaydon, along with the histogram approach is either too restrictive (small R) or requires too much storage (large R).

Unsupervised estimation [20,21,22,23,24,25] allows the estimate of (1.9) to be more general by providing a way to estimate parameters characterizing each ϕ_{ji} in $\{\phi_{ji}\}_{i=1}^{R}$ as well as the weighting coefficients.

Another approach [26,27,28] that adapts the ϕ_{ji}'s to the data is based on distribution free tolerance regions. Instead of defining a partition of $\mathcal{D}$ beforehand as in the histogram approach, a procedure

is given for defining the partition in terms of the training observations; then the distribution free techniques described in references [2,29] can be used.

Sebestyen [30,31] considers a method that is similar to the Parzen technique but uses limited storage. Training observations in close proximity with one another in $\mathcal{R}$ are lumped into an average observation. Sebestyen's estimate d.f. is in the form (1.9) where Φ_{ji} is a Gaussian d.f. having mean at the i^{th} average observation and variance related to the size of the region in which observations contribute to the average. $\hat{C}_{ji}$ is the relative frequency of observations in the region. The procedure does not have the properties that Parzen used in his convergence proof.

Specht [32] reduces the storage required in a utilization of the Parzen approach by expanding estimates in the form (1.10) into a Taylor series about a selected point in $\mathcal{R}$ and then retaining only the low order terms. The resulting truncated Taylor series is accurate only near the point of expansion. To obtain accuracy over the whole domain, the expansion should be carried out at each of sufficiently many points in the domain. A different set of coefficients must be stored for each expansion; thus the storage required would increase in proportion to the number of expansion points used.

When estimating d.f.'s, one must use care to choose a suitable estimation criterion. This is especially true if one is faced with the problem of estimating while being constrained with limited storage.

If the d.f.'s cannot be characterized by a number of parameters that will fit into the limited storage, then some information must be discarded. In this case, the criterion should not require accurate estimation where it is not needed because such accuracy is obtained at the expense of accuracy where it is needed. When the goal of the estimation is for the estimate d.f.'s to make good decisions if substituted for the actual d.f.'s in (1.4), it is reasonable that some measure of the quality of these decisions should be used as the estimation criterion. Since (1.4) involves a d.f. for each class, the estimation of one function should involve interaction with the estimation of the other function. With the exception of the K-nearest neighbor rule, the above d.f. estimation procedures do not have this property.

For other work related to computerized recognition, the reader is referred to the survey articles by Nagy [33], and Ho and Agrawala [34] which contain extensive lists of references.

1.4 The Approach

The d.f. estimation used in this report is essentially a histogram approach but with the partition periodically adapted to improve a measure of performance. Enough storage is assumed available to handle parameters associated with each interval in the partition. The supposition is that a number R of intervals too restrictive in the ordinary histogram approach may be adequate with the adaptive capability. This idea is suggested by the fact that a R-interval histogram d.f. estimation procedure is capable of giving an optimum

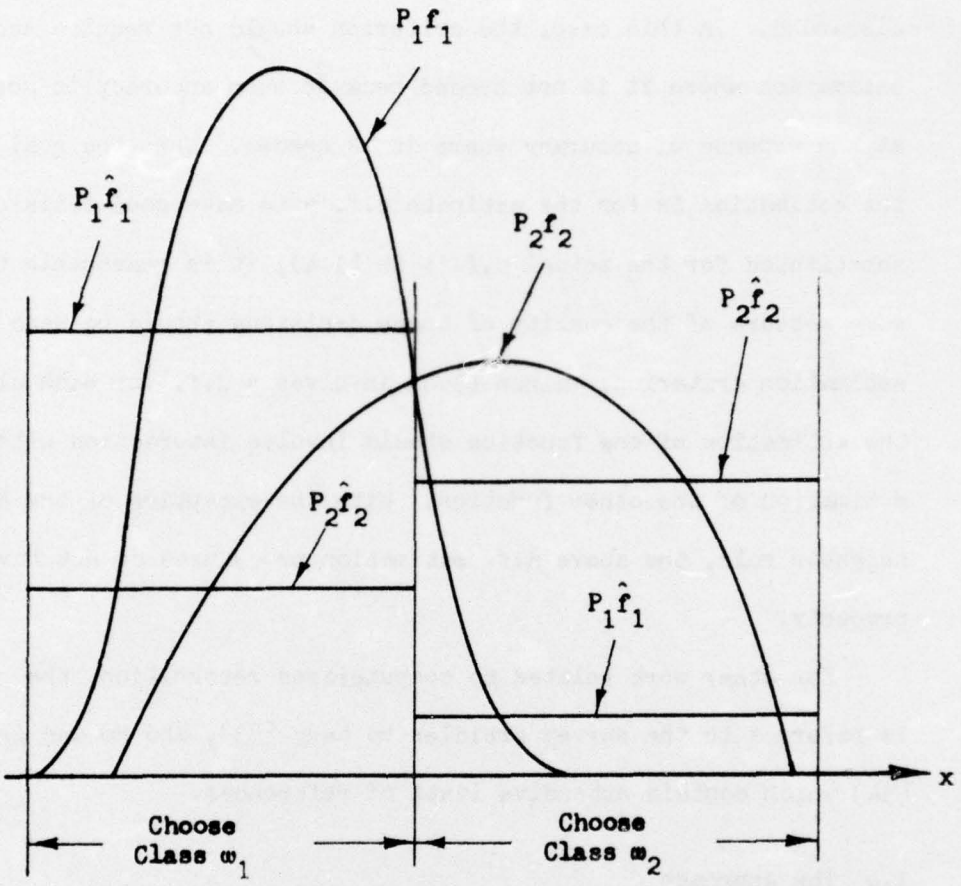


Figure 2. Two Interval Histogram Estimation Giving an Optimum Decision Rule.

decision rule provided the problem has fewer than R decision thresholds. Figure 2 illustrates a one threshold example optimally solved with a two interval histogram estimation of f_1 and f_2.

The framework or model within which the classification procedure operates is now described. Consider a partition I of the domain $\mathcal{D}$ into R intervals. Label these intervals $\mathcal{I}_I(1)$, $\mathcal{I}_I(2),\ldots,\mathcal{I}_I(R)$ and the interval widths $W_I(1)$, $W_I(2),\ldots,W_I(R)$. Define the probabilities $\rho^*_{Ij}(1)$, $\rho^*_{Ij}(2),\ldots,\rho^*_{Ij}(R)$, $j = 1,2$, by

$$\rho^*_{Ij}(i) \triangleq \int_{\mathcal{I}_I(i)} f_j(x)\,dx \qquad \begin{matrix} i = 1,\ldots,R \\ j = 1,2 \end{matrix} \qquad (1.12)$$

Although the ρ^*'s are unknown, any a priori knowledge concerning them is represented by the notation A. The set consisting of the first n training observations is denoted Y_n. Through the use of A and Y_n, the classification procedure obtains estimates $\hat{f}_j$ conditioned on I,A, and Y_n. In the remainder of this report, the partition I, the a priori knowledge A, and the training observations Y_n will be understood from the text and are omitted from the notation.

Given a partition, the estimate for f_j is

$$\hat{f}_j = \sum_{i=1}^{R} \frac{\hat{\rho}_j(i)}{W(i)}\, \phi_i \qquad j = 1,2 \qquad (1.13)$$

where ϕ_i is the indicator function for the i^{th} interval. $\hat{\rho}_j(i)$ is the expected value of a distribution on $\rho_j(i)$ which is a random variable describing the current uncertainty of $\rho^*_j(i)$. The a priori knowledge A, or in its absence the first few training observations,

are used to assign this distribution initially. It is updated with subsequent training observations through the use of Bayes Rule. By adjusting the variance of the initial distribution, its effect on the result can be made large or small as desired.

When the resulting estimates are used in place of f_1 and f_2 to obtain decision rule d, there can be no finer resolution of decision thresholds than the boundaries of the intervals comprising the partition. For this reason the capability of altering the partition is included in the model.

If the number R of intervals in the partition is greater than or equal to the number of decision thresholds plus one, then the model is capable of giving an optimum decision rule. An optimum decision rule is attained when all thresholds coincide with interval boundaries and when each interval is classified correctly through use of the estimate functions.

A general description of the approach used to satisfy condition (1.6) is now presented. The discussion follows the system flow diagram+ of Figure 3.

a) Initialization

Initially, a partition and a distribution on each $P_j(i)$ is assigned. This assignment is based on a priori knowledge about $P_j^*(i)$.

b) Updating

A set of supervised training observations is used to update the distribution on each $P_j(i)$ through use of Bayes Rule.

+This flow diagram corresponds to an actual implementation, the results of which are presented in Chapter IV.

c) Classification

The constants α and $1-\beta$ are allocated to the intervals and a set of R conditions, one for each interval, similar to condition (1.6) for the whole domain, is developed. These interval conditions taken together are sufficient for (1.6). The interval condition for each interval is checked independently of the others. A record is made of any interval whose interval condition is satisfied and of the class assigned to that interval; (such an interval is said to be classified). If all intervals are classified then processing is stopped with the statement that condition (1.6) is satisfied. Otherwise processing continues.

d) Adjust the Partition

The classification rate of an interval is defined as the total probability in the interval divided by the number of training observations required to classify it. The unclassified intervals are ranked in a priority table according to estimates of the maximum possible classification rates for the intervals. The maximization is with respect to interval width. The partition is adjusted by considering the intervals one at a time in the order that they appear in the priority table. An interval is either split into two intervals, combined with one of its adjacent intervals, or left unchanged according to a rule based on a measure of performance and the storage constraint. After partition adjustment, a priori knowledge is reassigned to the intervals. The process repeats as often as is necessary according to the flow diagram of Figure 3.

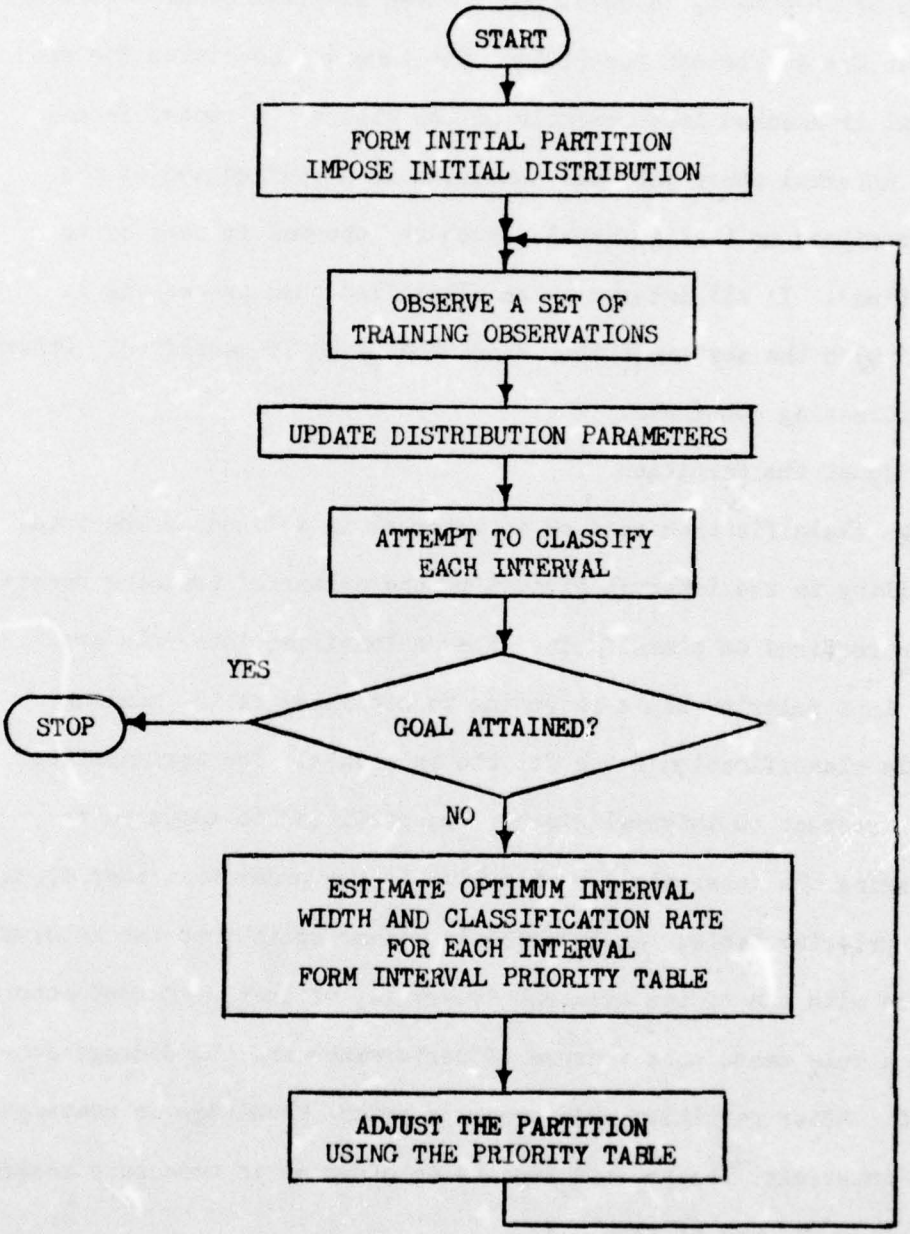


Figure 3. System Flow Diagram

1.5 Report Organization

Chapter II contains the details of the approach for a fixed partition. The initialization and updating of the distributions on the $P_j(i)$'s is discussed. The use of these distributions to obtain estimate d.f.'s and the subsequent use of the estimates to obtain a decision rule is described. Next, a set of interval conditions that is sufficient for condition (1.6) is derived.

Chapter III describes an ad hoc approach for altering the partition in order to arrive at the goal with limited storage and with fewer training observations.

Chapter IV contains computer simulated results. Experimental studies are included on the effects of tradeoffs between α and β of condition (1.6) and the total number of training observations required for satisfying it. Also studied are the effects of altering the Lipschitz constants, the number of intervals, and the number of training observations observed between times of making computations.

Chapter V contains suggestions for extending the approach to the multidimensional case via a technique that transforms the multidimensional problem into a 1-dimensional one. Possible uses for the mapping other than computerized recognition are discussed.

Chapter VI summarizes the results, their possible engineering application, and suggests ways in which they might be improved and extended.

CHAPTER II

SOLUTION FOR A FIXED PARTITION

2.1 Introduction

This chapter contains a description of the technique employed for a fixed partition I of the domain. Estimation of density functions and a decision rule are discussed. The difference between the probability of error using the estimated d.f.'s and that using the actual d.f.'s is expanded into a sum of difference probabilities where each difference probability corresponds to an interval in the partition. The objective is to achieve a specified confidence that the sum is less than a specified constant. A method that operates by considering each interval independently is developed for checking whether the confidence is attained.

2.2 Density Function Estimates

A piecewise constant estimate of the j^{th} class-conditional d.f. is

$$\hat{f}_j = \sum_{i=1}^{R} \frac{\hat{p}_j(i)}{W(i)} \Psi_i \qquad (2.1)$$

The random vector[+] $\underline{\rho_j} = (\rho_j(1), \rho_j(2), \ldots, \rho_j(R))^t$ has the $R-1$ variate Dirichlet density function

$$f(\underline{\rho_j}|\underline{m_j}) = \Gamma\Big(\sum_{i=1}^{R} m_j(i)\Big) \prod_{i=1}^{R} \frac{\rho_j(i)^{m_j(i)-1}}{\Gamma\big(m_j(i)\big)}, \quad 0 \le \rho_j(i) \le 1, \quad \sum_{i=1}^{R} \rho_j(i) = 1$$

$$= 0 \quad , \text{ otherwise} \tag{2.2}$$

assuming an a priori Dirichlet density function on $\underline{\rho_j}$ and subsequent training observations where $\underline{m_j} = (m_j(1), m_j(2), \ldots, m_j(R))^t$. Each $m_j(i)$ is obtained from training observations and a priori knowledge about $\rho_j(i)$ [35]. $\rho_j(i)$ has the beta (univariate Dirichlet) density function,

$$\beta\big(\rho_j(i)|\gamma_{j1}(i), \gamma_{j2}(i)\big) = \frac{\Gamma\big(\gamma_{j1}(i)+\gamma_{j2}(i)\big)}{\Gamma\big(\gamma_{j1}(i)\big)\Gamma\big(\gamma_{j2}(i)\big)} \rho_j(i)^{\gamma_{j1}(i)-1}\big(1-\rho_j(i)\big)^{\gamma_{j2}(i)-1}$$

$$0 \le \rho_j(i) \le 1$$

$$= 0 \quad , \text{ otherwise} \tag{2.3}$$

where

[+] t indicates transpose.

$$\gamma_{j1}(i) = m_j(i)$$

$$\gamma_{j2}(i) = \sum_{k \neq i} m_j(k) \tag{2.4}$$

The mean and variance of $\rho_j(i)$ are

$$E_j(i) = E[\rho_j(i) | \gamma_{j1}(i), \gamma_{j2}(i)] = \frac{\gamma_{j1}(i)}{\gamma_{j1}(i) + \gamma_{j2}(i)}$$

$$\text{Var}_j(i) = E\left[\left(\rho_j(i) - E_j(i)\right)^2 | \gamma_{j1}(i), \gamma_{j2}(i)\right]$$

$$= \frac{E_j(i)[1 - E_j(i)]}{\gamma_{j1}(i) + \gamma_{j2}(i) + 1} \tag{2.5}$$

If $\hat{\rho}_j(i) = E_j(i)$, then

$$\hat{f}_j = \sum_{i=1}^{R} \frac{\gamma_{j1}(i)}{[\gamma_{j1}(i) + \gamma_{j2}(i)]W(i)} \forall i$$

The components of $\underline{m_j}$ may not be consistent with a priori knowledge of the expected value and variance for each $\rho_j(i)$. For this reason and because each interval is to be considered independently, the Dirichlet d.f. is abandoned in favor of an independent beta d.f. on each $\rho_j(i)$. Then, it is consistent to constrain the γ's as follows:

$$\gamma_{j1}(i) = s_{j1}(i) + v_{j1}(i)$$

$$\gamma_{j2}(i) = s_{j2}(i) + v_{j2}(i) \tag{2.6}$$

where the s's and v's account respectively for a priori knowledge and training observations. The ability to use a priori knowledge is important for the partition changing technique developed in Chapter III. An "a priori d.f." on $P_j(i)$ is converted to an "a posteriori d.f." by using a Bayes iteration:

$$\beta\left(P_j(i) \mid \gamma_{j1}(i), \gamma_{j2}(i)\right) = \frac{\Pr\left(v_{j1}(i), v_{j2}(i) \mid P_j(i), s_{j1}(i), s_{j2}(i)\right) \beta\left(P_j(i) \mid s_{j1}(i), s_{j2}(i)\right)}{\int_0^1 \text{Numerator } dP_j(i)} \tag{2.7}$$

The iteration includes the information that out of

$$n_j = v_{j1}(i) + v_{j2}(i) \tag{2.8}$$

training observations from the j^{th} class, $v_{j1}(i)$ are in, and $v_{j2}(i)$ are out of the i^{th} interval.

Appendix A considers approaches for specifying the s's that characterize the a priori d.f. on $P_j(i)$. From (2.6), it is seen that enough training observations will eventually cause the effects of the s's to be negligible (provided each interval probability is greater than zero).

2.3 The Decision Rule

An estimate of the minimum probability of error decision rule is

$$d(x) = \text{Arg}\left[\max_{j=1,2} P_j \hat{f}_j(x)\right], \quad x \in \mathcal{X} \tag{2.9}$$

The difference between the probability of error when using d and the probability of error when using an optimum decision rule d_o is

$$\Pr(\mathcal{E}|d) - \Pr(\mathcal{E}|d_o) = \sum_{i=1}^{R} [\Pr(\mathcal{E}|i,d)\Pr(i|d) - \Pr(\mathcal{E}|i,d_o)\Pr(i|d_o)]$$

where $\Pr(\mathcal{E}|i,d)$ is the probability that d errors in classifying an observation in the i^{th} interval. The probability that an observation is in the i^{th} interval is Pr(i) and is independent of d. Thus,

$$\Pr(\mathcal{E}|d) - \Pr(\mathcal{E}|d_o) = \sum_{i=1}^{R} Q(i,d,d_o)$$

where

$$Q(i,d,d_o) = [\Pr(\mathcal{E}|i,d) - \Pr(\mathcal{E}|i,d_o)]\Pr(i)$$

The next section is devoted to obtaining a sufficient condition for the goal

$$\Pr[\Pr(\mathcal{E}|d) - \Pr(\mathcal{E}|d_o) \leq \alpha] \geq \beta$$

Then, computational techniques are developed for checking if this sufficient condition is satisfied for a given partition.

2.4 A Sufficient Condition

The following proposition gives a set of interval conditions (one for each interval in the partition) such that satisfaction of all of them is sufficient for Condition (1.6).

Proposition 1

Given:

a) Constants α and β such that

$$0 \leq \alpha \leq 1$$

$$0 \leq \beta \leq 1$$

b) Constants $\alpha(i) \geq 0$ and $\tau(i) \geq 0$, $i = 1, \ldots, R$, such that

$$\sum_{i=1}^{R} \alpha(i) = \alpha$$

$$\sum_{i=1}^{R} \tau(i) = 1 - \beta$$

Then the set of interval conditions

$$\Pr[Q(i,d,d_o) > \alpha(i)] < \tau(i) \qquad , i = 1, \ldots, R \qquad (2.10)$$

implies

$$\Pr[\Pr(e|d) - \Pr(e|d_o) \leq \alpha] \geq \beta$$

Proof

The set of conditions (2.10) implies that

$$\sum_{i=1}^{R} \Pr[Q(i,d,d_o) > \alpha(i)] < \sum_{i=1}^{R} \tau(i)$$

It follows that

$$\Pr\left[\bigcup_{i=1}^{R} \left(Q(i,d,d_o) > \alpha(i)\right)\right] < \sum_{i=1}^{R} \tau(i)$$

From de Morgan's laws

$$\Pr\left[\bigcup_{i=1}^{R} \left(Q(i,d,d_o) > \alpha(i)\right)\right] = \Pr\left[\left\{\bigcap_{i=1}^{R} \left(Q(i,d,d_o) \leq \alpha(i)\right)\right\}^{C}\right]$$

where the superscript "C" indicates complementation.

Then

$$\Pr\left[\bigcap_{i=1}^{R} \left(Q(i,d,d_o) \leq \alpha(i)\right)\right] \geq 1 - \sum_{i=1}^{R} \tau(i) = \beta$$

which implies that

$$\Pr\left[\sum_{i=1}^{R} Q(i,d,d_o) \leq \sum_{i=1}^{R} \alpha(i)\right] \geq \beta$$

The conclusion follows.

2.5 Classification Procedure for the i^{th} Interval

Consider just the i^{th} interval.

Define

$$U_j(i) = \frac{P_j \rho_j(i)}{W(i)}$$

The d.f. on $U_j(i)$ is

$$\beta^*\left(U_j(i) \mid \gamma_{j1}(i),\ \gamma_{j2}(i)\right) = \frac{W(i)}{P_j}\,\beta\left(\frac{W(i)U_j(i)}{P_j} \;\middle|\; \gamma_{j1}(i),\ \gamma_{j2}(i)\right) \tag{2.11}$$

Define

$$\mu_j(i) = EU_j(i) = \frac{P_j}{W(i)} \cdot \frac{\gamma_{j1}(i)}{\gamma_{j1}(i) + \gamma_{j2}(i)}$$

$$\sigma_j^2(i) = \text{Var } U_j(i) = \left(\frac{P_j}{W(i)}\right)^2 \cdot \frac{\gamma_{j1}(i)\ \gamma_{j2}(i)}{\left(\gamma_{j1}(i)+\gamma_{j2}(i)\right)^2\left(\gamma_{j1}(i)+\gamma_{j2}(i)+1\right)}$$

$a(i) = \text{Arg}\left[\max_{j=1,2} \mu_j(i)\right]$: corresponds to the class chosen by d in the i^{th} interval.

$b(i)$: corresponds to the class not chosen by d in the i^{th} interval

(2.12)

To avoid redundant notation a(i) and b(i) are denoted a and b when the i^{th} interval is understood.

The objective is to find a region V(i) in the $\left(U_a(i),\ U_b(i)\right)$ plane containing all points for which

$$Q(i,\ d,\ d_o) > \alpha(i)$$

is possible. Then the probability $Pr\left(V(i)\right)$ is an upper bound for the probability

$$Pr\left[Q(i,\ d,\ d_o) > \alpha(i)\right]$$

$Pr\left(V(i)\right)$ is obtained by integrating the d.f.* on $\left(U_a(i),\ U_b(i)\right)$ over points in V(i).

$$Pr\left[Q(i,\ d,\ d_o) > \alpha(i)\right] \leq Pr\left(V(i)\right)$$

$$= \int_{V(i)} \beta^*\left(U_a(i) \mid \gamma_{a1}(i), \gamma_{a2}(i)\right)\beta^*\left(U_b(i) \mid \gamma_{b1}(i), \gamma_{b2}(i)\right) dU_a(i) dU_b(i)$$

The i^{th} interval condition is satisfied if

$$Pr\left(V(i)\right) < \tau(i) \tag{2.13}$$

The region of integration V(i) is obtained as the intersection of two regions $\tilde{V}_1(i)$ and $\tilde{V}_2(i)$, each containing all points in the $\left(U_a(i),\ U_b(i)\right)$ plane for which $\left(Q(i,\ d,\ d_o) > \alpha(i)\right)$ is possible. Define the events

*The d.f. on $\left(U_a(i),\ U_b(i)\right)$ is the product of d.f.'s defined by (2.11) on $U_a(i)$ and $U_b(i)$ separately because those d.f.'s are obtained from independent samples.

$$V_1(i) = \left(Q(i, d, d_o) > 0\right)$$

$$V_2(i) = \left(Q(i, d, d_o) > \alpha(i)\right)$$

Because of the possible variation of the density functions f_1 and f_2 about their averages in the i^{th} interval, it is not possible in general to specify $V_1(i)$ and $V_2(i)$ as regions in the $\left(U_a(i), U_b(i)\right)$ plane. However, the following proposition gives regions $\tilde{V}_1(i)$ and $\tilde{V}_2(i)$ that contain $V_1(i)$ and $V_2(i)$ respectively. Note that* $V_2(i) \subset V_1(i)$ and thus $V_2(i) \subset \tilde{V}_1(i)$. Careful examination shows that $\tilde{V}_2(i)$ may contain points for which it is known that $Q(i, d, d_o) = 0$. Elimination of these points from the region of integration yields a smaller upper bound. Intersection of $\tilde{V}_2(i)$ with $\tilde{V}_1(i)$ to obtain $V(i)$ accomplishes the elimination of these points. Figure 4 illustrates with Venn diagrams the set relations involved.

The proof of the proposition uses an easily proved statement relating the range of variation of a density function f in an interval $\mathcal{J}$ to its average over the interval and an assumed Lipschitz condition.

Statement

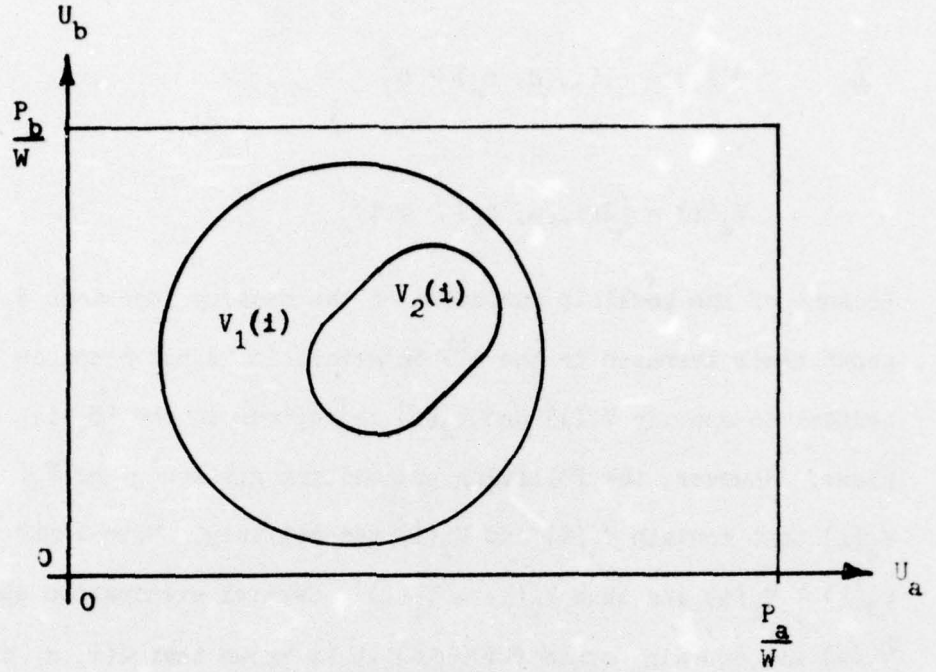
If 1) Interval $\mathcal{J}$ has width W

2) Density function f satisfies

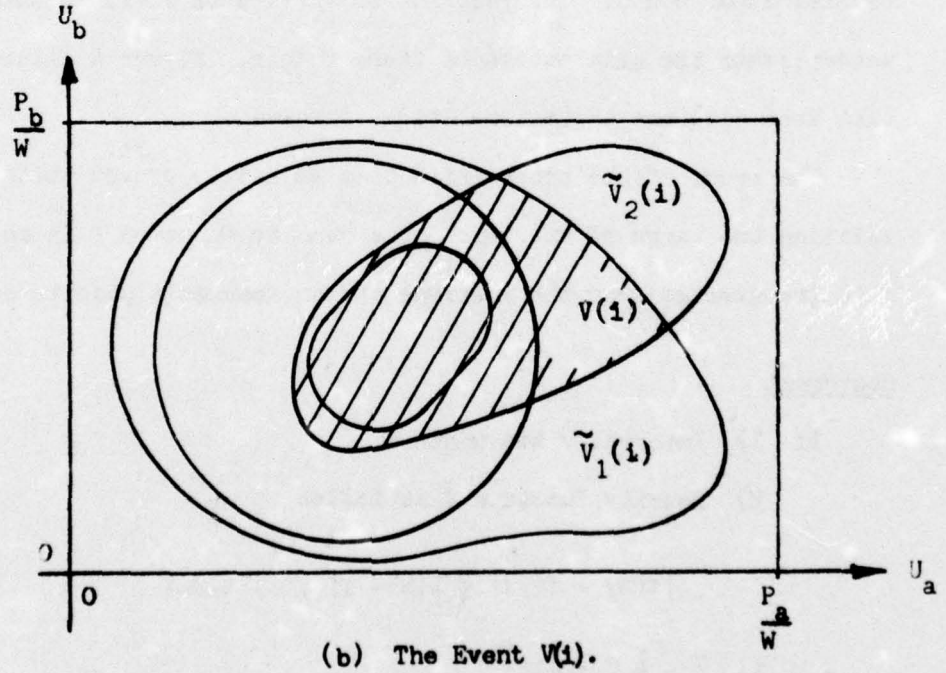
$$|f(x) - f(y)| \leq L|x - y| , \quad x,y \in \mathcal{J}$$

3) $\bar{f} = \frac{1}{W}\int_{\mathcal{J}} f(x)dx$

*Notation $V_2(i) \subset V_1(i)$ allows $V_2(i) = V_1(i)$.



(a) The Events $V_1(i)$ and $V_2(i)$.



(b) The Event V(i).

Figure 4. Set Relations Among $V_1(i)$, $V_2(i)$, $\tilde{V}_1(i)$, $\tilde{V}_2(i)$ and V(i)

then

$$\underset{x\in\mathcal{I}}{\text{Max}}\ f(x) \leq \bar{f} + \frac{LW}{2}$$

$$\underset{x\in\mathcal{I}}{\text{Min}}\ f(x) \geq \bar{f} - \frac{LW}{2} \qquad (2.14)$$

and if

$$\bar{f} < \frac{LW}{2}$$

then

$$\underset{x\in\mathcal{I}}{\text{Max}}\ f(x) \leq (2WL\bar{f})^{\frac{1}{2}}$$

$$\underset{x\in\mathcal{I}}{\text{Min}}\ f(x) \geq 0 \qquad (2.15)$$

Proposition 2*

Given the definitions

$$\overline{f_j} = \rho_j/W \quad , \quad j = 1,2$$

$$C_j = P_j L_j W/2 \quad , \quad j = 1,2$$

$$\rho = \alpha(i)/W$$

$$\delta = C_b \qquad \text{if } U_b \geq C_b$$

$$= [(4C_bU_b)^{\frac{1}{2}} - U_b] \qquad \text{if } U_b < C_b$$

*Because the i^{th} interval is understood, the "i" is dropped from the notation when confusion does not result.

Then

1) A region $\tilde{V}_1$ containing V_1 in the (U_a, U_b) plane is:

$$\tilde{V}_1 = (U_b > U_a - C_a - \delta) \tag{2.16}$$

and

2) A region $\tilde{V}_2$ containing V_2 in the (U_a, U_b) plane is:

$$\tilde{V}_2 = \left(U_b - \min_{j=1,2} [\max (0, U_j - C_j)] > \rho\right) \tag{2.17}$$

where it is understood that the definition of $\tilde{V}_1$ and $\tilde{V}_2$ includes intersection with

$$\left(0 \le U_a \le \frac{P_a}{W}\right) \cap \left(0 \le U_b \le \frac{P_b}{W}\right) .$$

Proof

Part 1

$$V_1 = \left(P_a f_a(x) < P_b f_b(x) \quad \text{for some } x \in \vartheta\right)$$

$$\subset \left(P_a \min_{x \in \vartheta} f_a(x) < P_b \max_{x \in \vartheta} f_b(x)\right)$$

From (2.14) and (2.15)

$$\min_{x \in \vartheta} f_a(x) \ge \bar{f}_a - \frac{C_a}{P_a}$$

$$\max_{x \in \vartheta} f_b(x) \le \bar{f}_b + \frac{\delta}{P_b}$$

Then

$$V_1 \subset (P_a \bar{f}_a - C_a < P_b \bar{f}_b + \delta) = (U_b > U_a - C_a - \delta) = \tilde{V}_1$$

Part 2

$$V_2 = \left(Q(i,\ d,\ d_o) > \alpha(i)\right)$$

$$= \left(\frac{Pr(i)}{W}\left[Pr(\mathcal{E}|i,d) - Pr(\mathcal{E}|i,d_o)\right] > \rho\right)$$

Note that

$$Pr(\mathcal{E}|i,\ d) = \frac{P_b \rho_b}{\sum_{j=1}^{2} P_j \rho_j} = \frac{WU_b}{Pr(i)}$$

$Pr(\mathcal{E}|i,d_o)$ can be expanded as

$$Pr(\mathcal{E}|i,d_o) = \int_{\mathcal{I}} Pr(\mathcal{E}|i,\ x,\ d_o) f(x|i,\ d_o)dx$$

$$= \int_{\mathcal{I}} \min_{j=1,2} \frac{[P_j f_j(x)]}{f(x)} \cdot f(x|i)dx$$

Note that

$$\min_{j=1,2} [P_j f_j(x)] \geq \min_{j=1,2} \left[\min_{x \in \mathcal{S}} P_j f_j(x)\right]$$

By (2.14) and (2.15)

$$\min_{x \in \mathcal{S}} [P_j f_j(x)] \geq \text{Max } (0,\ U_j - C_j)$$

such that

$$\Pr(\mathcal{E}|i,\ d_o) \geq \min_{j=1,2} \left[\text{Max } (0,\ U_j - C_j)\right] \int_{\mathcal{S}} \frac{f(x|i)}{f(x)}\, dx$$

$$= \frac{W}{\Pr(i)} \min_{j=1,2} \left[\text{Max } (0,\ U_j - C_j)\right]$$

Then

$$\frac{\Pr(i)}{W} [\Pr(\mathcal{E}|i,d) - \Pr(\mathcal{E}|i,d_o)] \leq U_b - \min_{j=1,2} [\text{Max } (0, U_j - C_j)]$$

and thus $V_2 \subset \tilde{V}_2$.

Figure 5 illustrates the region $\tilde{V}_1$ while Figure 6 illustrates two cases that result for $\tilde{V}_2$ depending on the relative sizes of ρ and C_b. The region V over which the density function on $(U_a,\ U_b)$ is to be integrated is obtained as the intersection of $\tilde{V}_1$ and $\tilde{V}_2$.

$$V = (U_b > U_a - C_a - \delta\ ) \cap \left(U_b - \min_{j=1,2}[\text{Max } (0,\ U_j - C_j)] > \rho\right)$$

(2.18)

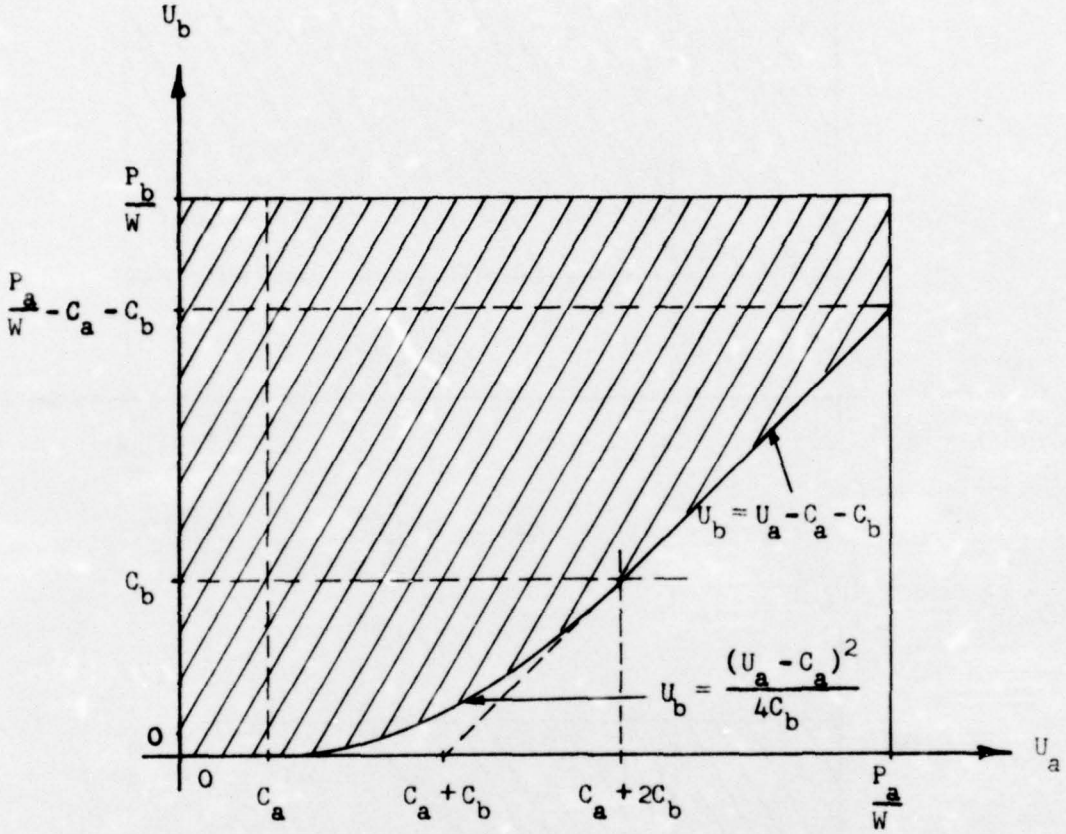
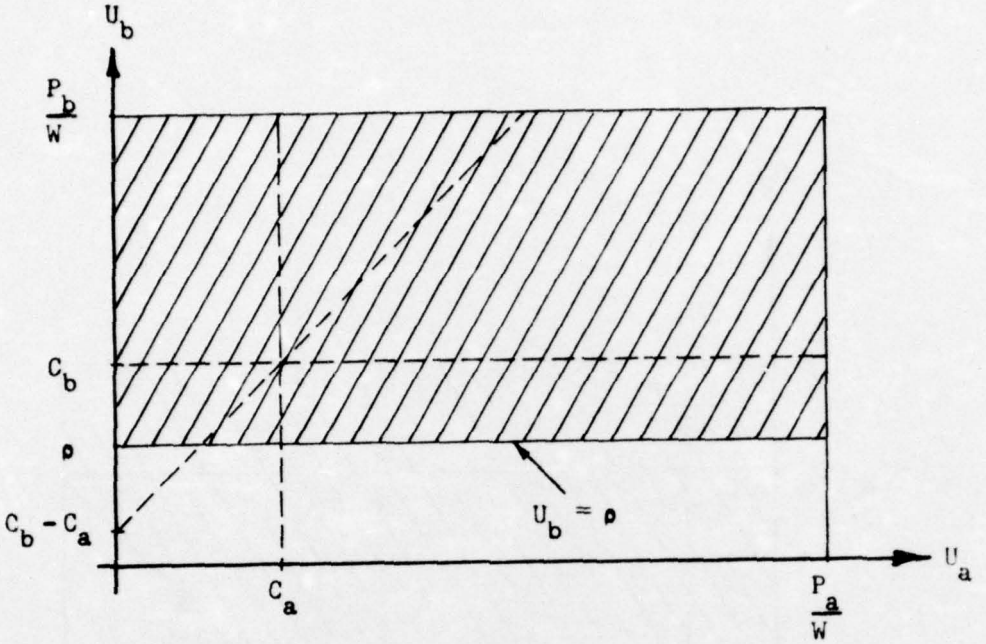
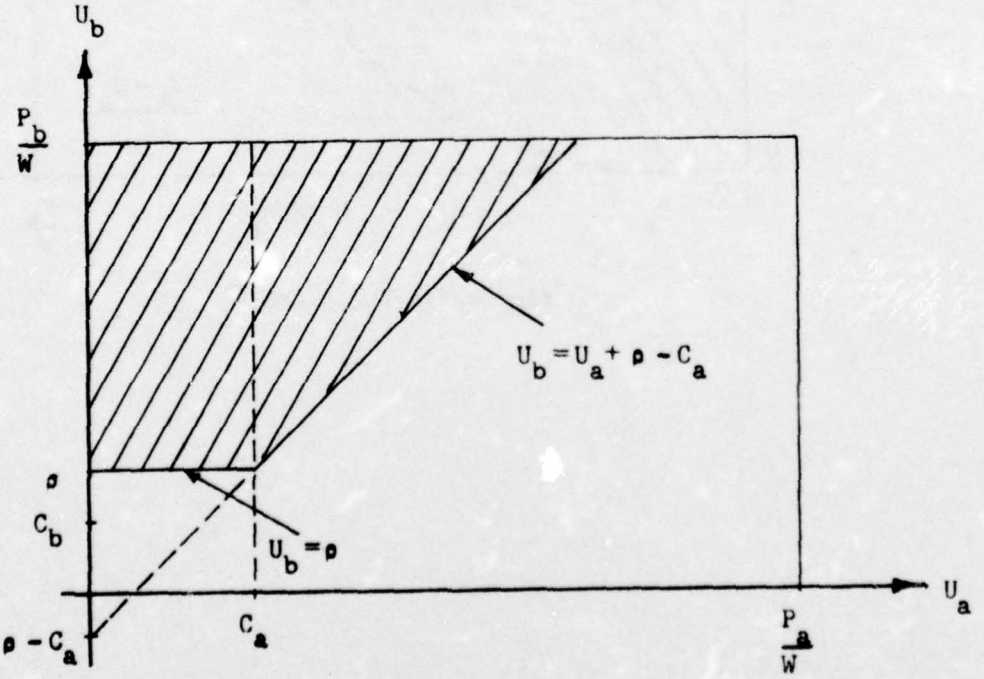


Figure 5. The Event $\tilde{V}_1$

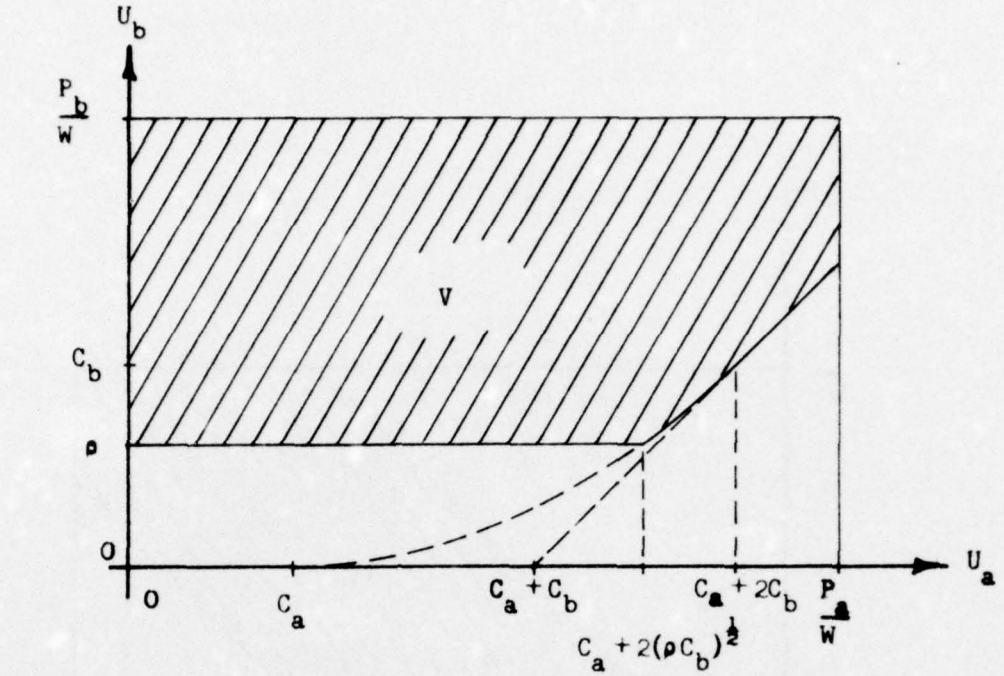


(a) $\rho < C_b$

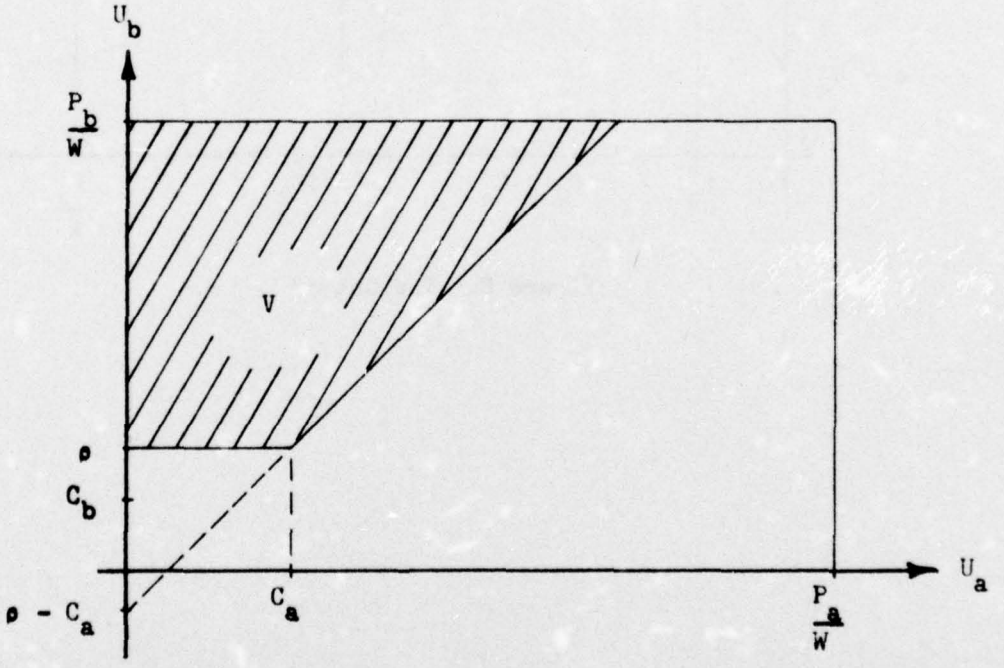


(b) $\rho \geq C_b$

Figure 6. The Event $\tilde{V}_2$



(a) $\rho < C_b$



(b) $\rho \geq C_b$

Figure 7. The Event V

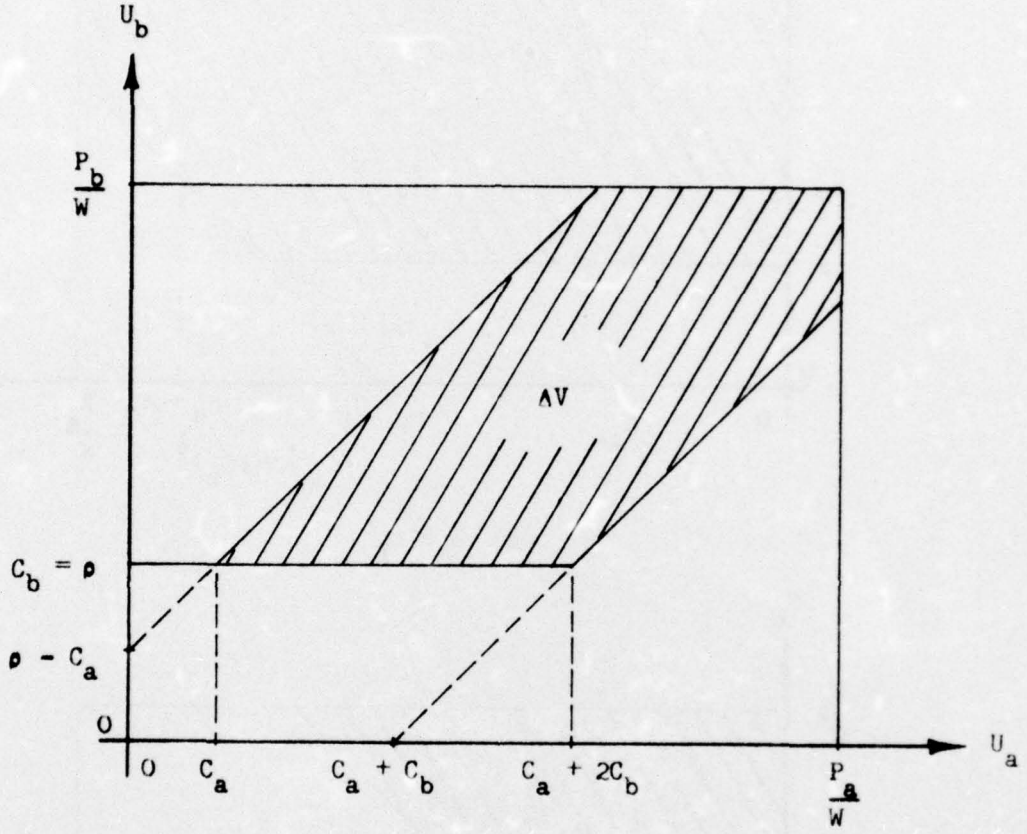


Figure 8. The Set ΔV

Figure 7 illustrates V for each of the two cases, $\rho < C_b$ and $\rho \geq C_b$. Note that if $\rho \geq C_b$, then $V = \bar{V}_2$. The transition from $\rho \geq C_b$ to $\rho < C_b$ is not smooth with respect to the region V. Figure 8 illustrates the increment ΔV included into V when $\rho - C_b$ is changed from slightly positive to slightly negative.

Note 1

Given

1) Scaled beta density functions $\beta^*\left(U_j | \gamma_{j1}, \gamma_{j2}\right)$ on U_j, $j = 1,2$, according to (2.11) where γ_{j1} and γ_{j2} are positive integers.
2) Definition $Be(\gamma_{j1}, \gamma_{j2}) = \dfrac{\Gamma(\gamma_{j1})\Gamma(\gamma_{j2})}{\Gamma(\gamma_{j1}+\gamma_{j2})}$

An upper bound for

$$Pr[Q(i,d,d_o) > \alpha(i)]$$

is obtained by integrating the d.f. $\beta^*(U_a|\gamma_{a1}, \gamma_{a2})\beta^*(U_b|\gamma_{b1}, \gamma_{b2})$ over the region V(i) and hence over any region containing V(i) in the (U_a, U_b) plane. By inspection of Figure 7 and by definition of β^*,

$$Pr\Big(V(i)\Big) \leq T(i)$$

$$= \int_{\rho}^{\frac{P_b}{W}} \frac{W}{P_b}\beta\left(\frac{WU_b}{P_b}\,|\gamma_{b1}, \gamma_{b2}\right) \int_0^{\text{Min}\left[U_b - q, \frac{P_a}{W}\right]} \frac{W}{P_a}\beta\left(\frac{WU_a}{P_a}\,\Big|\gamma_{a1}, \gamma_{a2}\right) dU_a\, dU_b$$

where

$$q = \rho - C_a \qquad \text{if } \rho \geq C_b$$

$$= -C_a - C_b \qquad \text{if } \rho < C_b$$

When the γ's are integers, Appendix B carries out the integration with the result:

$$T(i) = Be^{-1}(\gamma_{b1}, \gamma_{b2}) Be^{-1}(\gamma_{a1}, \gamma_{a2}) \sum_{j=0}^{\gamma_{a2}-1} \binom{\gamma_{a2}-1}{j} \frac{(-1)^j \left(\frac{P_b}{P_a}\right)^{\gamma_{a1}+j}}{(\gamma_{a1}+j)}$$

$$\cdot \sum_{v=0}^{\gamma_{a1}+j} \binom{\gamma_{a1}+j}{v} \left(-\frac{Wq}{P_b}\right)^{\gamma_{a1}+j-v} \sum_{k=0}^{\gamma_{b2}-1} \binom{\gamma_{b2}-1}{k} (-1)^k \cdot \frac{\left[\left(\frac{Wy_o}{P_b}\right)^{k+\gamma_{b1}+v} - \left(\frac{W\rho}{P_b}\right)^{k+\gamma_{b1}+v}\right]}{(k+\gamma_{b1}+v)}$$

$$+ 1 - Be^{-1}(\gamma_{b1}, \gamma_{b2}) \sum_{k=0}^{\gamma_{b2}-1} \binom{\gamma_{b2}-1}{k} (-1)^k \frac{\left(\frac{Wy_o}{P_b}\right)^{k+\gamma_{b1}}}{(k+\gamma_{b1})} \qquad (2.19)$$

where

$$y_o = \text{Min}\left[\text{Max}\left(\frac{P_a}{W} + q, \rho\right), \frac{P_b}{W}\right]$$

The computations for T(i) are time consuming and subject to accumulated error. A simplifying approximation is to approximate the d.f.'s on the U_j's with Gaussian d.f.'s having the same means and variances. This approximation is suggested by the fact that as its parameters get large while maintaining constant ratio, a beta d.f. converges pointwise to the Gaussian d.f. having the same mean and variance [36,37].

Note 2

Given

(1) Gaussian density functions

$$g(U_j|\mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi}\,\sigma_j}\, e^{-\frac{1}{2}\left(\frac{U_j-\mu_j}{\sigma_j}\right)^2}, \quad -\infty < x < \infty$$

on U_j, $j = 1,2$.

(2) The U_b intercept ξ_1 and the slope ξ_2 of a straight line

$U_b = \xi_1 + \xi_2 U_a$ supporting the region V(i) in the (U_a, U_b) plane.

An upper bound $\Lambda(i)$ for

$$\Pr\left[Q(i, d, d_o) \geq \alpha(i)\right]$$

is obtained by integrating the d.f.

$$g(U_a|\mu_a, \sigma_a^2)g(U_b|\mu_b, \sigma_b^2)$$

over the half-plane supported by

$$U_b = \xi_1 + \xi_2 U_a .$$

Thus

$$\Pr\big(V(i)\big) \leq \Lambda(i) = \iint_{U_b > \xi_1 + \xi_2 U_a} g(U_b|\mu_b,\sigma_b^2)g(U_a|\mu_a,\sigma_a^2)dU_a dU_b$$

Appendix C carries out the integration with the result:

$$\Lambda(i) = \Phi\left[\frac{-\xi_1 - \mu_a\xi_2 + \mu_b}{\left((\sigma_a\xi_2)^2 + \sigma_b^2\right)^{\frac{1}{2}}}\right] \tag{2.20}$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\, dy \qquad -\infty < x < \infty$$

Because of its simplicity, $\Lambda(i)$ of (2.20) is a more practical result than $T(i)$ of (2.19). In the remainder of this report $\Lambda(i)$ is used in place of $T(i)$, and the i^{th} interval condition is satisfied (approximately) if

$$\Lambda(i) < \tau(i) \tag{2.21}$$

The i^{th} interval is said to be classified if (2.21) holds; the whole domain is said to be classified if (2.21) holds for each interval. Because $\Lambda(i)$ is obtained as the integral over a region of integration that contains the one used for $T(i)$, the approximation $\Lambda(i)$ for $T(i)$ tends to be conservative. Appendix E contains comparisons of $\Lambda(i)$ and $T(i)$ for some special cases in which the regions of integration are identical. Good agreement is observed.

Λ is a function of the supporting line $U_b = \xi_1 + \xi_2 U_a$. The problem of minimizing Λ with respect to the parameters ξ_1 and ξ_2 is now considered. This minimization is subject to the constraint that the line $U_b = \xi_1 + \xi_2 U_a$ supports the region $V(i)$. If the mean (μ_a, μ_b) is in $V(i)$, Λ is given the value 1, and no minimization is attempted. In the following minimization, it is assumed that (μ_a, μ_b) is not in $V(i)$. Because Φ

is monotonically increasing in its argument, minimization of Λ is accomplished by minimizing the argument of Φ.

Case 1

$$\rho \geq C_b$$

From Figure 7b, it is clear that only lines through the point $(U_a, U_b) = (C_a, \rho)$ need be considered. For such lines the U_b intercept ξ_1 can be written in terms of the slope ξ_2 as

$$\xi_1 = \rho - C_a \xi_2$$

A straight-forward minimization of the argument of Φ in (2.20) with respect to ξ_2 subject to the constraint that ξ_2 is in the range $[0,1]$ leads to the value of $\Lambda(i)$ computed according to the flow diagram of Figure 9. The requirement ξ_2 in $[0,1]$ ensures that the line supports V(i).

Case 2

$$\rho < C_b$$

From Figure 7a, it is determined that $\Lambda(i)$ is minimized for a line through the point $(U_a, U_b) = (C_a + 2\sqrt{\rho C_b}\ , \rho)$ or for a line tangent to the quadratic portion of the boundary curve to V(i). Minimization of $\Lambda(i)$ with respect to lines through $(C_a + 2\sqrt{\rho C_b})\ , \rho)$ is accomplished similarly to the minimization for Case 1 except that ξ_1 is given in terms of ξ_2 by

$$\xi_1 = \rho - (C_a + 2\sqrt{\rho C_b})\xi_2$$

with ξ_2 constrained to the range $[0, \sqrt{\rho/C_b}]$. Minimization of $\Lambda(i)$ with

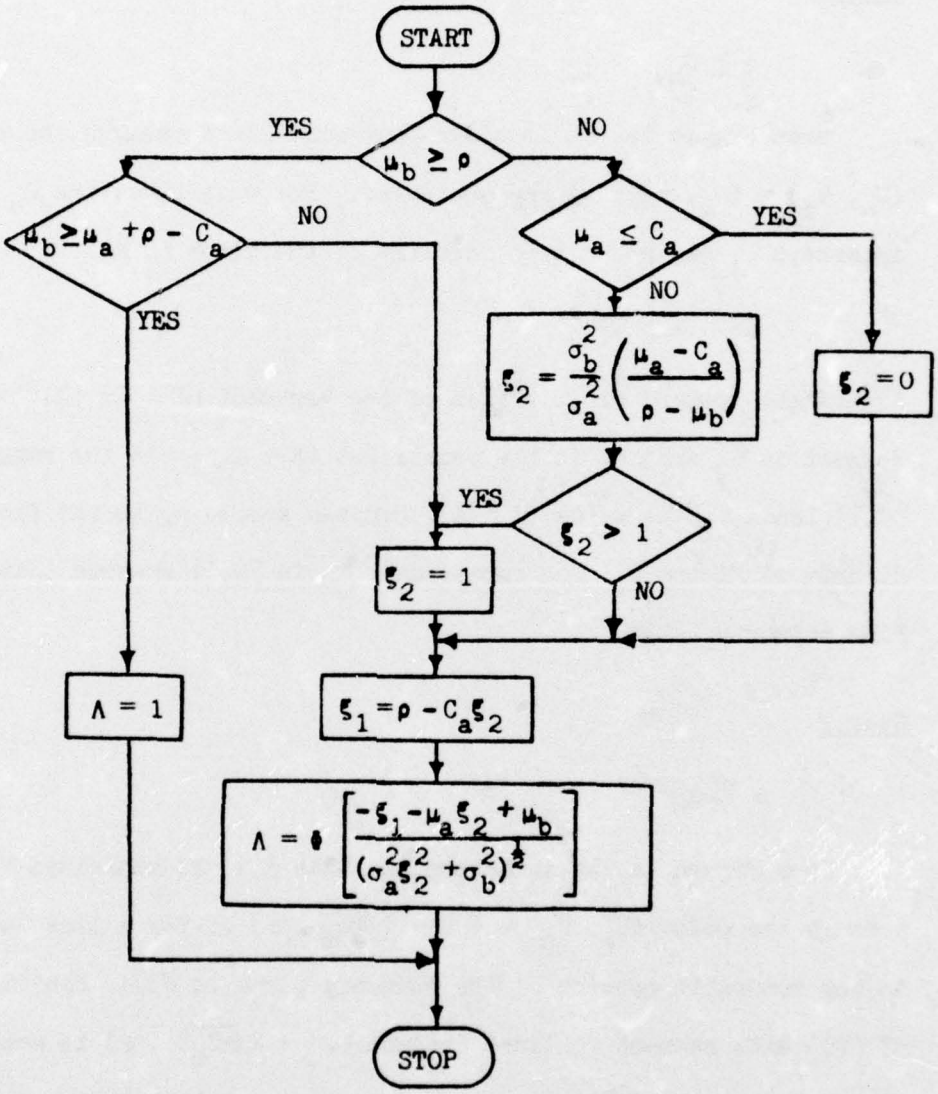


Figure 9. Flow Diagram, $\rho \geq C_b$.

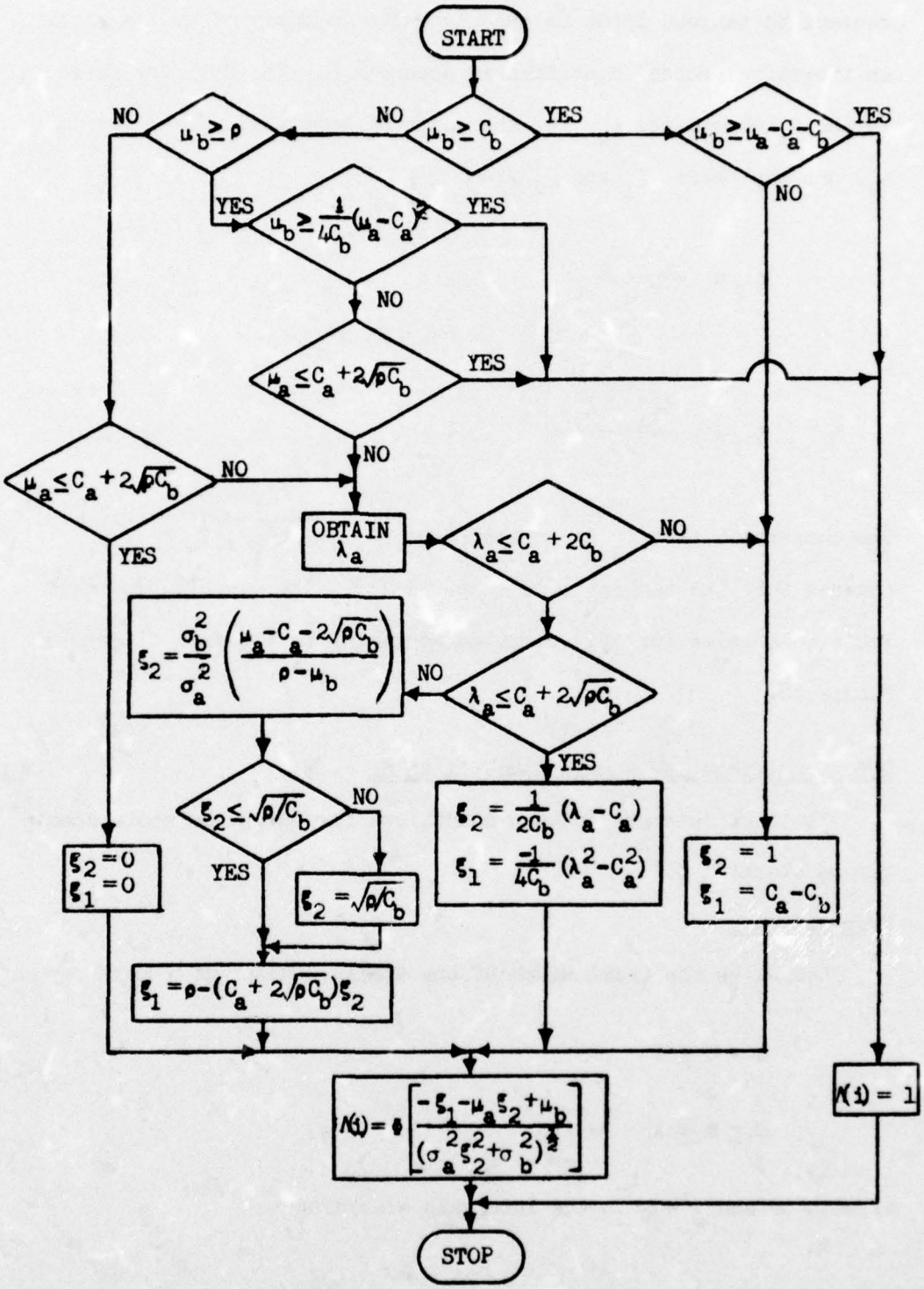


Figure 10. Flow Diagram, $\rho < C_b$

respect to tangent lines to the quadratic boundary of V(i) requires an iterative process described in Appendix D. The defining parameter is the U_a coordinate λ_a for the point of tangency. Given the value λ_a, one can obtain ξ_1 and ξ_2 from

$$\xi_2 = \frac{\lambda_a - C_a}{2C_b}$$

$$\xi_1 = -\frac{(\lambda_a^2 - C_a^2)}{4C_b}$$

The constraint that λ_a is in the range $[C_a + 2\sqrt{\rho C_b}\ ,\ C_a + 2C_b]$ ensures that the tangent line supports V(i). The overall procedure leads to a value for $\Lambda(i)$ computed according to the flow diagram of Figure 10.

2.6 Conditions for Domain Classification

It is of interest to know conditions for which the whole domain can be classified.

Proposition 3

Let W_T be the total width of the domain. Restrict α and β by

$$0 < \alpha \leq 1$$

$$0 \leq \beta < 1$$

Allocate α and $1 - \beta$ to the intervals according to

$$\alpha(i) = \frac{W(i)}{W_T}\,\alpha$$

$$\tau(i) = \frac{W(i)}{W_T}\,(1 - \beta)$$

and define ρ by

$$\rho = \frac{\alpha(i)}{W(i)} = \frac{\alpha}{W_T}$$

Let an R-interval partition of the domain be given with each interval width W(i), satisfying

$$0 < W(i) \leq W_{max}$$

where†

$$W_{max} < \text{Min}\left[\frac{2\rho}{\underset{j=1,2}{\text{Max}}\,(P_j L_j)}\ ,\ \frac{W_T}{2(1-\beta)}\ ,\ W_T\right]$$

Let n_j training observations from Class ω_j, $j = 1,2$, be used to form a decision rule d as discussed previously in this chapter.

Then:

$$\left(\frac{P_1^2}{n_1+2} + \frac{P_2^2}{n_2+2}\right)^{\frac{1}{2}} < \underset{i=1,\ldots,R}{\text{Min}} \left[\frac{2W(i)\left(\rho - \frac{W(i)}{2}\underset{j=1,2}{\text{Max}}\left(P_j L_j\right)\right)}{-\phi^{-1}\left(\frac{W(i)}{W_T}(1-\beta)\right)}\right] \tag{2.22}$$

† $W_{Max} < \dfrac{2\rho}{\underset{j=1,2}{\text{Max}}\,(P_j L_j)}$ implies $\rho > C_b$

implies

$$\Lambda(i) < \tau(i) \quad , \quad i = 1,2,\ldots,R$$

the requirement for classification of the domain.

Proof

Consider the i^{th} interval. By hypothesis, the case $\rho \geq C_b$ of the previous analysis applies. The event V(i) for that case is illustrated by the cross-hatched region of Figure 11. Each point (U_a, U_b) in the region defining the event V(i) satisfies

$$U_b > U_a + \rho - C_a \tag{2.23}$$

The line given by

$$U_b = U_a + \rho - C_a$$

supports the region V(i) and is one of those considered for the best such support in the computation of $\Lambda(i)$. If $\Lambda_1(i)$ is the integral of the approximating joint Gaussian d.f. over the half plane defined by (2.23), then

$$\Lambda(i) \leq \Lambda_1(i)$$

From (2.21), using $\xi_1 = \rho - C_a$ and $\xi_2 = 1$,

$$\Lambda_1(i) = \Phi\left[\frac{-\rho + C_a - \mu_a + \mu_b}{(\sigma_a^2 + \sigma_b^2)^{\frac{1}{2}}}\right]$$

Thus, satisfaction of

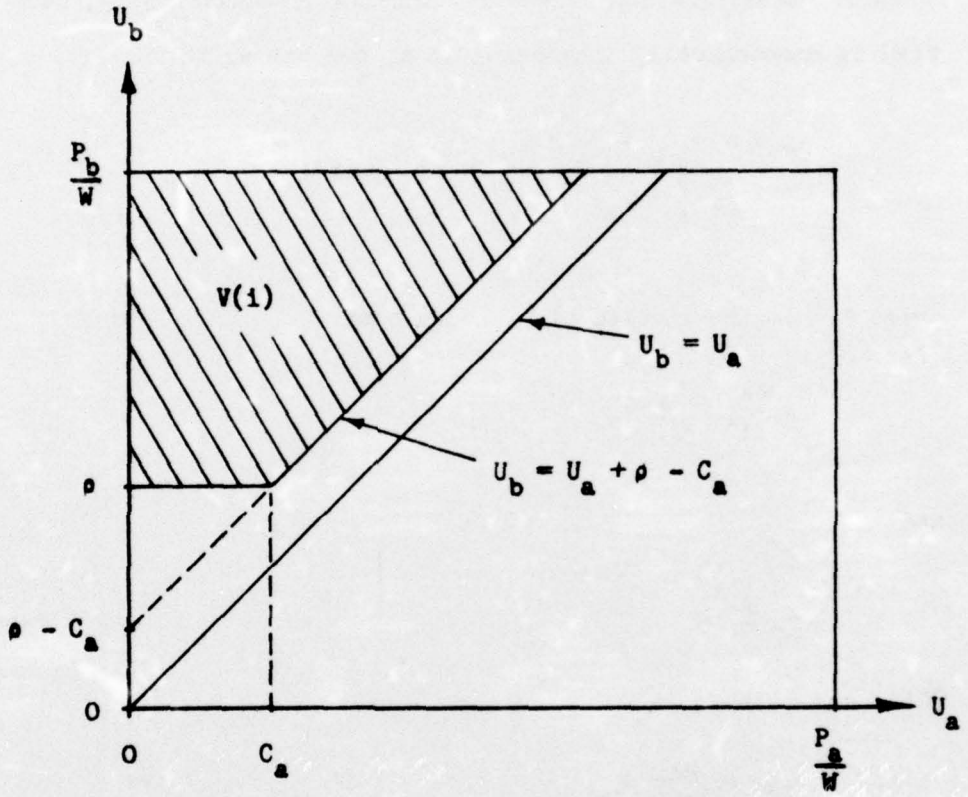


Figure 11. The Event V(i)

$$\Phi\left[\frac{-\rho + C_a - \mu_a + \mu_b}{(\sigma_a^2 + \sigma_b^2)^{\frac{1}{2}}}\right] < \tau(i)$$

assures classification of the i^{th} interval. Equivalently, because $\Phi(x)$ is monotonically increasing in x, one can write

$$\left[\frac{-\rho + C_a - \mu_a + \mu_b}{(\sigma_a^2 + \sigma_b^2)^{\frac{1}{2}}}\right] < \Phi^{-1}[\tau(i)] \tag{2.24}$$

where Φ^{-1} is the inverse of Φ. Note that

$$\underset{j=1,2}{\text{Max}}\ C_j \geq C_a$$

and

$$\mu_a \geq \mu_b$$

Then

$$\frac{-\rho + \underset{j=1,2}{\text{Max}}\ C_j}{(\sigma_a^2 + \sigma_b^2)^{\frac{1}{2}}} < \Phi^{-1}[\tau(i)]$$

implies (2.24). By hypothesis $\tau(i) < \frac{1}{2}$ so that $\Phi^{-1}[\tau(i)] < 0$. Rearranging gives (σ_a, σ_b is a permutation of σ_1, σ_2)

$$(\sigma_1^2 + \sigma_2^2)^{\frac{1}{2}} < \frac{-\rho + \underset{j=1,2}{\text{Max}}\ C_j}{\Phi^{-1}[\tau(i)]} \tag{2.25}$$

From (2.12)

$$\sigma_j^2 = \frac{\left(\frac{P_j}{W(i)}\right)^2 \left(\frac{Y_{j1}}{Y_{j1} + Y_{j2}}\right)\left(1 - \frac{Y_{j1}}{Y_{j1} + Y_{j2}}\right)}{(Y_{j1} + Y_{j2} + 1)}$$

But the product of two numbers that sum to 1 is bounded by $\frac{1}{4}$. Thus

$$\sigma_j^2 \leq \left(\frac{P_j}{2W(i)}\right)^2 \left(\frac{1}{Y_{j1} + Y_{j2} + 1}\right) = \left(\frac{P_j}{2W(i)}\right)^2 \frac{1}{(n_j + 2)}$$

with this bound on σ_j^2, the inequality

$$\frac{1}{2W(i)} \left(\frac{P_1^2}{n_1 + 2} + \frac{P_2^2}{n_2 + 2}\right)^{\frac{1}{2}} < \frac{(-\rho + \underset{j=1,2}{\text{Max}}\ C_j)}{\phi^{-1}[\tau(i)]}$$

implies (2.25). Appropriate substitutions give

$$\left(\frac{P_1^2}{n_1 + 2} + \frac{P_2^2}{n_2 + 2}\right)^{\frac{1}{2}} < \frac{2W(i)\left(\rho - \frac{W(i)}{2} \underset{j=1,2}{\text{Max}}\ (P_j L_j)\right)}{-\phi^{-1}\left[\frac{W(i)}{W_T}(1 - \beta)\right]} \tag{2.26}$$

Thus (2.22) implies that

$$\Lambda(i) < \tau(i) \quad , \quad i = 1, \ldots, R$$

It is interesting that one can specify - before taking any training observations - a satisfactory partition and the number of training observations that assure classification of the whole domain. Consider, for example, the special case in which:

$$P_1 = P_2 = \tfrac{1}{2}$$

$$L_1 = L_2 = L$$

$$W_T = 1$$

$$\alpha = 0.1$$

$$\beta = 0.9$$

$$W(i) = W, \quad i = 1, \ldots, R$$

$$n_1 = n_2$$

Then (2.22) simplifies to

$$n_j > 2\left[\frac{\Phi^{-1}[0.1W]}{W(0.4-WL)}\right]^2 - 2 \quad , \quad j = 1,2$$

where W must satisfy $0 < W < \frac{0.4}{L}$. The smallest n_j that satisfies this inequality is plotted in Figure 12 as a function of W for each of several L values.

Several observations concerning Proposition 3 can be made.

1) The numbers n_1 and n_2 required for satisfaction of (2.22) are generally very large. This is to be expected because the proposition states a result that does not use the values Y_{j1}, Y_{j2}. Regardless of these values the result is applicable. Suppose that training observations and hence Y_{j1}, Y_{j2}, are available for the i^{th} interval; hence μ_a and μ_b can be determined for the interval.

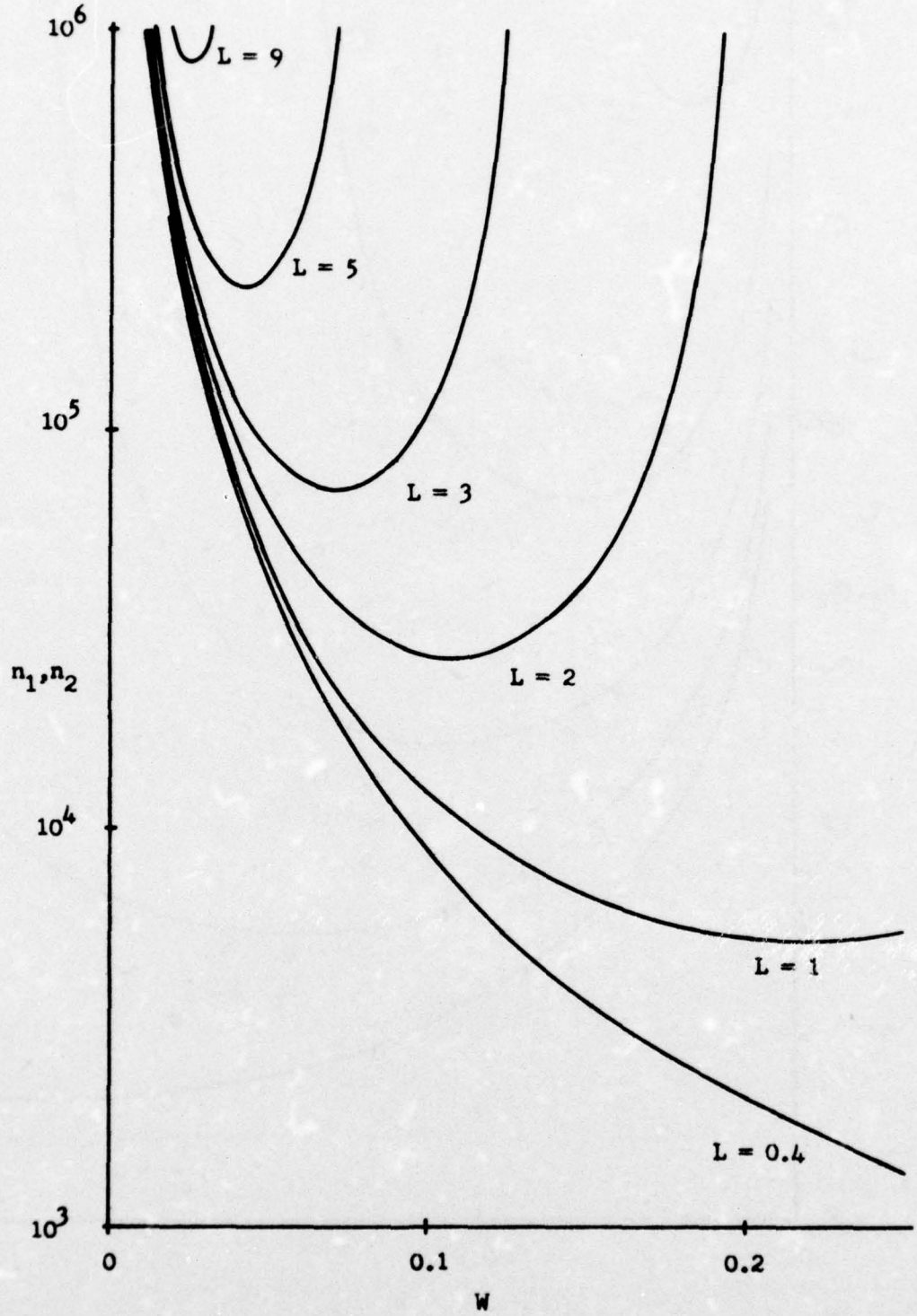


Figure 12. Training Observations Required Versus Interval Width

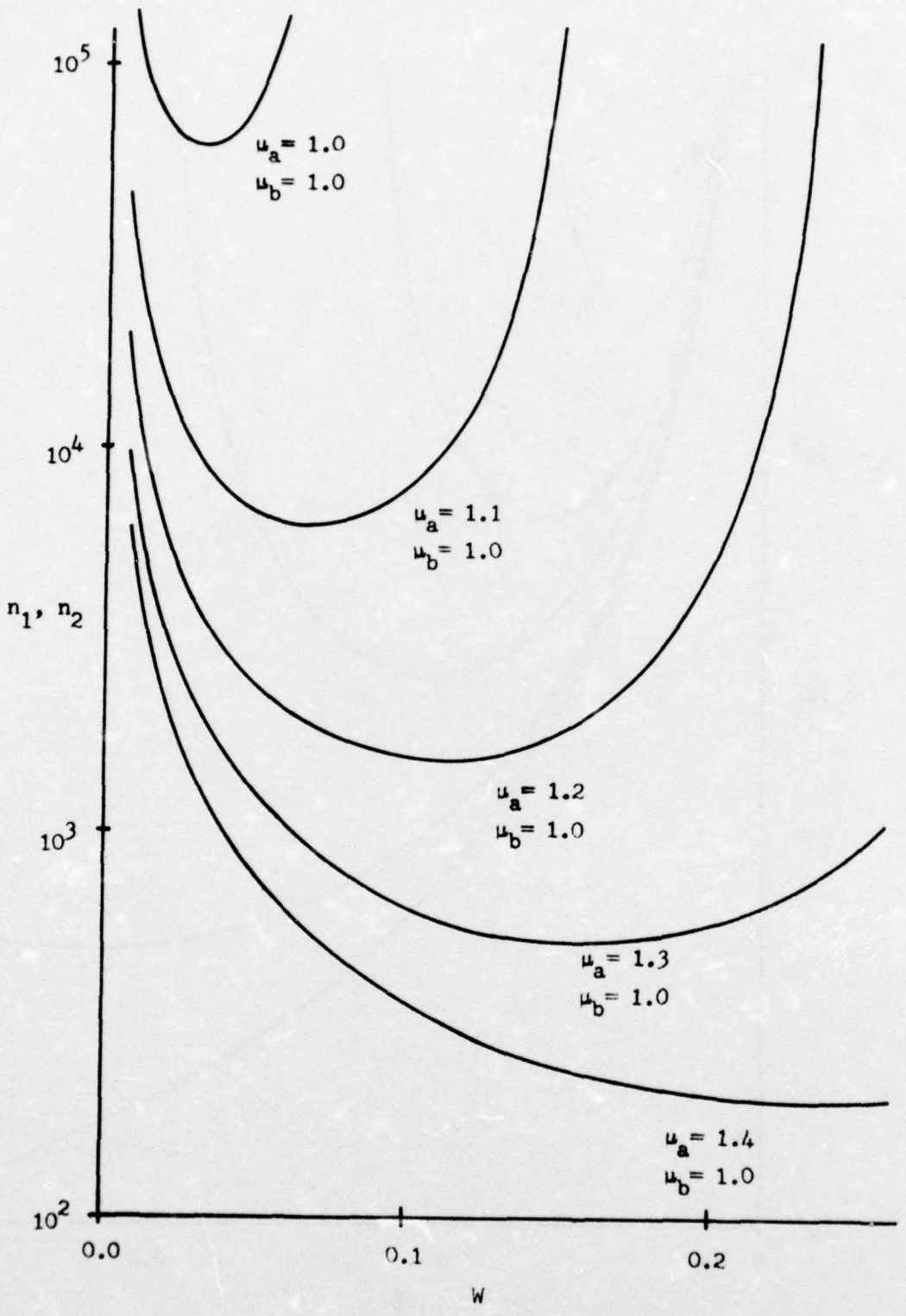


Figure 13. n_1, n_2 Versus W, (μ_a, μ_b Known, L_j= 5)

Suppose further that μ_a and μ_b remain constant as n_1, n_2, and W are varied. If $P_1 = P_2 = \frac{1}{2}$, $L_1 = L_2 = 5$, $\alpha = 0.1$, $\beta = 0.9$, and $n_1 = n_2$, then the *minimum* n_j required to classify the interval is plotted in Figure 13 for several μ_a, μ_b values. Note that n_j is much smaller when the parameters γ_{j1}, γ_{j2} can be used. The next chapter assumes μ_a and μ_b are constant over n_1, n_2, and W, so that estimates of the number of training observations required to classify the interval can be obtained. Adjustment in interval width is made based on these estimates.

2) Maximization of the right side of (2.22) with respect to the interval widths allows widths to be chosen that correspond to the smallest values n_1 and n_2 that satisfy (2.22) . Such "best" interval widths correspond to a "best" number of intervals. Hughes [26], using a mean recognition accuracy criterion, also arrives at a "best" number of intervals.

3) By requiring the interval widths to be less than or equal to W_{Max}, a minimum is placed on the number of intervals. It is possible that this minimum conflicts with the assumed storage constraint.

Also of interest is the rate at which the quantity

$$\Phi\left[\frac{-\beta + C_a - \mu_a + \mu_b}{(\sigma_a^2 + \sigma_b^2)^{\frac{1}{2}}}\right] = \Phi(Q)$$

changes with n, where

$$Q = \frac{-\rho + C_a - \mu_a + \mu_b}{(\sigma_a^2 + \sigma_b^2)^{\frac{1}{2}}}$$

$$= \frac{-\rho + C_a - \mu_a + \mu_b}{\left[\frac{\mu_a\left(\frac{P_a}{W} - \mu_a\right)}{n_a + 2} + \frac{\mu_b\left(\frac{P_b}{W} - \mu_b\right)}{n_b + 2}\right]^{\frac{1}{2}}}$$

Assuming that μ_a, μ_b are constant with n, that

$$n_a = P_a\, n \gg 2$$

$$n_b = P_b\, n \gg 2$$

and that

$$\frac{P_a}{W} \gg \mu_a$$

$$\frac{P_b}{W} \gg \mu_b$$

the rate of change in $\Phi(Q)$ is given by

$$\text{Rate} = \frac{\partial\, \Phi(Q)}{\partial\, n} = \frac{1}{2\sqrt{2\pi}}\; C\; \frac{1}{\sqrt{n}}\; e^{-\frac{1}{2}C^2 n}$$

where C is a constant given by

$$C = \left(\frac{W}{\mu_a + \mu_b}\right)^{\frac{1}{2}} \left(-\rho + C_a - \mu_a + \mu_b\right)$$

This chapter has discussed the classification procedure for a fixed R-interval partition of the domain. In Chapter III an ad hoc approach for adjusting the partition is described. The objective is to classify the whole domain with as few training observations as possible.

CHAPTER III

ALTERING THE PARTITION

3.1 Introduction

The classification procedure of Chapter II operates with a given partition. Generally, the partition can be adjusted to decrease the number of training observations required for classification of the whole domain. A desirable adjustment procedure would be one that minimized this number.

A hill climbing technique could be used to minimize an estimate of the number of training observations required for domain classification. Similarly, hill climbing techniques could be used to maximize an estimate of the divergence [52], an estimate of the information contained in an observation about its unknown class [39], or any other global measure of the separation of density functions. The hill climbing technique using the first criterion mentioned generally requires many intervals, while using the other criteria, it has not been shown to achieve satisfaction of condition 1.6.

This chapter describes an ad hoc partition adjustment procedure that operates by sequentially adjusting the widths of unclassified intervals in the order that they appear in a table called the priority table. The width of an interval under

consideration is adjusted to increase an estimate $\hat{r}$ of the interval's "classification rate";

$$\hat{r} = \frac{\hat{p}}{\hat{n}}$$

where $\hat{p}$ is an estimate of the mixture probability in the interval ($\hat{p} = P_1 \hat{\rho}_1 + P_2 \hat{\rho}_2$) and $\hat{n}$ is an estimate of the number of training observations required to classify the interval†. The estimated rate $\hat{r}$ is a reasonable performance measure in that it increases with $\hat{p}$ and decreases with $\hat{n}$. A possible disadvantage is that it is local (applies to one interval) as opposed to being global (applies to all intervals); i.e. partition adjustment using a global measure may result in a smaller estimated number of training observations required for classification. Application of a global technique would need to constrain the partition so that it allows classification of the intervals.

With suitable approximations (to be listed) $\hat{p}$, $\hat{n}$, and thus $\hat{r}$ can be written as functions of the interval width W' (in this chapter notation with a prime refers to variable quantities, whereas unprimed notation refers to observed quantities). $\hat{r}(W')$ can be maximized with respect to W'; the resulting maximum is denoted $\hat{r}_M$, and the interval width giving this maximum is denoted $\hat{W}_M$. The intervals are listed in the priority table in order of decreasing $\hat{r}_M$ values.

---

†The notation omits reference to a particular interval.

The following items constrain the adjustment procedure:

1) A change in the size of an interval influences the sizes and or number of other intervals. A change of an interval's size is not allowed if it affects an interval preceding it in the priority table.

2) Rules are stated that determine if an interval adjustment is made. An interval is adjusted either by splitting it into two intervals or by combining it with an adjacent interval. These types of adjustments allow for a reasonable amount of change at each adjustment stage and for larger changes over several adjustment stages.

3) No more than R intervals are allowed in the partition at any one time.

4) After a partition adjustment, the beta distributions† on the P's are reinitialized.

3.2 Rate Estimates

The estimate $\hat{r}(W')$ is obtained by first obtaining $\hat{p}(W')$ and $\hat{r}(W')$. The following simplifying approximations are useful. The quality of these approximations affects only the partition adjustment procedure and not classification based on a given partition.

---

†Note that even though Gaussian approximations are used for computations, all updating and reinitialization is done with beta distributions.

Approximations

1)† $\gamma'_{j1} \cong \gamma_{j1}\left(\frac{W'}{W}\right)\left(\frac{n'_j}{n_j}\right)$

Then

$$\mu'_j = \frac{P_j}{W'} \cdot \frac{\gamma'_{j1}}{(n'_j + 1)} = \frac{P_j}{W} \cdot \frac{\gamma_{j1}}{(n'_j + 1)} \cdot \left(\frac{n'_j}{n_j}\right)$$

$$\cong \frac{P_j}{W} \cdot \left(\frac{\gamma_{j1}}{n_j + 1}\right) = \mu_j$$

and

$$\sigma_j^{2'} = \frac{\mu'_j\left(\frac{P_j}{W'} - \mu'_j\right)}{(n'_j + 2)} \cong \frac{\mu_j\left(\frac{P_j}{W'} - \mu_j\right)}{n'_j + 2}$$

2)†† $n'_j \cong P_j\, n'$

3)††† $\left(\frac{P_j}{W'} - \mu_j\right) \cong \frac{P_j}{W'}$

Then

$$\sigma_j^{2'} \cong \frac{\mu_j}{W'}\frac{P_j}{n_j} \cong \frac{\mu_j}{W' n'}$$

---

† γ_{j1} is the number of training observations in the interval from the the j^{th} class. It is assumed that $n_j + 1 \cong n_j$.

†† The number of training observations from the 2 classes are assumed proportional to the a priori class probabilities.

††† Number of training observations in the interval is assumed small compared with the total number from the j^{th} class. Also subsequently assumed is $n'_j + 2 \cong n'_j$.

With these approximations, $\hat{p}(W')$ is easily computed as

$$\hat{p}(W') = W'(\mu_a + \mu_b) \tag{3.1}$$

Let α and $(1 - \beta)$ be allocated to the intervals according to

$$\alpha(i) = \frac{W'(i)}{W_T}\,\alpha$$

$$\tau(i) = \frac{W'(i)}{W_T}\,(1 - \beta) \tag{3.2}$$

Note that

$$\rho = \frac{\alpha(i)}{W'(i)} = \frac{\alpha}{W_T}$$

is constant with $W'(i)$. Computation of $\hat{r}(W'(i))$ proceeds for the i^{th} interval by using the above approximations in the i^{th} interval condition

$$\Phi\left[\frac{-\xi_1 - \mu_a\,\xi_2 + \mu_b}{((\sigma_a\,\xi_2)^2 + \sigma_b^2)^{\frac{1}{2}}}\right] < \tau \tag{3.3}$$

for suitable values of ξ_1 and ξ_2 (i has been dropped from the notation).

Taking the inverse gives

$$\frac{-\xi_1 - \mu_a\,\xi_2 + \mu_b}{((\sigma_a\,\xi_2)^2 + \sigma_b^2)^{\frac{1}{2}}} < \Phi^{-1}(\tau) \tag{3.4}$$

Using the variable quantities and the above approximations leads to

$$\frac{-\xi_1 - \mu_a \xi_2 + \mu_b}{\left(\frac{1}{n'W'}\right)^{\frac{1}{2}} \left(\mu_a \xi_2^2 + \mu_b\right)^{\frac{1}{2}}} < \Phi^{-1}\left(\frac{W'}{W_T}(1-\beta)\right)$$

or assuming that[†]

$$\frac{W'}{W_T}(1-\beta) < \tfrac{1}{2}$$

and[††]

$$\xi_1 + \mu_a \xi_2 - \mu_b > 0$$

The inequality becomes

$$n' > \left(\frac{\mu_a \xi_2^2 + \mu_b}{W'}\right)\left[\frac{-\Phi^{-1}\left(\frac{W'}{W_T}(1-\beta)\right)}{\xi_1 + \mu_a \xi_2 - \mu_b}\right]^2$$

The estimate $\hat{n}(W')$ is taken as

$$\hat{n}(W', \xi_1, \xi_2) = \left(\frac{\mu_a \xi_2^2 + \mu_b}{W'}\right)\left[\frac{-\Phi^{-1}\left(\frac{W'}{W_T}(1-\beta)\right)}{\xi_1 + \mu_a \xi_2 - \mu_b}\right]^2 \qquad (3.5)$$

[†]This is a constraint on W'.

[††]Assumes (μ_a, μ_b) is below the supporting line for V(i) in the (U_a, U_b) plane.

where ξ_1, ξ_2 have been included in the notation to indicate dependence on these quantities. From (3.1) and (3.5)

$$\hat{r}(W', \xi_1, \xi_2) = \frac{(\mu_a + \mu_b)}{(\mu_a \xi_2^2 + \mu_b)} \left[\frac{W'(\xi_1 + \mu_a \xi_2 - \mu_b)}{-\phi^{-1}\left(\frac{W'}{W_T}(1 - \beta)\right)} \right]^2$$

$$\text{if} \begin{cases} \frac{W'}{W_T}(1 - \beta) < \frac{1}{2} \\ \xi_1 + \mu_a \xi_2 - \mu_b > 0 \end{cases} \tag{3.6}$$

To obtain $\hat{r}_M$, the quantity $\hat{r}(W', \xi_1, \xi_2)$ should be maximized over all values ξ_1, ξ_2 for a line $U_b = \xi_1 + \xi_2 U_a$ that supports the region V(i) and over all interval widths W'. For simplification $\hat{r}(W', \xi_1, \xi_2)$ is maximized over W' for each of three sets of ξ_1, ξ_2 values, and then the maximum of these is chosen for $\hat{r}_M$. It is now assumed that

$$\frac{W'}{W_T}(1 - \beta) << \frac{1}{2}$$

Then for maximization of $\hat{r}(W', \xi_1, \xi_2)$ with respect to W', the quantity $-\phi^{-1}\left(\frac{W'}{W_T}(1 - \beta)\right)$ is considered to be approximately constant.

<u>Case 1</u> $(\xi_1, \xi_2) = (\rho, 0)$, $\rho > \mu_b$

$$\hat{r}(W', \rho, 0) = \left(\frac{\mu_a + \mu_b}{\mu_b}\right) \left[\frac{W'(\rho - \mu_b)}{-\phi^{-1}\left(\frac{W'}{W_T}(1 - \beta)\right)} \right]^2$$

Thus $\hat{r}(W',\rho,0)$ is maximum for W' as large as possible. To avoid problems with poor approximation accuracy for large W', the value $\hat{W}_M(\rho,0)$ is defined as the current width[†]

$$\hat{W}_M(\rho,0) = W$$

and

$\hat{r}_M(\rho,0)$ is defined by

$$\hat{r}_M(\rho,0) = \left(\frac{\mu_a + \mu_b}{\mu_b}\right)\left[\frac{W(\rho - \mu_b)}{\Phi^{-1}\left(\frac{W}{W_T}(1-\beta)\right)}\right]^2$$

The rules presented shortly for adjusting intervals encourage combining an interval with another when $\hat{W}_M \geq W$ which is true in this case.

Case 2 $(\xi_1,\xi_2) = (\rho - C_a',1)$, $\rho \geq C_b'$

or

$$(\xi_1,\xi_2) = \left(\rho - \frac{P_a L_a W'}{2},1\right),\ W' \leq \frac{2\rho}{P_b L_b}$$

$$\hat{r}(W',\rho - C_a,1) = \left[\frac{W'\left[W' - 2\left(\frac{\rho + \mu_a - \mu_b}{P_a L_a}\right)\right]\frac{P_a L_a}{2}}{\Phi^{-1}\left(\frac{W'}{W_T}(1-\beta)\right)}\right]^2$$

[†]This is ad hoc and another value for $\hat{W}_M(\rho,0)$ could be used.

$$\hat{W}_M(\rho - C_a,1) = \frac{\rho + \mu_a - \mu_b}{P_a L_a} \text{ if } \frac{\rho + \mu_a - \mu_b}{P_a L_a} < \frac{2\rho}{P_b L_b}$$

$$= \frac{2\rho}{P_b L_b} \text{ if } \frac{2\rho}{P_b L_b} \leq \frac{\rho + \mu_a - \mu_b}{P_a L_a}$$

$\hat{r}_M(\rho - C_a,1)$ is obtained by substituting $\hat{W}_M(\rho - C_a,1)$ for W' in $\hat{r}(W',\rho - C_a,1)$.

Case 3 $(\xi_1,\xi_2) = (- C_a' - C_b',1)$, $\rho < C_b'$

or

$$(\xi_1,\xi_2) = \left(- \frac{W'}{2} (P_a L_a + P_b L_b),1\right), \; W' > \frac{2\rho}{P_b L_b}$$

$$\hat{r}(W', - C_a - C_b,1) = \left[\frac{W'\left[W' - 2\left(\frac{\mu_a - \mu_b}{P_a L_a + P_b L_b}\right)\right]\left(\frac{P_a L_a + P_b L_b}{2}\right)}{\phi^{-1}\left(\frac{W'}{W_T}(1 - \beta)\right)}\right]^2$$

$$\hat{W}_M(- C_a - C_b,1) = \frac{\mu_a - \mu_b}{P_a L_a + P_b L_b} \text{ if } \frac{2\rho}{P_b L_b} < \frac{\mu_a - \mu_b}{P_a L_a + P_b L_b}$$

$\hat{r}_M(- C_a - C_b,1)$ is obtained by substituting $\hat{W}_M(- C_a - C_b,1)$ for W' in $\hat{r}(W',- C_a - C_b,1)$. The rate for Case 2 is at least as large as that for Case 3 when $\frac{2\rho}{P_b L_b} \geq \frac{\mu_a - \mu_b}{P_a L_a + P_b L_b}$, thus obviating the need to consider Case 3 in that event. $\hat{r}_M$ is set to the maximum of $\hat{r}_M(\rho,0)$, $\hat{r}_M(\rho - C_a,1)$, and $\hat{r}_M(- C_a - C_b,1)$ and $\hat{W}_M$ is the corresponding interval width. Similar computations for each of the intervals and a subsequent ranking according to $\hat{r}_M$ values gives the priority table.

3.3 Interval Operations

The unclassified intervals are considered sequentially in the order that they appear in the priority table. Suppose that attention is centered on the i^{th} interval of the priority table. Operations assumed available to change its size are:

1) Do not alter the interval.

2) Split the interval into two equal intervals.

3) Combine the interval with an adjacent interval.

If $\hat{W}_M \geq W$, an attempt is made to combine the interval with an adjacent interval; if $W \geq v\hat{W}_M$, an attempt is made to split the interval into two equal intervals; otherwise, no interval change is attempted. The constant v was arbitrarily chosen as 1.6 (values of v closer to one can cause too frequent interval changing). Without a constant $v > 1$, an interval might never be classified because of alternate splitting and combining from one set of training observations to the next.

Combining

The following is a list of conditions that must be satisfied before the i^{th} interval is combined with an adjacent interval.

a) The adjacent interval is unclassified and appears in a lower position of the priority table than the i^{th} interval.

b) $\hat{W}_M \geq W$ for the adjacent interval as well as for the i^{th} interval.

c) The adjacent interval is tentatively classified to the same class as the i^{th} interval; that is, "a" and "b" in μ_a and μ_b for the adjacent interval are the same as those for the i^{th} interval.

d) The estimated "classification rate" for the combined interval is greater than the sum of the rates for the component intervals considered separately. Before the combined interval rate can be obtained, parameters characterizing the interval are computed by a procedure discussed in the next section.

e) If both intervals adjacent to the i^{th} interval satisfy these conditions, then combining is performed with the adjacent interval giving the largest improvement in classification rate.

If combining takes place, then $\hat{W}_M$ for the combined interval is taken as the average of the $\hat{W}_M$ values for its component intervals. The combined interval takes the place of the i^{th} interval in the priority table and is processed again in exactly the same fashion. The adjacent interval that is combined is removed from the priority table.

Splitting

In order for the i^{th} interval to be split, the total number of intervals must be less than R. If not, then a search is made for an adjacent pair of intervals that can first be combined. The search is made in the following order.

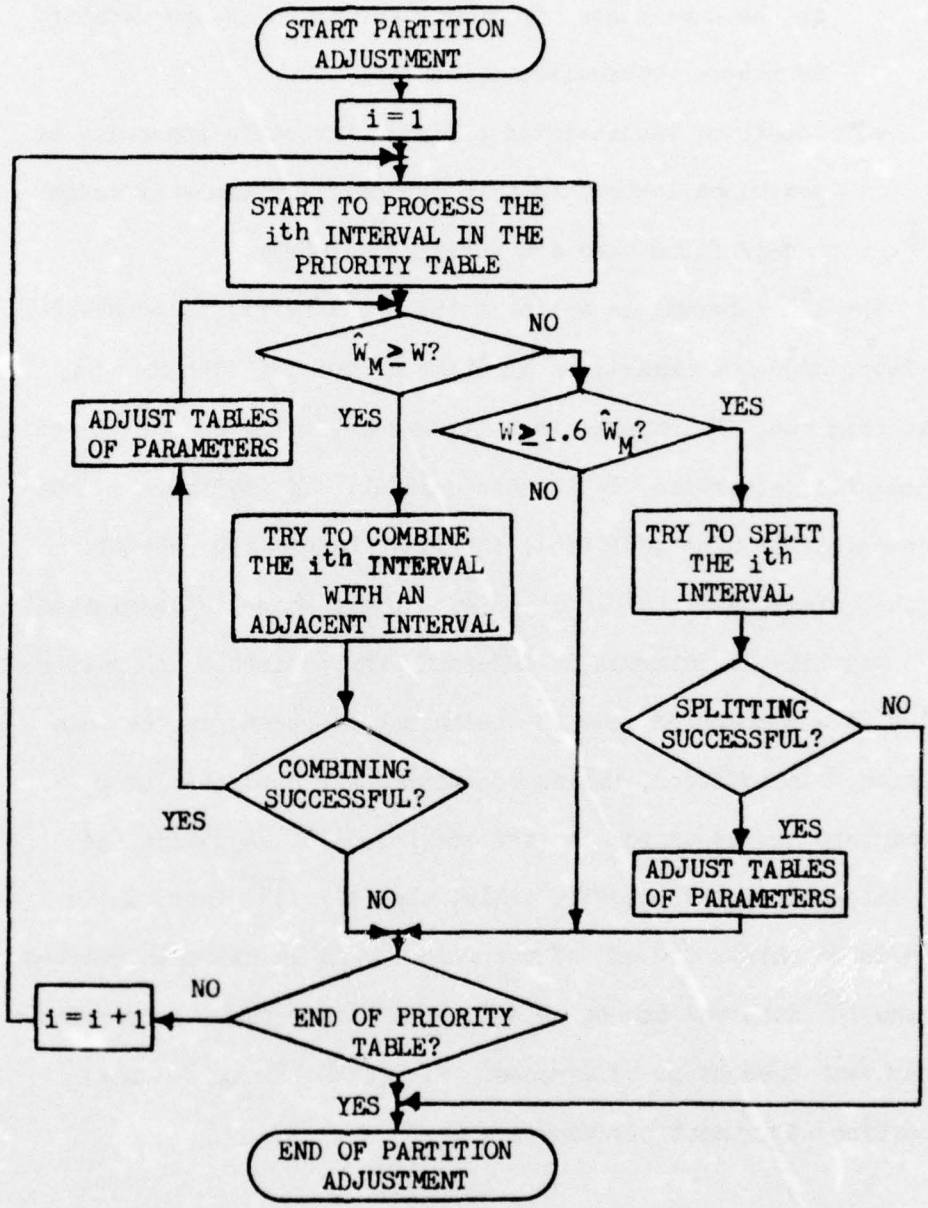


Figure 14. Flow Diagram of Partition Adjustment

a) A pair of adjacent classified intervals, classified to the same class, is sought and if found are combined together. Otherwise,

b) A pair of unclassified adjacent intervals appearing in positions lower than i in the priority table is sought and if found they are combined together.

The i^{th} interval is split if the total number of intervals is less, or can be made less by Items a) and b), than R. In that case the i^{th} interval is split and splitting is said to be successful; otherwise, it is unsuccessful. If splitting is not successful, then no additional interval changes are possible at that stage, and the partition adjustment phase is terminated.

Any time an interval is adjusted, the parameters characterizing it are computed from the technique discussed in the next section. In addition, tables containing the characterizing parameters are adjusted. At the conclusion of adjusting the i^{th} interval in the priority table, the $(i + 1)^{st}$ interval is considered unless the end of the priority table has been reached or the i^{th} interval cannot be split—in either case the partition adjustment process is terminated. Figure 14 summarizes the partition adjustment procedure.

3.4 Estimation of Parameters After a Partition Adjustment

Before the partition adjustment, a beta d.f. on each of the P's is known. After the partition adjustment, d.f.'s for those P's in un-altered intervals are the same as before the

adjustment. For an altered interval, however, a record of the number of training observations in the interval from each class is unavailable. This section derives characterizing parameters for beta d.f.'s on the P's after a partition adjustment, in terms of the expected values and variances of the P's in contributing intervals before the adjustment.

Interval Combining

Suppose that the i^{th} and $(i + 1)^{st}$ intervals are combined. Consider just the j^{th} class and let i and (i + 1) denote the i^{th} and $(i + 1)^{st}$ intervals respectively with no interval notation indicating the combined interval. Thus,

$$P = P(i) + P(i + 1)$$

$$EP = EP(i) + EP(i + 1)$$

An upper bound on the variance of P is

$$\text{Var } P = \text{Var } P(i) + \text{Var } P(i + 1) + 2(\text{Var } P(i)\text{Var } P(i + 1))^{\frac{1}{2}}$$

because for any random variables X and Y

$$\begin{aligned}\text{Var}(X + Y) &= E(X + Y)^2 - E^2(X + Y) \\ &= \text{Var } X + \text{Var } Y + 2\left[\frac{EXY - EXEY}{(\text{Var } X \text{ Var } Y)^{\frac{1}{2}}}\right](\text{Var } X \text{ Var } Y)^{\frac{1}{2}} \\ &\leq \text{Var } X + \text{Var } Y + 2(\text{Var } X \text{ Var } Y)^{\frac{1}{2}}\end{aligned}$$

The characterizing parameters of the beta d.f. on P for the combined interval are obtained from Equations (A.4) in Appendix A.

Interval Splitting

Again only the j^{th} class probabilities are considered with no notational mention of it. Suppose that an interval is split into the i^{th} and $(i + 1)^{st}$ intervals. P is the random variable for the interval probability before splitting; $P(i)$ and $P(i + 1)$ are random variables for the i^{th} and $(i + 1)^{st}$ interval probabilities after splitting. Because no information is available about the variation across the interval before splitting, the distributions on $P(i)$ and $P(i + 1)$ after splitting should be identical to each other; thus,

$$EP(i) = \frac{EP}{2} \tag{3.7}$$

The allocation of the probability P is not necessarily uniform across the interval. The worst case for this allocation is governed by the Lipschitz constant L for the j^{th} class-conditional d.f. Figure 15 illustrates a worst case allotment of P to the interval. The worst case occurs when the actual d.f. f for the j^{th} class has its maximum absolute slope L over the interval as shown in the figure. Then $P(i)$ is given by

$$P(i) = \frac{P}{2} + \epsilon \tag{3.8}$$

where ϵ for such a worst case is given by

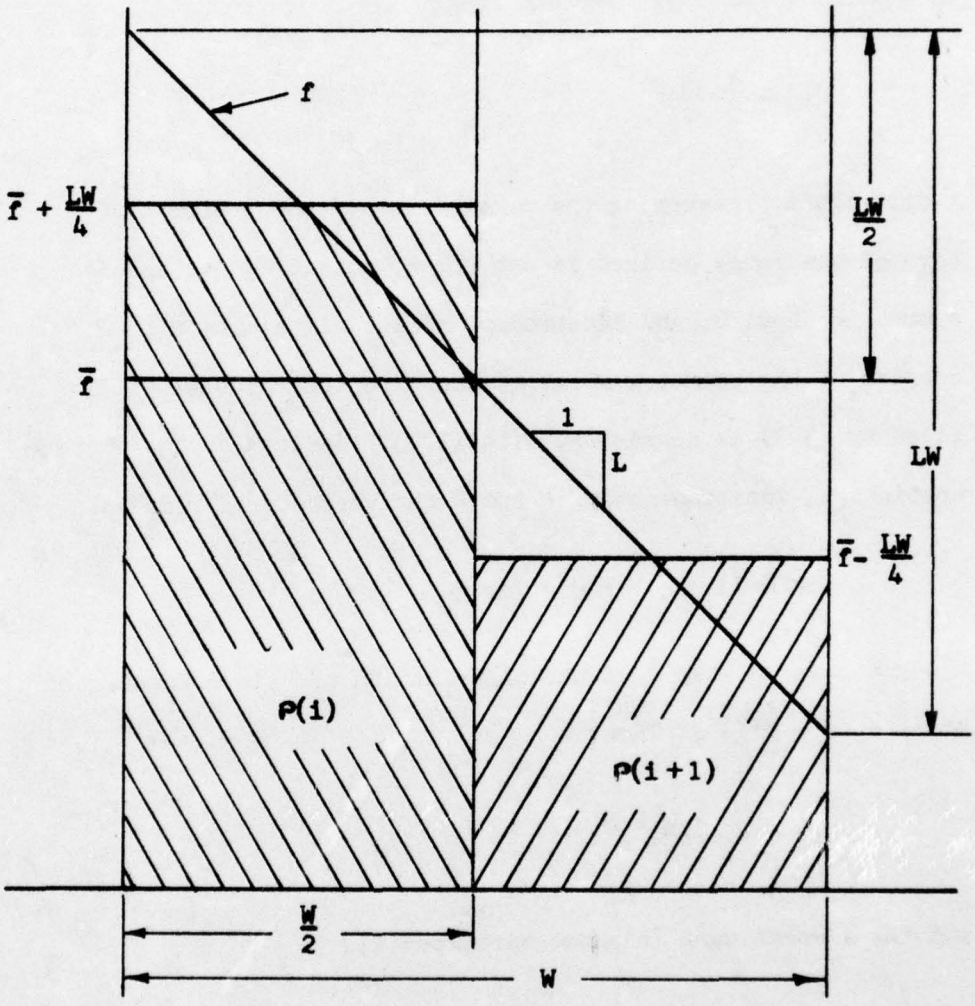


Figure 15. Variation of Probability in an Interval

$$\epsilon = (L\,\frac{W}{4})(\frac{W}{2})$$

$$= \frac{L}{2}\,(\frac{W}{2})^2$$

In general ϵ can vary over the range

$$|\epsilon| \leq \frac{L}{2}\,(\frac{W}{2})^2$$

A distribution governing the probability of occurrence of ϵ through the range defined is not known, but it can be assumed symmetric about 0, and independent of the distribution on ρ. Because of the symmetry about 0, the expected value of $\rho(i)$ given by (3.8) is consistent with (3.7). Because of the assumed statistical independence of ϵ and ρ the variance of $\rho(i)$ is

$$\text{Var}\ \rho(i) = \tfrac{1}{4}\ \text{Var}\ \rho + \text{Var}\ \epsilon$$

A worst case is when $\frac{1}{2}$ of the distribution of ϵ is concentrated at each of $\pm\ \frac{L}{2}(\frac{W}{2})^2$. Then

$$\text{Var}\ \epsilon = (\frac{L}{2}(\frac{W}{2})^2)^2$$

and for a worst case (highest variance)

$$\text{Var}\ \rho(i) = \tfrac{1}{4}\ \text{Var}\ \rho + \tfrac{1}{4}(L(\frac{W}{2})^2)^2 \qquad (3.9)$$

The characterizing parameters of the beta d.f. on $\rho(i)$ are obtained by using $E\rho(i)$ and Var $\rho(i)$ of (3.7) and (3.9) in Equations (A.4) of Appendix A.

CHAPTER IV

COMPUTER SIMULATED RESULTS

4.1 Introduction

This chapter contains results obtained by using an IBM 1130 computer system to generate and to process simulated data. The simulated data is generated using standard pseudo-random number generation techniques (see e.g. [40]). Processing follows the flow diagram of Figure 3 in Chapter I. An interval of a given R-interval partition of $\mathcal{A}$ is "classified" if condition (2.21) of Chapter II is satisfied. The intervals of an initial R-interval partition are defined using the first R - 1 training observations by the technique described in Appendix A. The partition is subsequently adjusted using the procedure of Chapter III. The i^{th} interval, if not "classified" by satisfaction of (2.21), can be "tentatively classified" to class ω_a by using (2.12). Thus, even if all intervals are not classified, tentative results are available until they are classified.

4.2 Allocation of α and $1 - \beta$ to the Intervals

Experimentally, it was found that assignment of $\alpha(i)$ and $\tau(i)$ according to

$$\alpha(i) = \frac{W(i)}{W_T}\,\alpha$$

$$\tau(i) = \frac{W(i)}{W_T}\,(1 - \beta) \qquad (4.1)$$

is not economical in terms of the number n of training observations required for domain classification. The assignment (4.1) is equivalent to

$$\alpha(i) = \frac{W(i)}{W_T^*}\,\alpha^*$$

$$\tau(i) = \frac{W(i)}{W_T^*}\,(1 - \beta)^* \qquad (4.2)$$

where α^* and $(1 - \beta)^*$ are the portions of α and $(1 - \beta)$ that have not been used for the classified intervals, and W_T^* is the cumulative length of the unclassified intervals. A significant reduction in n was experimentally observed with modification of (4.2),

$$\alpha(i) = u_1\,(\hat{p}_u)\,\frac{W(i)}{W_T^*}\,\alpha^*$$

$$\tau(i) = u_2\,\frac{W(i)}{W_T^*}\,(1 - \beta)^* \qquad (4.3)$$

where $u_1\,(\hat{p}_u)$ depends on an estimate $\hat{p}_u$ of the probability† in all unclassified intervals by the relation

---

†$\hat{p}_u$ is the sum of estimates of mixture probabilities in unclassified intervals.

$$u_1(\hat{p}_u) = t_1 + (1 - t_1)(1 - \hat{p}_u)^{t_2}$$

and u_2 is constant. u_2, t_1, and t_2 are experimentally chosen constants; the experimental examples subsequently described use

$$u_2 = 0.5$$
$$t_1 = 0.02$$
$$t_2 = 0.05/\alpha$$

The modification allots larger portions of α and $(1 - \beta)$ to the last regions classified, causing them to be classified with less difficulty. The observed decrease in n is attributed to this fact.

4.3 Study of a Particular Problem

Let f_1 and f_2 be truncated Gaussian d.f.'s given by

$$f_1(x) = K_1 e^{-\frac{1}{2}\left(\frac{x - 0.4}{0.1}\right)^2}, \quad 0 < x < 1$$
$$= 0, \quad \text{otherwise}$$

$$f_2(x) = K_2 e^{-\frac{1}{2}\left(\frac{x - 0.6}{0.1}\right)^2}, \quad 0 < x < 1$$
$$= 0, \quad \text{otherwise} \qquad (4.4)$$

where K_1 and K_2 are normalization constants included so that f_1 and f_2 integrate to 1, and let the problem parameters be

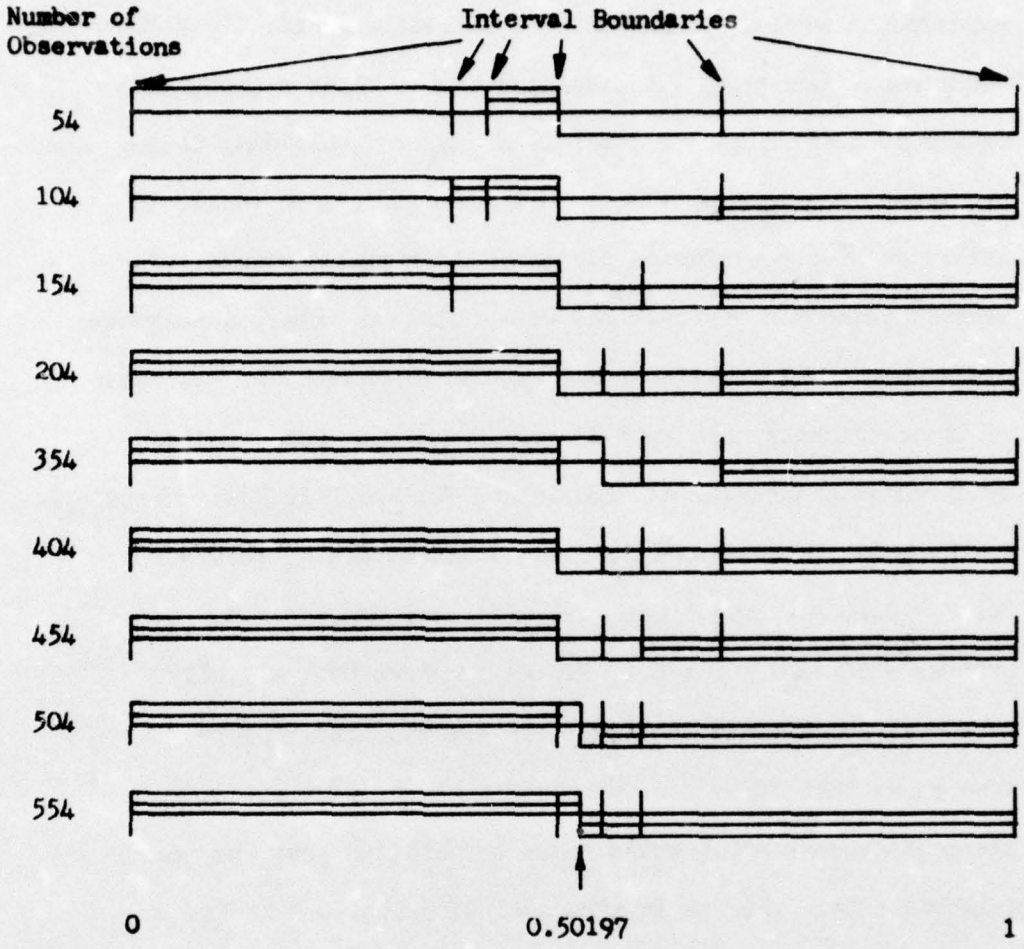
$$\alpha = 0.1$$
$$\beta = 0.9$$
$$P_j = 0.5,\ j = 1,2$$
$$L_j = 25,\ j = 1,2$$
$$R = 9 \qquad (4.5)$$

Before presenting results for R = 9, a graphical illustration of the partition changes for R = 5 is presented in Figure 16 for one experiment. In Figure 16, a single horizontal line indicates the corresponding interval is tentatively classified; a double horizontal line indicates the interval is classified. The lines being above or below the axis indicate Class ω_1 or Class ω_2 respectively. The result after 554 training observations is a classified domain with decision threshold at 0.50197 and error probability of 0.15870. This compares with an optimum threshold at 0.50000 and error probability of 0.15866.

For a comparison with well known parametric techniques (see e.g. [5]), assume it is known that f_1 and f_2 are Gaussian with standard deviation 0.1. Then the only unknown parameters are the means. It is easily shown that only four observations (two from each class) need be taken to satisfy the condition

$$\Pr[\Pr(\mathcal{E}|d) - \Pr(\mathcal{E}|d_o) \leq 0.1] \geq 0.9$$

if the d.f.'s are given by (4.4). The additional a priori knowledge drastically reduces the number of training observations required.



$Pr(\mathcal{E}|d) = 0.15870$ $Pr(\mathcal{E}|d_o) = 0.15866$

Figure 16. Example of Partition Changing

For comparison with a commonly used nonparametric technique, experiments were performed using a nearest neighbor classification technique. Each x in $\mathcal{D}$ is assigned to the class represented by the nearest member in a set of $n = n_1 + n_2$ training observations (n_j from Class ω_j, $n_1 = n_2$). Using the d.f.'s of (4.4), 100 experiments were performed for each of several n values. For each experiment n observations were taken, the nearest neighbor decision rule was obtained, and $Pr(\mathcal{E}|d)$ computed. An estimate of the confidence that $Pr(\mathcal{E}|d) - Pr(\mathcal{E}|d_o) < \alpha$ was obtained by dividing the number of experiments for which $Pr(\mathcal{E}|d) - Pr(\mathcal{E}|d_o) < \alpha$ by 100. The curves in Figure 17 illustrate n_j versus α for $\beta = 0.1, 0.2, 0.5, 0.8, 0.9$. For confidence $\beta = 0.9$ that $Pr(\mathcal{E}|d) - Pr(\mathcal{E}|d_o) < \alpha = 0.1$, Figure 17 shows that slightly more than 100 training observations are required. Figure 17 also shows that for α somewhat less than 0.1, say $\alpha = 0.05$, the confidence $\beta = 0.9$ would never be attained from the nearest neighbor rule. This is in agreement with the work of Fix and Hodges [55] who show that

$$Pr(\mathcal{E}|d) \underset{n_j \to \infty}{\longrightarrow} 0.225$$

or

$$Pr(\mathcal{E}|d) - Pr(\mathcal{E}|d_o) \underset{n_j \to \infty}{\longrightarrow} 0.225 - 0.159 = 0.066$$

when the nearest neighbor procedure is used on this problem.*

*Fix and Hodges also show that for the K nearest neighbor rule, the asymptotic difference $Pr(\mathcal{E}|d) - Pr(\mathcal{E}|d_o)$ decreases to zero as $K \to \infty$.

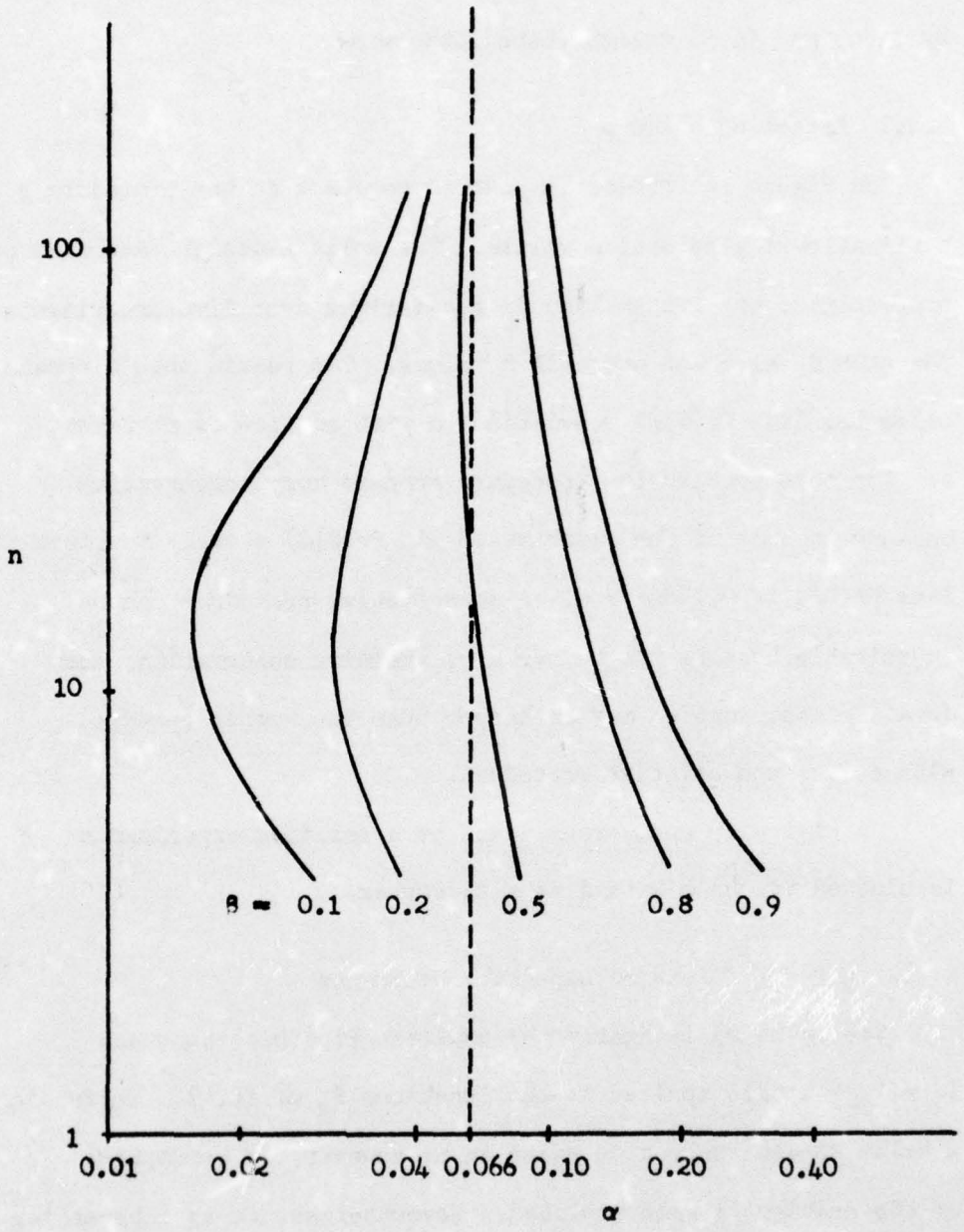


Figure 17. Tradeoffs Among n, α, and β, (Nearest Neighbor Classification)

The rest of this section is concerned with the problem defined by (4.4) and (4.5) unless stated otherwise.

4.3.1 Effect of α and β

In Figure 18 $Pr(\mathcal{E}|d)$ is plotted versus α at the procedure's termination for several β values. The solid lines are averages over five experiments; the broken line is the maximum over five experiments for each β value and over all β values. The result should remain below the line $Pr(\mathcal{E}|d) = 0.15866 + \alpha$ with confidence at least β. For this problem the procedure appears very conservative because in none of the experiments did $Pr(\mathcal{E}|d)$ closely approach the line $Pr(\mathcal{E}|d) = 0.15866 + \alpha$. A conservative procedure can be undesirable because the number n of training observations for domain classification may be larger than the number required with a less conservative procedure.

In Figure 19 an average value of n for five experiments is plotted versus α with β as a parameter.

4.3.2 Effect of Assumed Lipschitz Constants

The value 25 is nearly the smallest Lipschitz constant $L_1 = L_2 = L$ that applies to the functions f_j of (4.4). Decreasing L below 25 can cause a decrease in n; however, an assumption of the problem is then violated. Nevertheless, it is interesting that for the problem of (4.4) and (4.5), reduction of L causes a reduction of n without causing $Pr(\mathcal{E}|d)$ to exceed an acceptable limit $(Pr(\mathcal{E}|d_o) + \alpha)$.

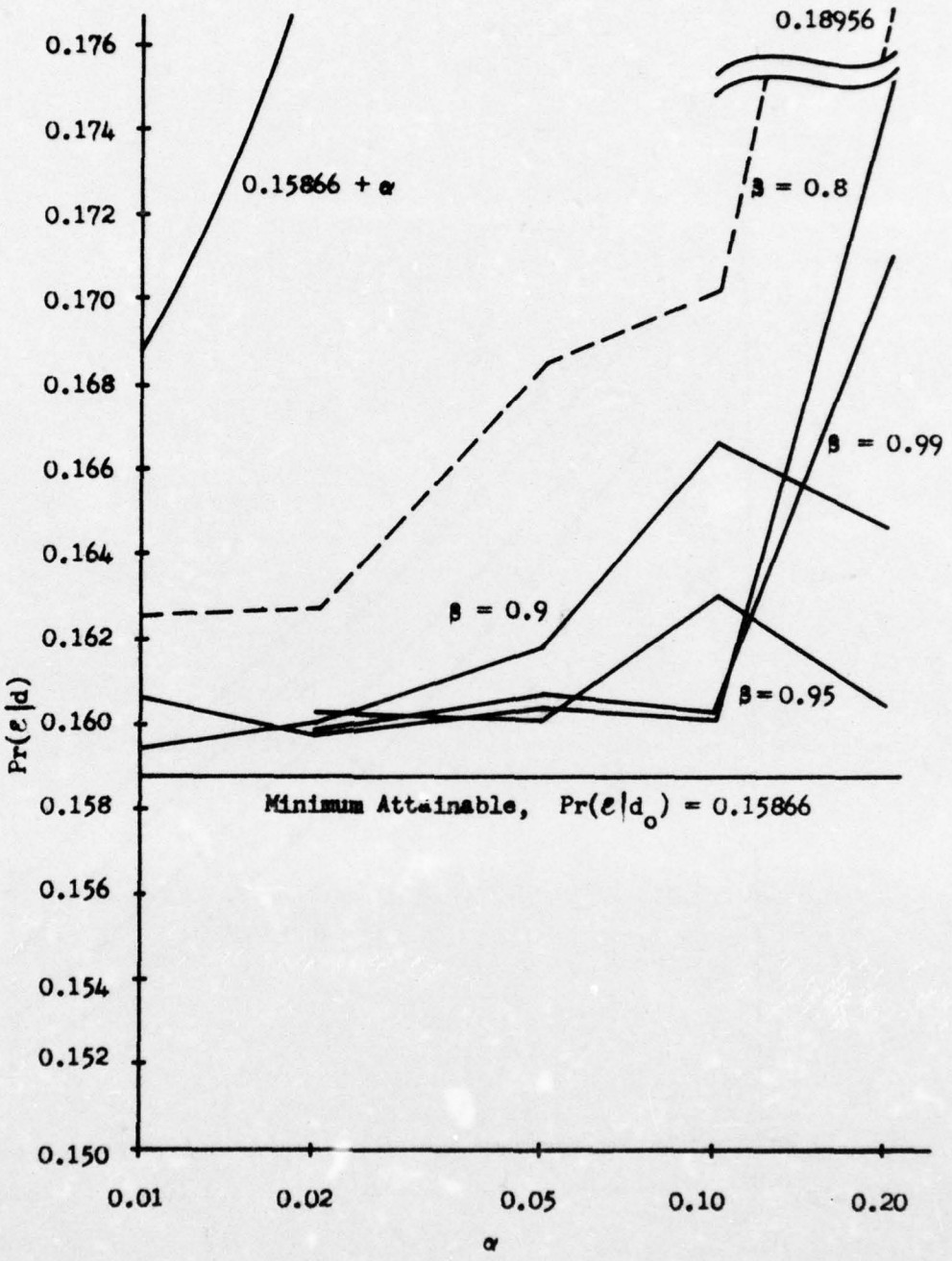


Figure 18. Pr(e|d) Versus α for Several β Values

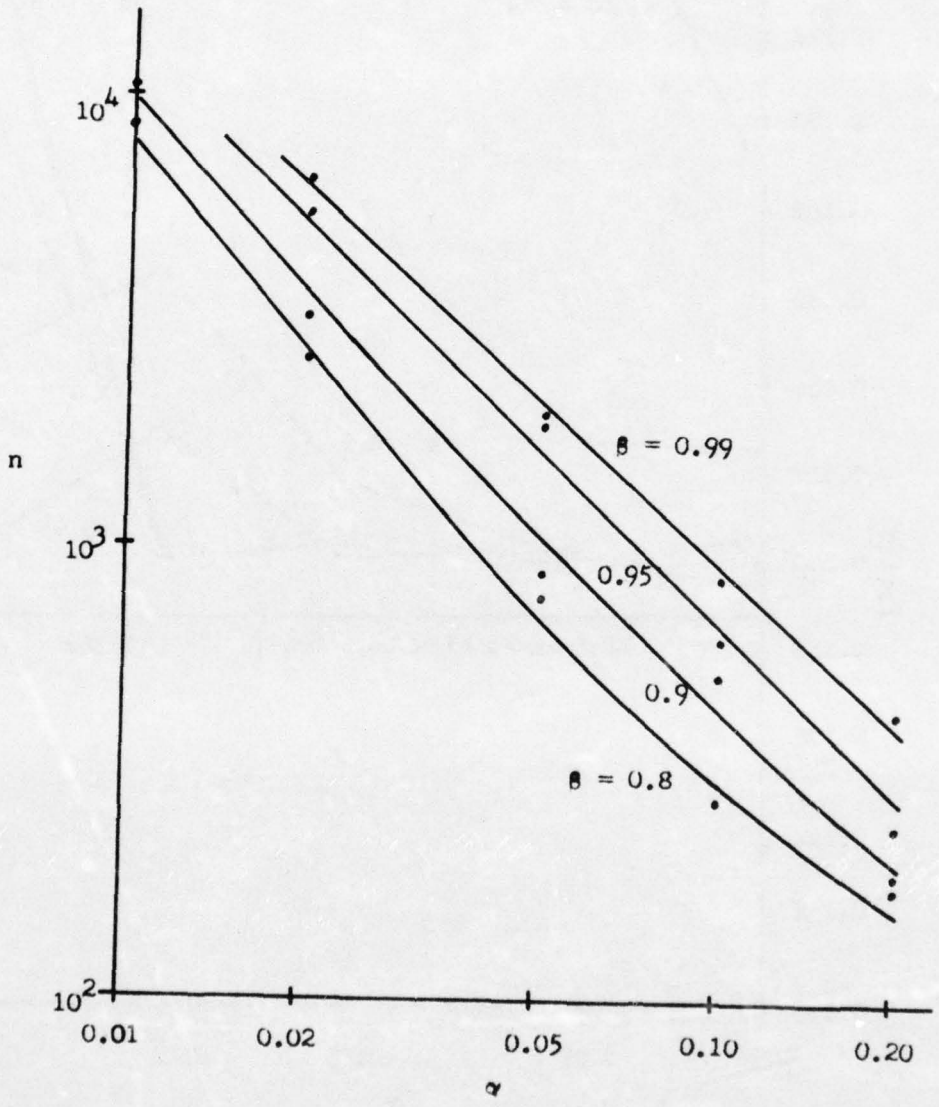
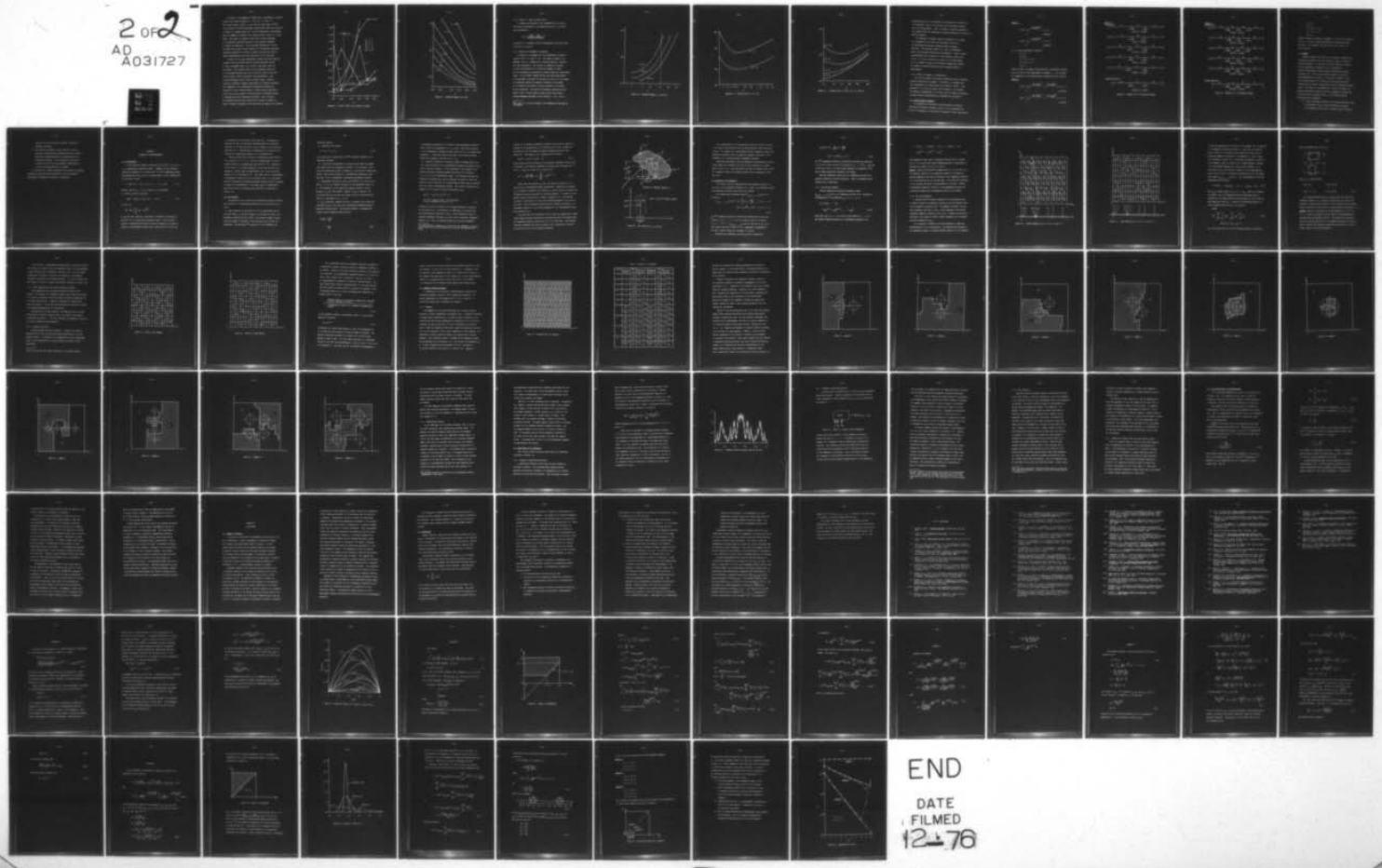


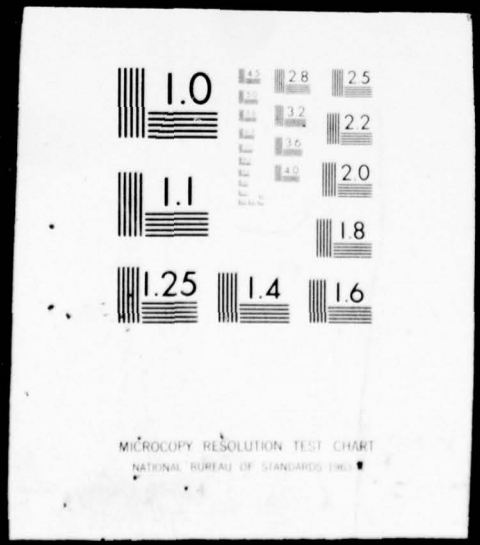
Figure 19. Tradeoffs Among n, α, and β

AD-A031 727 PURDUE UNIV LAFAYETTE IND SCHOOL OF ELECTRICAL ENGI--ETC F/G 9/3
A NONPARAMETRIC RECOGNITION PROCEDURE WITH STORAGE CONSTRAINT,(U)
AUG 69 E A PATRICK, F K BECHTEL F30602-68-C-0186
UNCLASSIFIED TR-EE69-24 NL

2 OF 2
AD A031727



END
DATE FILMED
12-76



In Figure 20 the maximum of $Pr(\mathcal{E}|d)$ over 5 experiments is plotted versus α for several values of L. Only for L = 1 and L = 2 did $Pr(\mathcal{E}|d)$ exceed o.15866 + α, and even for these cases, $Pr(\mathcal{E}|d)$ for only one of the five experiments exceeded that value for a given α. In Figure 21, average values of n for five experiments corresponding with the examples in Figure 20 are plotted versus α with L as a parameter. The results in Figure 21 show that a priori knowledge of the smallest applicable values for the Lipschitz constants is helpful in reducing n. For the problem considered it can be concluded from Figure 20 that violation of the smallest applicable Lipschitz constants by a factor as large as ten may not prevent domain classification such that condition (1.6) is satisfied.

A reason for the good experimental results even with Lipschitz constants that are smaller than the minimum applicable values is that the maximum slope of f_j occurs in just small parts of the domain. This suggests that a priori knowledge consisting of the maximum absolute value of the slope of $f_j(x)$ at each $x \in \mathcal{A}$ could be used to make the approach less conservative. Such a priori knowledge could be used to define "local" Lipschitz constants, different constants applicable for different intervals. Also suggested is the possibility of adaptively altering the constants for each interval based on current results; adaptation could occur with an operator interacting with a histogram display or automatically by an estimation procedure. Such a display or estimation procedure may require local storage of samples in order to obtain an estimate of the local rate of change of the density.

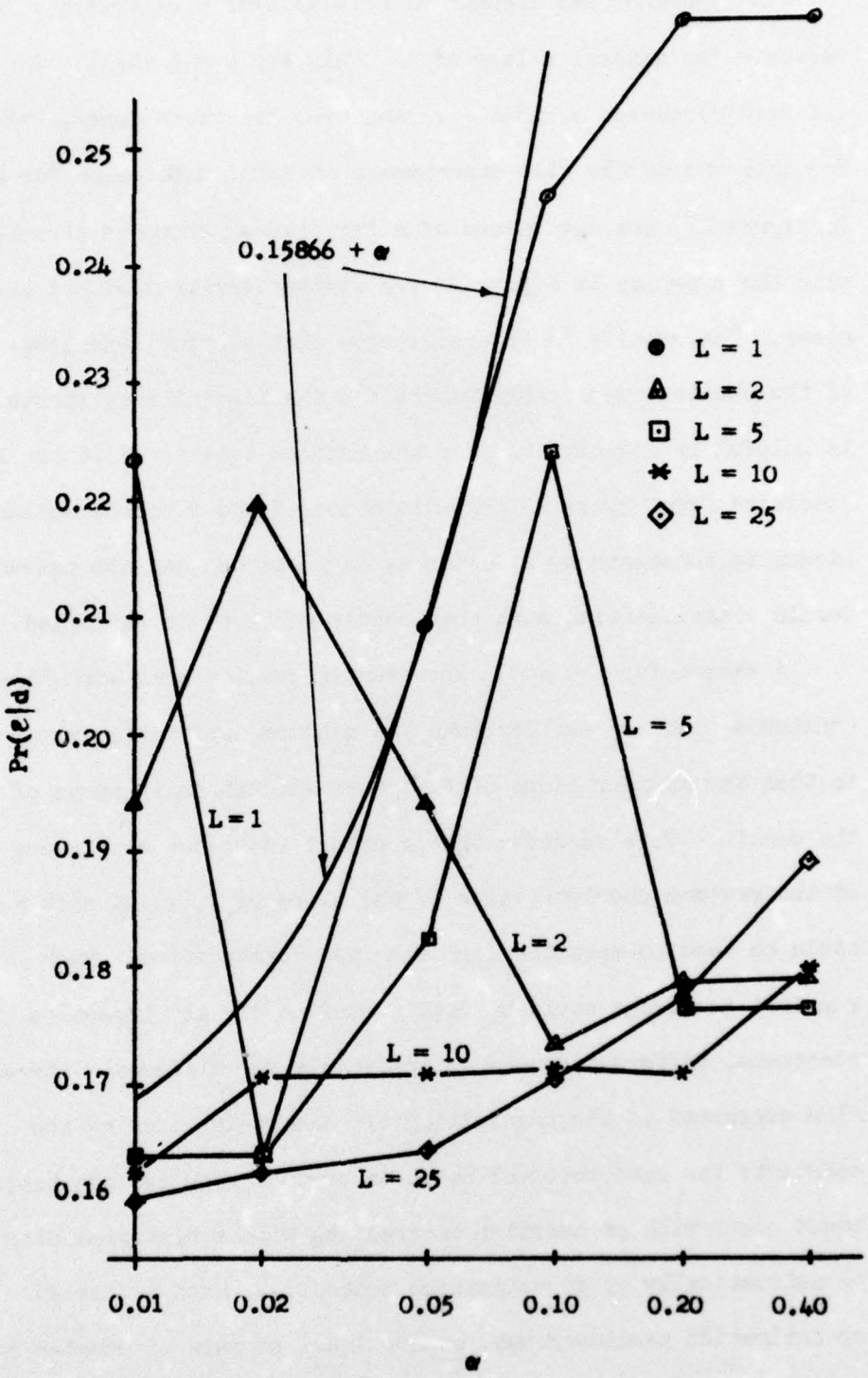


Figure 20. Pr(e|d) Versus α for Several L Values

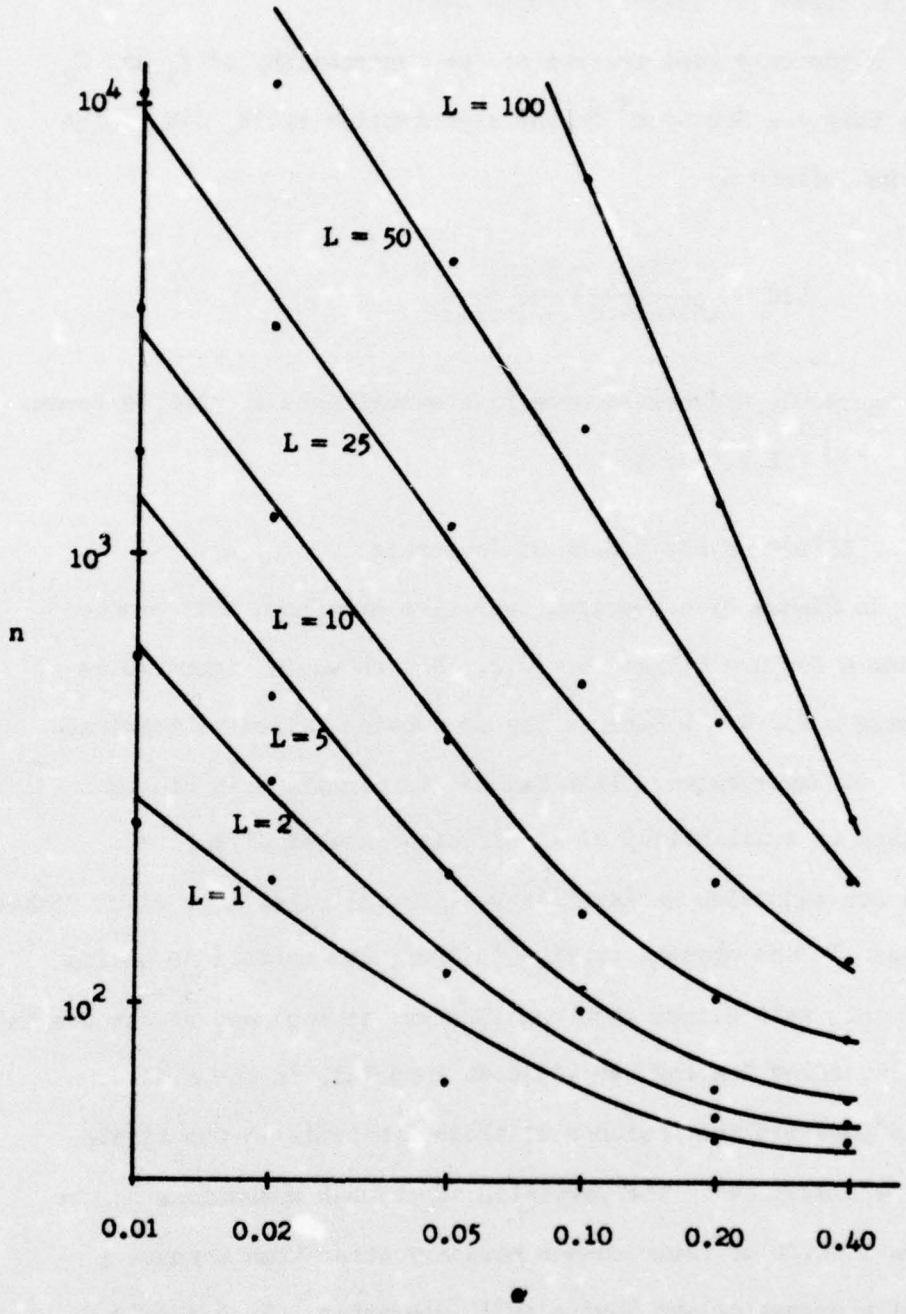


Figure 21. Tradeoffs Among n, α, and L

4.3.3 Effect of Signal to Noise Ratio

A commonly used measure of the separability of f_1 and f_2 when they are Gaussian† is the signal-noise ratio, S:N, which can be defined by

$$S:N = \left(\frac{\text{Mean}_1 - \text{Mean}_2}{\text{Standard Deviation}}\right)^2$$

In Figure 22, n averaged over five experiments is plotted versus L for S:N = 1,2, and 4.

4.3.4 Effect of the Number of Intervals

In Figure 23 n averaged over five experiments is plotted versus R for $\alpha = 0.1$ and $\alpha = 0.2$. Not shown in Figure 23 is average n for R = 4 because the processing failed to terminate for some experiments. This failure to terminate is caused by lack of availability of a sufficient number of intervals for adjusting unclassifiable interval sizes into classifiable sizes. In the present example failure with partitions having four intervals occurs when one interval at each end of the domain is classified leaving two adjacent intervals in the middle. It is possible that neither of these intervals in the middle can be classified. The partition adjustment operations do not allow a shift of their common boundary other than through a combine operation and then a split operation. Such a pair of

---

†When the d.f.'s are not Gaussian, this definition loses much of its appeal.

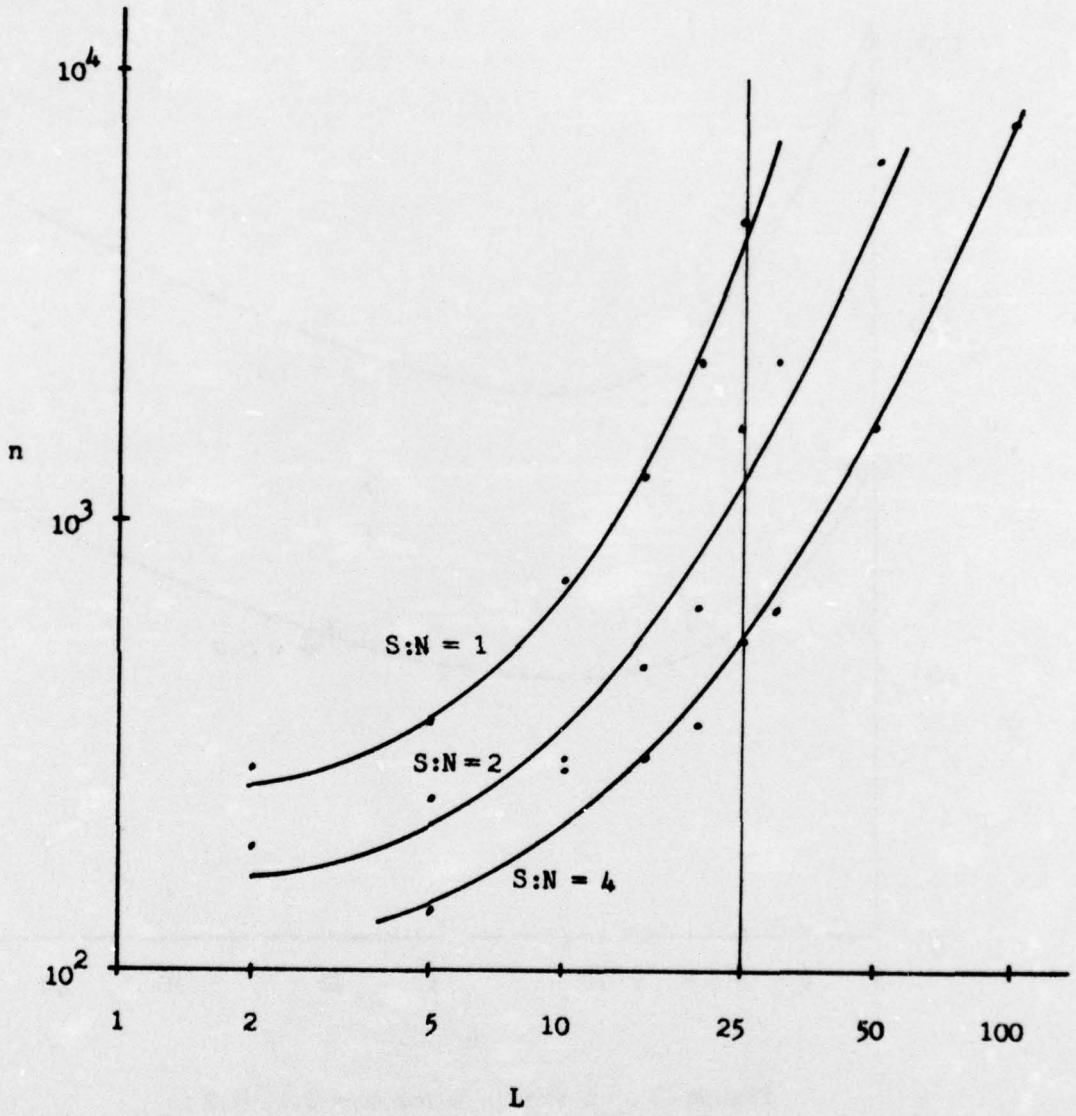


Figure 22. Tradeoffs Among n, L, and S:N

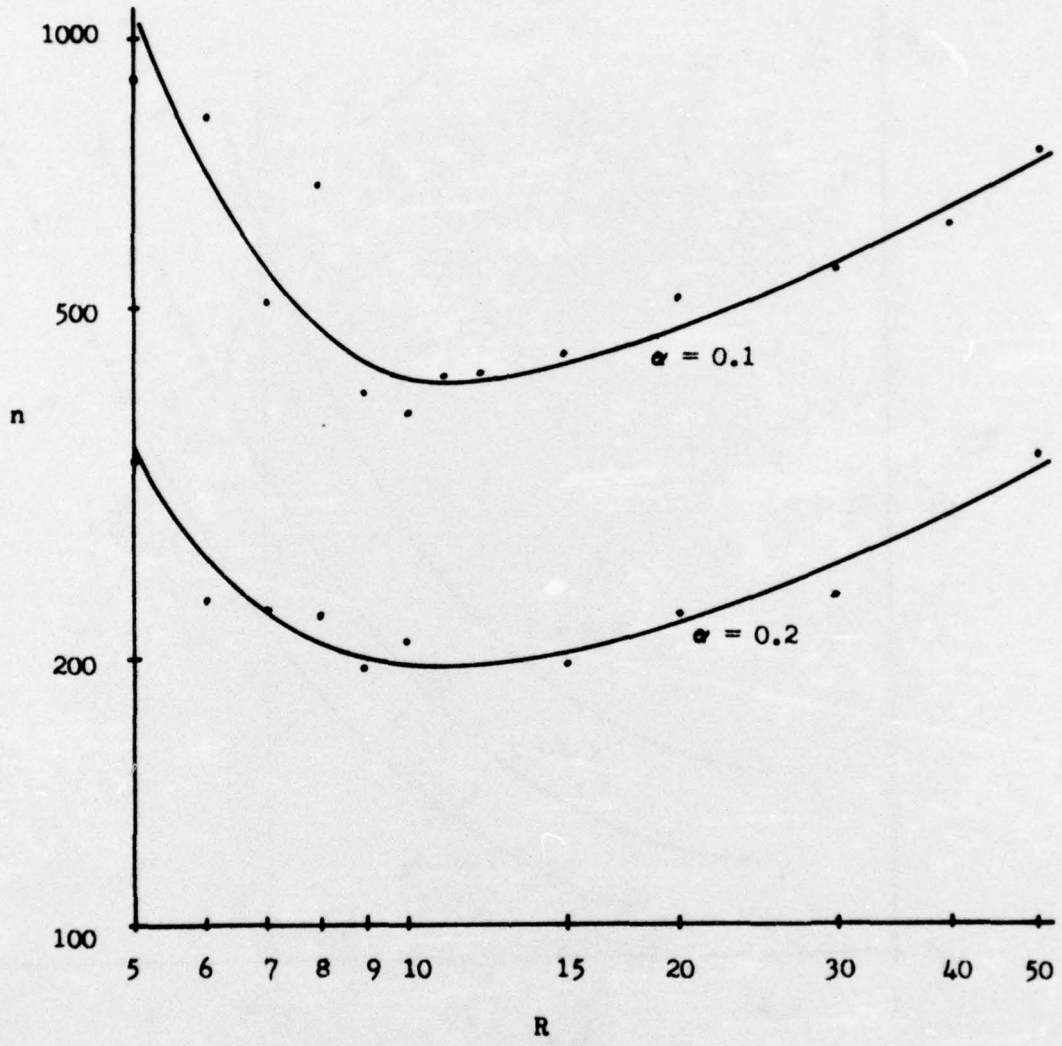


Figure 23. n Versus R for α = 0.1, 0.2

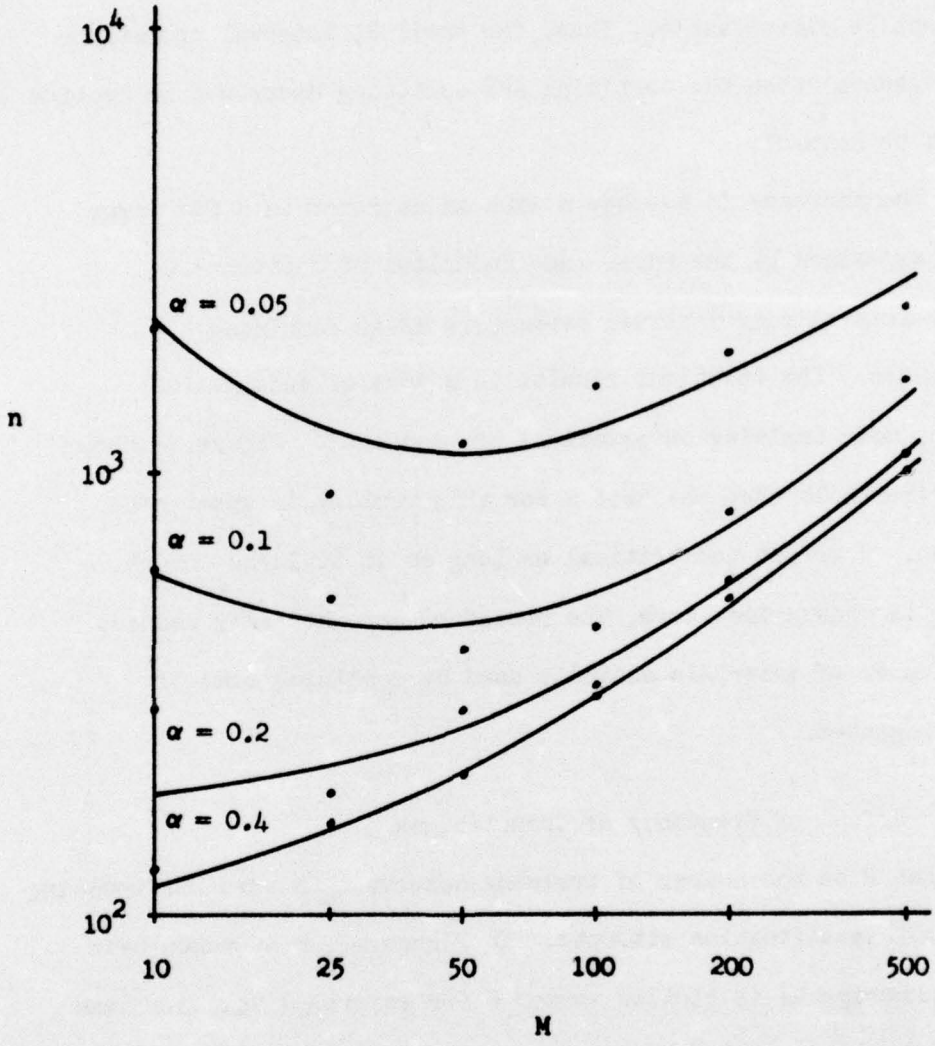


Figure 24. n Verses M for $\alpha = 0.05$, 0.1, 0.2, and 0.4

operations may not occur because of the conditions of Section 3.3 for combining. Even if it does occur, the resulting intervals may not be classifiable. Thus, for small R, interval operations more general than the combining and splitting described in Section 3.3 might be helpful.

The increase in average n with an increase in R for large R is explained by the worst case technique of Section 3.4 for reinitializing interval parameters after combining intervals. The technique results in a loss of information; hence, more training observations are required. Figure 23 shows experimentally that the best R for this problem is about nine or ten. R is not too critical as long as it is large enough. If it is chosen too large, the procedure automatically reduces the number of intervals actually used by combining some of them together.

4.3.5 Effect of Frequency of Computations

Let M be the number of training observations used for updating between classification attempts. In Figure 24 , n averaged over five experiments is plotted versus M for several α values. Some increase in n is noted for small and for large M. Not plotted, but perhaps as significant, is the fact that for large M processing is faster because computations are performed less frequently.

4.4 Multi-Threshold Examples

To illustrate the procedure for multi-threshold problems, including non Gaussian problems, results of five experiments for each of two examples are illustrated in Figure 25 and 26 respectively.

Example 1

$$f_1(x) = K_1 \, e^{-\frac{1}{2}\left(\frac{x - 0.5}{0.1}\right)^2}, \quad 0 < x < 1$$

$$= 0, \quad \text{otherwise}$$

$$f_2(x) = K_2 \, e^{-\frac{1}{2}\left(\frac{x - 0.5}{0.2}\right)^2}, \quad 0 < x < 1$$

$$= 0, \quad \text{otherwise}$$

(K_1, K_2 are normalization constants)

$\alpha = 0.2$

$\beta = 0.9$

$P_j = 0.5, \; j = 1,2$

$L_1 = 25, \; L_2 = 7$

$R = 13$

Figure 25 shows the domain classification at termination, $\Pr(\mathcal{E}|d)$, and n for each of five experiments for Example 1. Also included for comparison is the optimum domain classification and $\Pr(\mathcal{E}|d_o)$.

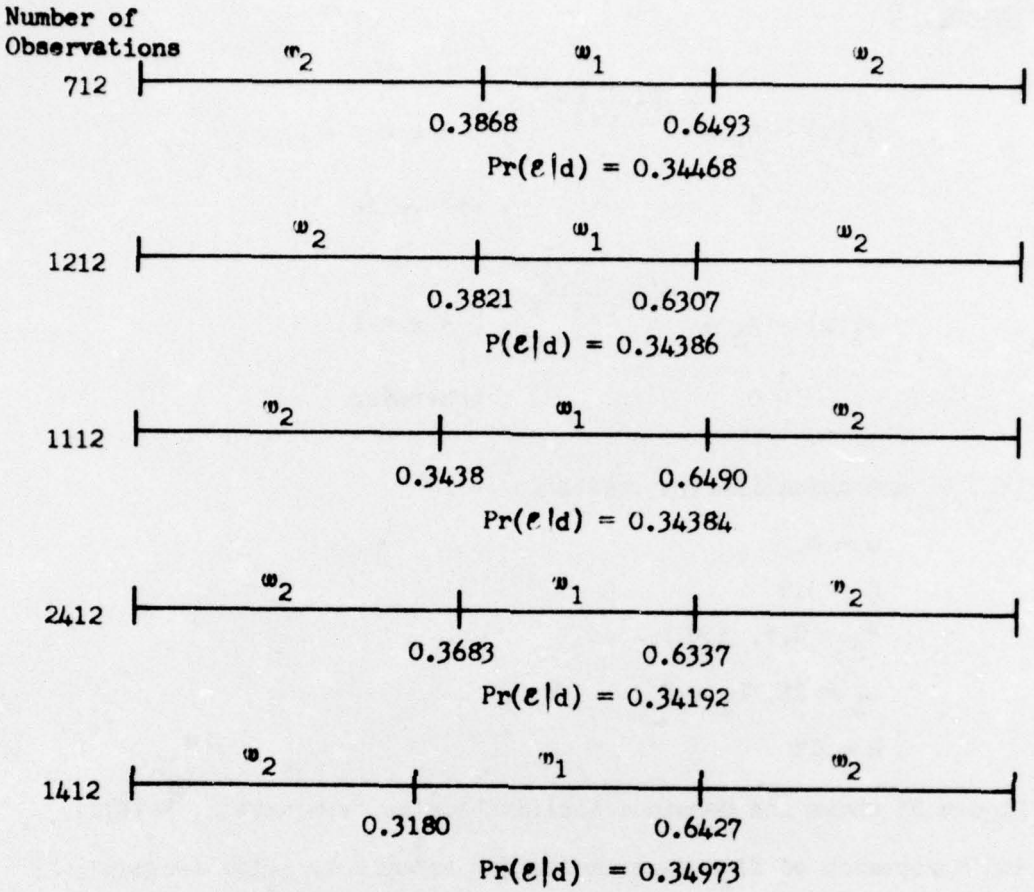
Example 2

$$f_1(x) = K_1 \left[e^{-\frac{1}{2}\left(\frac{x - 0.2}{0.05}\right)^2} + e^{-\frac{1}{2}\left(\frac{x - 0.6}{0.05}\right)^2} \right], \quad 0 < x < 1$$

$$= 0, \quad \text{otherwise}$$

$$f_2(x) = K_2 \left[e^{-\frac{1}{2}\left(\frac{x - 0.4}{0.05}\right)^2} + e^{-\frac{1}{2}\left(\frac{x - 0.8}{0.05}\right)^2} \right], \quad 0 < x < 1$$

$$= 0, \quad \text{otherwise}$$



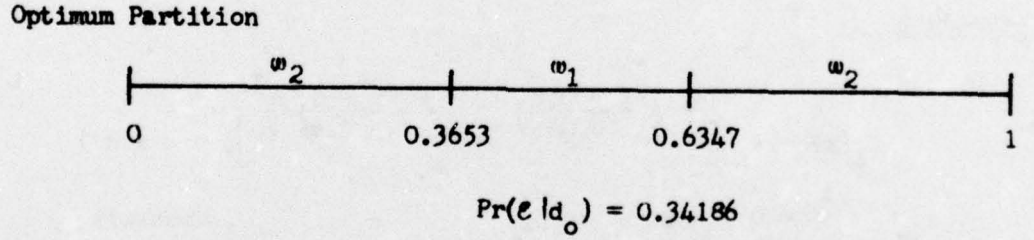


Figure 25. Results for a 2 Threshold Problem

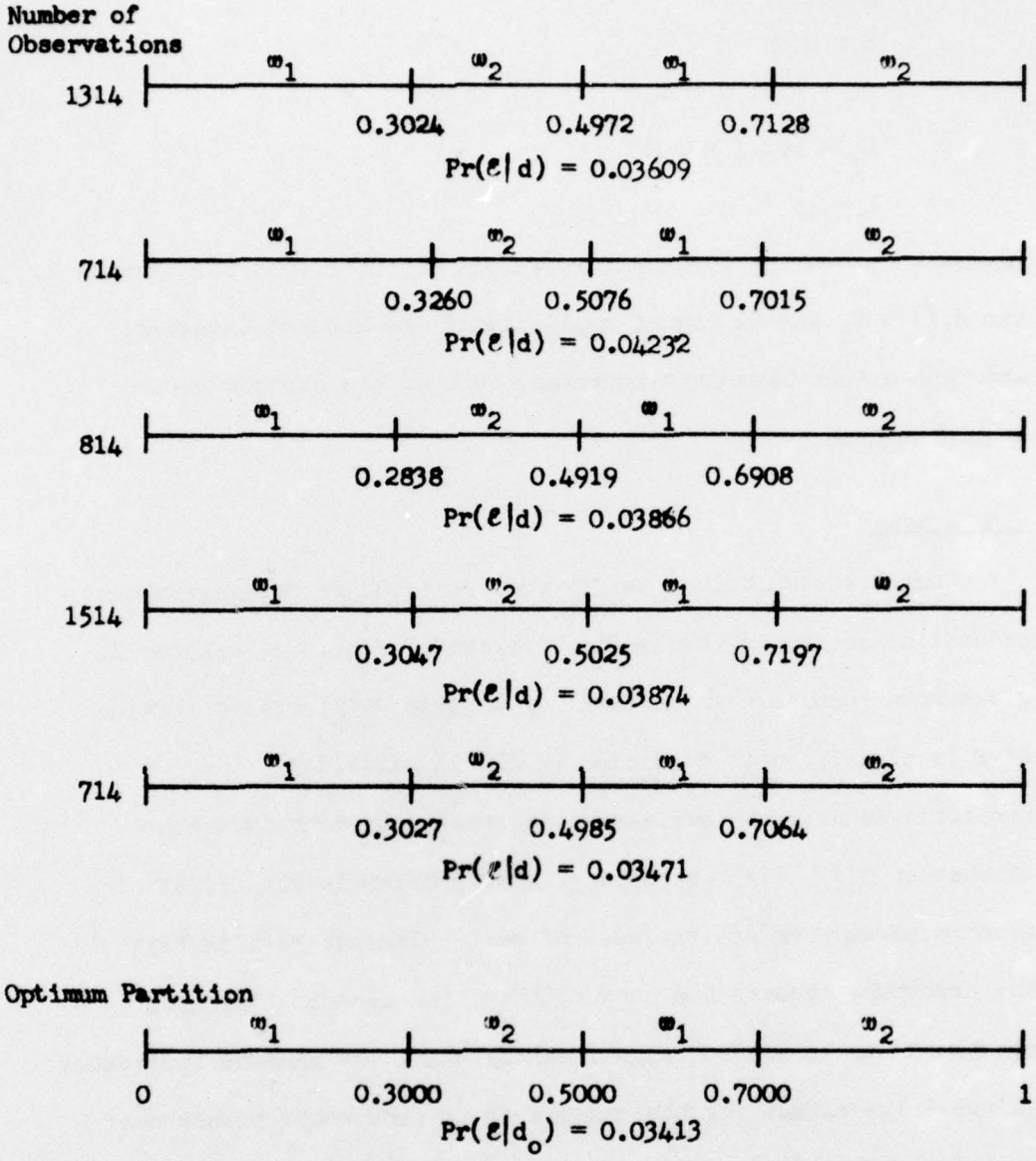


Figure 26. Results for a 3 Threshold Problem

$\alpha = 0.1$

$\beta = 0.9$

$P_j = 0.5,\ j = 1,2$

$L_j = 50,\ j = 1,2$

$R = 15$

Figure 26 shows the results for Example 2. Note that for Example 2, the d.f.'s f_1 and f_2 cannot realistically be assumed Gaussian, and thus a "non Gaussian" approach, such as the current one, should be used.

4.5 Summary

Computer simulations verify for the problems considered that processing according to the flow diagram of Figure 3, Chapter I gives good results. An interval of a given R-interval partition of $\mathcal{D}$ is classified if condition (2.21) is satisfied. Partition adjustments are made using the adjustment technique of Chapter III. First a one threshold problem is studied as problem parameters are varied. Of particular interest is that the procedure appears too conservative; the assumed Lipschitz constants can be reduced significantly below the minimum applicable values. The effect for the example is to reduce the number n of training observations required without increasing the probability of error above an acceptable value. The following possible modifications are suggested:

1) A priori knowledge consisting of the maximum absolute value of the slope of $f_j(x)$ at each $x \in \mathcal{D}$ might be available. Such knowledge could be used to define "local" Lipschitz

constants for the intervals—different constants for different intervals.

2) The Lipschitz constants for each interval could be adaptively altered—either interactively by an operator observing a histogram display or automatically by an estimation procedure. Such an approach could lead to a practical solution of the problem of obtaining the a priori knowledge required in 1) above.

It is noted that a drastic decrease in the number of training observations required can be obtained if a priori knowledge appropriate to parametric procedures is available.

CHAPTER V

EXTENSION TO MULTIDIMENSIONS

5.1 Introduction

In the preceding chapters computerized recognition is restricted to a 1-dimensional observation space. A mapping is now utilized to extend the procedure to a ℓ-dimensional (ℓ finite) observation space. The observation vector $\underline{x}$ is in a bounded domain $\mathcal{D}$ of an ℓ-dimensional vector space V^{ℓ} where

$$\mathcal{D} = \left\{\underline{x} = (x_1,\ldots,x_\ell) : 0 \le x_j < 1, \quad j=1,\ldots,\ell\right\} \tag{5.1}$$

Density functions f_j, $j=1,2$, defined on $\mathcal{D}$ are assumed to satisfy Lipschitz conditions,

$$|f_j(\underline{x}) - f_j(\underline{y})| \le L_j \, \|\underline{x} - \underline{y}\| \quad , \quad j=1,2 \tag{5.2}$$

for the norm

$$\|\underline{x} - \underline{y}\| = \Big(\sum_{i=1}^{\ell} (x_i - y_i)^2 \Big)^{\frac{1}{2}}.$$

In the previous chapters, a procedure is developed for adjusting a partition of a 1-dimensional observation space. In the current chapter, an appropriately defined one-to-one mapping is utilized to achieve a correspondence between sets in a partition of $\mathcal{D}$ and sets

in a partition of an interval of the real line. The mapping is defined such that the previously developed partition adjustment technique on the real line can be used to adjust the corresponding partition of $\mathcal{D}$. Alternatively, the mapping can be viewed as converting the ℓ-dimensional problem to a 1-dimensional one.

There has been recent interest in transforming data vectors in V^{ℓ} to vectors in $V^{\ell'}$, $\ell' < \ell$. One such transformation is used to display clusters of ℓ-dimensional data vectors in $V^{\ell'}$, especially for the case $\ell' = 2$ (a human operator then can view them for data analysis). Another type of transformation is a one-to-one map of regions in V^{ℓ} to intervals in $V^{\ell'}$. The former type of transformation is discussed in Section 5.6; in its present form, it is not applicable to the partition adjustment problem although it may be possible to modify the transformation. The latter type of transformation which will be used for adjusting the partition is discussed in Section 5.3.

5.2 The Approach

The approach used to convert the ℓ-dimensional partition adjustment problem to a 1-dimensional problem involves the following six steps:

1) Each dimension of the domain $\mathcal{D}$ in V^{ℓ} is partitioned into b^K intervals where b, a positive integer, is the base for some of the arithmetic computations that follow. The positive integer K determines the number of intervals in the partition and is called the complexity. The resulting $b^{K\ell}$ regions in V^{ℓ} are referred to as

elementary regions.

2) Similarly, the interval

$$\mathcal{R} = \{y : 0 \leq y < 1\} \tag{5.3}$$

of the real line is partitioned into $b^{K\ell}$ intervals referred to as elementary intervals.

3) A one-to-one transformation is defined which maps the elementary regions onto the elementary intervals. In this manner, a partitioned ℓ-dimensional domain is mapped to a partitioned 1-dimensional domain. Data vectors falling in a ℓ-dimensional elementary region also fall in its corresponding 1-dimensional elementary interval.

4) Approximate functions h_1 and h_2 are defined for f_1 and f_2 such that h_j, $j=1,2$, is constant over each of the elementary regions in $\mathcal{D}$. The constant h_j on any particular region is taken to be the average of f_j over that region. Because f_1 and f_2 satisfy (5.2), the partitioning can be made fine enough so that for practical purposes h_j is equivalent to f_j, $j=1,2$.

5) g_j, a piecewise constant function, is defined on the real line such that g_j and h_j are equal over corresponding elementary region - elementary interval pairs. The interior content or ℓ-dimensional volume of each elementary region given by

$$\text{Volume} = \left(\frac{1}{b^K}\right)^{\ell}$$

$$= \frac{1}{b^{K\ell}} \tag{5.4}$$

is equivalent numerically to the width of each elementary interval.* Thus, function g_j integrates to 1 if h_j does. Data vectors falling in the ℓ-dimensional observation space have the d.f. f_j or practically speaking h_j. These data vectors also fall on the real line where, practically speaking, they have the d.f. g_j.

6) Lipschitz conditions introducing a priori knowledge about the d.f.'s were utilized in the 1-dimensional recognition procedure considered in previous chapters. At the beginning of this chapter, (5.2) defines Lipschitz conditions assumed satisfied by the functions f_j, $j=1,2$, on the ℓ-dimensional domain. The following concerns the problem of utilizing the a priori knowledge contained in these conditions in such a way that the 1-dimensional procedure may be employed with the current ℓ-dimensional problem. This involves obtaining constants L_j^* to be used in defining constraints on g_j.

$$|g_j(x') - g_j(y')| \leq L_j^* |x' - y'|$$

for $x' \neq y'$ where x' and y' are mid-points of any 2 elementary intervals in $\mathcal{R}$. (5.5)

Equation (5.5) can be thought of as a "pseudo-Lipschitz" condition on the function g_j. The procedure for obtaining L_j^* requires that the transformation discussed in Item 3. above relates each pair of adjacent elementary intervals in $\mathcal{R}$ with a pair of adjacent elementary regions in $\mathcal{A}$. Then, the maximum change in g_j from any elementary

*The partitioning is assumed to be such that all elementary intervals are the same size and all elementary regions are the same size and shape.

interval to an adjacent elementary interval occurs when the function f_j changes at its maximum rate in the direction of the line joining the mid-points of the corresponding adjacent elementary regions (see Figure 27). The change in g_j is bounded by the relation:

$$|g_j(x') - g_j(y')| \leq L_j \|\underline{x} - \underline{y}\| \tag{5.6}$$

where $\underline{x}$ and $\underline{y}$ are the mid-points of the adjacent elementary regions that correspond to the adjacent elementary intervals whose mid-points are x' and y'. Using (5.5) gives L_j^* in terms of L_j:

$$L_j^* = L_j \frac{\|\underline{x} - \underline{y}\|}{|x' - y'|} = L_j \frac{\frac{1}{b^K}}{\frac{1}{b^{K\ell}}} = L_j b^{K(\ell - 1)} \tag{5.7}$$

Recall that the functions g_j are, for practical purposes, d.f.'s governing the 1-dimensional mapped observations. Treating the constants L_j^* as Lipschitz constants for functions g_j, one can use the 1-dimensional recognition procedure developed in previous chapters. It operates on the mapped training observations to obtain a solution in $\mathfrak{R}$. The solution consists of a partition of $\mathfrak{R}$ with each interval assigned to one class or the other. The ℓ-dimensional solution can be obtained by assigning each elementary region in $\mathfrak{D}$ to the class assigned to its corresponding elementary interval in $\mathfrak{R}$.

One could avoid the conversion in (5.7) above by treating the 1-dimensional mapped training observations as though they were the original data and assuming Lipschitz conditions on d.f.'s for this data. Such assumed Lipschitz conditions are open to question, but, inpractice, so are the ones given by (5.2) on the original functions.

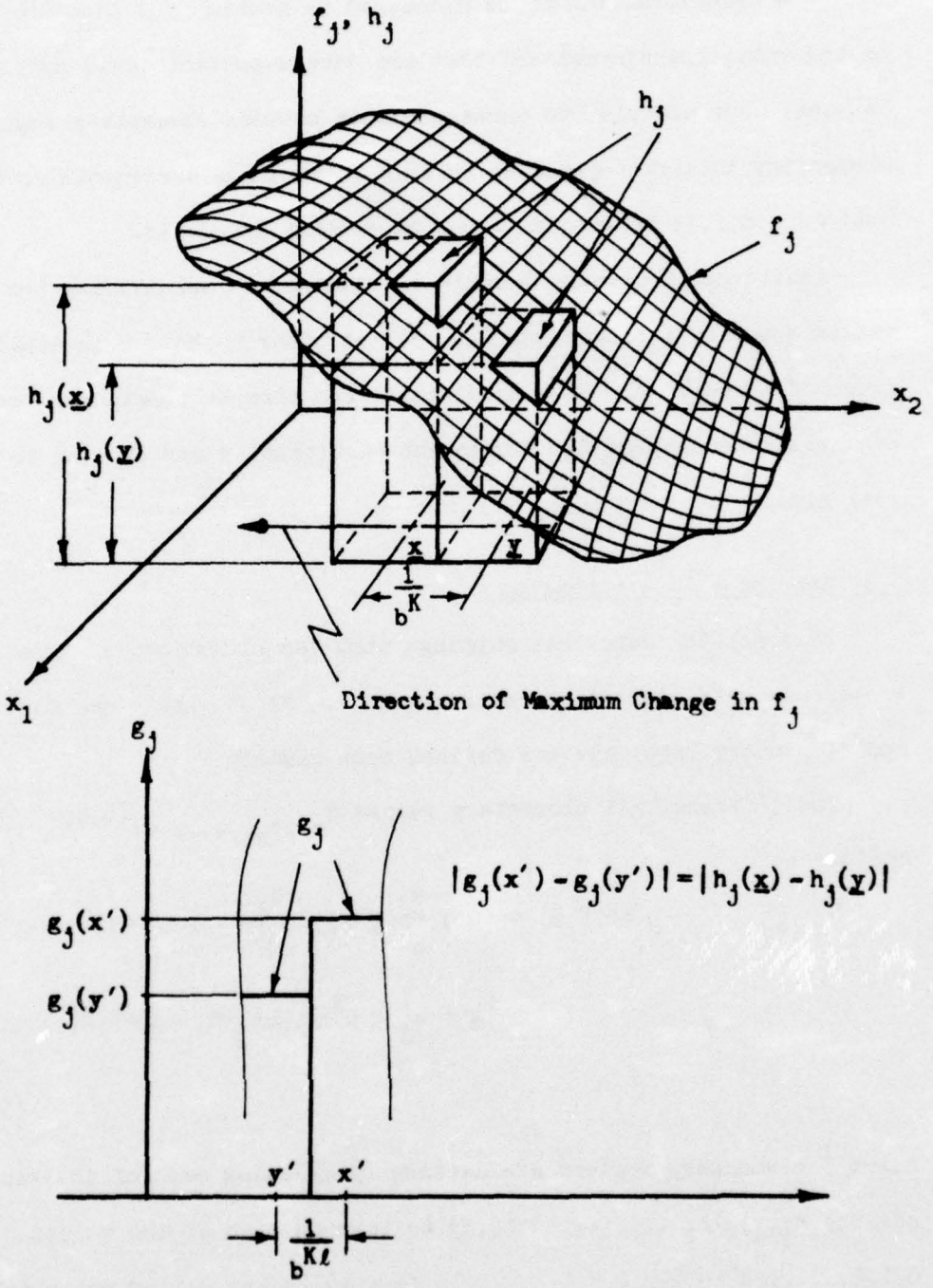


Figure 27. The Functions f_j, h_j, and g_j

The transformation to be discussed in Section 5.3 does not depend on the data; transformations that are data dependent could profitably be used. For example the correspondence between elementary regions and elementary intervals could be defined to minimize increments in the estimated d.f.'s between adjacent elementary intervals.

Additional improvement could be attained by subdividing the observation space by a clustering [23, 56] or other technique to isolate modes of the d.f.'s. Each subdivision can then be treated as the domain of a separate problem for subsequent partitioning and mapping to the real line.

5.3 Mapping to One Dimension

This section describes mappings that map elementary regions in $\mathcal{R}$ one-to-one onto elementary intervals in $\mathcal{Q}$. First, the elementary regions and elementary intervals are defined more clearly.

The ℓ-dimensional elementary region $S_{e_1,e_2,\ldots,e_\ell}(b,K,\ell)$ is defined by

$$S_{e_1,e_2,\ldots,e_\ell}(b,K,\ell) = \left\{\underline{x} : \frac{e_j}{b^K} \le x_j < \frac{e_j+1}{b^K}, \quad j=1,2,\ldots,\ell\right\}$$

$$= \left\{\underline{x} : e_j \le b^K x_j < e_j+1, \quad j=1,2,\ldots,\ell\right\} \tag{5.8}$$

All $b^{K\ell}$ elementary regions are defined by allowing each of the subscripts e_j, $j=1,\ldots,\ell$, in (5.8) to take on each of the values $0,1,2,\ldots,b^K-1$. $S_{e_1,e_2,\ldots,e_\ell}(b,K,\ell)$ is the set of all $\underline{x}$ in $\mathcal{R}$ that become identical if each of its ℓ components is expressed in the base b number system and truncated to K digits.

Similarly the elementary interval $S_e(b,K\ell)$ is defined by

$$S_e(b,K\ell) = \left\{y : \frac{e}{b^{K\ell}} \le y < \frac{e+1}{b^{K\ell}}\right\}$$

$$= \left\{y : e \le b^{K\ell}y < e + 1\right\} \qquad (5.9)$$

All $b^{K\ell}$ elementary intervals are defined by allowing the subscript e in (5.9) to take on each of the values $0,1,2, \ldots, b^{K\ell}-1$. $S_e(b,K\ell)$ is the set of all y in $\mathcal{R}$ that become identical if expressed in the base b number system and truncated to $K\ell$ digits.

Both the elementary regions and the elementary intervals are uniquely identified by their subscripts. Hence the mappings can be defined via the subscripts.

5.3.1 The Dovetail Mapping

Consider mapping the arbitrary elementary region $S_{e_1,e_2, \ldots, e_\ell}(b,K,\ell)$ to an elementary interval in $\mathcal{R}$. The base b representation of the subscript e_j, $j=1, \ldots, \ell$, is

$$e_j = \alpha_{j1}\alpha_{j2} \cdots \alpha_{jK}$$

$$= \alpha_{j1}b^{K-1} + \alpha_{j2}b^{K-2} + \ldots + \alpha_{jK}b^0 \qquad (5.10)$$

where each α_{ji}, $i=1, \ldots, K$, is one of the values $0,1, \ldots, b-1$. The Dovetail Mapping defines the corresponding subscript e by

$$e = \alpha_{11}\alpha_{21} \cdots \alpha_{\ell 1}\alpha_{12}\alpha_{22} \cdots \alpha_{\ell 2} \cdots\cdots \alpha_{1K}\alpha_{2K} \cdots \alpha_{\ell K}$$

$$= \alpha_{11}b^{K\ell-1} + \alpha_{21}b^{K\ell-2} + \alpha_{\ell K}b^{0} \qquad (5.11)$$

This mapping has been used in integration theory (see e.g. Wiener [41]). It is called the Dovetail Mapping because it interleaves or dovetails the digits of the e_j's to get e.

Example Figure 28 illustrates an example with $b=3$, $K=2$, and $\ell=2$. The ordering imposed on the elementary regions in $\mathcal{D}$ through the Dovetail Mapping by the natural ordering of the elementary intervals in $\mathcal{R}$ is illustrated with an ordering path. The dotted line portions of the ordering path denote discontinuities in the path. Several corresponding elementary regions in $\mathcal{D}$ and elementary intervals in $\mathcal{R}$ have been labeled with corresponding letters in the Figure.

5.3.2 The Column Mapping

A problem with the Dovetail Mapping is the discontinuous way in which it orders the elementary regions in $\mathcal{D}$ by the natural ordering of elementary intervals in $\mathcal{R}$. For example, in Figure 28, adjacent elementary intervals F_1 and G_1 in $\mathcal{R}$ correspond to the widely separated elementary regions F_2 and G_2 in $\mathcal{D}$. Because of the discontinuities, (5.7) cannot be used to obtain constants L_j^* for pseudo-Lipschitz conditions on g_j.

It is possible to modify the Dovetail Mapping to remove the discontinuities in its ordering path. By inverting the ordering of the elementary regions in suitably defined regions of $\mathcal{D}$, the mapping

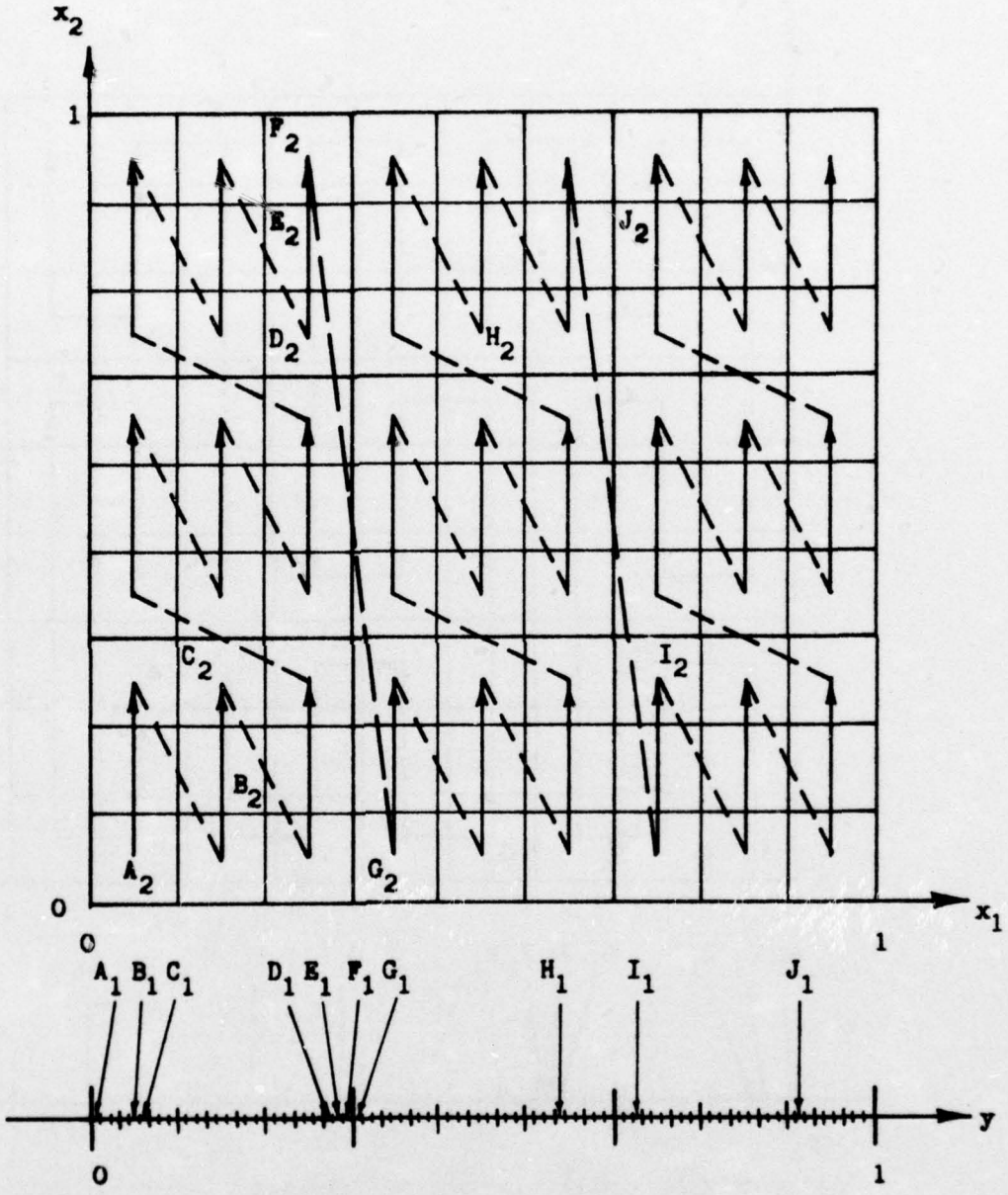


Figure 28. Dovetail Mapping for b = 3, k = 2, and ℓ = 2

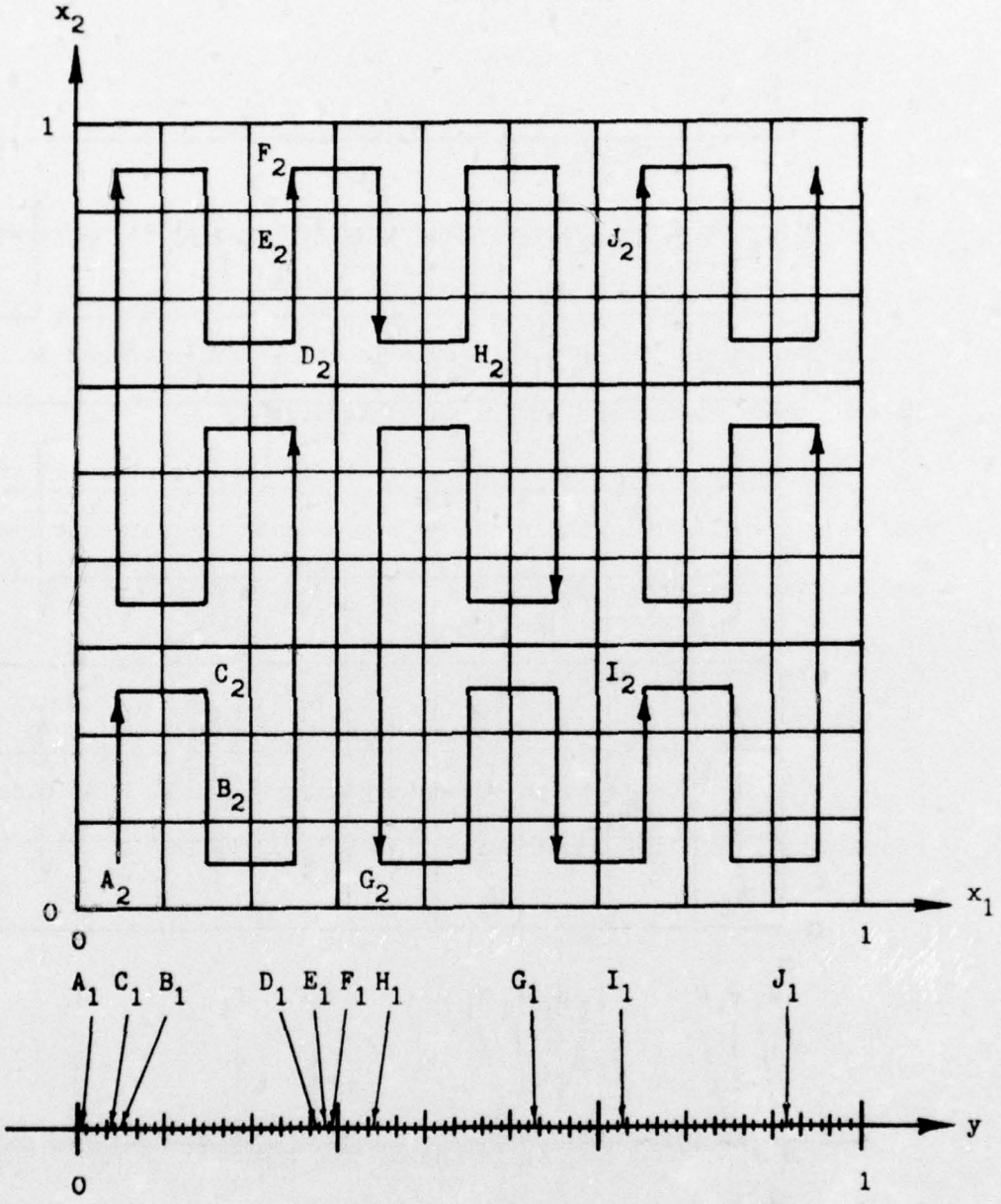


Figure 29. Column Mapping for b = 3, k = 2, and ℓ = 2

illustrated geometrically by Figure 28, for example, can be converted to the mapping illustrated by Figure 29. Note that the ordering path in Figure 29 is continuous which implies that each pair of adjacent elementary intervals in $\mathcal{R}$ corresponds to a pair of adjacent elementary regions in $\mathcal{D}$. The mapping illustrated in Figure 29 is an example of what is called here a "Column Mapping". Whereas b for the Dovetail Mapping can be any positive integer, it is restricted to be an odd positive integer for the Column Mapping. Again (5.10) is used to represent the subscripts e_j for the elementary region $S_{e_1,e_2,\ \ldots,\ e_\ell}(b,K,\ell)$. The Column Mapping is defined algebraically by writing,

$$e = \beta_{11}\beta_{21} \cdots \beta_{\ell 1}\beta_{12}\beta_{22} \cdots \beta_{\ell 2} \cdots\cdots \beta_{1K}\beta_{2K} \cdots \beta_{\ell K} \qquad (5.12)$$

Each β_{ji} is either equal to α_{ji} or $b-1-\alpha_{ji}$. Which it is depends on whether or not an ordering inversion as mentioned above is required for the region defined by those digits in (5.11) that are more significant than α_{ji}. A method for determining whether an order inversion should be made is now given.

Define

$$Q_{ji} = \sum_{m=0}^{i-1} \sum_{n=1}^{\ell} \alpha_{nm} - \sum_{m=0}^{i-1} \alpha_{jm} + \sum_{n=0}^{j-1} \alpha_{ni} \qquad (5.13)$$

$$\text{where } \alpha_{01} = \alpha_{10} = \alpha_{00} = 0$$

Q_{ji} is the sum of all α's in the following blocked in portions of

base b representations of the e_j's:

$$
\begin{array}{lll}
e_1 = & \boxed{\alpha_{11} \cdots \alpha_{1i}} & \cdots \alpha_{1K} \\
\vdots & \boxed{\vdots \qquad \vdots} & \vdots \\
e_j = & \alpha_{j1} \cdots \alpha_{ji} & \cdots \alpha_{jK} \\
\vdots & \boxed{\vdots \qquad \vdots} & \vdots \\
e_\ell = & \boxed{\alpha_{\ell 1} \cdots \alpha_{\ell i}} & \cdots \alpha_{\ell K}
\end{array}
$$

Then, β_{ji} is obtained from:

$$\beta_{ji} = \alpha_{ji} \qquad \text{if } Q_{ji} \text{ is even}$$

$$\beta_{ji} = b - 1 - \alpha_{ji} \qquad \text{if } Q_{ji} \text{ is odd} \tag{5.14}$$

The Column Mapping defines an ordering path that orders the elementary regions in $\mathcal{S}$ in exactly the same way as a curve in the sequence of curves defined by Moore [42], who shows for the 2-dimensional case that the limit curve is a space-filling curve (Peano [43]).

Example Figure 29 illustrates the Column Mapping for $b=3$, $K=2$, $\ell=2$. It shows the ordering path and pairs of corresponding sets (labeled with corresponding letters). By unbending the ordering path and carrying along with it the elementary regions in $\mathcal{S}$ through which it passes, the elementary regions are strung out in a line or column, hence the name Column Mapping.

The fact that a Column Mapping ordering path is continuous ensures that there is an adjacent pair of elementary regions in $\mathcal{A}$ corresponding to each adjacent pair of elementary intervals in R. The term "quasi-continuous"* is adopted to describe this property of the mapping (actually a property of the inverse mapping). It is this property that was required in order to convert the Lipschitz constants by using (5.7).

5.3.3 Other Mappings with the Quasi-Continuity Property

Other mappings having the quasi-continuity property can be defined. For example, the elementary regions can be ordered according to a curve in a sequence of curves giving the Hilbert realization [45] of a space filling curve. Figure 30 illustrates an ordering path that could result. In a recent paper [53] Butz has defined the Hilbert Curve Mapping algebraically for ℓ dimensions.

Starting with the Column Mapping, the dimensions can be ordered differently in different regions giving, for example, the mapping illustrated by Figure 31. For later reference, the mapping of Figure 31 is called a "modified Column Mapping."

5.3.4 A Mapping Criterion

Several mappings have been discussed. Of these, the Dovetail Mapping does not have the quasi-continuity property and is not considered further. A criterion is now suggested for use in determining which of the mappings with the quasi-continuity property is most appropriate.

---

*Butz [44] uses the term "quasi-continuous" in a similar context.

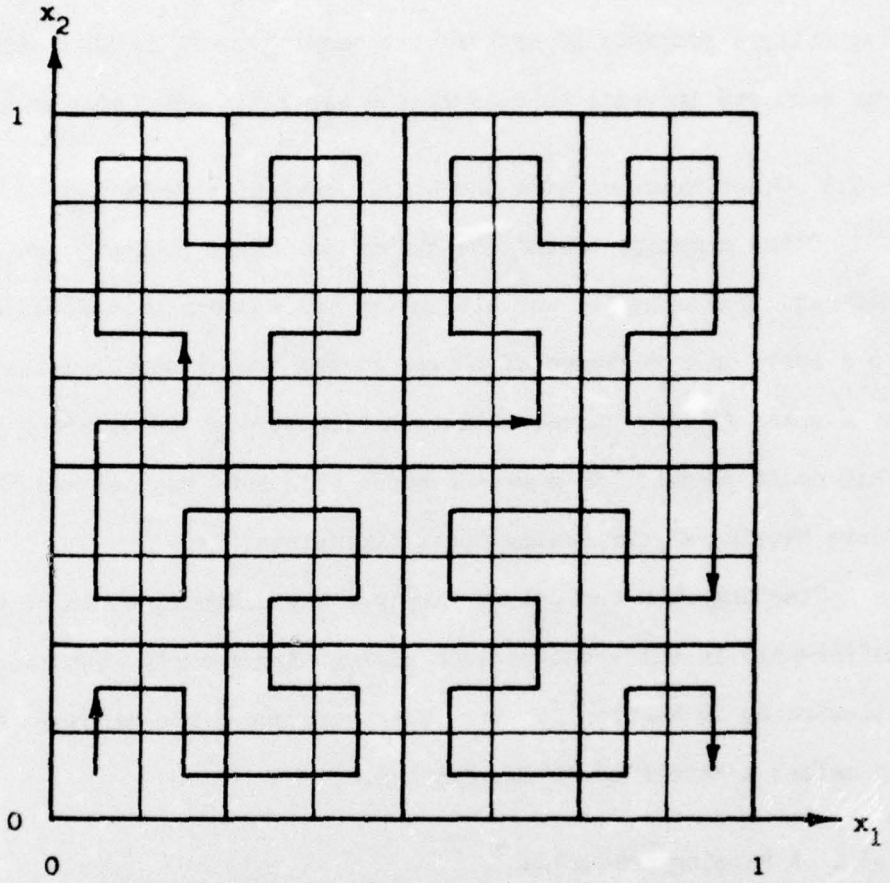


Figure 30. Hilbert Curve Mapping

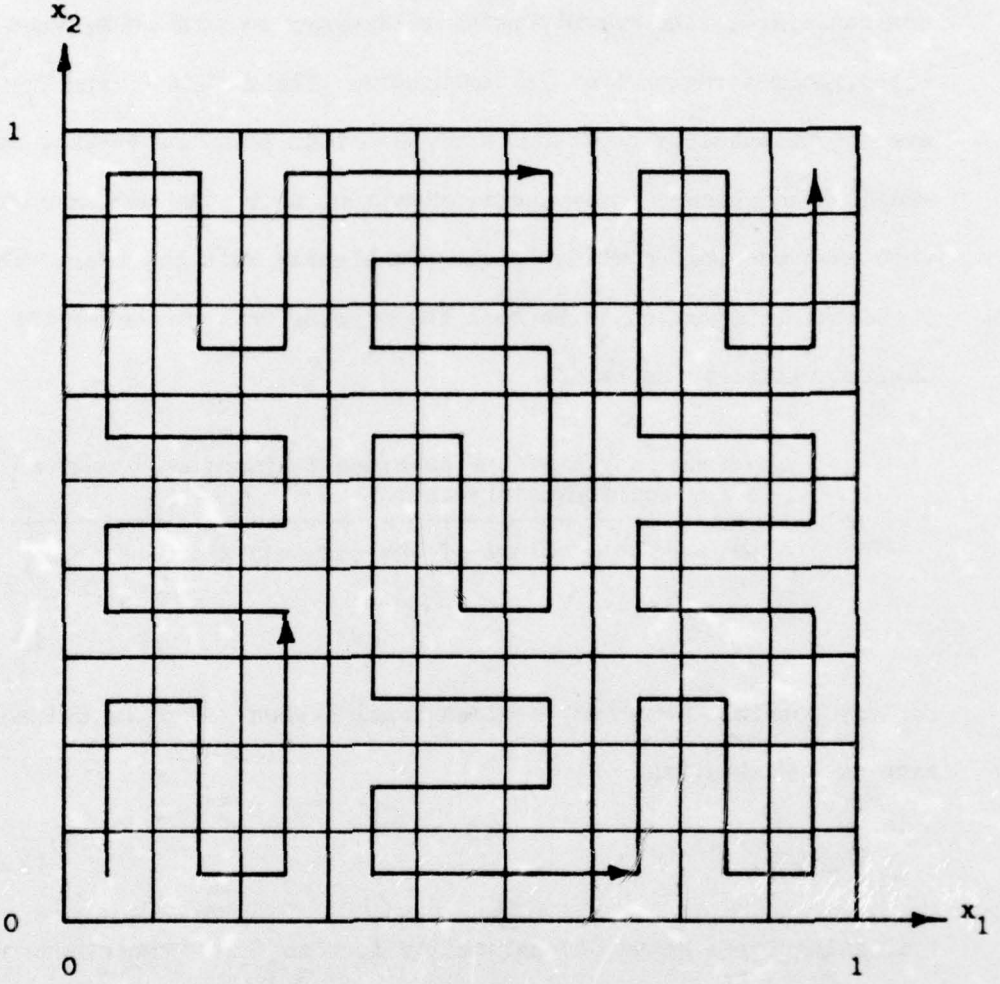


Figure 31. Modified Column Mapping

The 1-dimensional partition adjustment technique of Chapter IV automatically combines contiguous groups of elementary intervals in $\mathcal{R}$ together. Because of the quasi-continuity property of the mappings now considered, the corresponding elementary regions in $\mathcal{D}$, when combined, form a region that is contiguous. The d.f.'s f_1 and f_2 are approximated by constants over each such combined region. One would expect these constant approximations to be the most accurate when each possible combined region is tightly knit together. Then, a reasonable approach is to seek the mapping that minimizes the maximum value of the ratio

$$\Omega = \frac{\left(\begin{array}{l}\text{Maximum length of the combined } \ell\text{-dimensional region} \\ \text{in any coordinate direction.}\end{array}\right)^{\ell}}{\left(\begin{array}{l}\ell\text{-dimensional volume of the combined } \ell\text{-dimensional} \\ \text{region.}\end{array}\right)} \tag{5.15}$$

for any possible combined, ℓ-dimensional region. For the Column Mapping, Ω satisfies

$$\Omega \leq (2b)^{\ell - 1} \tag{5.16}$$

indicating for a given dimensionality ℓ, that Ω is independent of K, but that the base b should be chosen as small as possible. The smallest nontrivial odd base is 3 (recall that, for the Column Mapping, b must be odd). For this reason only base 3 is considered further for use with the Column Mapping. With $b = 3$ and $\ell = 2$ in (5.16), Ω is bounded by 6. The worst case for the Hilbert Curve Mapping of

Figure 30 would be the case in which four elementary regions in a line are combined. Ω from (5.15) is then bounded by 4. Similarly, Ω for the Modified Column Mapping of Figure 31 is bounded by the value 5.4. The examples discussed later in this chapter all use the Column Mapping. However, it is apparent that, based on the ratio Ω, the Hilbert Curve Mapping and the Modified Column Mapping merit further study.

5.4 Computer Simulated Results

To demonstrate the extension to multidimensions, several two-dimensional examples using the Column Mapping are presented. The mapping parameters for each example are $b = 3$, $K = 3$, and $\ell = 2$, giving the ordering path illustrated in Figure 32 .

5.4.1 Examples

The examples all use class-conditional d.f.'s that are either Gaussian or linear combinations of Gaussian d.f.'s. Though the procedure does <u>not</u> require Gaussian data, such data is easy to generate on the computer and, with linear combinations of Gaussian d.f.'s, is felt to represent, as well as any data, the type of problems to be handled. Table 1 lists the weighting coefficients, means and covariance matrices used for the components of the linear combination in each example. Any observation falling outside the domain $\mathcal{D}$ is rejected and a new one obtained. This truncation effect is minimal for the examples because of the placement of the component d.f.'s well within the boundaries of $\mathcal{D}$. A priori probabilities are assumed to be 0.5. The goal is to satisfy condition (1.6) with $\alpha = 0.1$ and $\beta = 0.9$. Figures 33

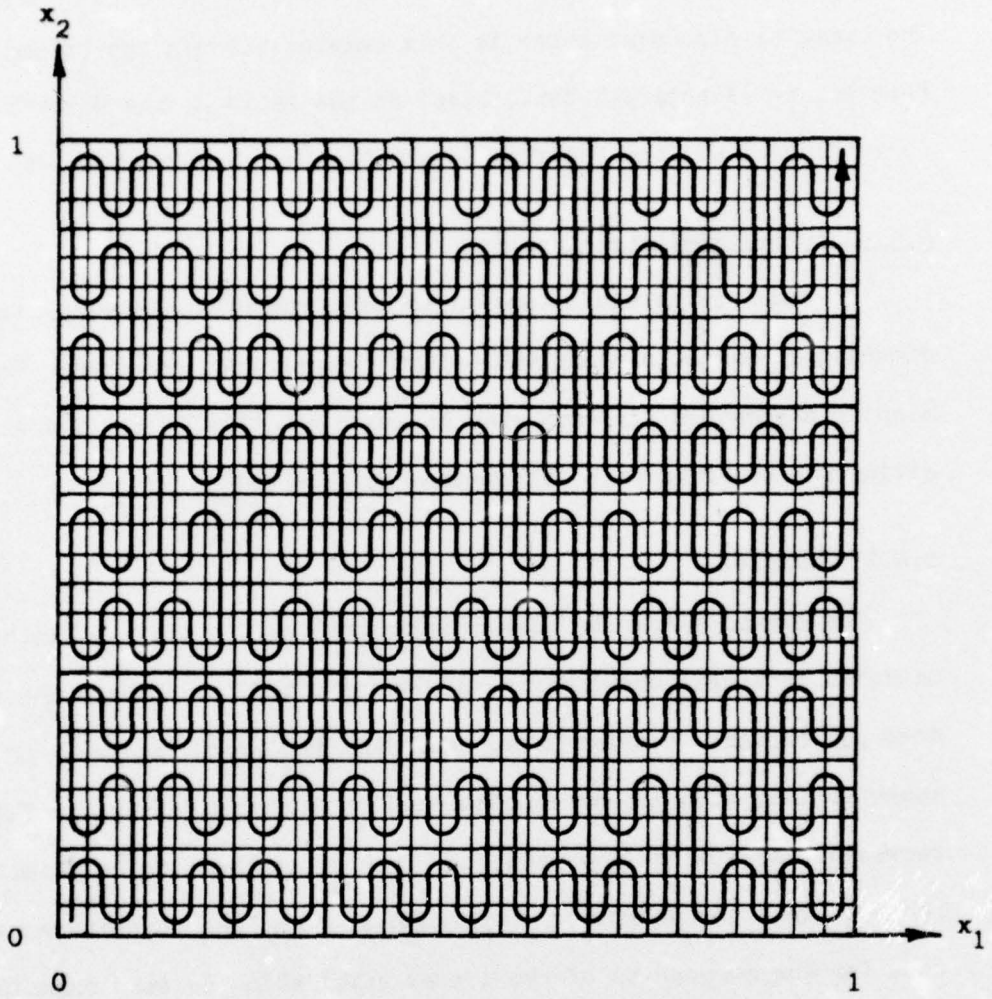


Figure 32. Ordering Path for Examples

Table 1. Definition of Examples

Example	Class ω_1			Class ω_2		
	Weighting Coef's	Means	Covariance Matrices	Weighting Coef's	Means	Covariance Matrices
1	1.0	0.35 0.50	0.01 0 0 0.01	1.0	0.65 0.50	0.01 0 0 0.01
2	1.0	0.50 0.35	0.01 0 0 0.01	1.0	0.50 0.65	0.01 0 0 0.01
3	1.0	0.394 0.394	0.01 0 0 0.01	1.0	0.606 0.606	0.01 0 0 0.01
4	1.0	0.394 0.606	0.01 0 0 0.01	1.0	0.606 0.394	0.01 0 0 0.01
5	1.0	0.50 0.50	0.01 0 0 0.01	1.0	0.50 0.50	0.04 0 0 0.04
6	1.0	0.50 0.50	0.0025 0 0 0.0025	1.0	0.50 0.50	0.04 0 0 0.04
7	Mode 1 0.5	0.20 0.50	0.0025 0 0 0.0025	Mode 1 0.5	0.40 0.50	0.0025 0 0 0.0025
	Mode 2 0.5	0.60 0.50	0.0025 0 0 0.0025	Mode 2 0.5	0.80 0.50	0.0025 0 0 0.0025
8	Mode 1 0.5	0.50 0.20	0.0025 0 0 0.0025	Mode 1 0.5	0.50 0.40	0.0025 0 0 0.0025
	Mode 2 0.5	0.50 0.60	0.0025 0 0 0.0025	Mode 2 0.5	0.50 0.80	0.0025 0 0 0.0025
9	Mode 1 0.5	0.35 0.35	0.01 0 0 0.01	Mode 1 0.5	0.35 0.65	0.01 0 0 0.01
	Mode 2 0.5	0.65 0.65	0.01 0 0 0.01	Mode 2 0.5	0.65 0.35	0.01 0 0 0.01
10	Mode 1 0.5	0.25 0.25	0.01 0 0 0.01	Mode 1 0.5	0.25 0.75	0.01 0 0 0.01
	Mode 2 0.5	0.75 0.75	0.01 0 0 0.01	Mode 2 0.5	0.75 0.25	0.01 0 0 0.01

through 42 illustrate the resulting assignments of regions to the two classes. A circle with radius one standard deviation is drawn about the center of each component to facilitate visualization of the results.

Figures 33 through 36 for Examples 1 through 4 show that the procedure succeeds for different arrangements of the class-conditional d.f.'s. Separation of the means in each case is three times the standard deviation. Figures 37 and 38 for Examples 5 and 6 illustrate the capability of the procedure to separate the space based solely on the dispersion of the distributions. Figures 39 through 42 for Examples 7 through 10 portray some bimodal results for cases in which regions assigned to the two classes are interleaved.

Similar to the one-dimensional case, it is found that significantly fewer training observations are required when smaller Lipschitz constants are used. For the examples illustrated, the constants L_j^* are approximately one-tenth the values computed by (5.7) from the smallest applicable Lipschitz constants for the d.f.'s. The effect on the boundary of using the smaller constants is not serious for these examples. However, it must be noted that an assumption of the problem is violated, and attainment of the goal is not verified. With large constants L_j^*, the solution is generally obtained only after very many training observations; however, it is observed that tentative classification of the domain usually settles down quickly to a reasonable result. Thus, another way to make the procedure more nearly practical is to

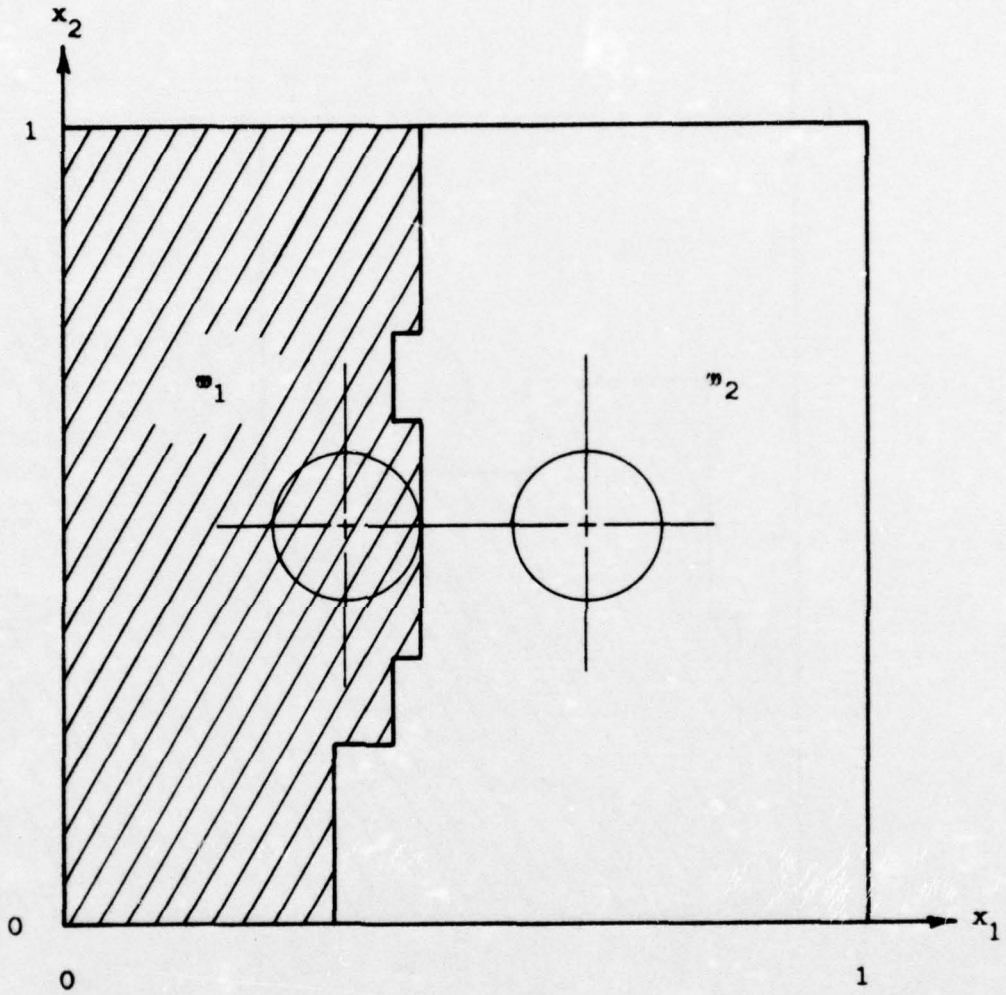


Figure 33. Example 1

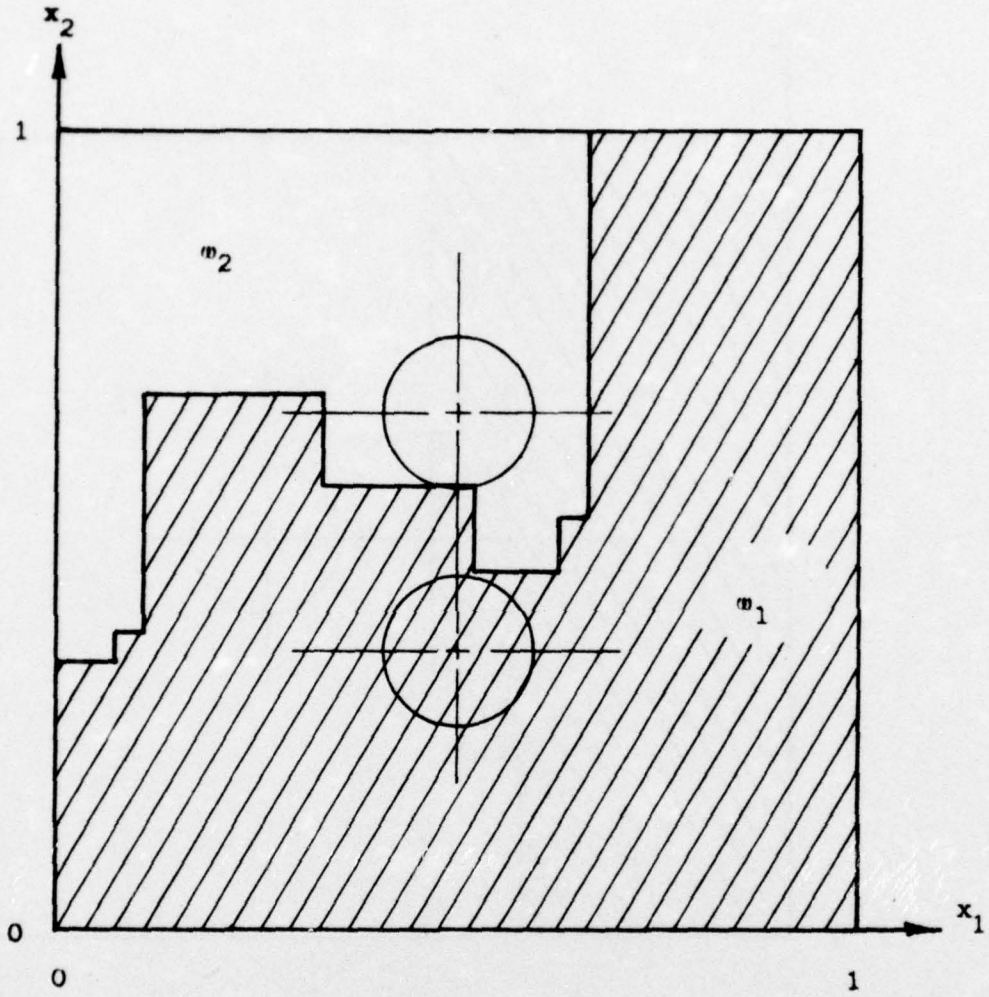


Figure 34. Example 2

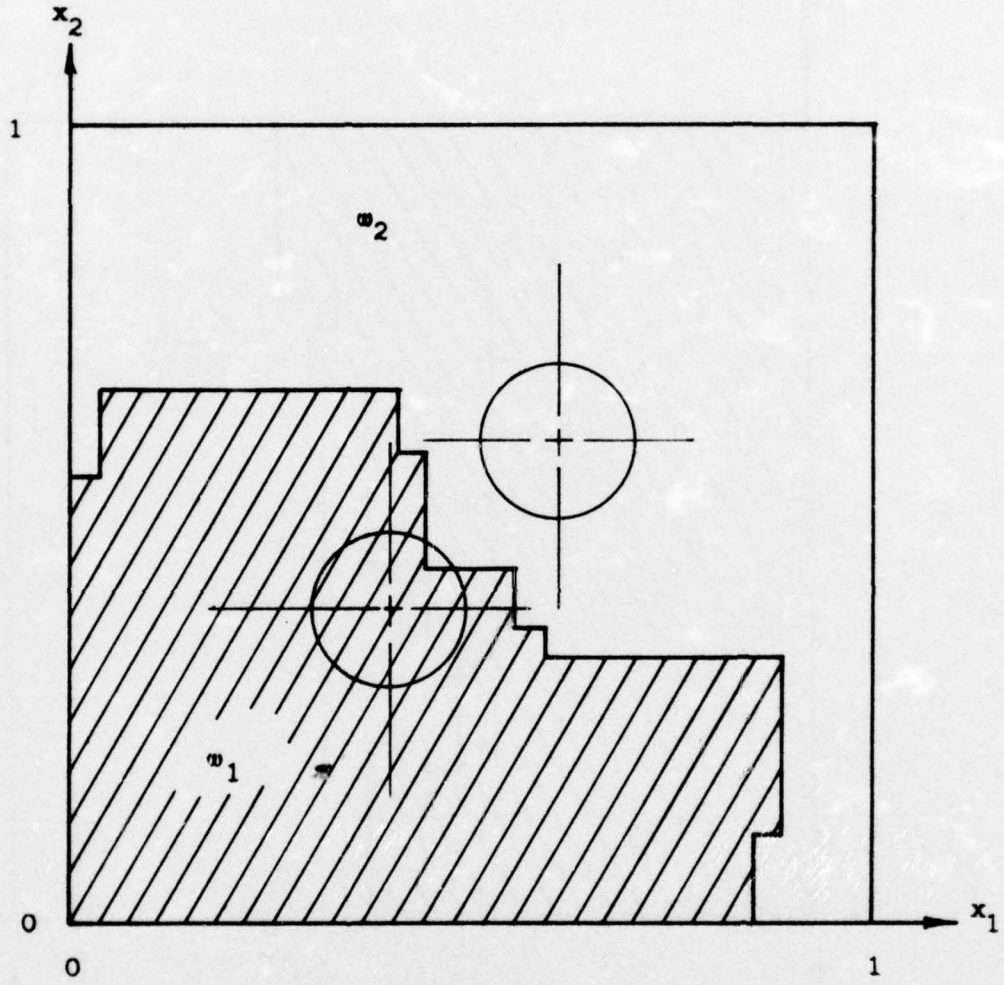


Figure 35. Example 3

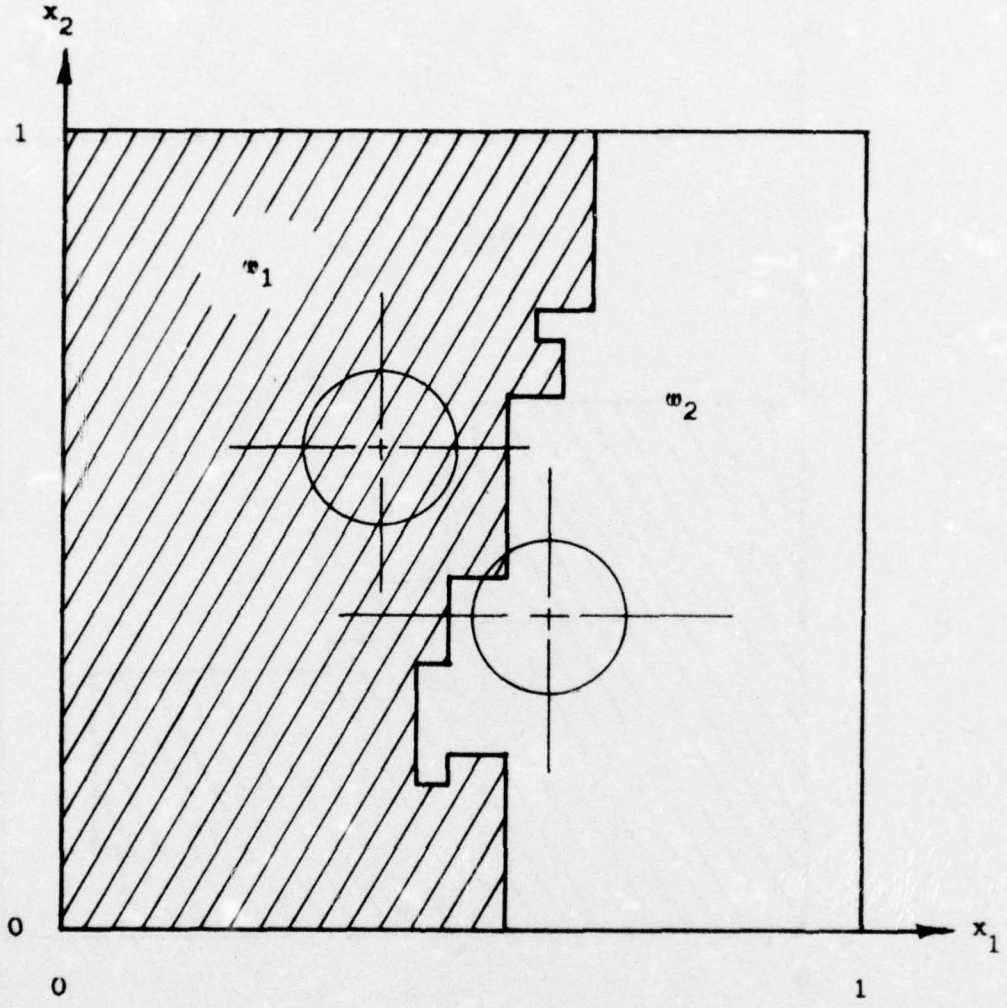


Figure 36. Example 4

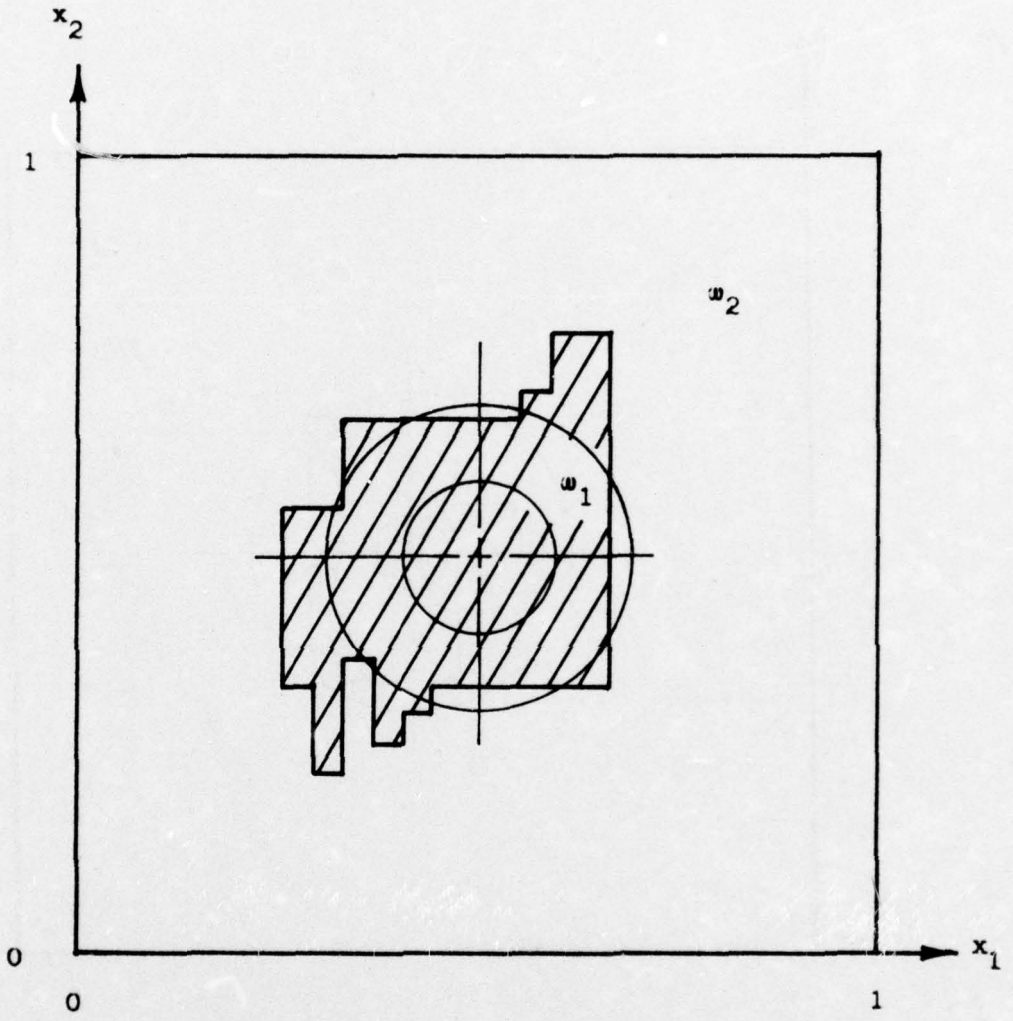


Figure 37. Example 5

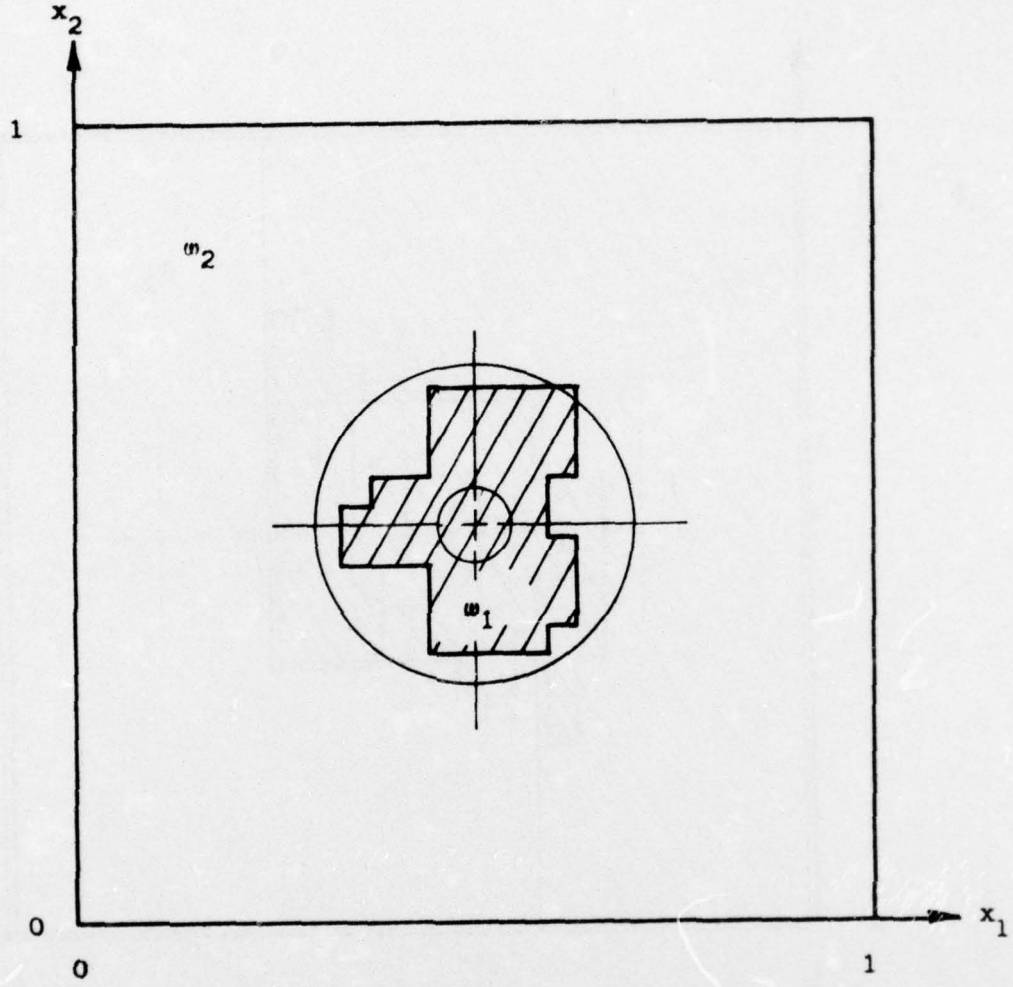


Figure 38. Example 6

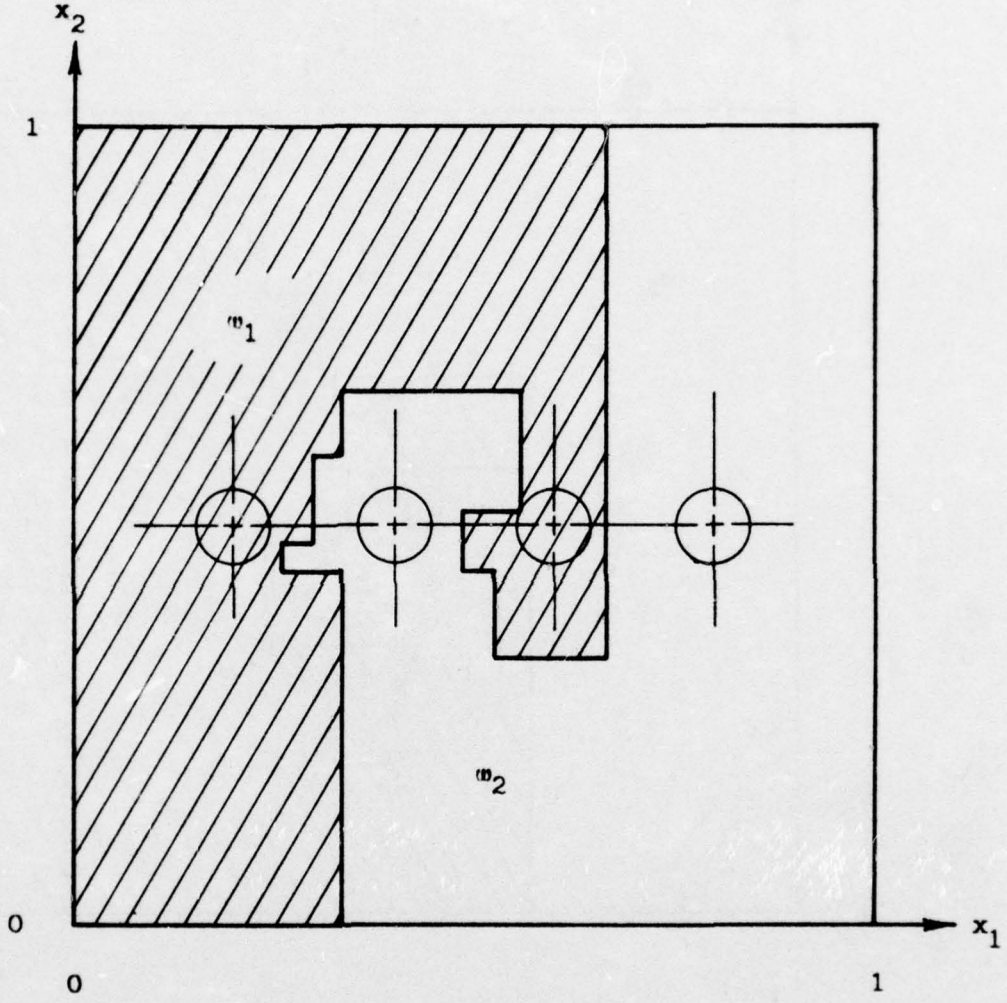


Figure 39. Example 7

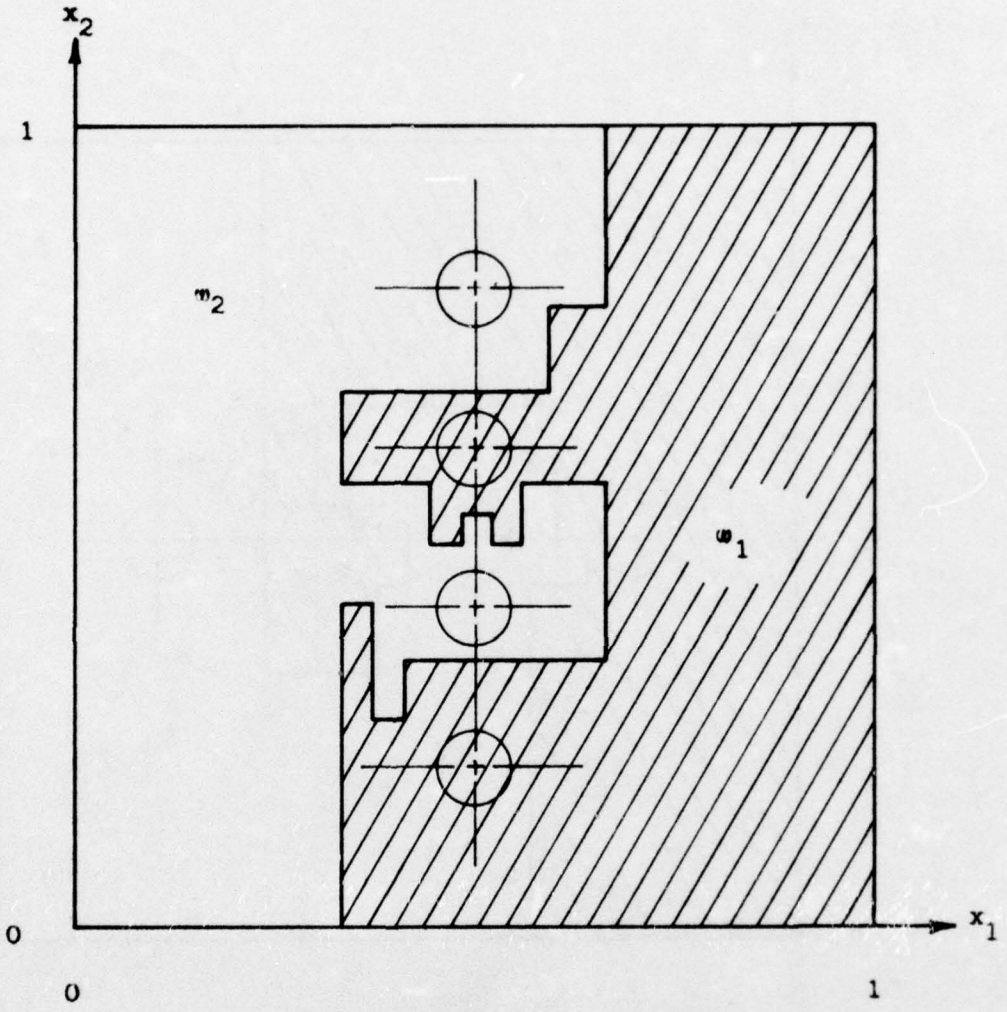


Figure 40. Example 8

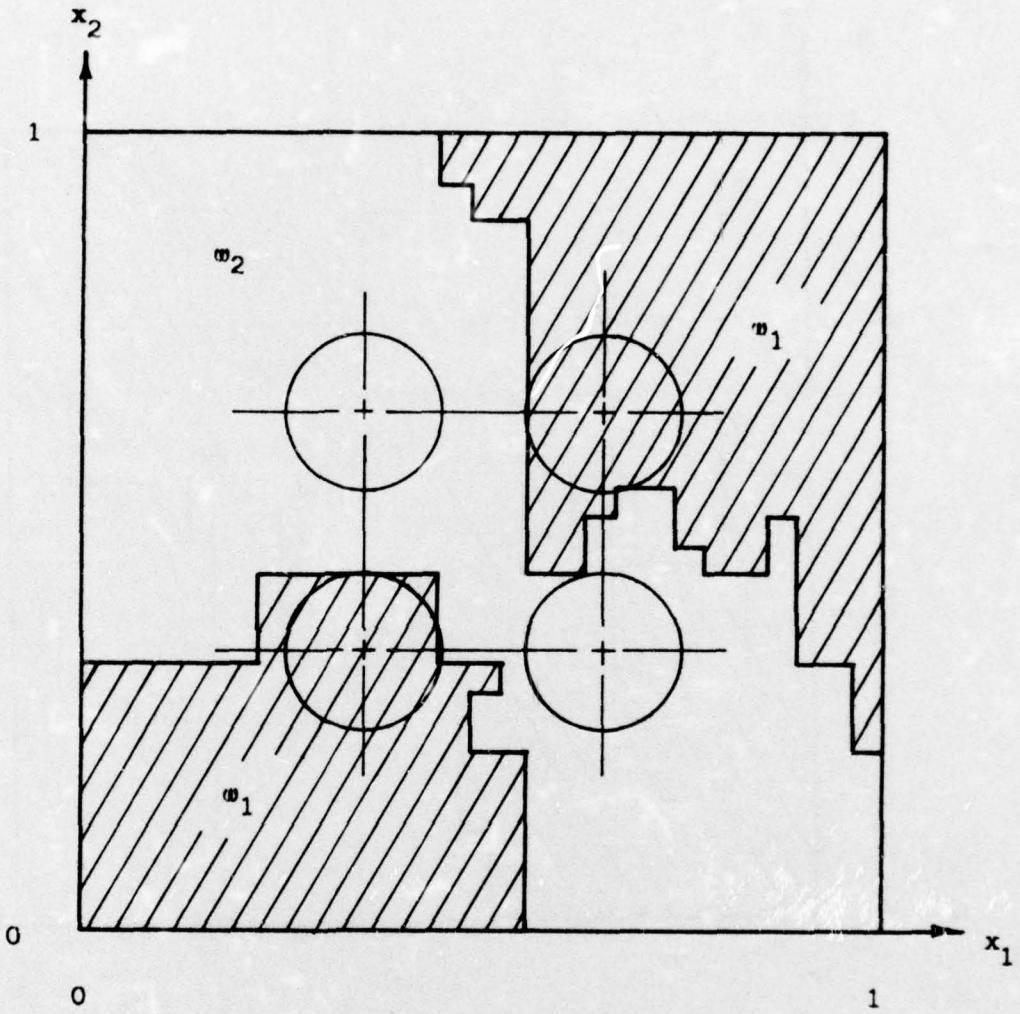


Figure 41. Example 9

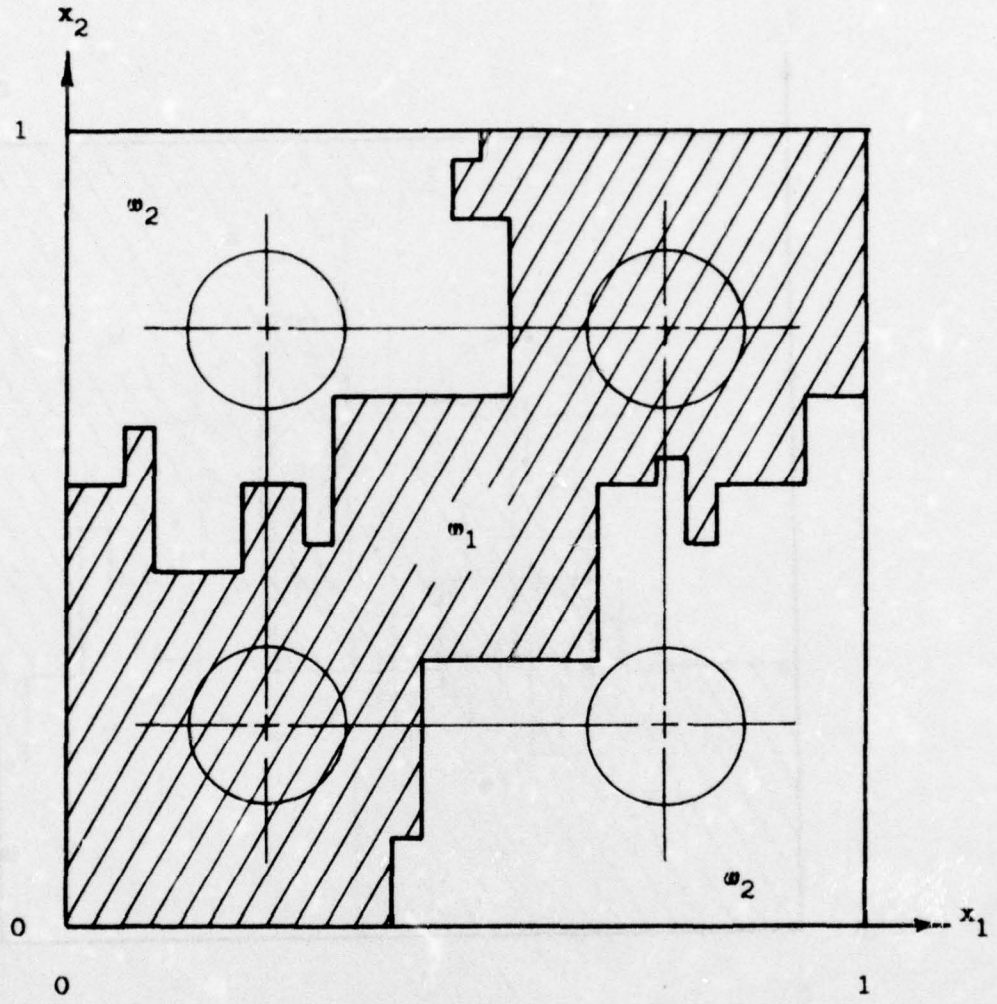


Figure 42. Example 10

use the tentative results when results are needed but to allow the system to continually process additional incoming training observations until the final solution is obtained. If large enough constants L_j^* have been used, then the final result can be trusted.

For these examples, the procedure terminates after approximately 1000 training observations. The maximum number of intervals allowed is 15 for the unimodal d.f. problems and 20 for those with bimodal d.f.'s.

5.4.2 Computational Aspects

As the complexity of the problem increases, that is, as the $K\ell$ product increases, some computational problems appear. For example, the IBM 1130 computer system used for the examples maintains accuracy to about five significant decimal digits. So long as the real number representation of an interval boundary* needs no more than five decimal digits accuracy, the ordinary arithmetic operations and storage techniques provided with the computer system can be employed. Five decimal digits corresponds roughly to ten ternary digits; thus, if a mapping using base 3 is employed, the $K\ell$ product is limited to about ten with ordinary operations of the IBM 1130 system. This corresponds at one extreme to a ten-dimensional problem with each dimension partitioned into three intervals, and at the other extreme, to a

---

*An interval boundary is identified by the elementary interval immediately to its right.

two-dimensional problem with each dimension partitioned into 243 intervals. For either case (or any intermediate case) an interval boundary (corresponding to a mapped region boundary) can be stored as an ordinary real number.

When $K\ell > 10$, other techniques must be employed. One approach is to employ a computer system with more storage in each computer word; however, at some critical $K\ell$ product for a given base b, the problem reappears. Another approach is to provide for the storage of each boundary in several words of storage. Such extended precision requires programs to handle the arithmetic operations involved. Increased computer time as well as increased storage (for the multiword interval boundaries) results. For the examples handled in this report, one word per boundary is used. All variables and the entire program are contained in the 16000, 16-bit word, main storage of the IBM 1130 computer system. Processing time for each of the two-dimensional examples is approximately five minutes.

5.5 Other Uses for the Mappings

This section briefly discusses other uses for the mappings described in Section 5.3.

5.5.1 Display of Real-Valued Functions

A real-valued function of more than one real variable is difficult to observe. The two-dimensional display surfaces generally used have the capability of displaying such a function defined on no more than one variable. When the domain is greater

than one-dimensional, various projections and sectional views can be used to gain a perspective of the function. Another approach is to first map the multidimensional domain to one dimension via one of the mappings described in Section 5.3. Then the function's one-dimensional equivalent can be displayed on a two-dimensional surface [46]. Figure 43 illustrates the resulting display for a bivariate Gaussian d.f. given by

$$f(\underline{x}) = \frac{1}{2\pi(0.25)^2} \exp\left[-\tfrac{1}{2} \sum_{i=1}^{2} \left(\frac{x_i - 0.5}{0.25}\right)^2\right]$$

where the mapping used is the Column Mapping with b = 3, K = 3, and $\ell = 2$.

Unless one is accustomed to observing bivariate Gaussian d.f.'s in the form displayed by Figure 43, the function represented there probably is unrecognizable as a transformed Gaussian d.f. For purposes of recognizing functions, the display has little value. It is for purposes of comparing functions that such a display can profitably be used. One application is to display the difference of two d.f.'s in order to get an idea of what has been called the "separability" of the two functions. Two d.f.'s are highly separable if the d.f. generating an observation can be identified from the observation's location in $\mathcal{A}$ with small probability of error.

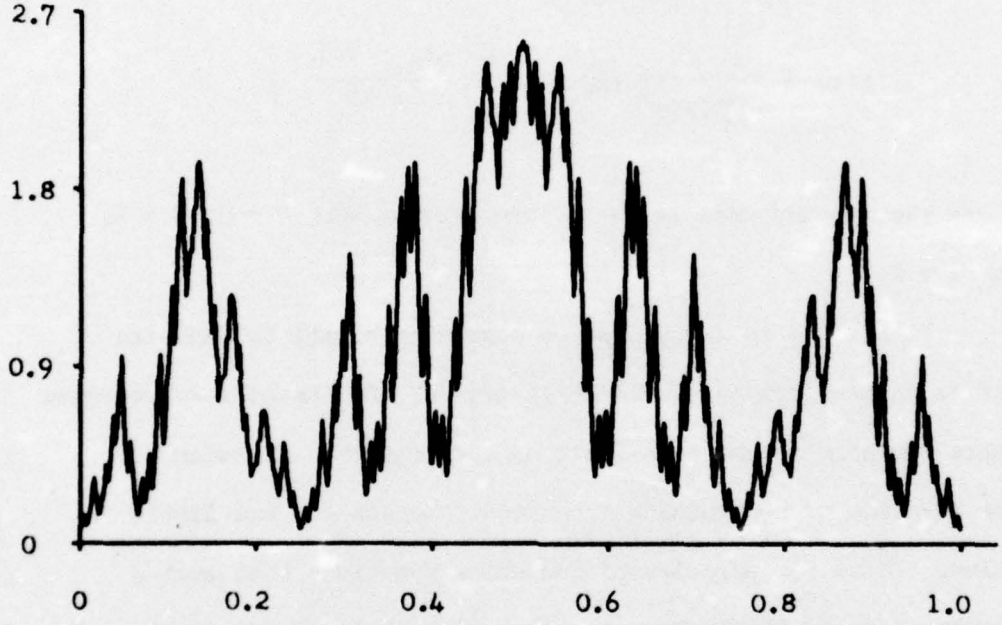


Figure 43. A Mapped Bivariate Gaussian Density Function

5.5.2 Parameter Sensitivity Studies

A display such as described in 5.5.1 can be used for parameter sensitivity studies. Suppose a system has a real-valued function output depending on M real variables $\alpha_1, \alpha_2, \ldots, \alpha_M$, as illustrated in Figure 44.

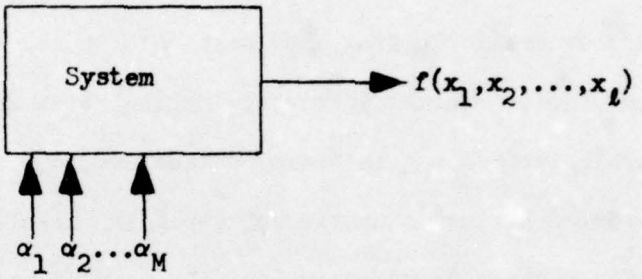


Figure 44. Study of a System's Input Parameters

Suppose that a known setting of these parameters produces a desired function $f(\cdot, \cdot, \ldots, \cdot)$. A problem is to adjust for a cheaper set of parameters without significantly degrading the function. The difference between the desired output and the output with adjusted parameters can be continually monitored as the parameters are adjusted. Such a use might not require the resolution of the difference function to be very large. In that case the result might be mapped back up to two dimensions

via the inverse* of a mapping with the "quasi-continuity" property, and the difference function displayed as intensity modulation.

The use of a color display can further enhance the usefulness of the mappings if quick interaction from the operator is desired. For example, while intensity is used to portray the difference function, color can be used to identify the region in $\mathcal{A}$ corresponding to any point on the display. Although this information is already available from the location of the point on the display, color enables its determination to be made more quickly. If particular regions in $\mathcal{A}$ are of interest, different colors can be reserved for use at their corresponding display points.

For another example consider the problem of representation of a d.f. as a linear combination of Gaussian d.f.'s. Suppose that an acceptable representation (perhaps from a histogram or some other estimation procedure) has been obtained, but that the number of parameters used is impracticably large. The difference between the acceptable representation and the linear combination of Gaussian d.f.'s can be mapped to the real line. A viewer controlling the parameters describing the linear representation can interact with the displayed result to find a set of parameters giving suitable agreement between the two representations. The representation problem just described also occurs in unsupervised estimation problems.

---

*The same formula for the inverse can be used as for the mapping itself. That is, the β's derived from the α's per (5.13) and (5.14) can be themselves processed by (5.13) and (5.14) as if they were the α's to get γ's. The resulting γ's are the original α's.

5.5.3 Data Reduction

Many data reduction schemes operate on real-valued functions of one real variable. They use techniques to reduce the redundancy in the function so that it can be represented with as few parameters as possible. The case in which the domain is multidimensional can be handled by mapping the domain to one dimension. For example, the television camera with its raster scan reduces a function of intensity on two dimensions to a function on one dimension*. Because each line in the raster traverses from one side of the picture to the other, the function cannot generally be well approximated by a constant for the length of the line. However, if the line were to wander around in a more or less tightly knit region such that the same area of the picture is covered, it is reasonable to assume that fewer changes in intensity will be encountered and hence a better chance for a satisfactory constant approximation exists. Using the same argument throughout the space leads to the conclusion that a mapping such as the Column Mapping can give a function of one variable that is generally characterizable with fewer parameters (at least when using a piecewise constant representation) than the ordinary raster scan type mapping. Hence, such mappings can be considered for use with data reduction schemes. Abend, Harley,

---

*Note that the conventional television raster scan is, except for the interleaving feature, a special case of the Dovetail Mapping of Section 5.3.1.

and Kanal [47] have considered the Hilbert Curve Mapping to account for spatial dependencies of random variables along the ordering path.

For purposes of data reduction, it may be advantageous to alter the mappings in a way that depends on the data; that is, to make the mappings interactive with the data. For example, instead of modifying the Column Mapping of Figure 29, by re-ordering the dimensions in different regions to obtain Figure 31, the ordering of the dimensions in a region could be made to depend on the function in that region. For the two-dimensional case (pictures), both orderings of the dimensions can be considered, and the one best satisfying some criterion, e.g. smoothness of the function along the resulting ordering path, can be chosen for the mapping.

5.5.4 Scanning for Regions with Specified Function Values

Butz [44] considers what he calls a "Finite Peano Mapping" which is essentially the inverse to the Column Mapping for base 3. From knowledge of properties of a function f defined on the domain $\mathcal{D}$, he searches for regions satisfying $f(\underline{x}) \leq 0$. Butz derives numerical bounds describing the quasi-continuity of the mapping. Then, from properties assumed satisfied by function f, an "implicitly exhaustive search" procedure can be used to find regions in $\mathcal{D}$ for which $f(\underline{x}) \leq 0$. Butz calls the search implicitly exhaustive because every point is accounted for without making computations at every point.

5.6 Other Extensions to Multidimensions

All the described mappings map regions in $\mathcal{R}$ one-to-one onto intervals in $\mathcal{R}$. Such mappings provide a way to extend the recognition techniques of previous chapters to multidimensions. In addition, they have other uses as discussed in Section 5.5.

Other mappings with the general purpose of reducing the dimensionality of data vectors have been defined in the literature. When the dimensionality is reduced to one, it is reasonable to consider these mappings as the means to extend the current work to multidimensions.

Mappings that operate only on the observations have been considered by Shepard and Carroll [48]. They map the set of n, ℓ-dimensional observations $\{\underline{y}_i\}_{i=1}^{n}$ to a set of n, ℓ'-dimensional observations $\{\underline{x}_i\}_{i=1}^{n}$ where $\ell' < \ell$. They strive to obtain this mapping so that an index

$$k = \frac{\sum\sum_{i \neq j} \frac{d_{ij}^2}{D_{ij}^2} W_{ij}}{C} \tag{5.17}$$

that measures continuity inversely is minimized. In (5.17) d_{ij} and D_{ij} are distances between the i^{th} and j^{th} observations as measured in the ℓ-dimensional and the ℓ'-dimensional spaces respectively. That is

$$d_{ij}^2 = \sum_{\xi=1}^{\ell} (y_{i\xi} - y_{i\xi})^2$$

$$D_{ij}^2 = \sum_{\xi=1}^{\ell'} (x_{i\xi} - x_{j\xi})^2 \qquad (5.18)$$

where $y_{j\xi}$ is the ξ^{th} component of observation $\underline{y}_i = (y_{i1},\ldots,y_{i\ell})$ and $x_{i\xi}$ is the ξ^{th} component of observation $\underline{x}_i = (x_{i1},\ldots,x_{i\ell'})$. In (5.17), W_{ij} given by

$$W_{ij} = \frac{1}{D_{ij}^2} \qquad (5.19)$$

is included to weigh the effect of the relation between the i^{th} and j^{th} observations less as the mapped distance between them is increased. The denominator C given by

$$C = \left[\sum_{i \neq j} \sum \frac{1}{D_{ij}^2}\right]^2 \qquad (5.20)$$

is included for normalization purposes. Without it, k could be made as small as desired by making each D_{ij} large. References [49,50] also consider mappings of this type.

Another mapping that maps only the observations is the "Chain Mapping" [23]. The Chain Mapping considers the observations sequentially—the next member in the sequence is the nearest neighbor to the current observation. An observation

is mapped from V^{ℓ} to the real line such that the distance to the previous member in the sequence is preserved.

An important use for mappings that operate only on the observations is to reduce their dimensionality so that they can be displayed. If preserved by the mapping, clustering information and other relations among the data can be learned visually from the display. Because of the inability to observe the data in the original multidimensional space, these relations could go unnoticed without the mapping. Applications include problems in radar and sonar. For example, signals from targets can be converted to ℓ-dimensional vectors, mapped to lower dimensional vectors, and observed. If the mapping has preserved cluster relationships, it may be possible to separate the data into two groups. Naming one group warheads and the other decoys could occur with additional information such as knowledge of the ratio of warheads to decoys.

A disadvantage of such mappings for the current work is the fact that they map only the observations and do not treat the rest of the space. The mapping of additional observations is handled by reprocessing the whole set with the additional ones appended. A way to avoid this problem would be to process just once an appropriate sized subset of the observations. The mapping at other points in the space could be defined by using an interpolation procedure. For example, a vector could be mapped to the real line so that the ratio of distances from the vector to its two nearest neighbors in the ℓ- dimensional

space is preserved and so that the mapped vector lies between the mapped nearest neighbors. The approach can be used for the current work but requires assumed Lipschitz conditions on d.f.'s for the mapped observations.

Another mapping that can be used is one proposed by Patrick and Fischer [51]. It is a linear transformation from the ℓ-dimensional to the ℓ'-dimensional space. The transformation is chosen to maximize a measure of separability between an estimated d.f. on transformed Class ω_1 observations and an estimated d.f. on transformed Class ω_2 observations. The d.f. estimates are of the Parzen [11] type. The measure of separability between these estimates is defined to be the square root of the integral of their difference squared. This mapping resembles the mapping proposed by Shepard and Carroll provided their mapping is first extended to the whole domain via an interpolation procedure. Both approaches depend only on the original training observations. Important differences are that Patrick and Fischer's mapping is linear and maximizes a measure of separability, whereas Shepard and Carroll's mapping is non-linear and minimizes an index that measures continuity inversely.

CHAPTER VI

CONCLUSIONS

6.1 Summary of Results

A procedure is described for determining a decision rule d for the 1-dimensional, 2 class, nonparametric, recognition problem with unknown class-conditional density functions. A priori probabilities are known, and the density functions are assumed to satisfy Lipschitz conditions with known Lipschitz constants. The procedure allows the achievement of a specified confidence that the probability of a recognition error when using d is within a specified constant of the minimum attainable probability of recognition error. A fixed storage constraint is imposed.

Histogram estimates of the unknown density functions using a R-interval partition of the domain are obtained from a sequence of training observations. These estimates are used to define the decision rule d. The specified confidence is achieved by achieving a similar confidence for each interval in the partition. During training, the partition (always restricted to R intervals or less) is altered in an effort to improve a measure of performance. Histogram estimation of the density functions procedes based on the new partition but makes use of information obtained while using the old one. A proposition presents requirements to achieve a specified

confidence for a fixed partition; however, there are no theoretical results showing achievement of the confidence when the partition is adjusted. Experimental results for several one-dimensional examples are presented that demonstrate achievement of the desired confidence when the partition is adjusted with R selected somewhat larger than the number of decision thresholds. These experimental results use training observations from densities that are linear combinations of Gaussian densities. The results indicate that Lipschitz constants smaller than the minimum applicable values give improved performance (a decrease in the number of training observations required without increasing the probability of error above acceptable limits). The explanation is that the density functions of the examples satisfy Lipschitz conditions with smaller constants in some intervals than in others. This suggests the possibility of supplying different Lipschitz constants for different intervals, perhaps through an operator interacting with a histogram display or automatically by an estimation technique. The recognition results would then be based on assumptions of the density functions satisfying "local" Lipschitz conditions with the supplied constants.

Extension of the procedure to ℓ-dimensional observation vectors is achieved using a transformation; this transformation maps elementary regions in a partition of the ℓ-dimensional observation space one-to-one onto elementary intervals in a partition of a one-dimensional domain. One-dimensional mapped versions of the ℓ-dimensional training observations are then used in the one-dimensional procedure.

The recognition procedure may have engineering application to problems for which storage is limited but many training observations are available. One possible example is a recognition system built for long life space vehicles which are weight and hence storage limited.

6.2 Extensions

Assumed Lipschitz conditions on the density functions f_j allow bands of uncertainty to be placed about the averages of the functions in each interval. The bands are statistically described by distributions on the averages where the distributions are obtained from training observations. The classification procedure of Chapter II uses these statistically described bands.

Statistically described bands of uncertainty can be obtained using a priori knowledge other than Lipschitz conditions on the density functions. For example, one could directly assume bands of uncertainty about the averages in each interval. More generally, one could assume bands of uncertainty about the approximation

$$\hat{f}_j = \sum_{t=1}^{m} C_{jt} \phi_t$$

in an interval involving more terms than just the average of f_j. The functions ϕ_t for simplicity would be orthonormal. The bands of uncertainty would be described statistically by distributions on the parameters $\{C_{jt}\}$ where the distributions are obtained from the training observations.

A priori knowledge consisting of bounds S_j on the variation of the d.f.'s has been considered. The variation is intuitively appealing because it provides a measure of the absolute value of the derivative averaged over the domain. In addition such knowledge allows for discontinuous d.f.'s. Instead of computing the band of uncertainty by $|\bar{f}_j - f_j(x)| \leq L_j W/2$ for an interval, the band can be computed by $|\bar{f}_j - f_j(x)| \leq S_j$. For example, d.f.'s used in the experimental work of Chapter IV have maximum derivatives equal to 25 but variation equal to 8. However, the band of uncertainty computed from the variation does not decrease with interval width as required for the classification of some intervals. Thus the procedure could not in general use bounds on the variation of f_j for classification of all intervals. Such bounds could profitably be used in those intervals for which it is known that $S_j \leq L_j W/2$.

The recognition procedure is extended to ℓ dimensions via a transformation that essentially converts the ℓ-dimensional problem into a one-dimensional problem. The transformation approach is desirable because

1) The one-dimensional techniques can be used.
2) A partition of the ℓ-dimensional domain can be altered by altering the corresponding partition of a one-dimensional domain.
3) The bookkeeping operation involved in storing the partition is simplified by storing the equivalent one-dimensional partition.

Disadvantages of the approach for handling the ℓ-dimensional problem and suggestions for their relief are

a) A partition of the ℓ-dimensional domain has definite restrictions imposed by the transformation. It is desirable that the transformation tends to form partitions with tightly knit regions. The Hilbert Curve Mapping illustrated in Chapter V for two dimensions is better in this regard than the transformation used. The implementation of the Hilbert Curve Mapping should thus give improved results.

b) The act of transforming the problem to one dimension causes neighborhood information between neighboring observations in ℓ dimensions to be lost. The effect is that more training observations are required than if the solution were carried out solely in ℓ dimensions. A way to decrease this effect is to account for the neighborhood information before performing the transformation. For example, a cluster of observations could be placed about each training observation. Their mapped equivalents in one dimension, if treated as mapped training observations carry the neighborhood information with them. This operation can be interpreted as smoothing the data before mapping. Another solution is to carry out the entire analysis in ℓ dimension; this involves developing ℓ-dimensional procedures for use with regions in a partition of the ℓ-dimensional domain. Techniques for a ℓ-dimensional

region would be similar to the techniques for a one-dimensional interval except that interval width would be replaced with maximum distance across the region. The required storage and partition adjustment would be the primary difficulties.

Experimental results in Chapter IV indicate that for a one-dimensional Gaussian example, the nonparametric procedure can require over one hundred times as many observations as the optimum Gaussian procedure to achieve equal performance. The reason is that the nonparametric procedure does not utilize the a priori knowledge that the density function is Gaussian. This can be an advantage when the density function is not Gaussian; on the other hand it is desirable to have provision for using a priori knowledge should it be available.

Gaussian approximations were used for the beta densities to simplify the integration over a region V in the (U_a, U_b) plane. For small n a numerical or a Monte Carlo integration method could be used. The latter method can be accomplished by generating ordered pairs of observations from statistically independent beta distributions. The first coordinate is generated according to $\beta^*(U_a|\gamma_{a1},\gamma_{a2})$ and the second according to $\beta^*(U_b|\gamma_{b1},\gamma_{b2})$. The relative frequency with which the result (U_a, U_b) occurs in V is an estimate for Pr(V). The coordinate U_j is easily generated by setting $U_j = \frac{P_j\rho_j}{W}$ with ρ_j generated according to $\beta(\rho_j|\gamma_{j1}, \gamma_{j2})$. ρ_j is taken as the $(\gamma_{j1})^{th}$ smallest outcome from a sequence of $\gamma_{j_1} + \gamma_{j_2} - 1$ observations of a uniform distribution on the interval $[0,1]$. This approach is

impractical for large $n_j = \nu_{j_1} + \nu_{j_2} - 1$ because of the large number of uniformly distributed observations required.

This report discusses the two class recognition problem. Generalization of the procedure to multiclasses has not been accomplished. One way to deal with the multiclass problem using the two class procedure is to lump classes into two disjoint groups. The group chosen can be split into two disjoint groups, etc., so that finally the chosen group consists of just one class.

LIST OF REFERENCES

[1] Nilsson, Nils J., Learning Machines, McGraw Hill, New York, 1965.

[2] Wilks, S. S., Mathematical Statistics, John Wiley and Sons, New York, 1963.

[3] Weiss, Lionel, Statistical Decision Theory, McGraw Hill, New York 1961.

[4] Fu, K. S., and E. G. Henrichon, Jr., "On Nonparametric Methods for Pattern Recognition," Purdue University, School of Electrical Engineering, TR-EE 69-19, August, 1968.

[5] Abramson, N., and D. Braverman, "Learning to Recognize Patterns in a Random Environment," IRE International Symposium on Information Theory, Vol. IT-8, pp. s58-s63, July, 1962.

[6] Keehn, D. G., "Learning the Mean Vector and Covariance Matrix of Gaussian Signals in Pattern Recognition," Stanford Electronics Laboratories, TR 2003-6, February, 1963.

[7] Aizerman, M. A., E. M. Braverman, and L. I. Rozonoer, "The Probability Problem of Pattern Recognition Learning and the Method of Potential Functions," Avtomatika i Telemekhanika, Vol. 25, No. 9, September, 1964.

[8] Tsypkin, Ya Z., "Use of the Stochastic Approximation Method in Estimating Unknown Distribution Densities from Observations," Avtomatika i Telemekhanika, Vol. 27, No. 3, March, 1966.

[9] Kashyap, R. L., and C. C. Blaydon, "Estimation of Probability Density and Distribution Functions," IEEE Transactions on Information Theory, Vol. IT-14, July, 1968.

[10] Rosenblatt, M., "Remark on Some Nonparametric Estimates of a Density Function," Annals of Mathematical Statistics, Vol. 27, 1956.

[11] Parzen, E., "On Estimation of a Probability Density Function and Mode," Annals of Mathematics, Vol. 33, 1962.

[12] Murthy, V. K., "Nonparametric Estimation of Multivariate Densities with Applications," Multivariate Analysis, Edited by P. R. Krishnaiah, Academic Press, New York, 1966.

[13] Whittle, P., "On the Smoothing of Probability Density Functions," Journal of the Royal Statistical Society, Series B, Vol. 20, No. 2, 1958.

[14] Watson, G. S., and M. R. Leadbetter, "On the Estimation of the Probability Density," Annals of Mathematical Statistics, Vol. 34, 1963.

[15] Cooper, G. R., and J. A. Tabaczynski, "Estimation of Probability Density and Distribution Functions," Purdue University School of Electrical Engineering, TR-EE 65-15, August, 1965.

[16] Cover, T. M., and Hart, P. E., "Nearest Neighbor Pattern Classification," IEEE Transaction on Information Theory, Vol. IT-13, January, 1967.

[17] Loftsgaarden, D. O., and C. P. Quesenberry, "A Nonparametric Estimate of a Multivariate Density Function," Annals of Mathematical Statistics, Vol. 36, 1965.

[18] Patrick, E. A., and F. P. Fischer II, "A Generalization of the K-Nearest Neighbor Decision Rule," presented at the 1969 International Joint Conference on Artifical Intelligence, May, 1969.

[19] Hart, P. E., "The Condensed Nearest Neighbor Rule," IEEE Transactions on Information Theory, Vol. IT-14, May, 1968.

[20] Patrick, E. A., and J. C. Hancock, "Nonsupervised Sequential Classification and Recognition of Patterns," IEEE Transactions on Information Theory, Vol. IT-12, No. 3, pp. 362-372, July, 1966.

[21] Yakowitz, S. J., and J. D. Spragins, "A Characterization Theorem on the Identifiability of Finite Mixtures," presented at the 1966 International Communications Conference and Ann. Math Statistics, Vol. 39, February, 1968.

[22] Cooper, D. B., and P. W. Cooper, "Nonsupervised Adaptive Signal Detection and Pattern Recognition," Information and Control, Vol. 7, No. 3, pp. 416-444, September, 1964.

[23] Patrick, E. A., and F. P. Fischer II, "Cluster Mapping with Experimental Computer Graphics," Proceedings of the Symposium on Computer Processing in Communications, New York, N. Y., April, 1969.

[24] Patrick, E. A., "Concepts of an Estimation System, an Adaptive System, and a Network of Adaptive Estimation Systems," IEEE Transactions on Systems Science and Cybernetics, Vol. SSC-5, No. 1, January, 1969.

[25] Patrick, E. A., "On a Class of Unsupervised Estimation Problems," IEEE Transactions on Information Theory, Vol. IT-14, No. 3, May, 1968.

[26] Patrick, E. A., "Distribution Free, Minimum Conditional Risk Learning Systems," Purdue University School of Electrical Engineering Technical Report TR-EE66-18, November, 1966.

[27] Patrick, E. A. and Fischer II, F. P., "Introduction to the Performance of Distribution Free Minimum Conditional Risk Learning Systems," Purdue University School of Electrical Engineering Technical Report No. TR-EE67-12, July, 1967.

[28] Anderson, T. W., "Some Nonparametric Multivariate Procedures Based on Statistically Equivalent Blocks," Multivariate Analysis, Edited by P. R. Krishnaiah, Academic Press, New York, 1966.

[29] Fraser, D. A. S., Nonparametric Methods in Statistics, John Wiley and Sons, New York, 1957.

[30] Sebesteyen, George S., "Pattern Recognition by an Adaptive Process of Sample Set Construction," IRE Transactions on Information Theory, Vol. 178, No. 5, September, 1962.

[31] Sebestyen, George S., Decision Making Processes in Pattern Recognition, The MacMillan Company, New York, 1962.

[32] Specht, D. F., "Generation of Polynomial Discriminant Functions for Pattern Recognition," Stanford Electronic Laboratories, TR 6764-5, May, 1966.

[33] Nagy, George, "State of the Art in Pattern Recognition," Proceedings IEEE, Vol. 56, No. 5, May, 1968.

[34] Ho, Yu-Chi and Agrawala, Ashok K., "On Pattern Classification Algorithms-Introduction and Survey," Proceedings IEEE, Vol. 56, No. 12, December 1968.

[35] Spragins, J. D., "Reproducing Distributions for Machine Learning," Stanford Electronics Laboratories Technical Report No. 6103-7, November, 1963.

[36] Cramer, H., Mathematical Methods of Statistics, Princeton University Press, 1946.

[37] Rao, C. Radhakrishna, Linear Statistical Inference and Applications, John Wiley and Sons, New York, 1965.

[38] Hughes, G. F., "On the Mean Accuracy of Statistical Pattern Recognizers" IEEE Transactions on Information Theory, Vol. IT-14, January, 1968.

[39] Lebo, J. A. and Hughes, G. F., "Pattern Recognition Preprocessing by Similarity Functionals," Proceedings of National Electronics Conference, 1966.

[40] IBM 1130 Subroutine Library, Form C 26-5929-4, 1966.

[41] Wiener, Norbert, The Fourier Integral and Certain of Its Applications, Dover Publications, New York, pp. 16-17, 1933.

[42] Moore, E. H., "On Certain Crinkly Curves," Transactions of American Mathematical Society, Vol. 1, 1900.

[43] Peano, G., "Sur une coube, que remplit toute une aire plane," Mathematische Annalen, Vol. 36, 1890.

[44] Butz, A. R., "Space Filling Curves and Mathematical Programming," Information and Control, Vol. 12, 1968.

[45] Hilbert, D., "Ueber die stetige abbildung einer linie auf ein flachenstuck," Mathematische Annalen, Vol. 38, March, 1891.

[46] Patrick, E. A., Anderson, D. R., and Bechtel, F. K., "Mapping Multidimensional Space to One-Dimension for Computer Output Display," IEEE Transactions on Computers, Vol. C-17, No. 10, October, 1968.

[47] Abend, K., Harley, T. J., and Kanal, L. N., "Classification of Binary Random Patterns," IEEE Transactions on Information Theory, Vol. IT-11, No. 4, October, 1965.

[48] Shepard, R. N. and Carroll, J. D., "Parametric Representation of Nonlinear Data Structures," Multivariate Analysis, Edited by P. R. Krishnaiah, Academic Press, New York, 1966.

[49] Sammon, J. W., Jr., "A Nonlinear Mapping for Data Structure Analysis," IEEE Transactions on Computers, Vol. C-18, No. 5, May, 1969.

[50] Calvert, T. W., "Projections of Multidimensional Data for Use in Man-Computer Graphics," Electrical Engineering Department, Carnegie-Mellon University, Pittsburgh, Pennsylvania.

[51] Patrick, E. A., and F. P. Fischer, II, "Nonparametric Feature Selection," to appear in IEEE Transactions on Information Theory, September, 1969.

[52] Kullback, Solomon, Information Theory and Statistics, John Wiley and Sons, New York, 1959.

[53] Butz, Arthur R., "Convergence with Hilbert's Space Filling Curve," Journal of Computer and System Sciences, No. 3, pp. 128-146, May, 1969.

[54] Patrick, E. A., and J. P. Costello, "Unsupervised Estimation and Processing of Unknown Signals," Purdue University School of Electrical Engineering Technical Report, TR-EE 69-18, June 1969.

[55] Fix, E., and J. L. Hodges, "Discriminatory Analysis," Project Number 21-49-004, Report Number 11, USAF School of Aviation Medicine, Randolph Field, Texas, August, 1952.

[56] Ball, G. H., "Data Analysis in the Social Sciences: What About the Details?", Fall Joint Computer Conference, AFIPS Proceedings, Vol. 27, Part I, 1965.

APPENDIX A

The object of this appendix is to discuss methods for specifying the characterizing parameters in a beta d.f. on $P_j(i)$

$$f(P_j(i)|s_{j1}(i),s_{j2}(i)) = \frac{\Gamma(s_{j1}(i) + s_{j2}(i))}{\Gamma(s_{j1}(i))\Gamma(s_{j2}(i))} P_j(i)^{s_{j1}(i) - 1} (1 - P_j(i))^{s_{j2}(i) - 1} \tag{A.1}$$

such that a priori knowledge about $P_j(i)$ is accounted for. $P_j(i)$ represents uncertainty in $P_j^*(i)$, the probability of an observation from class ω_j falling in the i^{th} interval of a given R-interval partition of [0,1].

First, consider the case where no a priori knowledge is available. The characterizing parameters, $s_{j1}(i)$, $s_{j2}(i)$, can be chosen as

$$s_{j1}(i) = 1$$
$$s_{j2}(i) = R - 1 \tag{A.2}$$

s's so chosen are consistent with s's replacing γ's specified by Equations (2.4), provided the $P_j(i)$'s are described jointly by the R - 1 variate Dirichlet d.f. having 1's for parameters. Then, each allowed set of $P_j(i)$'s is equally likely, a condition sometimes said to correspond to no a priori knowledge. Another method of

defining the s's without benefit of a priori knowledge on the $P_j(i)$'s is to use the first R - 1 training observations to specify the initial partition. $s_{j1}(i)$ is chosen as the ratio of the i^{th} interval width to the width of the smallest interval containing the i^{th} interval but having boundaries defined by observations from class ω_j. For these computations, observations from both classes are assumed to exist at the end points 0 and 1. The parameter t_j is specified to be the number of class ω_j observations from the first R - 1 training observations.

Then $s_{j2}(i)$ is given by

$$s_{j2}(i) = t_j + 1 - s_{j1}(i) \tag{A.3}$$

The maximum value for $s_{j1}(i)$ is one. In practice $s_{j1}(i)$ is guaranteed positive by discarding for interval forming purposes training observations that cause ties.

If a priori knowledge is available in the form of "a priori training observations" [35], (fictitious observations that might be obtained based on what is known about the $P_j(i)$'s), their numbers can be added to the appropriate s's.

Now suppose that a priori knowledge consists of the expected value $E_j(i)$ and variance $Var_j(i)$ for each $P_j(i)$. Such knowledge can be included by solving Equation (2.5) for $\gamma_{j1}(i)$ and $\gamma_{j2}(i)$ (in this case $s_{j1}(i)$ and $s_{j2}(i)$).

$$s_{j1}(i) = E_j(i)\left[\frac{E_j(i)(1 - E_j(i))}{Var_j(i)} - 1\right]$$

$$s_{j2}(i) = (1 - E_j(i))\left[\frac{E_j(i)(1 - E_j(i))}{Var_j(i)} - 1\right] \qquad (A.4)$$

The resulting relations among $E_j(i)$, $Var_j(i)$, $s_{j1}(i)$, and t_j are illustrated by Figure 45. t_j is related to $s_{j1}(i)$ and $s_{j2}(i)$ by (A.3). Requirements on the a priori values for $E_j(i)$ and $Var_j(i)$ are

$$0 < E_j(i) < 1$$

$$\frac{E_j(i)(1 - E_j(i))}{Var_j(i)} \geq 2 \qquad (A.5)$$

These requirements ensure that t_j is non-negative (t_j can be interpreted as a number of a priori training observations), and that $s_{j1}(i)$ and $s_{j2}(i)$ are positive (a requirement for parameters characterizing any beta d.f.).

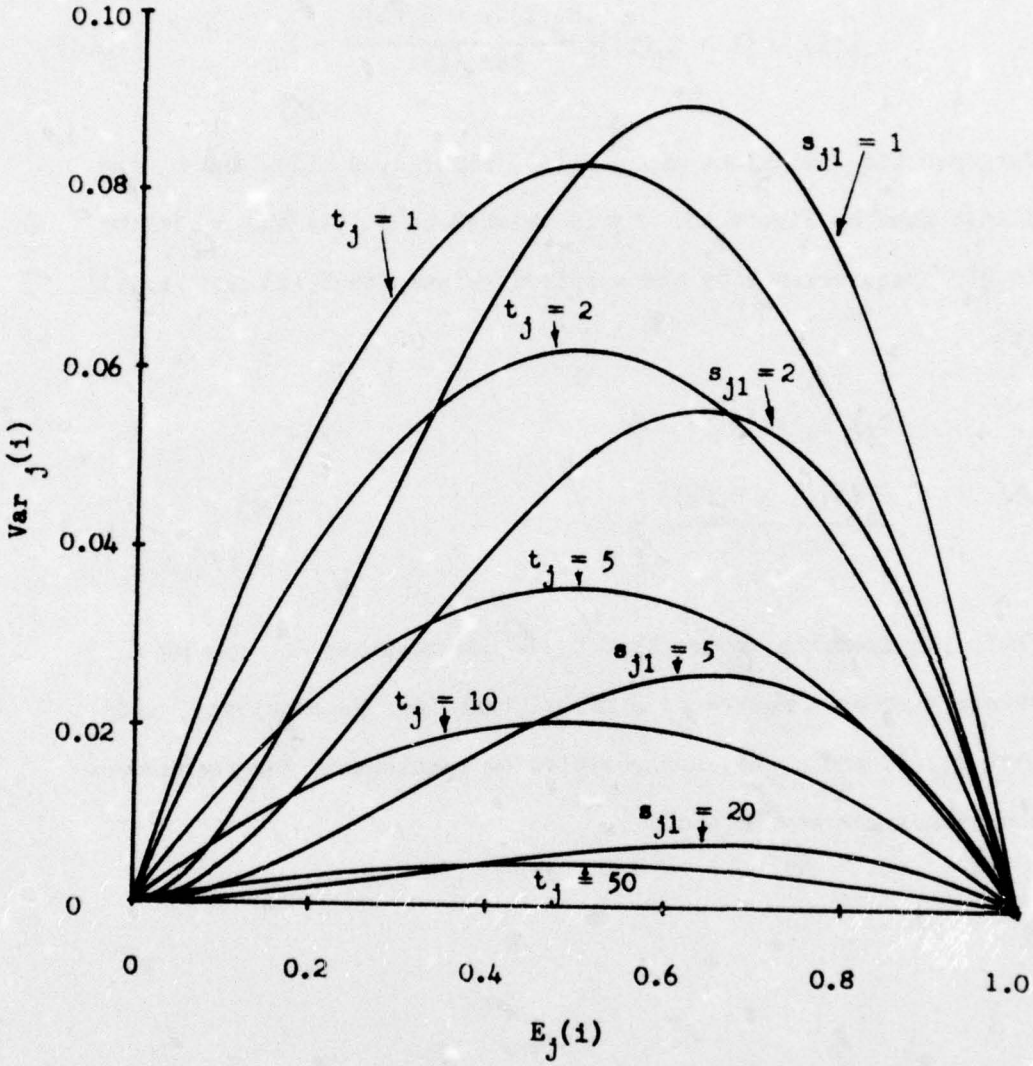


Figure 45. Relations Among $E_j(i)$, $Var_j(i)$, $s_{j1}(i)$, and t_j

APPENDIX B

The quantity

$$T = \int_0^{A_b} \frac{1}{A_b} \beta\left(\frac{U_b}{A_b} \mid \gamma_{b1}, \gamma_{b2}\right) \int_0^{r_o} \frac{1}{A_a} \beta\left(\frac{U_a}{A_a} \mid \gamma_{a1}, \gamma_{a2}\right) dU_a dU_b \qquad (B.1)$$

is evaluated in this appendix. In (B.1),

$$r_o = \text{Min}\left[U_b - q,\ A_a\right] \qquad (B.2)$$

A_a, A_b, and ρ are finite, positive, real constants, q is a real constant satisfying $q \leq \rho$, and γ_{b1}, γ_{b2}, γ_{a1}, and γ_{a2} are finite positive integer constants. $\beta(z|\alpha_1,\alpha_2)$ is defined by:

$$\beta(z|\alpha_1,\alpha_2) = Be^{-1}(\alpha_1,\alpha_2) z^{\alpha_1 - 1}(1-z)^{\alpha_2 - 1}$$

$$\text{if } 0 \leq z \leq 1$$

$$= 0$$

$$\text{otherwise} \qquad (B.3)$$

where

$$Be(\alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1)\,\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)} \qquad (B.4)$$

The limits of integration in (B.1) define the region of the (U_a, U_b) plane illustrated in Figure 46.

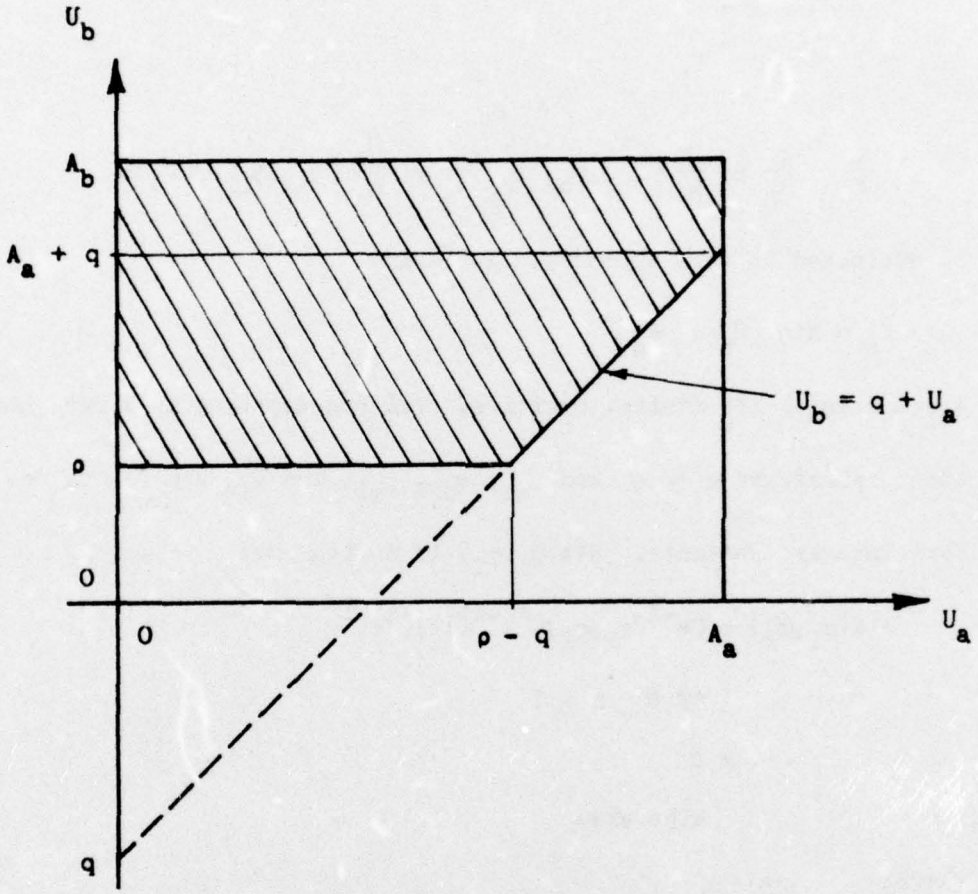


Figure 46. Region of Integration

Define:

$$Z = \int_0^{r_o} \frac{1}{A_a}\ \beta\left(\frac{U_a}{A_a}\ |\ \gamma_{a1},\gamma_{a2}\right) dU_a \tag{B.5}$$

Let $x = \frac{U_a}{A_a}$. Then:

$$Z = \int_0^{r_o/A_a} \beta(x|\gamma_{a1},\gamma_{a2})dx$$

$$= \int_0^{r_o/A_a} Be^{-1}(\gamma_{a1},\gamma_{a2})x^{\gamma_{a1}-1}(1-x)^{\gamma_{a2}-1}dx$$

$$= Be^{-1}(\gamma_{a1},\gamma_{a2})\int_0^{r_o/A_a} x^{\gamma_{a1}-1}\sum_{j=0}^{\gamma_{a2}-1}\binom{\gamma_{a2}-1}{j}(-1)^j x^j\,dx$$

$$= Be^{-1}(\gamma_{a1},\gamma_{a2})\sum_{j=0}^{\gamma_{a2}-1}\binom{\gamma_{a2}-1}{j}(-1)^j\ \frac{(r_o/A_a)^{\gamma_{a1}+j}}{(\gamma_{a1}+j)} \tag{B.6}$$

Substituting Z into (B.1) gives:

$$T = \int_\rho^{A_b} \frac{1}{A_b}\beta\left(\frac{U_b}{A_b}\ |\gamma_{b1},\gamma_{b2}\right)\left[Be^{-1}(\gamma_{a1}\gamma_{a2})\sum_{j=0}^{\gamma_{a2}-1}\binom{\gamma_{a2}-1}{j}(-1)^j \cdot \frac{(r_o/A_a)^{\gamma_{a1}+j}}{(\gamma_{a1}+j)}\right] dU_b \tag{B.7}$$

T can be written as the sum:

$$T = T_1 + T_2 \tag{B.8}$$

with T_1 and T_2 given by:

$$T_1 = \int_{o}^{y_o} \frac{1}{A_b} \beta\left(\frac{U_b}{A_b} \mid \gamma_{b1}, \gamma_{b2}\right) Be^{-1}(\gamma_{a1}, \gamma_{a2}) \sum_{j=0}^{\gamma_{a2}-1} \binom{\gamma_{a2}-1}{j} (-1)^j \cdot \frac{\left(\frac{U_b - o}{A_a}\right)^{\gamma_{a1}+j}}{(\gamma_{a1}+j)} \, dU_b$$

$$T_2 = \int_{y_o}^{A_b} \frac{1}{A_b} \beta\left(\frac{U_b}{A_b} \mid \gamma_{b1}, \gamma_{b2}\right) dU_b \qquad (B.9)$$

where

$$y_o = \text{Min}\left[\text{Max}(A_a + q, o), A_b\right] \qquad (B.10)$$

Let $y = \frac{U_b}{A_b}$. Then T_1 and T_2 become:

$$T_1 = \int_{o/A_b}^{y_o/A_b} Be^{-1}(\gamma_{b1}, \gamma_{b2}) \, Be^{-1}(\gamma_{a1}, \gamma_{a2}) \sum_{k=0}^{\gamma_{b2}-1} \binom{\gamma_{b2}-i}{k} (-1)^k y^{k+\gamma_{b1}-1} \cdot \sum_{j=0}^{\gamma_{a2}-1} \binom{\gamma_{a2}-1}{j} (-1)^j \left(\frac{A_b}{A_a}\right)^{\gamma_{a1}+j} \frac{\left(y - \frac{q}{A_b}\right)^{\gamma_{a1}+j}}{(\gamma_{a1}+j)} \, dy$$

$$T_2 = \int_{y_o/A_b}^{1} Be^{-1}(\gamma_{b1}, \gamma_{b2}) \sum_{k=0}^{\gamma_{b2}-1} \binom{\gamma_{b2}-1}{k} (-1)^k y^{k+\gamma_{b1}-1} \, dy \qquad (B.11)$$

The expansion

$$\left(y - \frac{q}{A_b}\right)^{\gamma_{a1}+j} = \sum_{v=0}^{\gamma_{a1}+j} \binom{\gamma_{a1}+j}{v} y^v \left(\frac{-q}{A_b}\right)^{\gamma_{a1}+j-v} \qquad (B.12)$$

can be used in (B.11), the integration performed, and T_1 and T_2 summed. The result is:

$$T = Be^{-1}(\gamma_{b1},\gamma_{b2})Be^{-1}(\gamma_{a1},\gamma_{a2}) \sum_{j=0}^{\gamma_{a2}-1} \binom{\gamma_{a2}-1}{j}(-1)^j \frac{\left(\frac{A_b}{A_a}\right)^{\gamma_{a1}+j}}{(\gamma_{a1}+j)}$$

$$\cdot \sum_{v=0}^{\gamma_{a1}+j} \binom{\gamma_{a1}+j}{v}\left(\frac{-q}{A_b}\right)^{\gamma_{a1}+j-v} \sum_{k=0}^{\gamma_{b2}-1} \binom{\gamma_{b2}-1}{k}(-1)^k \frac{\left[\left(\frac{y_o}{A_b}\right)^{k+\gamma_{b1}+v} - \left(\frac{\rho}{A_b}\right)^{k+\gamma_{b1}+v}\right]}{(k+\gamma_{b1}+v)}$$

$$+ 1 - Be^{-1}(\gamma_{b1},\gamma_{b2}) \sum_{k=0}^{\gamma_{b2}-1} \binom{\gamma_{b2}-1}{k}(-1)^k \frac{\left(\frac{y_o}{A_b}\right)^{k+\gamma_{b1}}}{(k+\gamma_{b1})} \qquad (B.13)$$

where y_o is obtained from (B.10).

APPENDIX C

Evaluate the integral

$$A = \iint_{U_b > \xi_1 + \xi_2 U_a} \frac{1}{\sqrt{2\pi}\,\sigma_a} e^{-\frac{1}{2}\left(\frac{U_a - \mu_a}{\sigma_a}\right)^2} \cdot \frac{1}{\sqrt{2\pi}\,\sigma_b} e^{-\frac{1}{2}\left(\frac{U_b - \mu_b}{\sigma_b}\right)^2} dU_a\, dU_b \qquad (C.1)$$

Let

$$y_a = \frac{\sigma_b}{(\sigma_b^2 + \sigma_a^2 \xi_2^2)^{\frac{1}{2}}}\left(\frac{U_a - \mu_a}{\sigma_a}\right) + \frac{\sigma_a \xi_2}{(\sigma_b^2 + \sigma_a^2 \xi_2^2)^{\frac{1}{2}}}\left(\frac{U_b - \mu_b}{\sigma_b}\right)$$

$$y_b = \frac{-\sigma_a \xi_2}{(\sigma_b^2 + \sigma_a^2 \xi_2^2)^{\frac{1}{2}}}\left(\frac{U_a - \mu_a}{\sigma_a}\right) + \frac{\sigma_b}{(\sigma_b^2 + \sigma_a^2 \xi_2^2)^{\frac{1}{2}}}\left(\frac{U_b - \mu_b}{\sigma_b}\right) \qquad (C.2)$$

Then

$$A = \iint_{y_b > \frac{\xi_1 + \mu_a \xi_2 - \mu_b}{(\sigma_b^2 + \sigma_a^2 \xi_2^2)^{\frac{1}{2}}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} y_a^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} y_b^2} dy_a\, dy_b \qquad (C.3)$$

or

$$\Lambda = \Phi\left(\frac{-\xi_1 - \mu_a \xi_2 + \mu_b}{(\sigma_b^2 + \sigma_a^2 \xi_2^2)^{\frac{1}{2}}}\right) \qquad (C.4)$$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} y^2} \, dy$

APPENDIX D

This appendix presents a numerical procedure for the minimization of Λ,

$$\Lambda = \Phi(-Q) \tag{D.1}$$

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\, dy, \quad -\infty < x < \infty$$

$$Q = \frac{\xi_1 + \mu_a \xi_2 - \mu_b}{((\sigma_a \xi_2)^2 + \sigma_b^2)^{\frac{1}{2}}}$$

$$\xi_1 = -\frac{1}{4C_b}(\lambda_a^2 - C_a^2)$$

$$\xi_2 = \frac{1}{2C_b}(\lambda_a - C_a)$$

with respect to λ_a. The constants μ_a, μ_b, σ_a, σ_b, C_a, and C_b are all positive. In addition, it is known that

$$\mu_b < \frac{(\mu_a - C_a)^2}{4C_b}$$

$$\mu_a > C_a \tag{D.2}$$

Because $\Phi(-Q)$ is strictly decreasing in Q, Λ is minimized by maximizing Q. Q can be written in terms of λ_a as:

$$Q = \frac{-\frac{1}{4C_b}(\lambda_a^2 - C_a^2) + \frac{\mu_a}{2C_b}(\lambda_a - C_a) - \mu_b}{\left[\left(\frac{\sigma_a}{2C_b}\right)^2 (\lambda_a - C_a)^2 + \sigma_b^2\right]^{\frac{1}{2}}} \qquad (D.3)$$

Set the derivative of Q with respect to λ_a to zero.

$$\frac{\partial Q}{\partial \lambda_a} = \left\{\left[\left(\frac{\sigma_a}{2C_b}(\lambda_a - C_a)\right)^2 + \sigma_b^2\right]^{\frac{1}{2}}\left(\frac{\mu_a - \lambda_a}{2C_b}\right)\right.$$

$$- \left[\frac{-1}{4C_b}(\lambda_a^2 - C_a^2) + \frac{\mu_a}{2C_b}(\lambda_a - C_a) - \mu_b\right]\left[\left(\frac{\sigma_a}{2C_b}(\lambda_a - C_a)\right)^2 + \sigma_b^2\right]^{\frac{1}{2}}$$

$$\left.\cdot \left(\frac{\sigma_a}{2C_b}\right)^2 (\lambda_a - C_a)\right\} \Big/ \left[\left(\frac{\sigma_a}{2C_b}(\lambda_a - C_a)\right)^2 + \sigma_b^2\right] = 0$$

or

$$\left[\left(\frac{\sigma_a}{2C_b}\right)^2 (\lambda_a - C_a)^2 + \sigma_b^2\right]\left(\frac{\mu_a - \lambda_a}{2C_b}\right)$$

$$- (\lambda_a - C_a)\left(\frac{\sigma_a}{2C_b}\right)^2 \left[\frac{-1}{4C_b}(\lambda_a^2 - C_a^2) + \frac{\mu_a}{2C_b}(\lambda_a - C_a) - \mu_b\right] = 0$$

Collecting powers of $(\lambda_a - C_a)$ gives:

$$\frac{(\lambda_a - C_a)^3}{4C_b} - (\lambda_a - C_a)\left[\mu_b - \frac{2C_b \sigma_b^2}{\sigma_a^2}\right] - (\mu_a - C_a)\frac{2C_b \sigma_b^2}{\sigma_a^2} = 0$$

(D.4)

(D.4) is a cubic in $(\lambda_a - C_a)$ and is difficult to solve analytically. However, the desired root may be obtained by using the following numerical technique. Define $g(\lambda_a)$ to be the left side of (D.4), and rearrange to get

$$g(\lambda_a) = (\lambda_a - C_a)\left[\frac{(\lambda_a - C_a)^2}{4C_b} - \mu_b\right] + \frac{2C_b\,\sigma_b^2}{\sigma_a^2}(\lambda_a - \mu_a) \tag{D.5}$$

Note from (D.2) that

$$g(C_a) = 2C_b\,\frac{\sigma_b^2}{\sigma_a^2}(C_a - \mu_a) < 0$$

$$g\left(C_a + 2\sqrt{\mu_b\,C_b}\right) = \frac{2C_b\,\sigma_b^2}{\sigma_a^2}\left(C_a + 2\sqrt{\mu_b\,C_b} - \mu_a\right) < 0$$

$$g(\mu_a) = (\mu_a - C_a)\left[\frac{(\mu_a - C_a)^2}{4C_b} - \mu_b\right] > 0$$

$$g(\lambda_a) > 0, \text{ for } \lambda_a > \mu_a \tag{D.6}$$

Conditions (D.6) together with the fact that the inflection point for g is at $\lambda_a = C_a$, guarantee a unique root of $g(\lambda_a)$ in the interval $(C_a + 2\sqrt{\mu_b\,C_b},\ \mu_a)$. Further, no root of $g(\lambda_a)$ exists for $\lambda_a > \mu_a$ or for λ_a in the interval $(C_a,\ C_a + 2\sqrt{\mu_b\,C_b})$. Physical considerations of the problem show that the root sought corresponds to a relative minimum for Λ.

Let $\lambda_a(t)$ be the root obtained at the t^{th} stage of a Newton's iterative procedure. Then $\lambda_a(t+1)$ is obtained from $\lambda_a(t)$ by:

$$\lambda_a(t+1) = \lambda_a(t) - \frac{g(\lambda_a(t))}{g'(\lambda_a(t))} \tag{D.7}$$

The initial value is chosen as

$$\lambda_a(0) = \mu_a \tag{D.8}$$

The process is stopped when

$$\frac{|\lambda_a(t) - \lambda_a(t+1)|}{C_b} < 0.001 \tag{D.9}$$

and the solution is taken to be:

$$\lambda_a = \lambda_a(t+1) \tag{D.10}$$

APPENDIX E

In this appendix an experimental comparison is made of the quantities T and Λ given by

$$T = \int_0^A \frac{1}{A}\,\beta\left(\frac{U_b}{A}\middle|\gamma_{b1},\ \gamma_{b2}\right)\int_0^{U_b} \frac{1}{A}\,\beta\left(\frac{U_a}{A}\middle|\gamma_{a1},\ \gamma_{a2}\right)dU_a\ dU_b \quad \text{(E.1)}$$

and

$$\Lambda = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_b}\, e^{-\frac{1}{2}\left(\frac{U_b - \mu_b}{\sigma_b}\right)^2} \int_{-\infty}^{U_b} \frac{1}{\sqrt{2\pi}\,\sigma_b}\, e^{-\frac{1}{2}\left(\frac{U_a - \mu_a}{\sigma_a}\right)^2} dU_a dU_b \quad \text{(E.2)}$$

for some particular values of the parameters γ_{a1}, γ_{a2}, γ_{b1}, and γ_{b2}. In (E.2) the values μ_a, μ_b, σ_a, and σ_b are given in terms of γ_{a1}, γ_{a2}, γ_{b1}, γ_{b2}, and A by

$$\mu_a = A\left(\frac{\gamma_{a1}}{\gamma_{a1} + \gamma_{a2}}\right)$$

$$\mu_b = A\left(\frac{\gamma_{b1}}{\gamma_{b1} + \gamma_{b2}}\right)$$

$$\sigma_a = A\left(\frac{\gamma_{a1}\,\gamma_{a2}}{(\gamma_{a1} + \gamma_{a2})^2\,(\gamma_{a1} + \gamma_{a2} + 1)}\right)^{\frac{1}{2}}$$

$$\sigma_b = A\left(\frac{\gamma_{b1}\,\gamma_{b2}}{(\gamma_{b1} + \gamma_{b2})^2\,(\gamma_{b1} + \gamma_{b2} + 1)}\right)^{\frac{1}{2}} \quad \text{(E.3)}$$

The function β is defined by Equation (2.3). The region of integration for T is the cross-hatched region of (U_a, U_b)-plane illustrated in Figure 47.

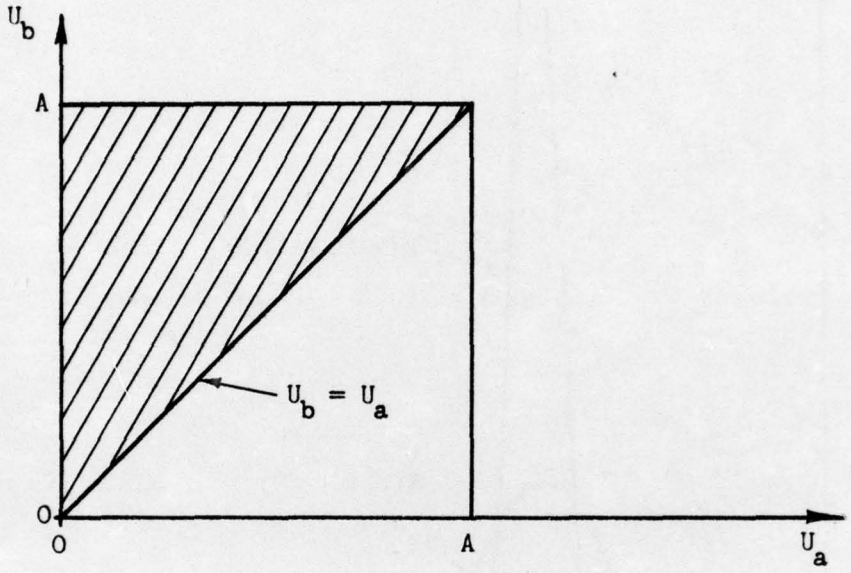


Figure 47. Region of Integration

For Λ, the region is the whole half plane above the line $U_b = U_a$. Since the product $\beta(\frac{U_b}{A}|\gamma_{b1},\gamma_{b2})\beta(\frac{U_a}{A}|\gamma_{a1},\gamma_{a2})$ is zero for all pairs (U_a, U_b) outside the cross-hatched region and above the line $U_b = U_a$, the region of integration for T may be considered the same as that for Λ. Note that (E.2) is obtained from (E.1) by replacing the beta d.f.'s with Gaussian d.f.'s having the same means and variances. Figure 48 shows the beta d.f. $\beta(x|Np,Nq)$

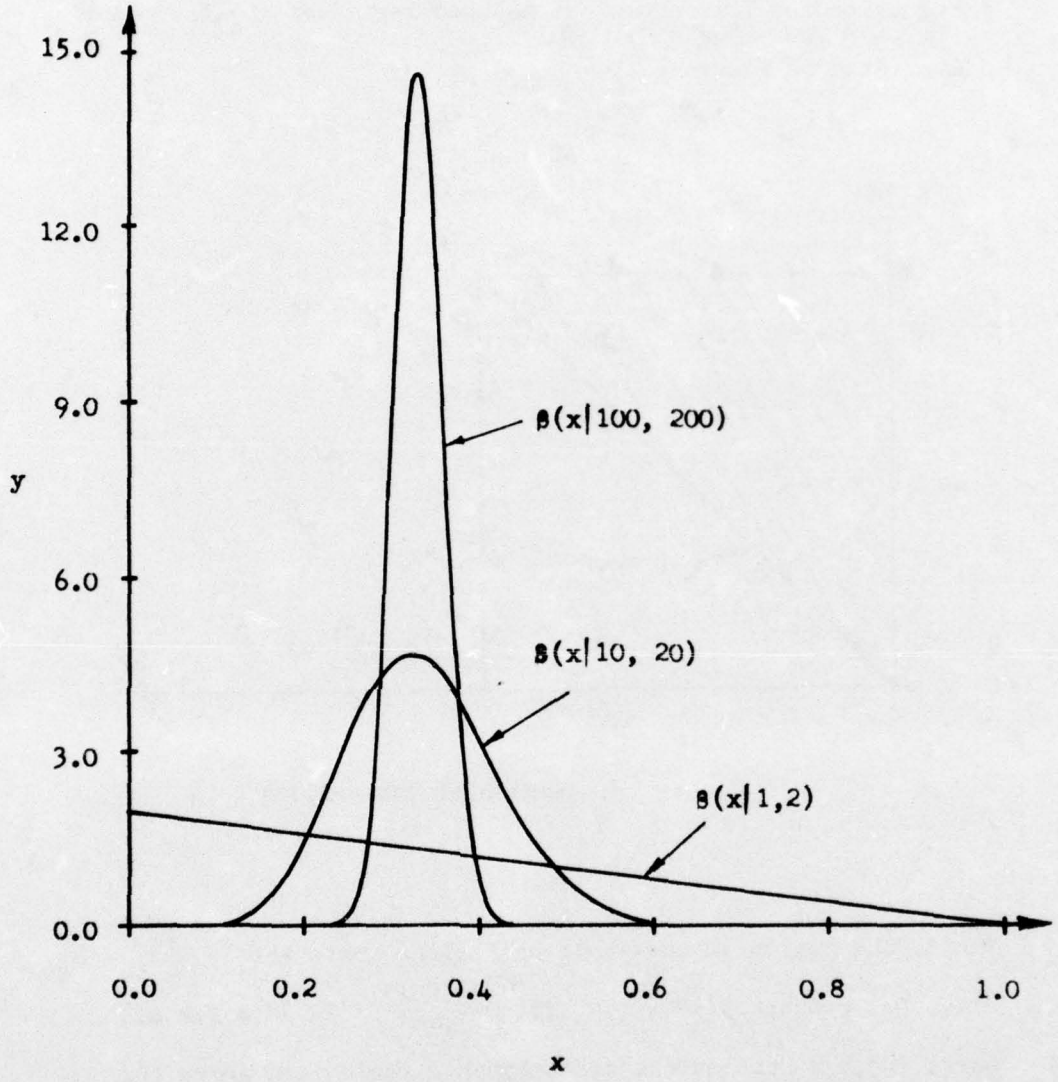


Figure 48. Sequence of Beta d.f.'s

for $p = 1$, $q = 2$, and three values of N; $N = 1, 10, 100$. It illustrates the convergence of a sequence of beta d.f.'s to a Gaussian d.f. as the parameters get large while maintaining constant ratio. This fact is proven in References [36,37].

Replacing ρ and q with 0, r_o with U_b, and A_a, A_b, and y_o with A, the result given by (B.13) can be used to evaluate (E.1)

$$T = Be^{-1}(\gamma_{b1},\gamma_{b2})Be^{-1}(\gamma_{a1},\gamma_{a2}) \sum_{j=0}^{\gamma_{a2}-1} \binom{\gamma_{a2}-1}{j}(-1)^j \left(\frac{1}{\gamma_{a1}+j}\right)$$

$$\cdot \sum_{k=0}^{\gamma_{b2}-1} \binom{\gamma_{b2}-1}{k}(-1)^k \left(\frac{1}{k+\gamma_{b1}+\gamma_{a1}+j}\right)$$

$$+ 1 - Be^{-1}(\gamma_{b1},\gamma_{b2}) \sum_{k=0}^{\gamma_{b2}-1} \binom{\gamma_{b2}-1}{k}(-1)^k \left(\frac{1}{k+\gamma_{b1}}\right)$$

or

$$T = Be^{-1}(\gamma_{b1},\gamma_{b2})Be^{-1}(\gamma_{a1},\gamma_{a2}) \sum_{j=0}^{\gamma_{a2}-1} \binom{\gamma_{a2}-1}{j}$$

$$\cdot (-1)^j \frac{Be(\gamma_{b1}+\gamma_{a1}+j,\gamma_{b2})}{(\gamma_{a1}+j)} \qquad (E.4)$$

where the identity

$$Be^{-1}(\gamma_1,\gamma_2) \sum_{k=0}^{\gamma_2-1} \binom{\gamma_2-1}{k}(-1)^k \left(\frac{1}{k+\gamma_1}\right) = 1 \qquad (E.5)$$

resulting from the beta density function integrating to one has been used.

From Appendix C, Λ is given by

$$\Lambda = \Phi\left(\frac{\mu_b - \mu_a}{(\sigma_a^2 + \sigma_b^2)^{\frac{1}{2}}}\right)$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\, dy, \quad -\infty < x < \infty$$

Let

$$Q = \frac{\mu_a - \mu_b}{(\sigma_a^2 + \sigma_b^2)^{\frac{1}{2}}} \tag{E.6}$$

From (E.3), Q becomes

$$Q = \frac{\dfrac{\gamma_{a1}}{\gamma_{a1} + \gamma_{a2}} - \dfrac{\gamma_{b1}}{\gamma_{b1} + \gamma_{b2}}}{\left[\dfrac{\gamma_{a1}\,\gamma_{a2}}{(\gamma_{a1} + \gamma_{a2})^2\,(\gamma_{a1} + \gamma_{a2} + 1)} + \dfrac{\gamma_{b1}\,\gamma_{b2}}{(\gamma_{b1} + \gamma_{b2})^2\,(\gamma_{b1} + \gamma_{b2} + 1)}\right]^{\frac{1}{2}}} \tag{E.7}$$

So that γ_{a1}/γ_{a2} and γ_{b1}/γ_{b2} are constant, let γ_{a1}, γ_{a2}, γ_{b1}, and γ_{b2} be written in terms of the constants p_a, q_a, p_b, q_b, and the variable N as follows:

$$\begin{aligned} \gamma_{a1} &= Np_a \\ \gamma_{a2} &= Nq_a \\ \gamma_{b1} &= Np_b \\ \gamma_{b2} &= Nq_b \end{aligned} \tag{E.8}$$

T and $\Lambda = \Phi(-Q)$ are computed for three different examples.

Example 1

$p_a = 1,\ q_a = 9$

$p_b = 1,\ q_b = 9$

$N = 1,2,\ldots,10$

Example 2

$p_a = 2,\ q_a = 8$

$p_b = 1,\ q_b = 9$

$N = 1,2,\ldots,10$

Example 3

$p_a = 3,\ q_a = 7$

$p_b = 1,\ q_b = 9$

$N = 1,2,\ldots,10$

The locations of the means in the (U_a, U_b) plane of the distributions for these three examples are shown in Figure 49.

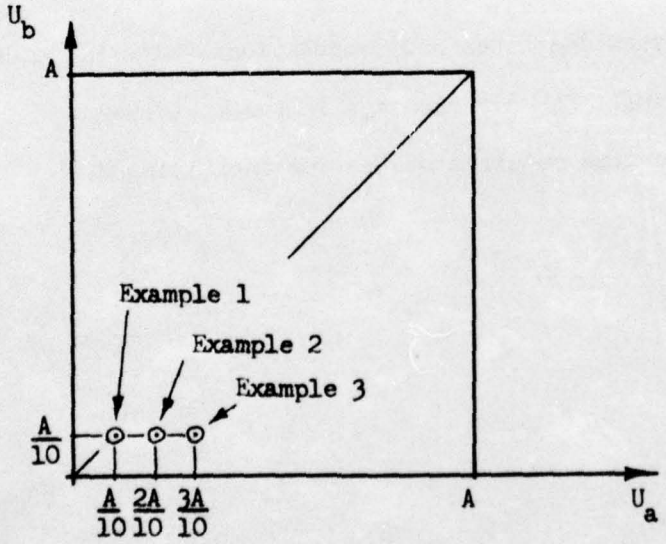


Figure 49. Distribution Means for Examples

The resulting T and Λ are plotted in Figure 50 as functions of N. The circles represent values for T while the triangles represent values for Λ. These computations were made with Purdue University's CDC 6500 digital computer using single precision. No special computational tricks were employed other than the performing of all possible factorial cancellation in the expression for T. Several pertinent facts are worth noting.

1) For these examples, close agreement between T and Λ for all except the higher values of N is observed.
2) Lack of agreement between T and Λ for large N is due to computer inaccuracies resulting from accumulated error in the many arithmetic operations necessary to compute T.
3) Computational time for T is approximately $1\frac{1}{2}$ minutes for each of the three examples. Computational time for Λ is relatively negligible.
4) For T, accuracy decreases and computational time increases as N increases. For Λ, no change in computational accuracy and time required occurs for increasing N.

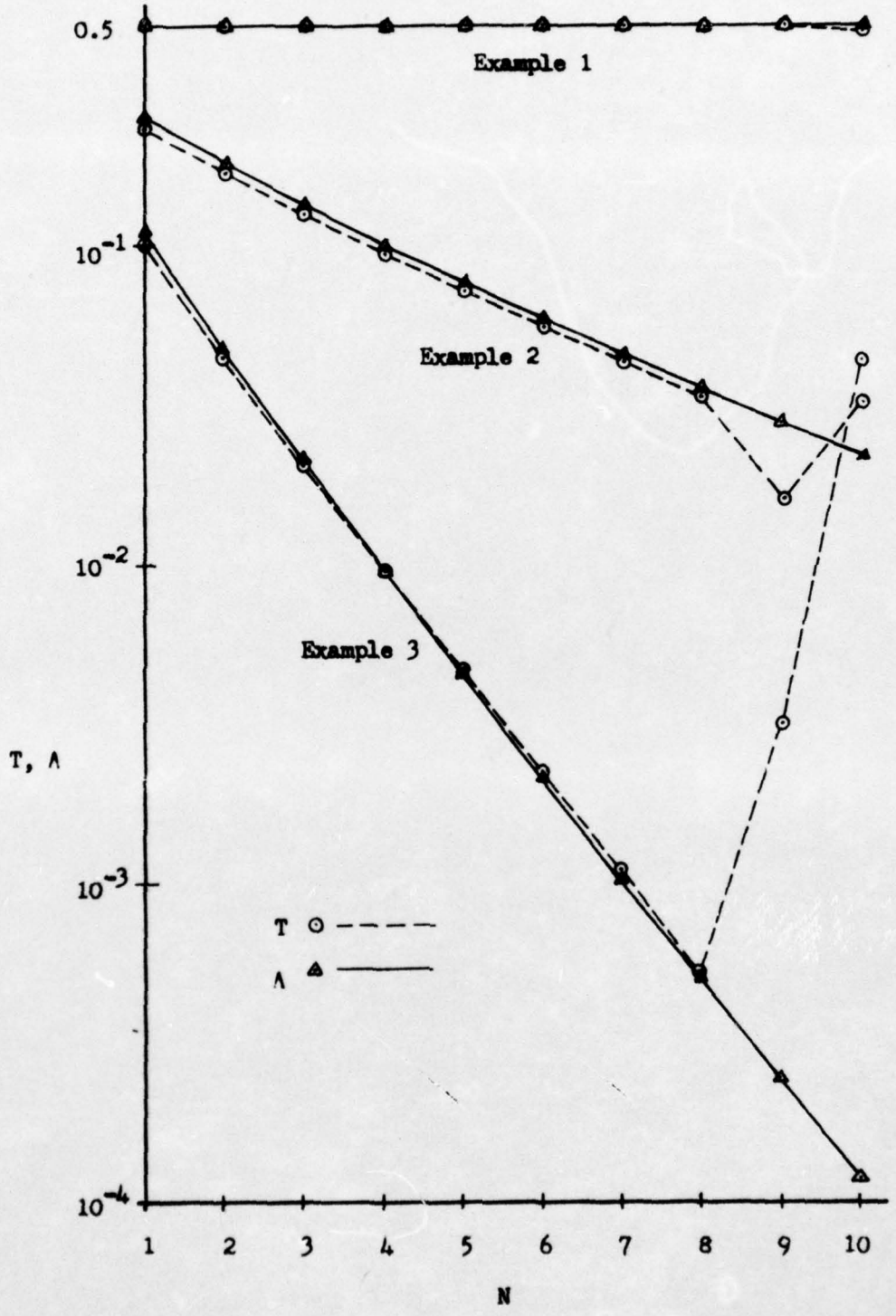


Figure 50. Comparison of T and A