COMPUTER RECOGNITION OF DIGITIZED IMAGE PATTERNS

by

John Cary Bellamy II

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ELECTRICAL ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSTIY OF ARIZONA

1971

THE UNIVERSITY OF ARIZONA

GRADUATE COLLEGE

I hereby recommend that this dissertation prepared under my direction by John Cary Bellamy II

entitled Computer Recognition of Digitized Image Patterns

be accepted as fulfilling the dissertation requirement of the degree of Doctor of Philosophy in Electrical Engineering

Frederick J. Hill — Dissertation Director — 25 May 1971 — Date

After inspection of the final copy of the dissertation, the following members of the Final Examination Committee concur in its approval and recommend its acceptance:*

John L. Kennedy	May 25, 1971
D. L. Swett	5/24/71
S. P. Thompson	5/24/71
John R. Wait	5/24/71
Victor H. Bellamy	5/23/71
Joseph F. Forbes	6/15/71

*This approval and acceptance is contingent on the candidate's adequate performance and defense of this dissertation at the final oral examination. The inclusion of this sheet bound into the library copy of the dissertation is evidence of satisfactory performance at the final examination.

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his judgement the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: John C. Bellamy

ACKNOWLEDGMENTS

The author is most grateful to Dr. P.H. Bartels of the Optical Sciences Center and Microbiology department at the University of Arizona for supplying a topic, applications, and numerous suggestions for the research leading to this dissertation. The author is indebted to both his major advisor Dr. F.J. Hill and Dr. Bartels for reading the preliminary drafts and providing the suggestions responsible for whatever readibility the reader may find.

The United States Air Force is also greatly appreciated for supplying through the Office of Aerospace Research under Project THEMIS the much needed research assistantship and computer time.

TABLE OF CONTENTS

		Page
	LIST OF ILLUSTRATIONS	vi
	LIST OF TABLES	vii
	ABSTRACT	viii
1.	INTRODUCTION	1
	Definitions and Notation	5
2.	PROGRAM DEVELOPMENT	10
	Theory	10
	Feature Selection	11
	Distance Measures	13
	Decision Functions	14
	Implementation	19
	Parametric Methods	19
	Nonparametric Methods	24
	Success Measures	28
3.	PROGRAM DESCRIPTION	30
	Subroutine EXTRACT	30
	Subroutine DSELECT	32
	Subroutine LRATIO	35
	Subroutine MVA	37
	Program DISTILL	38
	Auxiliary Subroutines	42
4.	SPURIOUS INFORMATION	46
	Spurious Information Expected from MVA	48
	Spurious Information Expected from EXTRACT	51
	Spurious Information Expected from Composite Functions	54
5.	FEATURE GENERATION ROUTINES	59
	Pre-Processing	61
	Subroutine HISTGM	62
	Subroutine ARRATIO	65
	Subroutine CELLID	66

Page

Subroutine POLAR 67
Auxiliary Subroutines 67

6. RESULTS 69

Processing Procedures 69
Normal Human Lyphocytes vs Lymphocytes from Asymptomatic Chronic Lymphocytic Leukemia 70
Normal Endometrium vs Hyperplasia of the Endometrium 71
Normal Endometrium vs Well Differentiated Adenocarcinoma of the Endometrium . . . 75
Normal Lymphocytes vs Lymphocytes from Lymphocytic Leukemia 77

7. CONCLUSIONS 79

APPENDIX A: THEOREMS AND DERIVATIONS 86

Derivation of Threshold for Minimum Risk Classification 86
Derivation of Canonical Form 87
Derivation of Equation (3-1) 88
Theorem 1 89
Corollary 90
Theorem 2 91
Divergence Between Two Distributions with Equal Variances 92

APPENDIX B: SPURIOUS INFORMATION SAMPLES 95

REFERENCES 101

LIST OF ILLUSTRATIONS

Figure		Page
1.	Picture of Normal Endometrial Cell	2
2.	Digitized Image of Normal Endometrial Cell	2
3.	Comparison of Marginal Distributions with Distributions of Optimum Linear Combination for Two Bivariate Normal Classes	12
4.	Two Bivariate Normal Distributions and the Corresponding Optimum Linear Combination for Discrimination	23
5.	Sequence of Decision Boundaries Chosen to Partition the Training Sets	26
6.	Flow Chart of Subroutine EXTRACT	33
7.	Flow Chart of Subroutine DSELECT	36
8.	Flow Chart of DISTILL	39
9.	Typical Processing Steps of DISTILL	41
10.	Spurious Information Expected from MVA for 4, 6, 8, and 10 Variables	50
11.	Spurious Information Expected from an Optimized Linear Combination of 10 Independent Variables	52
12.	Spurious Information Expected from Composite Decision Functions on Independent Variables	55
13.	Effects of Cross-Correlation on Spurious Information Expected from Composite Decision Functions	57
14.	Flow Chart of Program REDUCE	60
15.	Histogram of Array Values in Figure 2	63
16.	Polar Conversion and Orientation of Array Image in Figure 2	63

LIST OF TABLES

Table		Page
1.	Descriptor Blocks and Sample Values Obtained from Image of Figure 2	64
2.	Results of LNRM vs ACLL	72
3.	Results of NRME vs HYPE and NRME vs WDAE	74
4.	Results of NRML vs LLLE	78
5.	Spurious Information from 4 Variables Using MVA	95
6.	Spurious Information from 6 Variables Using MVA	96
7.	Spurious Information from 8 Variables Using MVA	96
8.	Spurious Information from 10 Variables Using MVA	97
9.	Spurious Information from 10 Independent Variables Using EXTRACT with Optimization	98
10.	Spurious Information from 4 Blocks of 10 Independent Variables Using EXTRACT with Optimization on Each Block and on Each of the Composite Variables	99
11.	Spurious Information from 4 Blocks of 10 Independent Variables Using EXTRACT with Optimization on Each Block and MVA on the Composite Variables	100

ABSTRACT

A system of Fortran programs has been developed that incorporate various statistical pattern recognition techniques to process and classify digitized image arrays. The principal application has involved biological cells where an automated determination of cell type and state is desired. Using an array of extinction values produced by a scanning microscope, the programs process the information in the array values in such a way as to determine a best estimate of which class is represented by the cell image.

Each digitized image array is initially processed to extract certain descriptors of the image contained in the array. These descriptors form a feature vector which describes certain statistical properties of the image.

Supervised learning is incorporated into the program system in that two populations of feature vectors called training sets are systematically compared to extract the discriminatory information and determine a classification procedure. The class represented by a feature vector not in the training sets can then be determined by applying the recognition procedure developed by supervised learning.

Although the process of generating a feature vector is application dependent, the supervised learning portion

of the program is general enough to be applicable to any statistical pattern recognition problem. The supervised learning procedures process all components of a feature vector in an identical manner, thus ignoring any physical relationships that may exist.

Whenever a classification procedure is determined by the supervised learning programs, certain measures of success are calculated to evaluate the performance expected of the classification procedure on unknown image types. The significance of these measures is analyzed in two ways. First, the success measures are evaluated when the supervised learning procedures are applied to an unsolvable problem, an unsolvable problem being one in which each training set is derived from an identical class. Any differences in the training sets then result only from the randomness of the variables. Second, a classification procedure is applied to a set of cells for which the type is known but were not included in the training sets. The performance of the classification procedure on these cells is then compared with the predicted performance.

CHAPTER 1

INTRODUCTION

This dissertation describes the application of digital data processing to one aspect of cytology--the classification of cell types. The determination of cell types facilitates a clinical diagnosis of a patient's state of health. Of particular importance in forming a judgment are the presence and frequency of occurrence of various cell types. The automated cytodiagnostic system as described here is not intended to replace the skilled cytopathologist, but to augment the available information upon which diagnostic decisions are made.

A computer is particularly adept at determining the statistical properties of a large array of numbers. Thus a computer can assess statistical information that is not easily assimilated into a visual impression. The human observer, on the other hand, can more easily evaluate other useful information such as the arrangement of neighboring cells (diathesis). Figure 1 is a picture of a typical cell that might be presented to a computer for classification. The texture, or granularity of the cell contains much discriminatory information (Bartels et al.,

Figure 1. Picture of Normal Endometrial Cell

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28	32	0	0	0	23	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	40	62	76	129	134	107	62	24	0	23	0	23
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	79	79	84	60	91	99	68	115	123	38	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	24	101	96	95	76	67	67	63	71	107	173	123	46	23	24
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	65	76	79	81	82	81	56	51	79	82	111	77	63	40	24
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	65	90	79	102	91	51	42	84	77	90	95	110	60	38	51	23
0	0	0	0	0	0	0	0	0	0	0	0	0	0	37	84	82	85	110	105	101	79	158	90	82	110	143	68	45	56	29
0	0	0	0	0	0	0	0	0	0	0	0	0	23	70	70	96	134	144	137	129	88	93	99	88	140	126	90	107	87	26
0	0	0	0	0	0	0	0	0	0	0	0	0	32	79	95	138	165	151	88	81	77	67	93	110	107	102	77	95	95	24
0	0	0	0	0	0	0	0	0	0	0	0	29	48	87	98	146	120	148	165	144	123	113	98	101	126	104	65	54	57	0
0	0	0	0	0	0	0	0	0	0	24	31	26	57	104	67	96	88	134	163	132	85	109	98	90	116	74	68	65	49	0
0	0	0	0	0	0	0	0	24	34	43	31	31	54	88	82	120	98	87	81	107	99	79	57	76	104	65	79	82	0	0
0	0	0	0	0	0	0	0	28	0	31	32	45	38	104	143	138	121	95	74	77	87	95	57	93	112	71	85	45	0	0
0	0	0	0	0	0	0	26	38	42	42	42	40	34	35	101	155	110	91	110	93	74	71	99	132	96	76	49	0	0	0
0	0	0	0	0	0	0	23	28	35	38	31	31	35	29	34	37	113	84	102	124	104	98	143	151	87	31	0	0	0	0
0	0	0	0	0	0	0	24	28	32	31	31	28	34	31	32	32	93	112	171	158	113	82	101	68	0	0	0	0	0	0
0	0	0	0	0	0	26	31	35	32	31	37	31	32	32	37	37	35	70	105	81	76	40	0	0	0	0	0	0	0	0
0	0	0	0	0	23	34	35	29	32	34	32	29	34	35	31	34	32	34	26	24	23	0	0	0	0	0	0	0	0	0
0	0	23	29	0	28	26	28	35	42	31	31	24	28	31	32	38	23	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	29	28	31	23	24	31	32	28	31	35	32	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	23	23	26	26	28	32	28	40	37	32	34	45	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	23	0	23	28	26	31	38	48	40	35	37	35	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	23	32	35	29	32	32	34	32	23	29	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	24	23	23	24	45	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	26	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2. Digitized Image of Normal Endometrial Cell

1970). These textural properties are easily evaluated by processing the digitized image of the cell as shown in Figure 2.

One particular advantage of automated cytodiagnosis is that objective quantitative definitions of cell types can be established whereas the cytopathologist uses subjective qualitative impressions that are not easily verbalized and defined. Automation makes universal standards possible through the rapid and more efficient exchange of information and experience.

Another attractive feature of an automated cytodiagnostic system is that a digitized image can be transmitted from a remote site to a computer center and evaluated. Thus an expert does not have to be available to examine the material itself.

The problem is approached by applying statistical pattern recognition techniques to digitized images of biological cell types. A digitized description of a cell can be readily produced by a scanning microscope where the numerical data is readily processed by a computer (Wied et al., 1968). A typical digitized image is shown in Figure 2. In this rectangular array of numbers each element represents an extinction value at the corresponding x-y coordinate of the cell. Extinction values ev are determined as:

$$ev = \log_e \left(\frac{\text{incident light intensity}}{\text{transmitted light intensity}} \right) \qquad (1\text{-}1)$$

Thus larger ev values correspond to points in the scan with higher optical densities.

A procedure is desired whereby the digitized representation of a cell is processed to determine its type. With the above application in mind the methodology of pattern recognition has been incorporated in a system of programs designed to determine such a procedure. Supervised learning techniques are used in that representative images of two populations (i.e., normal and abnormal) are systematically compared to extract the discriminatory information and determine the classification procedure.

The system of programs as described in Chapter 3 can be applied to a variety of pattern recognition problems.[1] For this reason the problem is formulated in a general sense although the particular approach described in Chapter 2 as the solution was chosen with the cell recognition problem in mind. Basically, the techniques are applicable to any problem in which the pattern measurements indicate significant variability within each class according to some type of a continuous probability distribution.

---

1. The programs have been used successfully to distinguish between urban and rural landscapes using digitized representations of aerial photographs.

Definitions and Notation

The categories or classes of objects to be recognized are considered to belong to a set $C = \{c_1, c_2, \ldots, c_{nc}\}$. Where needed, the a priori probability of each class c_k is denoted by $p(c_k)$. For the most part discrimination between only two classes c_1 and c_2 is considered. However, the results can be readily applied to the multiclass problem by using a sequence of pairwise comparisons.

The numerical information about a pattern is represented by a vector $\underline{x} = (x_1, x_2, \ldots, x_n)$ in n-space denoted by E^n. The $\underline{x}$ is allowed to assume only certain values in E^n depending on the pattern class represented. Associated with each class c_k there is an n-variate probability distribution $f_k(\underline{x})$ which is a mathematical description of the variability within class k. This distribution is unknown or at best estimated.

The selection of a pattern can now be characterized as a sequence of two random samplings. The first selects a class c_k from the alternatives in C and the second samples a vector $\underline{x}$ according to $f_k(\underline{x})$.

The learning phase (determination of a classification procedure) is that of finding a mapping d from E^n to C. A classification procedure is derived from the information contained in a set T_k of N_k sample patterns from each class k called training sets. The procedure so derived

classifies the training sets in a manner that is in some sense optimal with respect to a set D of permissible mappings. If each training set contains pattern samples that are typical representatives of their respective classes a meaningful classification procedure results.

In the above case, the significant differences between two large populations are determined from comparatively small subsets of each. In general it is tacitly assumed that the subsets are representative. A common method of determining the consistency of a newly derived procedure is to classify another set of known patterns that is not included in the training sets. The statistical significance that can be assigned to a classification procedure considering the training sets alone appears to be a better founded measure for such a determination. A method assessing this significance is described in Chapter 4.

For a given decision function d the probability of misclassification PE(d) can be determined as:

$$PE(d) = p(c_1) \int_{X_2} f_1(\underline{x})\, d\underline{x} + p(c_2) \int_{X_1} f_2(\underline{x})\, d\underline{x} \qquad (1\text{-}2)$$

where $X_1 = \{\underline{x} : d(\underline{x}) = c_1\}$

$X_2 = \{\underline{x} : d(\underline{x}) = c_2\}$

$X_1 \cup X_2 = E^n$

The criterion to be optimized here is the minimization of PE while the a priori probabilities $p(c_1)$ and $p(c_2)$ are assumed equal. The mapping d_{opt} minimizing PE over all possible mappings $d \in D$ then is the desired decision function. Appendix A shows how to generalize the classification decision to allow unequal a priori probabilities and to be optimal in a Bayes sense.

To facilitate the discussion of the processing steps involved, a decision function is considered as the composition of two functions g and h:

$$d = h \circ g \qquad (1\text{-}3)$$

where g maps $\underline{x} \in E^n$ to $\underline{x}' \in E^m$ $\quad m < n$

and h maps $\underline{x}' \in E^m$ into C

Functionally both g and h serve the purpose of information compression--they transfer the information contained in many variables into a set of fewer variables. The transformation g and h differ, however, in the manner of their selection. The determination of g represents feature extraction in that the raw data is preprocessed to produce a feature vector. The problem of determining what features to measure is in general dependent on the application and left to human intuition. Swonger (1968) and Uhr (1968) are among those who have considered generalizing the feature

generation process. The dimensionality of the problem considered here where a cell scan can consist of more than 2000 points prohibits the application of their methods.

The second mapping h represents the decision making process. Various parametric and non-parametric techniques have been programmed into a generalized procedure for determining a classification procedure. The techniques as described in Chapter 2 examine the statistics of two training sets of feature vectors and develop the best classification procedure of the form prescribed. If the $\underline{x}'$ samples do not contain all or most of the discriminatory information contained in the original variables $\underline{x}$ no amount of effort in optimizing h can produce an effective classification procedure. Hence, the selection of g is critical to the success of any recognition attempt.

In general g is determined by selecting all properties that the user feels may provide some discriminatory information. Which features actually do contain useful information is determined by the program. It is not advisable, however, to generate a larger set of features than is considered absolutely necessary. As the number of features increases, so does the likelihood of extracting misleading information. Randomness alone can produce seemingly useful sample values of at least one variable containing no inherent discriminatory information. For

this reason Chapter 4 examines the results of tests used to determine the effect of dimensionality and size of training set on the probability of extracting spurious information.

The set of features used in the cell classification problem is described in Chapter 5. The distribution, arrangement, and general appearance of chromatin particles in the cell nucleus offer important diagnostic clues to the pathologist. Therefore, properties assessing primarily the texture or granularity of the cell images were selected. Since the feature vectors can be well represented as typical or mean values disturbed by noise, the problem is well suited to the type of analysis described in the next chapter.

CHAPTER 2

PROGRAM DEVELOPMENT

This chapter describes the theory behind a system of programs developed to determine the classification procedure represented by the mapping h in equation (1-3). The pattern recognition techniques used to determine a suitable h are general enough to be applied to any statistical recognition problem. For this reason no assumptions are made as to the nature of the $\underline{x}'$ variables other than that they are in some way probabilistically distributed. The actual derivation of the $\underline{x}'$ (selection of the mapping g) is, in general, application dependent and therefore not considered here. The derivation of the $\underline{x}'$ in the cytodiagnostic application is described in Chapter 5. In some cases the features or components of $\underline{x}'$ are variables selected directly from the raw data. For this reason and to retain ease of notation in this chapter, the $\underline{x}'$ are referred to simply as $\underline{x} = (x_1, x_2, \cdots x_m)$.

Theory

Sample values of $\underline{x}$ from a particular class c_k are distributed according to the m-variate probability density function (pdf) $f_k(\underline{x})$. A training set for class k is a

set T_k consisting of N_k vectors representing pattern measurements obtained by sampling $f_k(\underline{x})$. If the pdf's are known a priori the classification problem can be solved using classical decision theory and no learning or training techniques are required. Therefore, the pdf's are assumed to be unknown and are at best estimated when needed.

Feature Selection

The determination of a classification procedure becomes increasingly more complicated as the dimensionality m increases. For this reason it is often desirable to reduce the dimensionality of the problem before establishing the classification procedure. A survey of techniques reducing dimensionality is provided by Prabhu (1970). The function g in (1-3) generally involves dimensionality reduction, but since the features selected to describe a pattern are often of a general nature, some of the variables may contain no discriminatory information for a specific problem. To select only the useful variables can be difficult. A variable may not provide discriminatory information by itself but in conjunction with one or more other variables may contribute to highly reliable classifications. Figure 3 depicts two bivariate classes whose variables have widely overlapping marginal distributions but

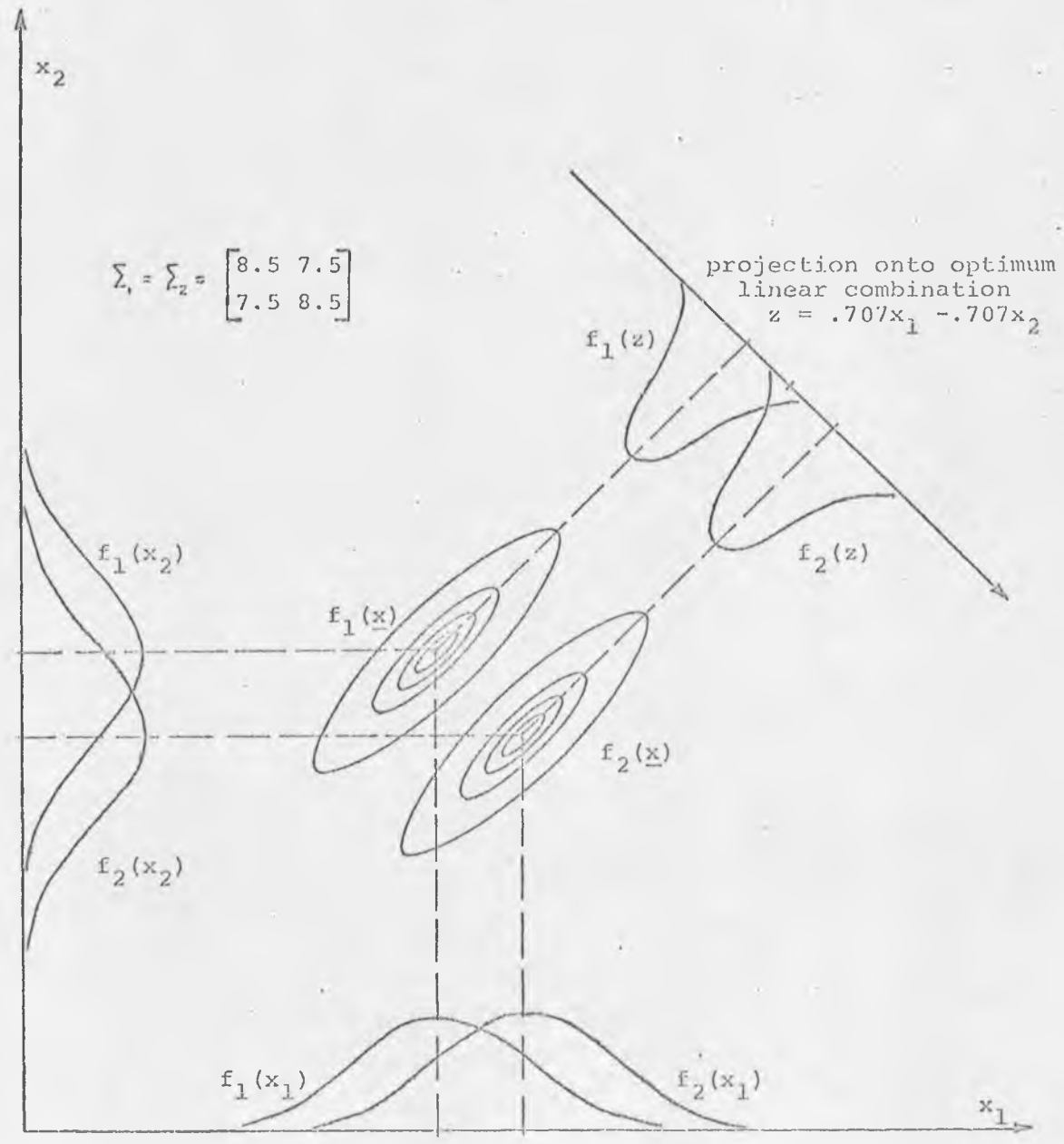

Figure 3. Comparison of Marginal Distributions with Distributions of Optimum Linear Combination for Two Bivariate Normal Classes

nevertheless, highly discriminatory information is provided when they are combined to form the variable z as shown.

Dimensionality reduction is also attained by forming a smaller set of new variables as combinations of the old ones. When combinations are formed, it is desirable to have all the discriminatory information of the old variables contained in the new variables. Sometimes a single linear combination of the given variables contains all the discriminatory information available. Such is the case for the linear combinations shown in both Figures 3 and 4. Various techniques such as principal component analysis (Anderson, 1958) and Karhunen-Loeve expansions (Fukunaga and Koontz, 1970) determine best sets of linear combinations containing as much of the information in the original variables as desired. These procedures generally overcome the shortcomings of individual variable selection. However, they utilize the entire set of original variables. Thus, the computational difficulties induced by spaces of high dimensionality are not alleviated and individual variable selection is still desirable.

Distance Measures. -- Variables may be selected on an individual basis by evaluating some sort of distance measure between the distributions of both classes. Variables with measures indicating a significant difference in

their distributions are retained. Various distance or information measures have been proposed, some of which are discussed by Swonger (1968) and Fukunga with Krile (1969). The program system uses two distance measures which are incorporated into separate processing routines to select variables.

Another distance measure J(1,2) as defined in (2-1) is used as a performance index of a decision function's application to a particular problem. The information measure J(1,2) is also called the divergence as defined by Kullback (1959).

$$J(1,2) = \tfrac{1}{2}[I(1:2) + I(2:1)]$$

$$= \tfrac{1}{2}\left[\int f_1(\underline{x}) \log_e \left(\frac{f_1(\underline{x})}{f_2(\underline{x})}\right) d\underline{x} + \int f_2(\underline{x}) \log_e \left(\frac{f_2(\underline{x})}{f_1(\underline{x})}\right) d\underline{x}\right] \qquad (2\text{-}1)$$

I(1:2) represents the mean information for discrimination of class 1 against class 2 while I(2:1) represents the counterpart for class 1.

Decision Functions

A common technique of solving a pattern recognition problem first estimates the pdf's and then applies statistical decision theory. The assumption is made that the estimated distributions do not differ greatly from nature. One method of estimating a pdf is to assume the

distribution has a known form but unknown parameters. This parametric method uses the training set T_k to estimate the parameter values of $f_k(\underline{x})$. The most common assumption is that the $\underline{x}$ are distributed normally and the training sets are used to estimate the mean and covariance matrix for each class.

Nonparametric techniques can also be used to estimate the pdf's. One of the simplest methods is to use empirical relative frequencies, i.e., histograms of the values observed in the training sets. A major drawback of using histograms is that comparatively large training sets are required (Hughes, 1968). Another technique approximates a general distribution function as a sum of subfunctions. For example, Sebestyen (1962) considers the use of sums of normal distributions to approximate an arbitrary distribution.

When the pdf's are approximated by means as indicated above a statistical model of the feature generation process is produced. This model provides a mapping from the pattern space to the feature space under the associated pdf. The desired discriminant function h can then be realized by selecting the class most likely to produce the sample in question. Maximum likelihood (ml) classification is defined by (2-2) while the extension to the more

general minimum risk (Bayes) classifier is given in Appendix A.

$$ml(\underline{x}) = c_k \quad \text{such that } \hat{p}(c_k)\hat{f}_k(\underline{x}) \geq \hat{p}(c_j)\hat{f}_j(\underline{x}) \quad \text{for every } c_j \in C \quad k \neq j \tag{2-2}$$

Where $\hat{p}(c_k)$ is an estimate of $p(c_k)$ and

$\hat{f}_k(\underline{x})$ is an estimate of $f_k(\underline{x})$

When only two classes are considered the above classification can be implemented using a likelihood ratio (lr) as:

$$ml(\underline{x}) = c_2 \quad lr \geq 1 \qquad = c_1 \quad lr < 1 \qquad \text{where } lr = \frac{\hat{p}(c_2)\hat{f}_2(\underline{x})}{\hat{p}(c_1)\hat{f}_1(\underline{x})} \tag{2-3}$$

The probability $pe(\underline{x})$ of a sample $\underline{x}$ being misclassified by the above classification procedure is:

$$pe(\underline{x}) = \frac{f_2(\underline{x})p(c_2)}{f(\underline{x})} \qquad lr < 1 \tag{2-4}$$

$$= \frac{f_1(\underline{x})p(c_1)}{f(\underline{x})} \qquad lr \geq 1$$

where $f(\underline{x}) = p(c_1)f_1(\underline{x}) + p(c_2)f_2(\underline{x})$

The expected error probability as defined in (1-2) is:

$$PE(ml) = \int_{E^m} pe(\underline{x})f(\underline{x})d\underline{x} \tag{2-5}$$

If the estimated pdf's $\hat{f}_k(\underline{x}) = f_k(\underline{x})$ for all k the integrand in (2-5) is minimized everywhere. Thus ml classification is optimal over all possible procedures when the pdf estimates are sufficiently close to the actual distributions. In general, however, it is difficult to estimate $f_k(\underline{x})$ accurately unless T_k is quite large, or unless some assumptions can be made about the form of the pdf.

An alternative to the assuming of certain forms for the pdf's is to postulate a form for the classifying function h.[1] Thus a recognition problem is solved by choosing a particular discriminant function from a set of preselected functions. A discriminant function chosen in this way can be assumed optimal only with respect to the preselected set. Often, however, nothing more can be asked of a design based on limited information.

The linear discriminant (ld) function as defined in equation (2-6) is one of the most easily implemented and often used.

$$\begin{aligned} ld(\underline{x}) &= c_1 \quad z \leq 0 \\ &= c_2 \quad z > 0 \end{aligned} \tag{2-6}$$

---

1. When certain forms for the probability distributions are assumed a form for the discriminant function is often implied. The optimal decision boundary, e.g., for two multivariate normal distributions is a hyperconic in E^m.

$$\text{where} \quad z = w_o + \sum_{i=1}^{m} w_i x_i = w_o + \underline{w} \cdot \underline{x}^T$$

$$\underline{w} = (w_1, w_2, \cdots, w_m)$$

A linear discriminant function is optimal over all possible discriminant functions when the variables representing each of two pattern classes are normally distributed with equal covariance matrices (Nilsson, 1965). This situation occurs when both classes can be characterized by typical or mean vectors disturbed by the same gaussian noise--as is the case in the classical communication channel.

In many practical problems the optimal decision surface can not be represented by a hyperplane. For this reason a generalized recognition procedure should be capable of implementing nonlinear decision surfaces. A polynomial discriminant function is a generalization of the linear discriminant function capable of producing an arbitrarily complex decision surface. In practice, however, the number of coefficients to be evaluated is prohibitively large for all but the lowest degree polynomials. Specht (1968) provides an effective algorithm for calculating the polynomial coefficients using approximations of the pdf's derived from the training sets.

Implementation

The following paragraphs describe the parametric and nonparametric processing steps that have been incorporated into the generalized procedure for determination of a classification algorithm. Although some of the most general techniques of pattern recognition are included, the particular methods were selected with the cytodiagnostic application in mind.

Parametric Methods

An attractive feature of maximum likelihood classification is that a quantitative measure of confidence is provided for each classification. The likelihood ratio defined in (2-3) is a direct measure of the probability of a misclassification. Thus, thresholds can be easily established in any application in which action should be taken only when relatively safe classifications are obtained. The value of a discriminant function also provides a measure of confidence for a classification. Such a measure is, however, only qualitative and therefore not easily adapted to minimum risk (Bayes) type decisions. For this reason maximum likelihood classification is used whenever the use of the estimated distributions seems justified. The pdf's are approximated by normal distributions whose parameters are estimated from the appropriate training sets. If an ml

classification is to be made using a single variable z (usually a linear combination of the variables in $\underline{x}$), the distribution for class k is approximated by:

$$\hat{f}(z) = \frac{1}{(2\pi)^{\frac{1}{2}} s_k} e^{-\frac{1}{2}(z-u_k)^2/s_k^2} \qquad (2\text{-}7)$$

where u_k is the sample mean and s_k^2 is the sample variance of the z values occurring in T_k.

When an ml classification is to be made using a set of variables $\underline{x} = (x_1, x_2, \cdots x_m)$, the multivariate approximation as given in (2-8) is used.

$$\hat{f}_k(\underline{x}) = \frac{1}{(2\pi)^{m/2} |S_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{u}_k) S_k^{-1} (\underline{x}-\underline{u}_k)^T} \qquad (2\text{-}8)$$

where $\underline{u}_k = (u_{k1}, u_{k2}, \cdots u_{km})$ is the sample mean vector and $S_k = [(s_k)_{ij}]$ is the sample covariance matrix of the $\underline{x}$ values occurring the T_k.

No attempt has been made to approximate the distributions by any other means. In cytodiagnostic applications training sets are often small enough to preclude using anything but a parametric assumption and normality seemed the most reasonable form. A chi-square test has been incorporated into the program to check the closeness of fit to normality for any variable. When the normality

assumption appears unjustified a discriminant function can be used to determine a classification.

The normality assumption has also been used to find a useful linear combination of given variables. The linear combination can be employed as a new composite variable z used in further processing or as a discriminant for classification purposes. Since the set of variables is replaced by the single variable z, the dimensionality of the classification decision has been reduced. Although a single linear combination may not contain all the discriminatory information contained in the replaced variables, the single best linear combination can be found. Sammon (1970) provides a method to extend the procedure to find a best pair of linear combinations. The linear combination as defined in (2-9) represents the optimal discriminant function when the variables in $\underline{x}$ of both classes are independent and identically distributed except for different means. That is $(s_1)_{ii} = (s_2)_{ii}$ for all i and $(s_1)_{ij} = (s_2)_{ij} = 0$ for all i, j with $i \neq j$.

$$z = w_o + \sum_{i=1}^{m} w_i x_i = w_o + \underline{w} \cdot \underline{x}^T \qquad (2\text{-}9)$$

where $w_i = \dfrac{u_{2i} - u_{1i}}{s_i^2}$ $\qquad i = 1, 2, \cdots m$

s_i^2 = the sample variance of x_i for both c_1 and c_2.

A parameter r_i as defined in (2-10) can be used as a difference measure between $\hat{f}_1(x_i)$ and $\hat{f}_2(x_i)$. Although the measure is not the best when the variance of x_i in c_1 differs greatly from the variance of x_i in c_2, r_i directly relates to the probability of correct classification when the variances are equal. Thus r_i can be used as a measure of the discriminatory information contained in the variable x_i alone.

$$r_i = w_i s_i \tag{2-10}$$

Figure 4 displays the equal probability density contours of two independent bivariate normal distributions with equal covariances. Also shown are the projections onto the optimal linear combination.

An optimal linear combination can be defined as in (2-11) for more general covariance matrices (Nilsson, 1965).

$$w^T = S^{-1}\underline{\Delta u}^T \tag{2-11}$$

where $S = S_1 = S_2$ and $\underline{\Delta u} = (\Delta u_1 \text{ -- } \Delta u_2)$ $\Delta u_i = u_{2i} - u_{1i}$

The linear combination in (2-11) is not incorporated into the program for two reasons. First, the more general combination is more difficult to evaluate since covariances between variables must be estimated in addition to the variances of the variables themselves. Secondly, a nonparametric

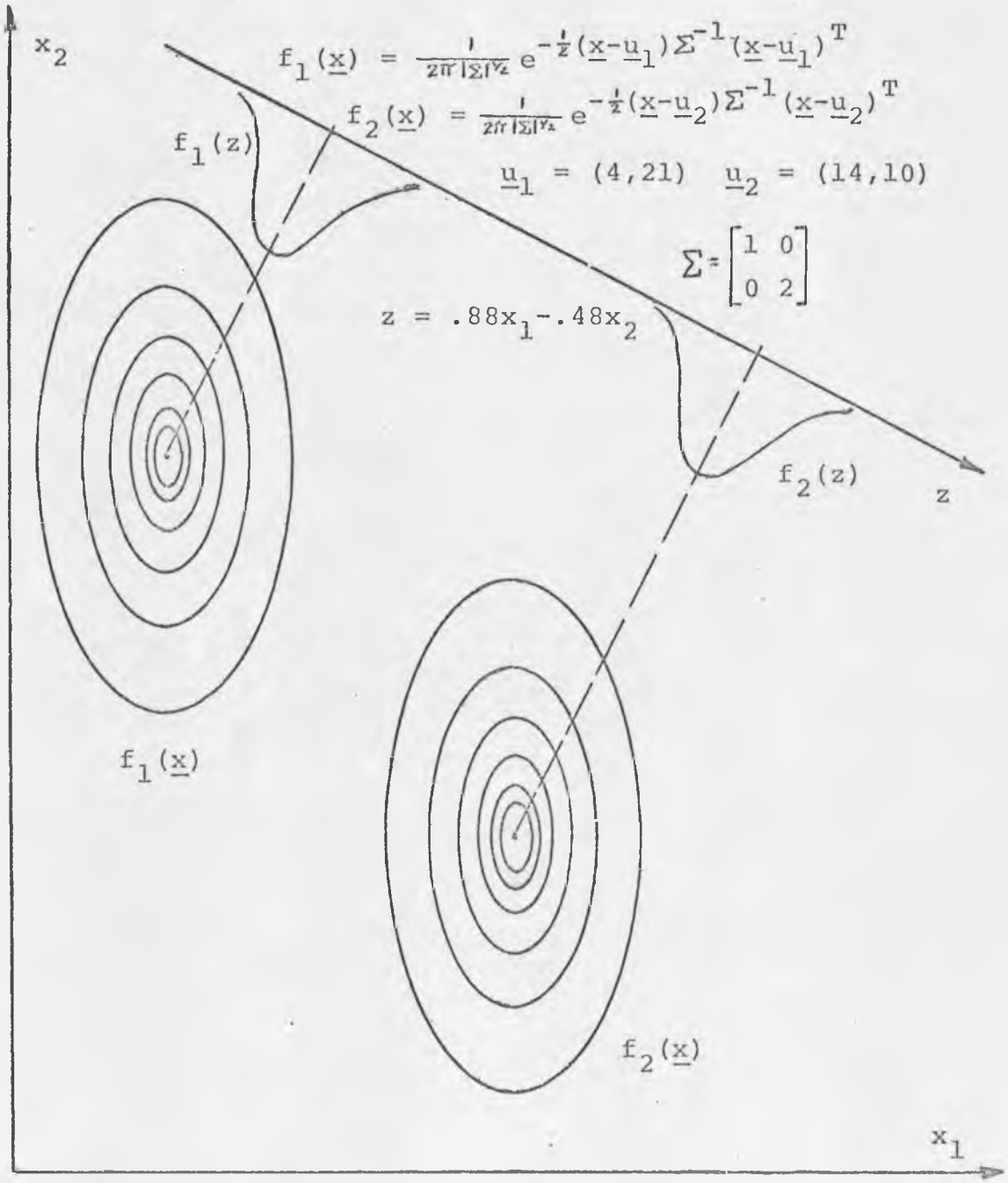

Figure 4. Two Bivariate Normal Distributions and the Corresponding Optimum Linear Combination for Discrimination

routine described below can be used to select relatively uncorrelated variables making the more general analysis unnecessary. In addition, an optimization procedure that finds the best linear combination of arbitrarily distributed variables has been incorporated into the program. Thus, the optimal linear combination derived from the general covariance matrix representing both classes can be determined as a special case of another technique.

Nonparametric methods

A nonparametric procedure has been incorporated into the system of programs that simultaneously selects useful variables and determines a partitioning procedure for the sample space. The algorithm is closely related to the method published by Henrichon and Fu (1969). This algorithm can serve two purposes. First, the procedure selects relatively uncorrelated variables that provide some separation of the training sets. The variables so selected can then be processed by other methods to determine a classification procedure. Secondly, since the separations are determined by partitioning the pattern space into regions containing samples from only one training set, the partitions can be used directly for classification purposes.

The algorithm proceeds by first determining a ranked list of the sample values for each variable and each class. Corresponding lists for each variable are compared in both ascending and descending order. The next step is to determine how many patterns from a given training set are encountered in each ranked list before a pattern from the other training set is found. Thus, if T_1 contains five sample patterns with values of x_i greater (lower) than any value of x_i in T_2, those five patterns can be classified without error by setting a threshold for x_i at the greatest (lowest) value of x_i in T_2. The variable yielding the highest number of error-free classifications is selected and the threshold establishes the first decision boundary. This variable is considered as providing the most information for error-free classification. Patterns that are classified by this decision are removed from the ranked lists and the procedure is repeated, operating on the reduced subsets. In this manner a sequence of decision boundaries is established that separate the training sets by considering one variable at a time. The set of variables so selected are those containing the most information for error-free classification. Since the classifiable patterns are removed at each step, only relatively uncorrelated variables are in the set. Figure 5 shows two bivariate

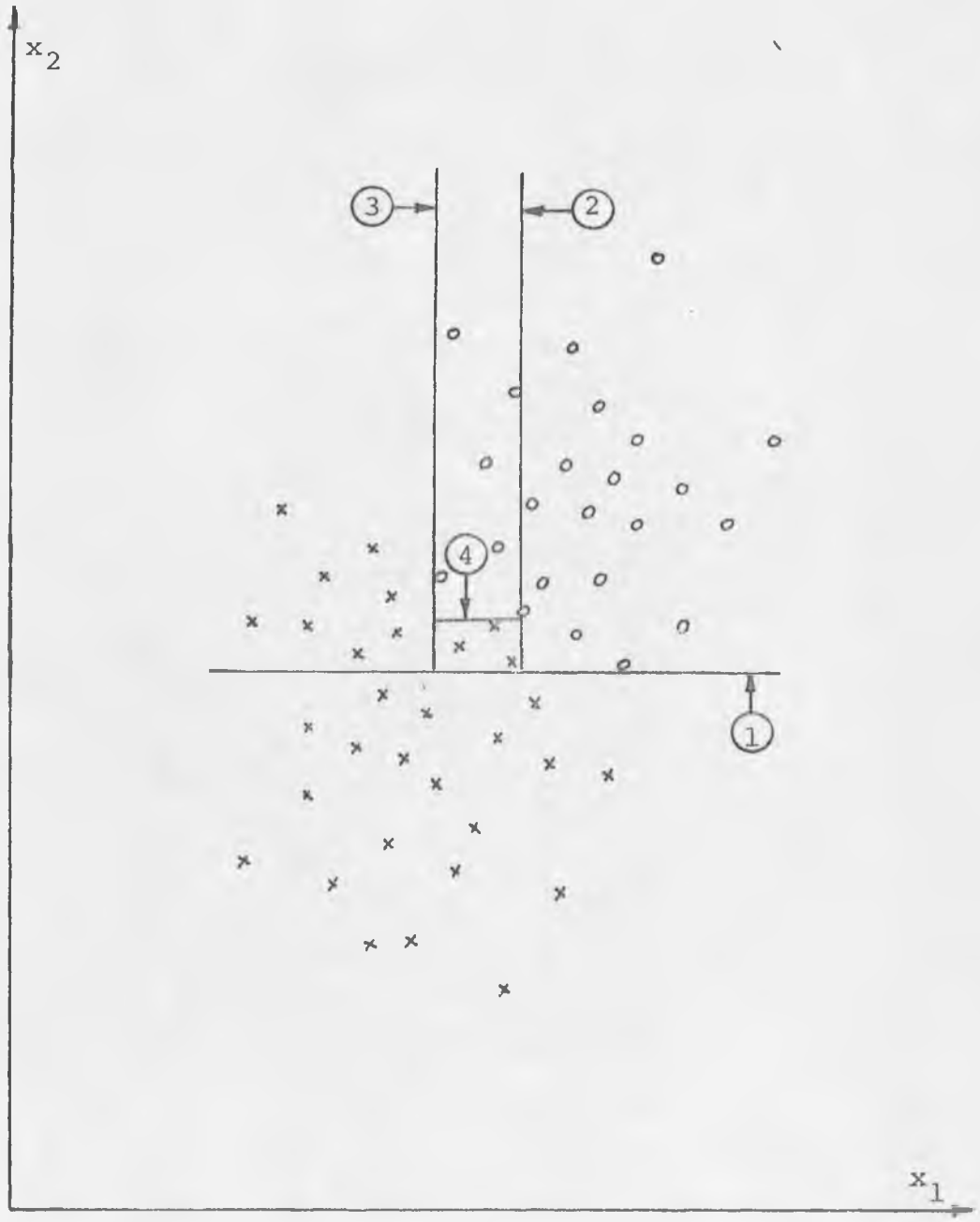

Figure 5. Sequence of Decision Boundaries Chosen to Partition the Training Sets

groups and the sequence of decision boundaries that would be established by the above procedure.

Occasionally an outlying point from one training set can prevent an otherwise useful variable from being selected. Provisions have therefore been included to allow a controlled number of errors. Thus, the search for a decision boundary can pass by an isolated value in one class to classify a large number of patterns in the other class. An additional decision boundary is established only when it classifies a specified percentage of patterns correctly. Thus, when a region of overlap between the training sets is encountered the procedure terminates.

An optimization procedure has been incorporated into the program to generate a variable z as the best linear combination of the variables in $\underline{x}$. The criterion to be optimized is one of maximizing the average log of the likelihood ratio J(1, 2) as defined in (2-1). If z is normally distributed with equal variances for class 1 and class 2, this optimal criterion also minimizes the probability of error when z is used as a discriminant.

J(1, 2) is maximized by varying the weight vector $\underline{w}$ until no improvement is obtainable. Since the effectiveness of any sequential optimization procedure is dependent on a good starting point, $\underline{w}$ is initially established by the values defined in (2-9). The linear

combination defined in (2-9) does in fact maximize J(1, 2) when the pattern vectors $\underline{x}$ of both classes are normally distributed with equal diagonal covariance matrices.

Success Measures

Whenever a classification procedure is established, the program determines a measure of success by applying the algorithm to the training sets. The percentage of errors is one measure that is always given. Another indication of the classification procedure's effectiveness is the average log of the likelihood ratio J(1, 2). This measure is determined whenever maximum likelihood is used as the decision criterion, J(1, 2) is approximated from the sample ratios as:

$$\hat{J}(1,2) = \tfrac{1}{2}[\hat{I}(1;2) + \hat{I}(2:1)]$$

$$= \tfrac{1}{2}\left[\frac{1}{N_1}\sum_{T_1} \log_e\left(\frac{\hat{f}_1(\underline{x})}{\hat{f}_2(\underline{x})}\right) + \frac{1}{N_2}\sum_{T_2} \log_e\left(\frac{\hat{f}_2(\underline{x})}{\hat{f}_1(\underline{x})}\right)\right] \tag{2-12}$$

where $\hat{f}_k(\underline{x})$ is defined in (2-8).

$\hat{J}(1, 2)$ is nothing more than a sample of the divergence defined in (2-1). Rather than perform the integration indicated in (2-1), actual sample averages are used for two reasons. First, integration is more difficult and time consuming especially in multidimensional spaces. Second,

the actual likelihood ratios give a better indication of what to expect in future classifications when the normal approximations do not accurately represent the true distributions.

Both measures of success are derived from the same patterns used to determine the classification procedure. Thus, a classification procedure may be tailored to the training sets to produce high measures of success even though only limited success is possible with other samples from the sample populations. Consequently, the measure may provide misleading information as to the effectiveness of the recognition algorithm. Chapter 4 contains the results of using simulated pattern spaces to estimate the amount of spurious information that may arise from various combinations of processing steps.

CHAPTER 3

PROGRAM DESCRIPTION

The supervised learning procedures described in the previous chapter have been incorporated into a system of Fortran programs. This chapter describes the basic procedures employed by the various routines in deriving the over all decision function h of equation (1-3). The main program DISTILL controls the processing procedure as established by a set of control parameters. These parameters allow a variety of procedures to be used. The principal subroutines used to determine a classification procedure are described in the following sections.

Subroutine EXTRACT

Subroutine EXTRACT is designed to process two sets of feature vectors serving as training sets, and to establish a new variable as a linear combination of the given variables. Initially, a linear combination as defined in (2-9) is established. Also determined is a normalized vector of relative weights $\underline{r} = (r_1, r_2, \cdots r_n)$ where the r_i are defined in (2-10). If desired, those variables contributing the least information to the linear combination can be discarded as follows. First, the

relative weights are ranked in descending order. Second, the ranked weights are examined to determine the smallest set whose sum of weights squared exceeds a prescribed value. This prescribed threshold (typically 0.9) establishes the percentage of information in the complete linear combination retained by the reduced set. When PWMIN is set to zero, the entire linear combination is retained except for variables with identical distributions in both classes. These are usually degenerate in that all sample values are identical. Third, those variables not represented in the set of largest relative weights are discarded. Reduction of the variable set is of particular use when the optimization procedure described below is used.

The next step of EXTRACT, if desired, is to refine the weight vector using an optimization procedure. When the control parameter LINOPT is positive, subroutine OPT9 is called. OPT9 incorporates a Fletcher-Powell optimization routine (Huelsman, 1968) to define successively better and better weight vectors. The best weight vector is the one that maximizes the sample divergence $\hat{J}(1,2)$ as defined in (2-12). Since the effectiveness of optimization procedures can be greatly improved by reducing the dimensionality of the problem, previous variable selection is desirable.

Variables can be selected by using the relative weights as described above or another routine (DSELECT) described below can be called prior to the call of EXTRACT to select the most useful variables. In any event a limit can be set on the number of components to be optimized. Whenever OPT9 is to be called using more than the limited number of descriptors, only the descriptors corresponding to the largest relative weights are retained.

The last step of EXTRACT is to determine the values of the composite variables defined by the newly derived linear combination. The entire set of given variables is then replaced by the new variable. A flow chart of EXTRACT is provided in Figure 6.

Subroutine DSELECT

Subroutine DSELECT incorporates the variable selection and partitioning procedures described in the non-parametric methods section in the preceeding chapter. The first step of DSELECT is to generate ranked lists of descriptor values for class 1 and class 2. The results are stored in separate arrays for each class. Next, subroutine SELECT is called to examine the arrays of ranked descriptors. SELECT determines the number of outlying values in class 1 for each variable. The highest number encountered and corresponding variable index and threshold are stored. Next,

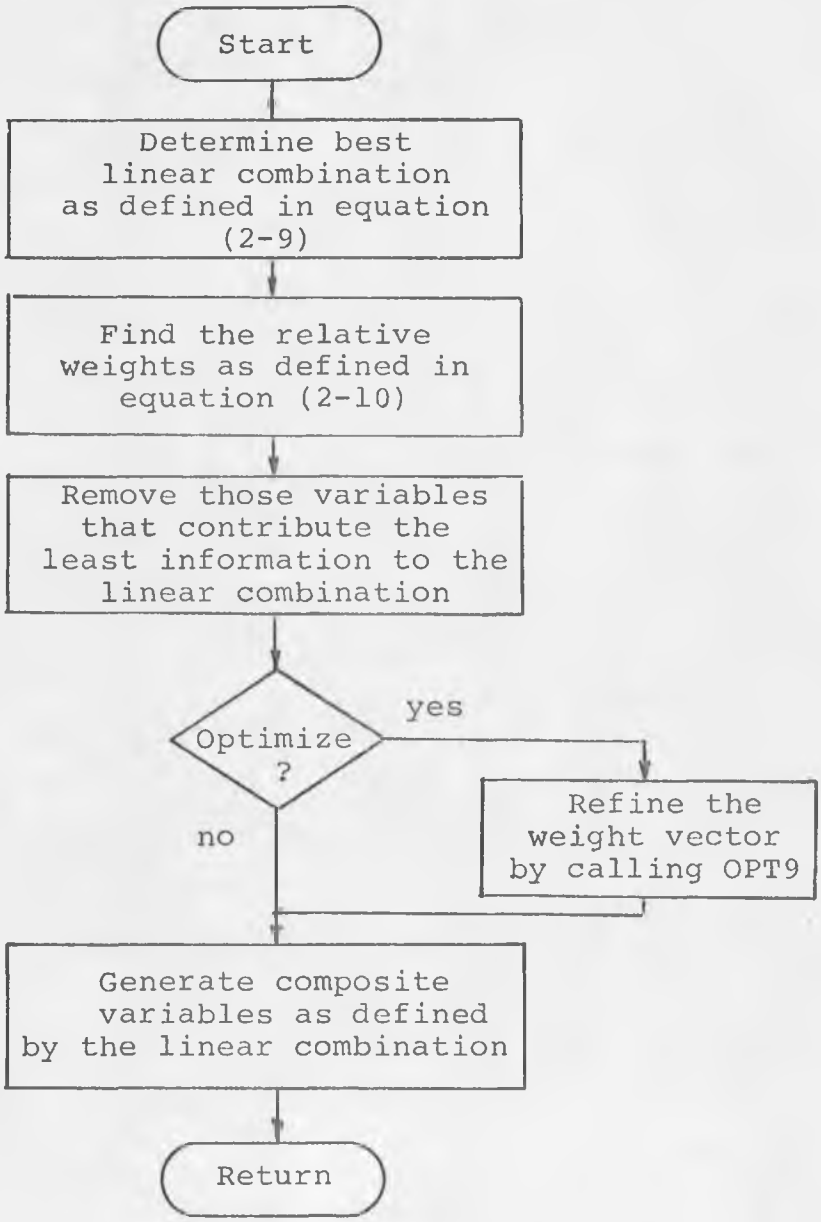

Figure 6. Flow Chart of Subroutine EXTRACT

all variables are examined to determine which one has the most outlying values for class 2. The highest number so determined is compared to that obtained for class 1. The highest overall number of outlying values determines which variable is selected and used to establish a decision boundary. The results of this first pass are stored in arrays. The descriptor values corresponding to the points classified by the decision boundary are removed from the ranked lists. The lists are then examined in another pass to determine the next decision boundary. This process is repeated until all patterns are classified, or until all remaining patterns occur in a region of prohibitive overlap between pattern classes.

No variable is selected unless the corresponding decision boundary classifies a certain percentage of points correctly. Subroutine SELECT establishes a parameter value as a certain percentage of the points in the class to be examined. Only when the number of outliers in that class exceeds this value is a variable selected.

Another parameter value is established in DSELECT as the minimum ratio of correctly to incorrectly classified points that a decision boundary must satisfy. This parameter is initially set to a value large enough to allow no errors. When SELECT cannot find a high enough percentage of outliers, the minimum ratio is reduced. Thus, a

controlled number of errors are permitted so that a decision boundary can pass by a few isolated points of one class. The minimum ratio is repeatedly reduced until a suitable decision boundary is established or until a prescribed minimum is reached. When the minimum ratio is reduced to its lower bound, the search for decision boundaries terminates. At this point the pattern classes are determined to be overlapping. A flow chart of DSELECT is provided in Figure 7.

Subroutine LRATIO

Subroutine LRATIO determines the likelihood ratio for all univariate descriptor values prescribed in the calling statement. LRATIO first calls subroutine PARAM to determine for both classes the parameters of the normally distributed model defined in (2-7). Using the distribution models so derived the likelihood ratio of each descriptor value is determined. Also determined is the average of the logs of the likelihood ratio and the percentage of errors. The average log represents a univariate case of the sample divergence defined in (2-12).

Normally all results are printed, but a control parameter has been incorporated to facilitate the optimization procedure described in EXTRACT. The optimization may call LRATIO many times in refining the weight vector. Since these intermediate calls are of no interest

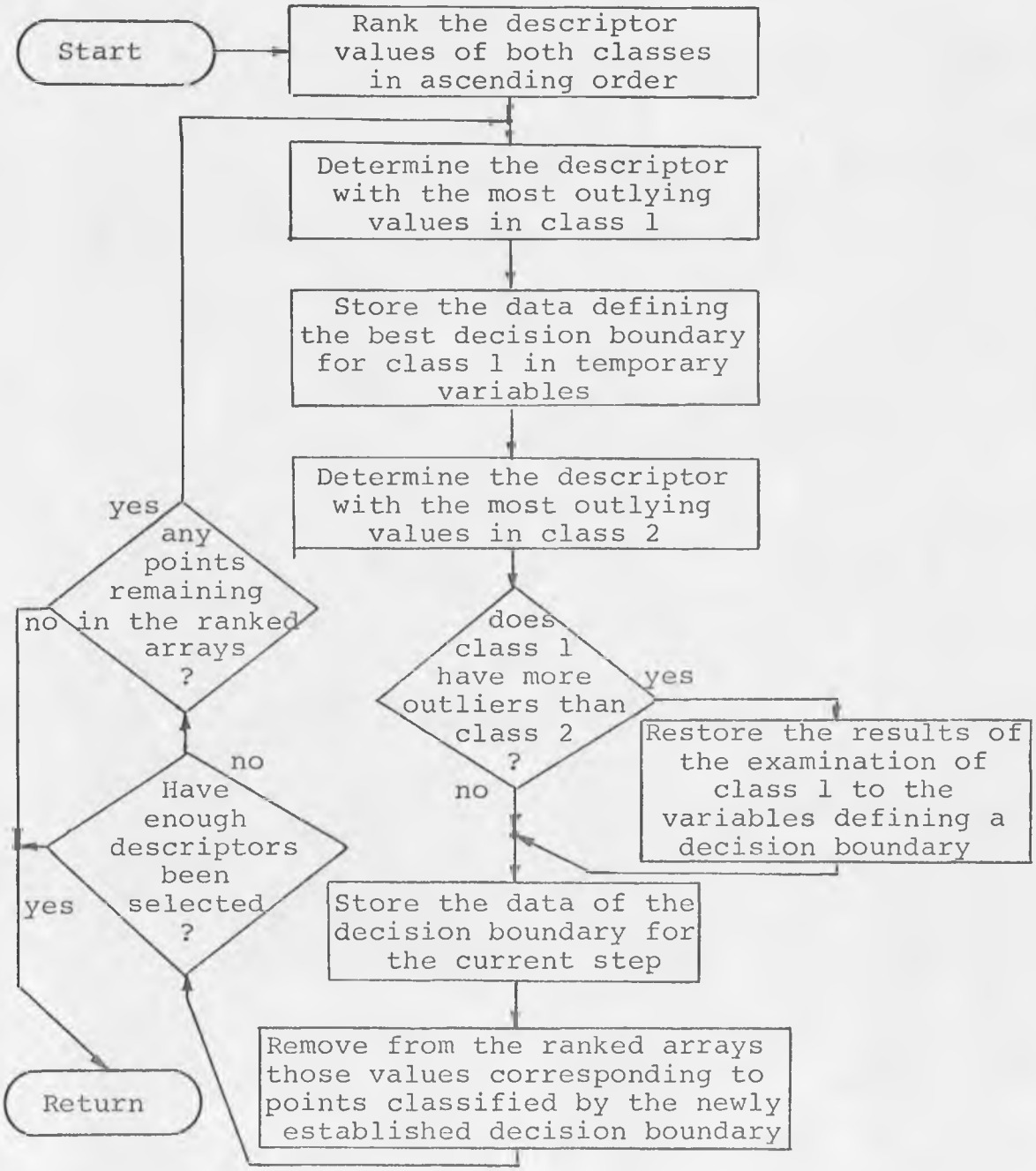

Figure 7. Flow Chart of Subroutine DSELECT

to the user, the results need not be printed. When the control parameter is less than or equal to zero, only the processing necessary to determine the average log is performed and no results are printed.

The last step of LRATIO calls subroutine CHISQ when prescribed by a control parameter. CHISQ determines the goodness of fit to a normal distribution for the distribution of sample values.

Subroutine MVA

Subroutine MVA evaluates the likelihood ratio of all vectors specified in the calling statement. Thus, MVA performs the same function for multivariate sample sets as LRATIO does for univariate sample sets. MVA begins by finding for each class the parameters of the multivariate model as defined in (2-8). These multivariate models are then used to estimate the likelihood ratios of each sample vector.

Rather than find the inverse of each covariance matrix as indicated in (2-8) to determine a likelihood ratio, a diagonalization procedure is used as follows. Subroutine MLEW is called to determine normalized eigenvectors and eigenvalues of each covariance matrix. The eigenvectors are divided by the square root of their respective eigenvalues. The matrix of normalized eigenvectors

then represents a linear transformation of the given variables to a new set of independent variables with an identity covariance matrix. The exponent of the normal distribution function can be easily evaluated as the Euclidean distance between the transformed sample vector and the transformed mean vector. The log to the base 10 of the likelihood ratio may then be determined by evaluating equation (3-1). A detailed derivation of (3-1) appears in Appendix A.

$$\log_{10}(l) = \frac{1}{4.605}\{r_{12} + d_1^2 - d_2^2\} \qquad (3\text{-}1)$$

where $d_k^2 = (\underline{z}_k - \underline{v}_k) \cdot (\underline{z}_k - v_k)^T \qquad k = 1,2$

$\underline{z}_k$ is the transformed sample vector and $\underline{v}_k$ is the transformed mean vector.

Program DISTILL

Program DISTILL is the main program designed to control the sequence of processing steps effected by calls to the previously discussed subroutines. A set of control parameters have been established to determine which subroutine is called at each of a series of steps. These basic steps are indicated by the flow chart of Figure 8. Each of the steps in Figure 8 represent a call to subroutine STEP. STEP examines the control parameters passed in the calling statement to determine which subroutines to call.

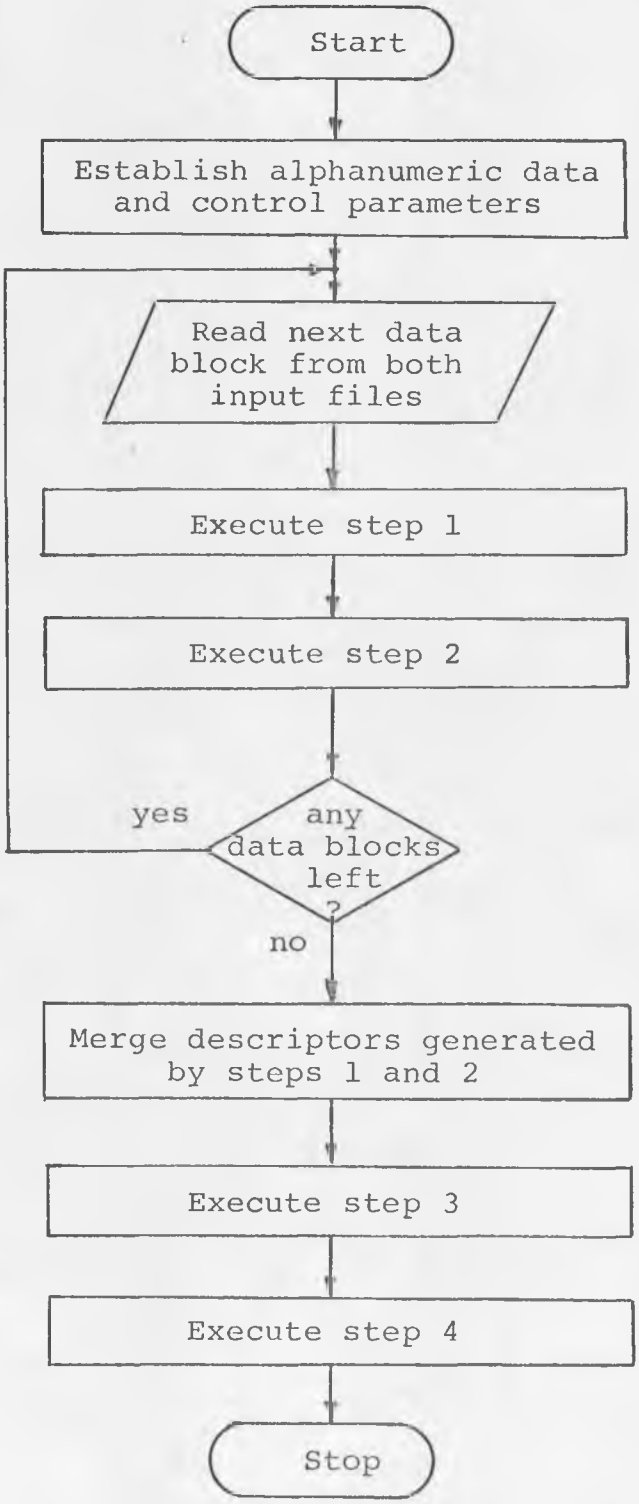

Figure 8. Flow Chart of DISTILL

In this manner, at any step any processing procedure can be used as seems appropriate to the particular problem. A generally useful sequence of steps is established by preset values of the control parameters. This particular sequence of processing steps is indicated in the flow chart of Figure 9.

DISTILL begins by reading new values as desired for those control parameters that are changed from their preset values. The entire set of parameters established to control the sequence of steps is then printed. DISTILL then reads a feature vector data block for each class and processes them with two calls of subroutine STEP. Additional data blocks are read and processed by the same procedures until the entire file of each class is read. After all data blocks have been processed, the results obtained from the separate blocks are merged into one data block for each class. One or two additional calls of STEP can then be used to determine the classification procedure. Two calls of STEP are allowed in processing each block of data so that DSELECT can be used to select variables and reduce the dimensionality before calling EXTRACT or MVA.

Various options have been incorporated into DISTILL to provide auxiliary information and processing of the data being processed. These options are as follows:

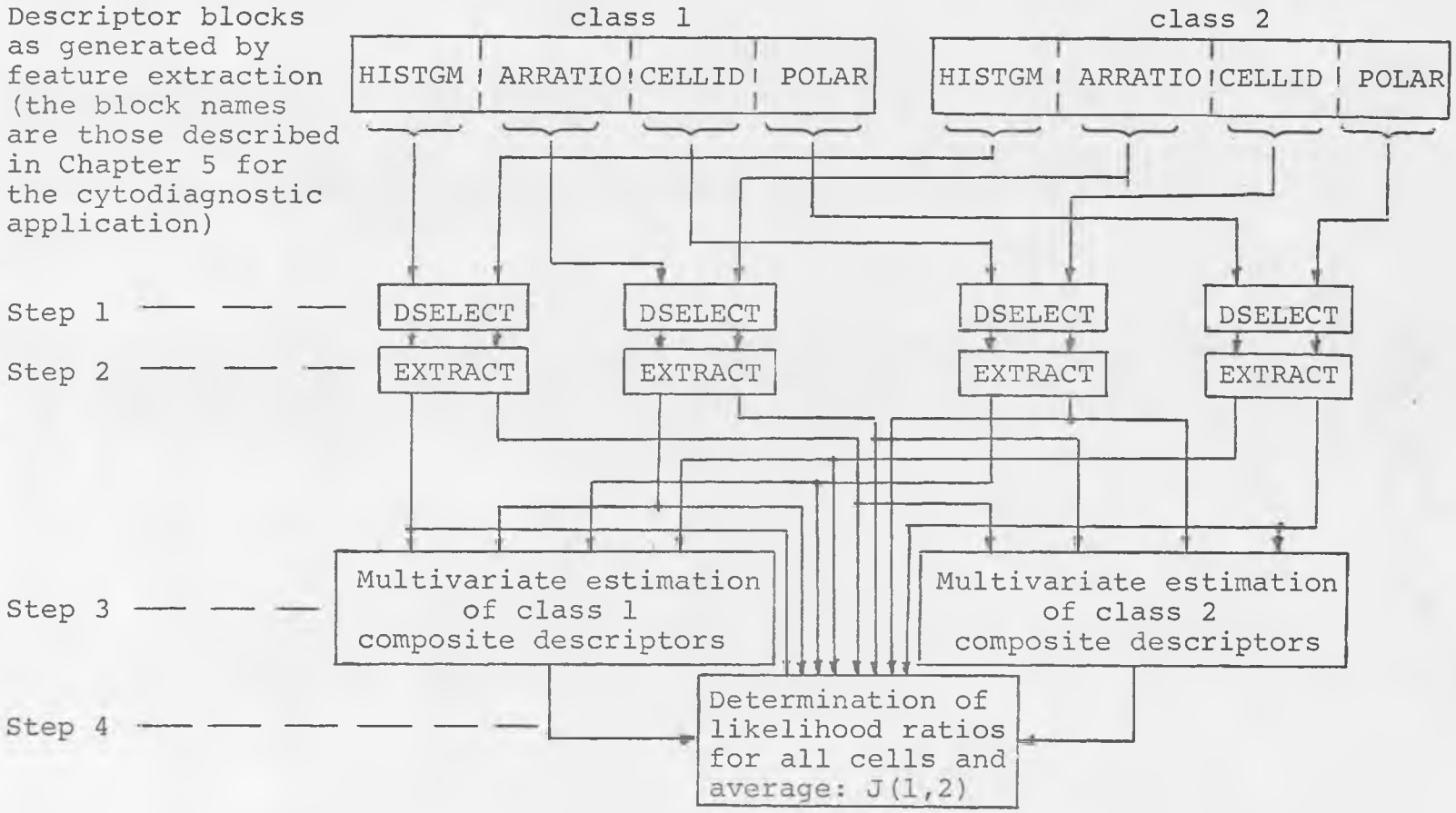

Figure 9. Typical Processing Steps of DISTILL

1. Removal of certain vectors in the data blocks as previously prescribed in an array.

2. Removal of certain variables in the data blocks as read from cards.

3. Printing of each input data block when desired.

4. Determination of likelihood ratios for each individual variable in the input data block when desired.

5. Printing the results of steps 1 and 2 as contained in the merged data block when desired.

6. Punching the results of steps 1 and 2 when desired.

7. Determination of likelihood ratios for the variables produced by any intermediate step when desired.

8. Punching the classification procedure when desired.

Auxiliary Subroutines

This section describes the basic function of the subroutines called by the previous programs in performing their prescribed operations.

1. Subroutine ANLYZ is called by OPT9 to evaluate the sample divergence of the composite variables defined by the weight vector W.

2. Subroutine CHISQ is called by LRATIO whenever it is desirable to determine how closely a sample distribution matches a normal distribution. The output of CHISQ

is the probability that the evaluated test statistic could be as large or larger under the normal hypothesis.

3. Subroutine GRAD is called by OPT9 to evaluate numerically the gradient of the function being optimized. In this case the sample divergence of a composite descriptor is being maximized.

4. Subroutine LLIK is called by MVA to evaluate the likelihood ratios of each vector in a prescribed set. LLIK first transforms the variables into a set of independent variables. LLIK then determines the euclidean distance between each transformed vector and the transformed mean of each class. The likelihood ratio for each vector is then determined as in equation (3-1).

5. Subroutine LINCOMB determines a composite descriptor value CD as a combination of the variables in VECT. The linear combination as established in the weight vector W is determined as

$$CD = \sum_{J=1}^{NCOORD} W(J)VECT(J)$$

6. Subroutine MLEW is called by MVA to determine the eigenvalues and normalized eigenvectors of the covariance matrix of each class.

7. Subroutine MOVE packs a set of selected descriptors into the leftmost columns of the data vector arrays. MOVE is called by DISTILL, EXTRACT, and DSELECT when certain variables are to be removed from consideration.

8. Subroutine MXMN1 is a matrix multiplication routine called by OPT9.

9. Subroutine OPT9 is a Fletcher-Powell optimization subroutine called by EXTRACT to optimize a weight vector. OPT9 is a modified version of the routine written by Dr. L. P. Huelsman for the GOSPEL optimization package.

10. Subroutine PARAM determines the sample means and covariance of the two sets of descriptor values prescribed in the calling statement. When the calling statement prescribes the same descriptor set, the sample mean and variance of that variable is determined.

11. Subroutine RANK1 is used by DSELECT to rank a prescribed set of descriptor values in ascending order.

12. Subroutine READDF is called by DISTILL to read a block of feature vector data from tape. READDF has the capability of removing vectors from the input file as prescribed by an array.

13. Subroutine REMOVE is called by DSELECT to remove those points in the ranked lists that are classified by a newly established decision boundary.

14. Subroutine SELECT is called by DSELECT to determine the maximum number of outliers that can be found in class 1. When DSELECT reverses the class 1 and class 2 calling parameters, SELECT finds the number of outliers in class 2.

15. Subroutine STEP is used to determine the processes involved at each step defined in DISTILL. The processing subroutines are called as determined by the values of two parameters passed by DISTILL. The first parameters determine whether or not to call subroutine EXTRACT or subroutine DSELECT. The second parameter effects calls to LRATIO or MVA as desired.

16. Subroutine WRTDF is called by DISTILL to print a prescribed block of sample vectors.

CHAPTER 4

SPURIOUS INFORMATION

Ideally, a supervised learning procedure such as DISTILL is designed to implement should recognize which variables or combinations of variables provide the desired discriminatory information and ignore any spurious information arising solely from the variability of the samples. That is, the ideal supervised learning procedure should be capable of detecting the true differences between two signals even though they are imbedded in noise. However, any parameter and consequently any classification procedure that is estimated from samples of a random variable is itself a random variable. Thus, a newly derived classification procedure may or may not be useful depending on the supervised learning procedure's susceptibility to noise. Generally speaking, the more information a supervised learning procedure tries to extract from training sets, the more susceptible the procedure is to noise.

The real test of any classification procedure is the performance attained when the procedure is applied to an extensive set of patterns with known classifications. However, an estimate of a classification procedure's

effectiveness is desirable before such exhaustive tests are made. In addition, large sample sets are sometimes infeasible, particularly in cytodiagnostic applications. Thus, it is advantagous to estimate the effectiveness of a classification procedure using only the training sets.

In this chapter the amount of spurious information is estimated as it may arise when measured random variables contain no inherent discriminatory information. That is, any differences in the samples can be due only to noise. The results are essential for the establishment of significance levels for the measure of success attained by a newly derived classification procedure.

Spurious information arising from identical normal distributions only is considered here. Other distributions capable of producing higher amounts of spurious information will undoubtedly be encountered in the practical application of these programs. The results presented here serve as a lower bound on the measure of success that a decision function must achieve before being considered meaningful.

Two sets of multivariate random variables are synthetically generated from identical distributions. These sets of feature vectors are then used as training sets by DISTILL to determine a classification procedure, and subsequently the average likelihood of a classification between training sets. Since the training sets here are simulated

to represent identical patterns, the average log of the likelihood, or sample divergence J(1,2) should be small. J(1,2) can not be expected to be zero, however, unless the sample size is infinite. For each of several processing procedures in DISTILL, a number of tests were made on separate synthetic data sets to determine empirically the distribution of spurious information and the effects of sample size and dimensionality on the spurious information extracted from synthetic training sets. The results of each run are tabulated in Appendix B. The mean value of the spurious information in each case is presented in graphic form in this chapter. The number of samples is insufficient to obtain a useful estimate of the variance. Appendix B, however, can be consulted for an indication of the variability in spurious information samples.

Spurious Information Expected from MVA

The first investigation involved the determination of the amount of spurious information that MVA can introduce into a classification decision. The question asked here is: to what extent does the decision function using a multivariate normal model tailor itself to the training sets? Each test involved estimating the parameters of a multivariate normal model for each training set. The

parameters so derived were used by DISTILL to determine the average likelihood of the sample values in the training sets. Since the estimated parameters are random variables, the models of each test differed. The averages of the observed sample divergences are shown in Figure 10 for synthetic vectors of 4, 6, 8, and 10 components.

To determine the significance of a newly derived classification procedure based only MVA, Figure 10 should be consulted. The observed divergence must be greater than the average divergence shown in Figure 10 for dimensionality and sample size of the problem, to assume a difference in the populations exists. Additional significance can be attributed to the classification procedure if the sample divergence is graeter than all corresponding test values in Appendix B.

The curves of Figure 10 were determined from models with varying amounts of correlation between variables. The results do not depend on cross-correlation as long as no value of unity occurs. Appendix A contains a proof that the distribution of spurious information depends only on the dimensionality and sample size of a multivariate normal model.

Cross-correlation of unity in effect reduces the dimensionality. Similarily, degenerate variables (those with zero variance) also reduce the dimensionality. When

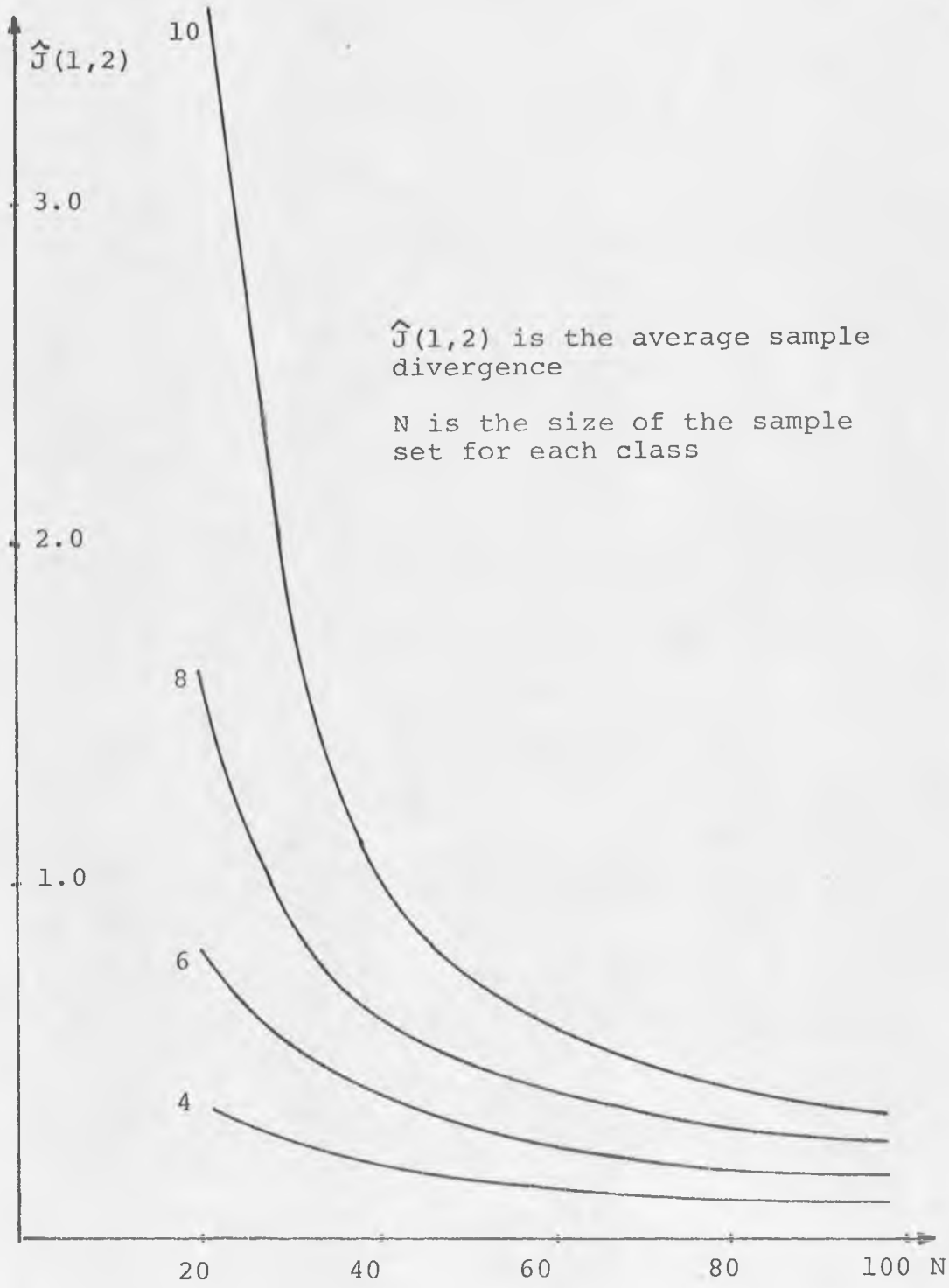

Figure 10. Spurious Information Expected from MVA for 4, 6, 8, and 10 Variables

either of these conditions exist, as in some cytodiagnostic cases, the effective dimensionality must be determined to ascertain the expected value of the spurious information.

Spurious Information Expected from EXTRACT

The amount of spurious information that EXTRACT is capable of introducing into a classification decision was determined as well. As in the previous case, a series of test runs were made comparing two sample sets obtained from identical normal distributions. The optimization procedure of EXTRACT was used to obtain the maximum sample divergence between the sample sets of 10 independent variables. Also considered were the effects of correlation and no optimization on the expected spurious information.

The first tests were made using uncorrelated variables derived from independent normal distributions. The average values of the spurious information are presented graphically in Figure 11 while the individual results are tabulated in Appendix B.

Correlation between the variables was observed to reduce the empirical spurious information in apparent contradiction to a proof in Appendix A that the distribution of maximum spurious information is independent of cross-correlation. The average spurious information decreased about 25% when the average cross-correlation increased to

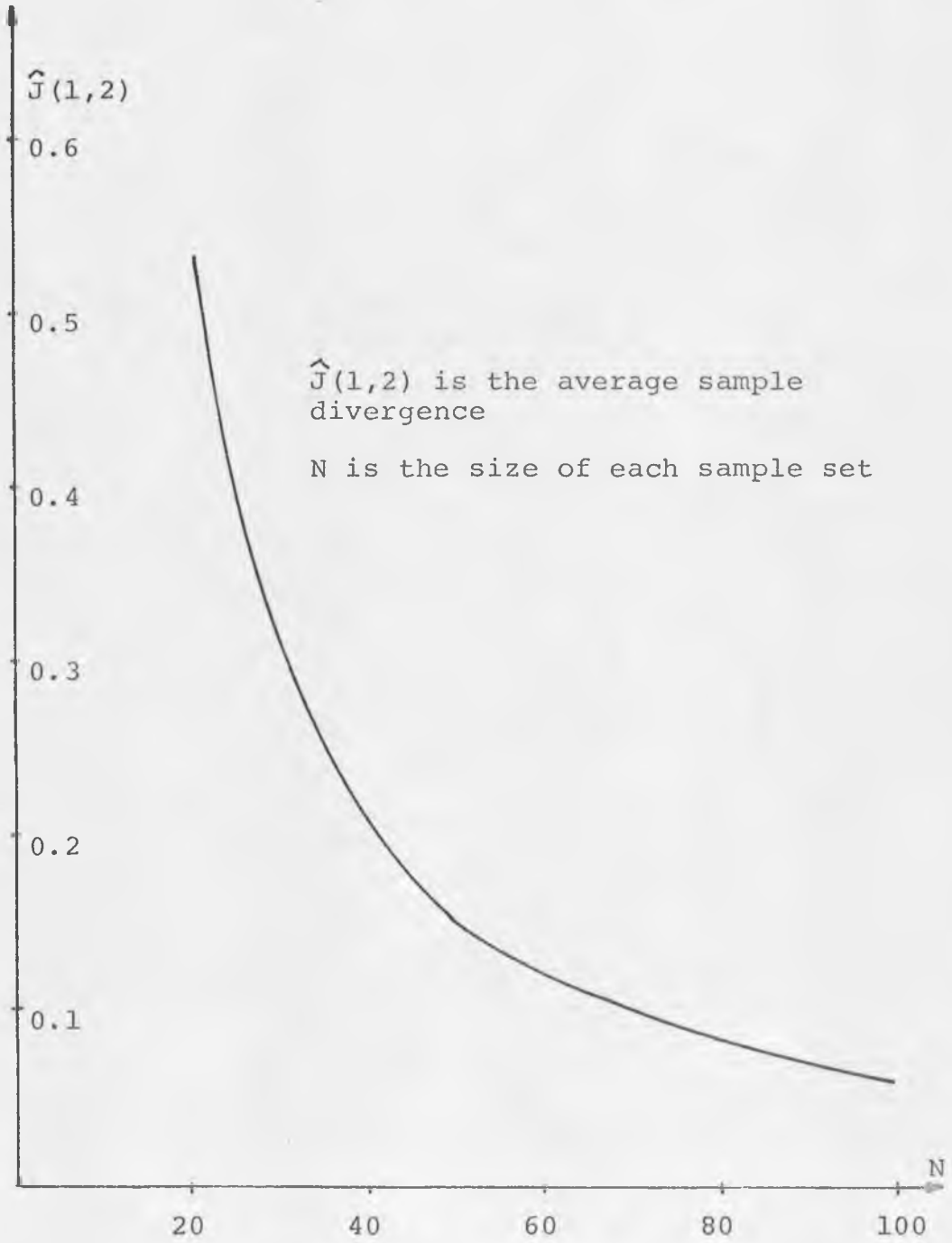

Figure 11. Spurious Information Expected from an Optimized Linear Combination of 10 Independent variables

0,34. Allowing more iterations in the optimization routine did significantly increase some of the sample values, but others appeared as though they could not be increased by any number of iterations. This is explained by the fact that in some cases the optimization procedure finds a local maximum instead of the desired overall maximum. Since the optimization procedure uses a starting point that assumes all variables to be uncorrelated, the starting point is far from the true optimum when high correlations are present.

A limited number of tests were made using EXTRACT without the optimization procedure but the results have not been included. In these cases a marked dependence on cross-correlation is observed since no optimization retains the starting point obtained by assuming variable independence. An extensive set of curves would thus be needed to indicate the dependency on correlation. In general, the spurious information obtained with optimization ranged from two to eight times that obtained without optimization as the average cross-correlation increased from 0.0 to 0.5. Some of the effects of cross-correlation on unoptimized linear combinations can be seen in the spurious information curves for composite functions in Figure 13.

Regardless of correlation, the curve in Figure 11 represents the maximum values of the spurious information

to be expected from a linear combination of normally distributed variables. Thus, a linear combination of 10 variables can be assumed to provide useful discrimination when the sample divergence is sufficiently higher than that indicated in Figure 11.

Spurious Information Expected from Composite Functions

A third set of tests were made on sets of data to which the procedures EXTRACT-EXTRACT and EXTRACT-MVA were applied. In both instances EXTRACT was applied to each of four data blocks producing a set of four composite descriptors. In the first case EXTRACT was again applied to the composite variables while the second case called MVA to estimate the divergence between the composite sets. EXTRACT was used with and without the optimization for both cases. The two resultant curves are provided in Figure 12. Each descriptor block contained 10 variables with no cross-correlation. Thus, the observed sample divergence is the maximum spurious information to be expected. The use of one of the above strategies on similar sized training sets can be considered meaningful only when the observed divergence is sufficiently larger than that indicated in Figure 12.

Additional tests were made to determine the effects of correlation on unoptimized composite decision functions

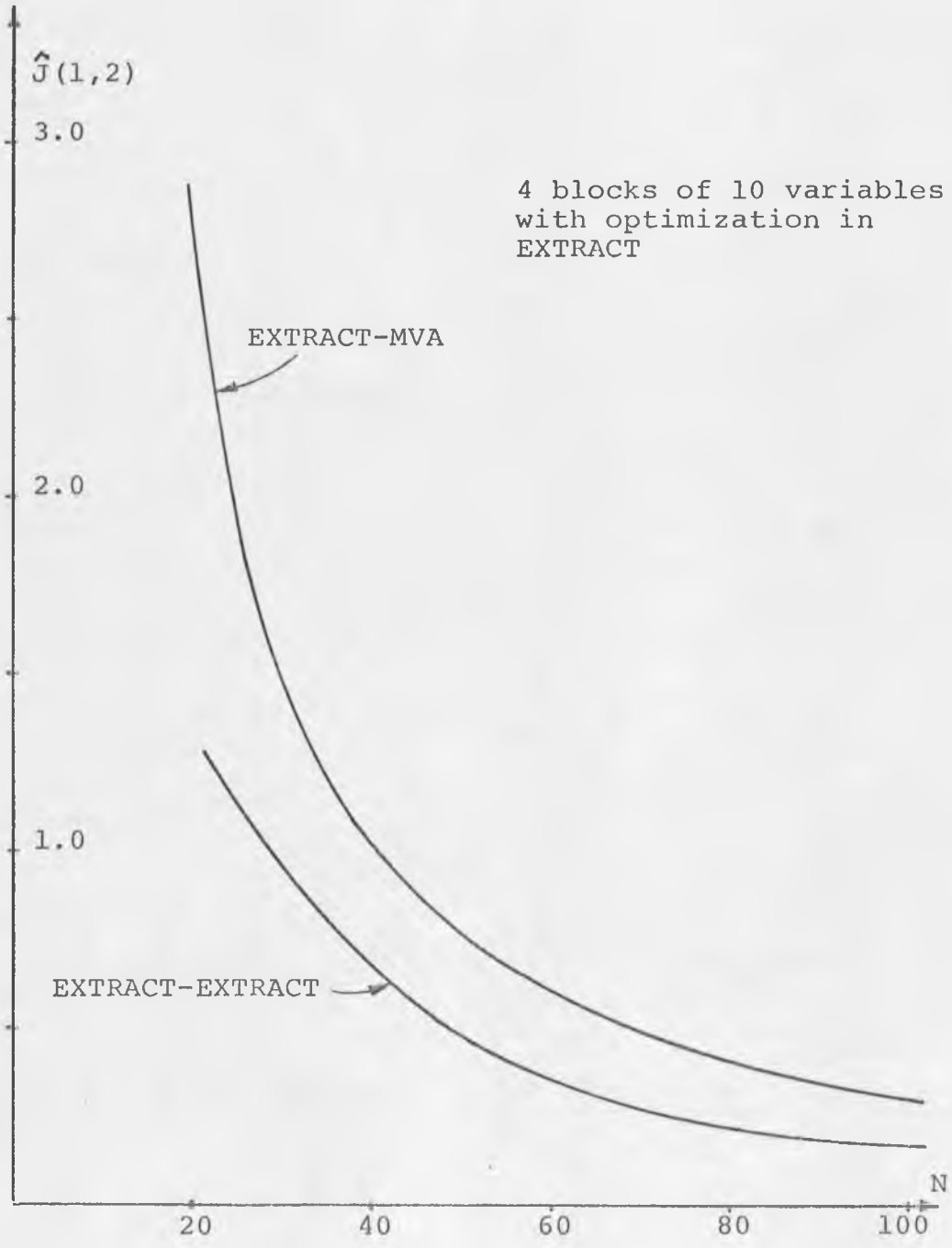

Figure 12. Spurious Information Expected from Composite Decision Functions on Independent Variables

as described above. A 40 dimensional covariance matrix with an average cross-correlation of .337 was used to generate synthetic sample vectors for both training sets. Each sample vector was broken into four sub-vectors of 10 variables each. Corresponding sub-vectors were stored in four separate blocks in each class to simulate the data blocks that are typically processed by DISTILL in the cytodiagnostic problem. The procedures EXTRACT-EXTRACT and EXTRACT-MVA were applied to the synthetic data to estimate the expected spurious information. The results of these tests and similar tests for uncorrelated variables are presented in Figure 13. In each case the optimization procedure of EXTRACT was not called.

Only a limited number of tests were made using subroutine DSELECT because the indicated results did not warrant the increased processing time. Using DSELECT to select the best six descriptors prior to using the optimization in EXTRACT did not change the false information significantly. When more than six descriptors are processed by EXTRACT, a call to the optimization procedure reduces the dimensionality to six. Since the dimensionality is reduced by the same amount regardless of whether or not DSELECT is used, no significant effect should be expected. Using DSELECT prior to calling MVA can, however, reduce

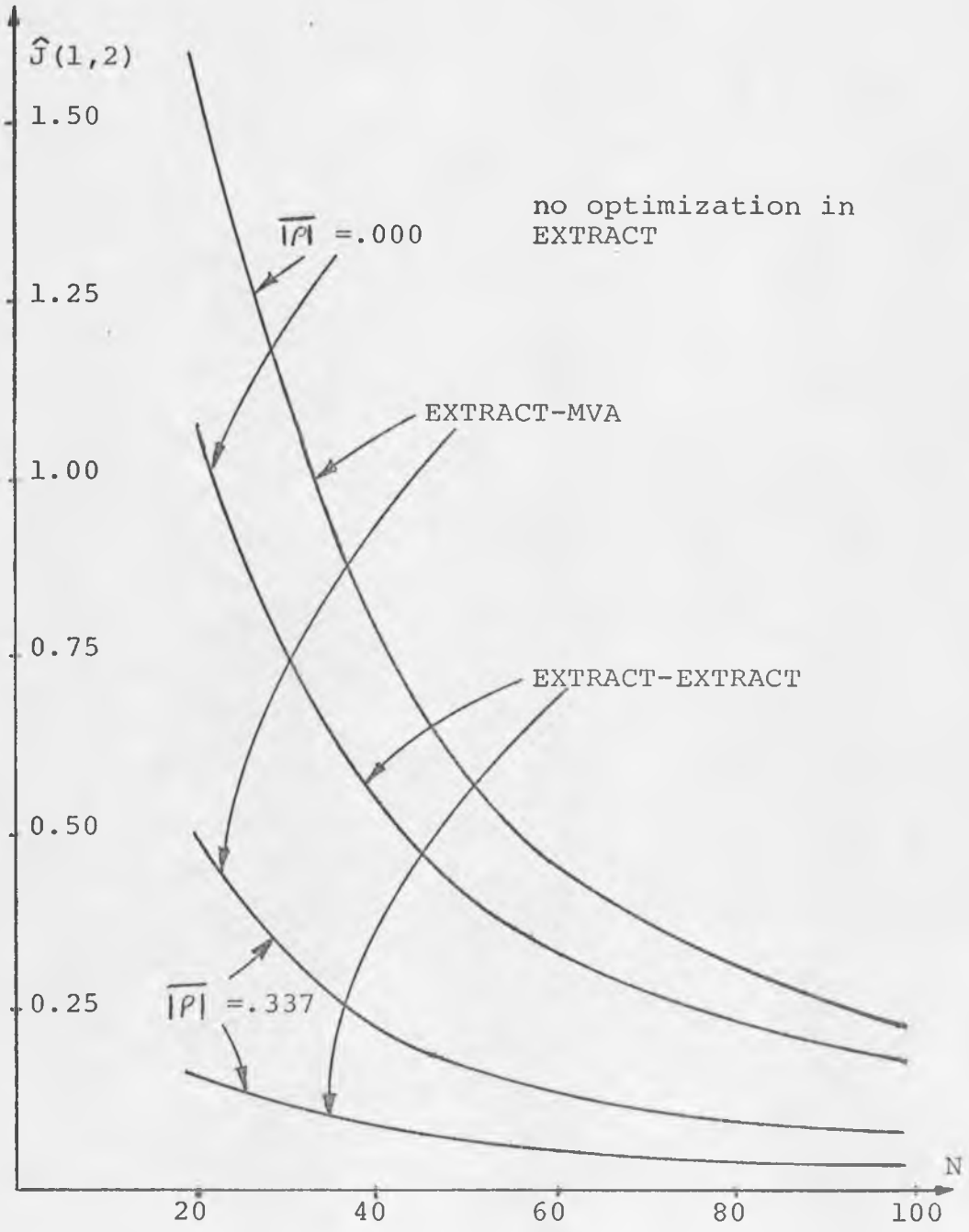

Figure 13. Effects of Cross-Correlation on Spurious Information Expected from Composite Decision Functions

the spurious information greatly. The curves in Figure 10 show marked decreases in the spurious information as the dimensionality decreases.

CHAPTER 5

FEATURE GENERATION ROUTINES

The first step in determining a classification procedure between two types of image arrays involves a feature extraction process. Functionally, this process is equivalent to the transformation g in equation (1-3). This chapter describes certain features that have been implemented to describe various statistical properties of images produced by scanning biological cell types. The features are extracted by a set of four subroutines: HISTGM, ARRATIO, CELLID, and POLAR.

A main program called REDUCE has been written to call the above subroutines and provide various clerical and pre-processing operations. A flow chart of REDUCE is provided in Figure 14. The main purpose of REDUCE is to read a set of image arrays from tape and call the analytical subroutines that extract the image descriptors. A set of array parameters control which output files the subroutines use to store their respective results. Upon completion of the program, these files are copied to magnetic tape for permanent storage. In this manner, each subroutine generates a separate block of data. The corresponding data blocks of two pattern classes are processed by DISTILL to

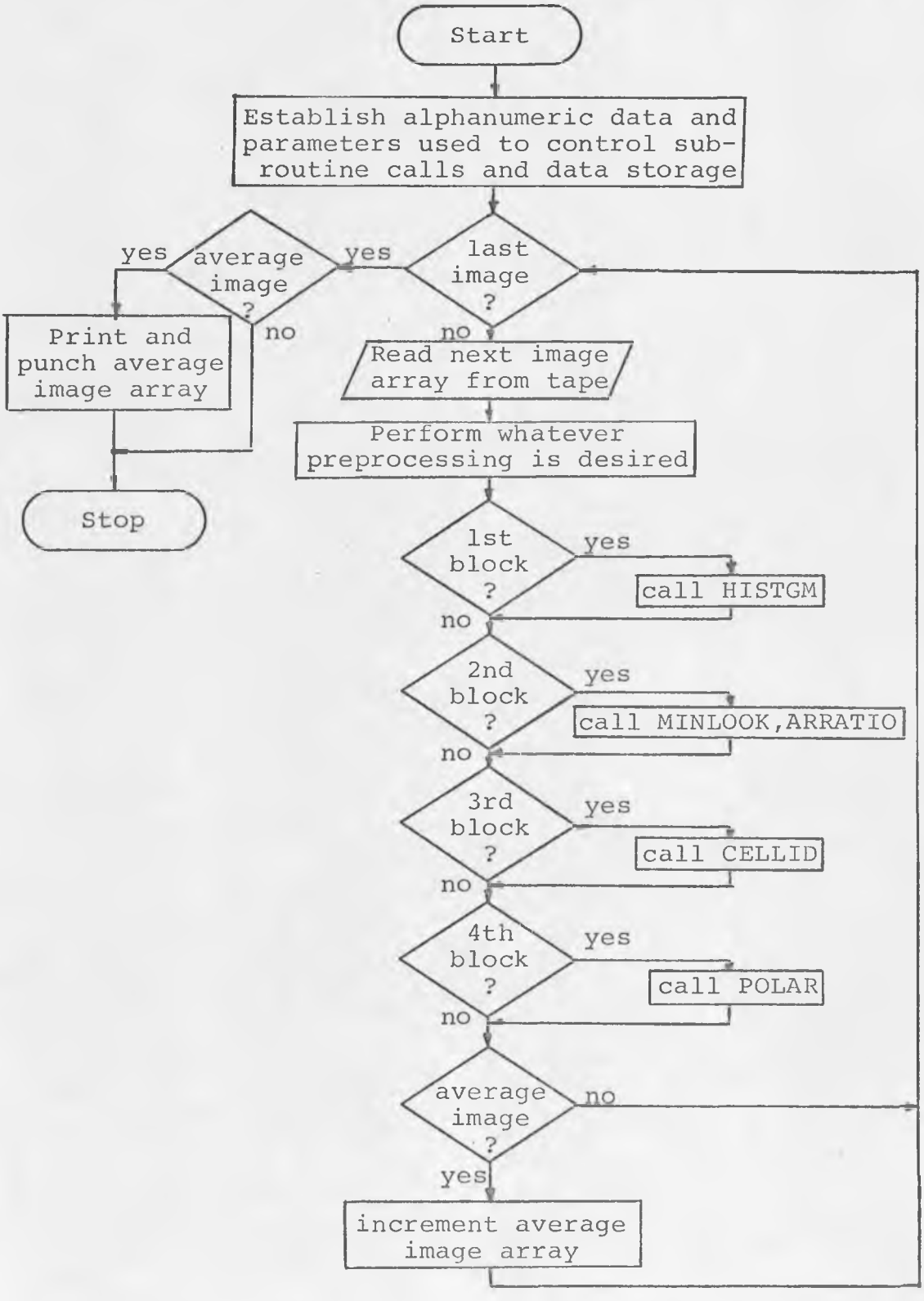

Figure 14. Flow Chart of Program REDUCE

determine which variables or combinations of variables show consistent differences between the training sets.

Pre-Processing

If desired, each image array to be processed can be compressed into a smaller array by averaging blocks of neighboring points together. This option facilitates an evaluation of the scanning step size necessary to effect adequate discrimination in a particular problem. Another pre-processing option normalizes each image array so that the maximum value of the array is a prescribed value. The use of this option removes the effects of variable staining intensity on the cell preparations. Normalization was also used to pre-process the terrain photographs so as to remove the effects of variations in photographic processing and exposure.

Two other options of REDUCE provide printing of input images and generation of the point by point mean extinction values. The array of mean values is determined by calculating the point by point averages of oriented image arrays. Each array (cell) is oriented by first finding the array's center of mass. Then the largest protuberance of non-zero array values (cytoplasm) is rotated about the center to a fixed position. When the

oriented images are averaged in this way a mean contour averaged over the population is produced.

Subroutine HISTGM

Subroutine HISTGM is the first feature extracting routine called by REDUCE. HISTGM is designed to determine the histogram of extinction values occurring in the image array. Each histogram value HIST(I) represents the probability that a randomly selected point of the image array has an extinction value falling in a respective histogram interval. The histogram of extinction values produced by the digitized image of Figure 2 is shown in Figure 15. HISTGM also determines the area, the total mass, and the average mass of the image. The area RN is determined as the number of array values greater than a prescribed threshold. The mass RITE is determined as the summation of all array values. The average mass (extinction value) RIAE is merely the mass divided by the area.

HISTGM writes the values of the histogram array HIST for each image onto the file designated by REDUCE. The 17 values of the HIST array in Figure 15 are tabulated in Table 1. The values of RN, RITE, and RIAE are written on a separate file by subroutine ARRATIO.

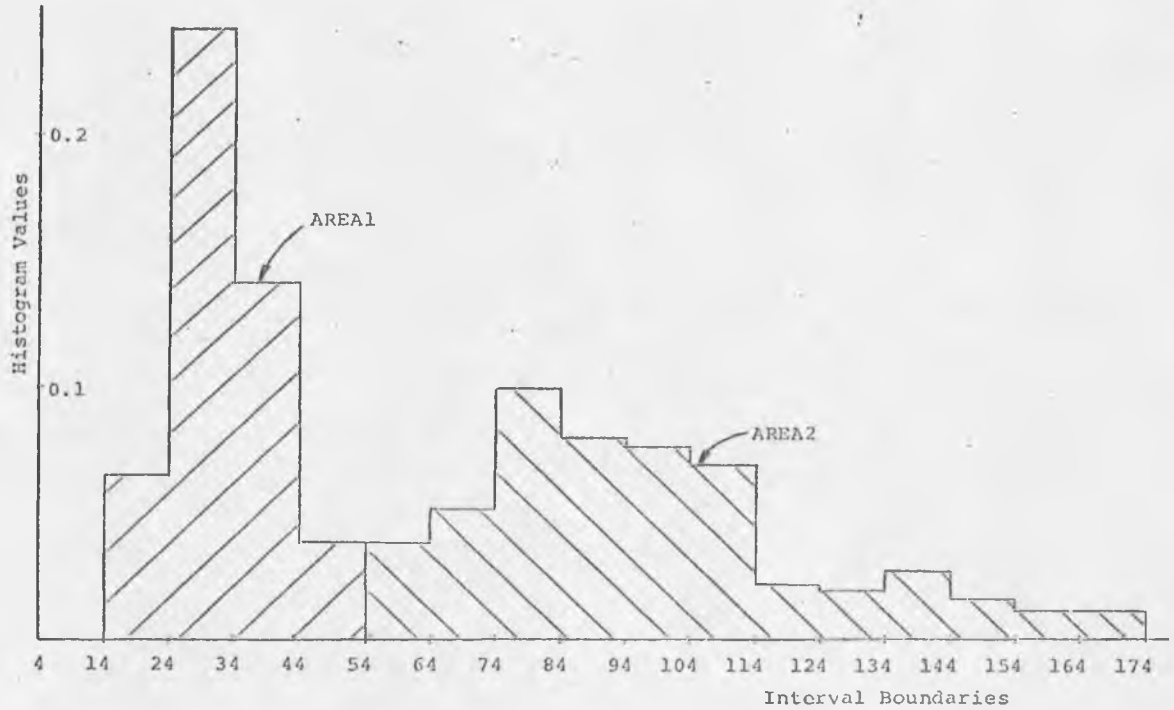

Figure 15. Histogram of Array Values in Figure 2

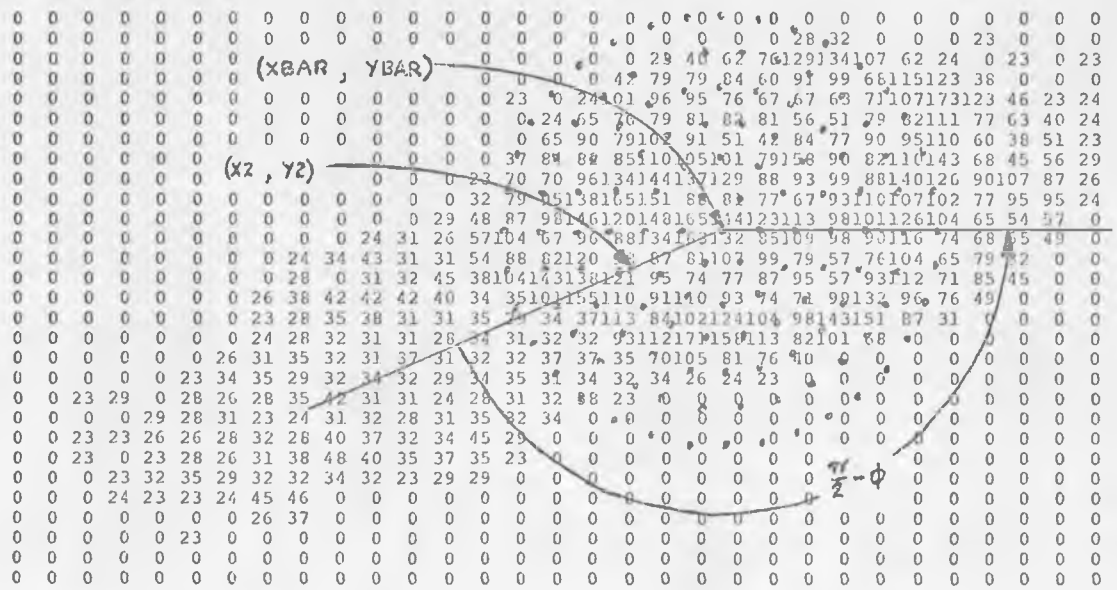

Figure 16. Polar Conversion and Orientation of Array Image in Figure 2

Table 1. Descriptor Blocks and Sample Values Obtained from Image of Figure 2

HISTGM[a]		ARRATIO		CELLID[b]		POLAR	
HIST(1)	.000	RN	369	XTRANS	336	PAVG(1)	143
HIST(2)	.065	RITE	24104	ROWNF1(1,1)	.054	PAVG(2)	130
HIST(3)	.241	RIAE	65	ROWNF1(1,2)	.110	PAVG(3)	103
HIST(4)	.141	AREA1	.485	ROWNF1(1,3)	.012	PAVG(4)	103
HIST(5)	.038	AREA2	.515	ROWNF1(1,4)	.009	PAVG(5)	102
HIST(6)	.038	ARATIO	1.1	ROWNF1(1,5)	.000	PAVG(6)	100
HIST(7)	.052			ROWNF1(1,6)	.000	PAVG(7)	90
HIST(8)	.098			ROWNF1(2,2)	.229	PAVG(8)	60
HIST(9)	.079			ROWNF1(2,3)	.051	PAVG(9)	40
HIST(10)	.076			ROWNF1(2,4)	.027	PAVG(10)	32
HIST(11)	.068			ROWNF1(2,5)	.009	PAVG(11)	19
HIST(12)	.022			ROWNF1(2,6)	.000		
HIST(13)	.019			ROWNF1(3,3)	.054		
HIST(14)	.027			ROWNF1(3,4)	.119		
HIST(15)	.016			ROWNF1(3,5)	.024		
HIST(16)	.011			ROWNF1(3,6)	.006		
HIST(17)	.011			ROWNF1(4,4)	.080		
				ROWNF1(4,5)	.092		
				ROWNF1(4,6)	.042		
				ROWNF1(5,5)	.009		
				ROWNF1(5,6)	.042		
				ROWNF1(6,6)	.033		

a. The intervals used by HISTGM are defined in Figure 15.

b. The intervals used by CELLID are: 3-28, 29-55, 56-80, 81-105, 106-130, and 131-174.

Subroutine ARRATIO

Subroutine ARRATIO in conjunction with subroutine MINLOOK determines for each array: the number of points with relatively high extinction values (area of nucleus) and the number of points with relatively low extinction values (area of cytoplasm). The procedure begins by calling subroutine MINLOOK which examines the histogram values generated by HISTGM. MINLOOK is designed to locate a local minimum positioned between two modes. In general, a cell image exhibits numerous high extinction values in the nucleus and much lower extinction values in the cytoplasm. Thus the histogram of extinction values is typically bimodal and the interior minimum separates histogram values in the cytoplasm from histogram values in the nucleus. As shown in Figure 15, the interior minimum of the example histogram occurs at HIST(6). The extinction values corresponding to the interval of the minimum establish a threshold between nucleus and cytoplasm. One image area AREA1 (area of cytoplasm) is determined by ARRATIO as the percentage of points with extinction values less than the threshold. Another area AREA2 (area of nucleus) is the percentage of points with extinction values greater than the threshold. AREA1 and AREA2 are indicated by cross-hatched portions in the example in Figure 15. A third descriptor ARATIO is determined by ARRATIO as AREA1/AREA2.

ARRATIO writes the six descriptors RN, RITE, RIAE, AREA1, AREA2, and ARATIO onto a file designed by REDUCE. The values of these descriptors as determined for the example array in Figure 2 are presented in Table 1.

Subroutine CELLID

Subroutine CELLID generates a matrix ROWNF1 containing probabilities of transition between extinction intervals for adjacent image points. A set of boundaries is used to define intervals into and from which the transitions are counted. The array ROWNF1 is determined by scanning rows of the image array and incrementing ROWNF1(I,J) whenever an image value in the Ith interval is adjacent to an image value in the Jth interval. A variable XTRANS is also determined as a count of the total number of transitions between the defined intervals. When each raw number of transitions is divided by XTRANS, ROWNF1 becomes a histogram of the transitions between neighboring image points.

Upon completion of the processing, XTRANS and the upper diagonal of ROWNF1 are written onto the file designated by REDUCE. The values of XTRANS and the ROWNF1 array determined from the example image of Figure 2 are presented in Table 1.

Subroutine POLAR

Subroutine POLAR is used to convert the rectangular image array into a polar image array. Subroutine CENTER is first called to find the center of the image in the rectangular image array. Next, the points of the polar array are evaluated by stepping at angular increments about the center for a range of radii. Interpolation between the four nearest rectangular points is used to determine each polar extinction value. Figure 16 shows the rectangular array of Figure 2 along with the corresponding polar representation. While evaluating the polar array, POLAR also determines the averages PAVG(I) for a constant radius I. When all polar averages have been calculated, POLAR writes the PAVG array onto the file designated by REDUCE. The PAVG array values as determined for the example are listed in Table 1.

Auxiliary Subroutines

The following paragraphs describe the subroutines used by REDUCE to perform various clerical and preprocessing operations.

1. Subroutine INTIZE is called to establish the parameters used to control the processing operations that follow. INTIZE also reads a heading card and other alphanumeric data to be stored on the data files generated by

the feature extraction routines. When all parameters and alphanumeric data have been established, the information is recorded on the print file.

2. Subroutine CENTER is used to determine an image scan's center of mass and the geometric center. The center of mass is determined by finding average moments about the index axes. The geometric center is determined as the average distance from the index axes. Figure 16 shows the center of mass (XBAR, YBAR) and the geometric center (X2,Y2) as determined for the array of Figure 2.

3. Subroutine ROTATE is used to orient the image in one array while copying it into another array. The center of mass of the image in the first array is placed at the center of the second array. The image is rotated $\pi/2 - \phi$ radians where ϕ is the angle between the I axis and the line drawn from the center of mass to the geometric center. The angle through which the array of Figure 2 is rotated for orientation is shown in Figure 16.

4. Subroutine COMPRES is used to reduce the size of the image array by averaging the points of successive blocks into one point.

CHAPTER 6

RESULTS

The system of computer programs described in the previous chapters have been applied to a variety of cell recognition problems of clinical interest. This chapter presents the results of applying various processing procedures to each of the problems. The results of each test are compared with the spurious information that can arise in comparable problems to determine if meaningful results were obtained. When possible, the classification procedures have been applied to a set of cell images not in the training sets (object set) but of the same types to evaluate the consistency of each algorithm.

The sample image sets used in each problem were processed by REDUCE to generate the descriptor blocks produced by the four subroutines HISTGM, ARRATIO, CELLID, and POLAR. REDUCE pre-processed each image array with normalization to remove the effects of staining density in the decision process.

Processing Procedures

DISTILL used as many as 10 processing sequences for each problem. In each sequence an identical procedure was

applied to the four descriptor blocks separately. The intermediate results obtained from the separate blocks were merged and then processed further to determine the complete decision function. These composite procedures are listed below where the dotted line indicates the separation between procedures applied to individual blocks and procedures applied to merged data. Each of the five sequences listed were used with and without optimization in EXTRACT thus providing 10 separate investigations.

1 DSELECT
EXTRACT
MVA

2 DSELECT
EXTRACT
DSELECT
EXTRACT

3 EXTRACT
MVA

4 EXTRACT
EXTRACT

5 EXTRACT
DSELECT

Normal Human Lymphocytes vs Lymphocytes from Asymptomatic Chronic Lymphocytic Leukemia

The first problem application involved the determination of a procedure to classify the image of an unknown lymphocyte as either normal (LNRM) or as asymptomatic chronic lymphocytic leukemia (ACLL). Feature vectors derived from 46 LNRM and 50 ACLL cell images were processed

by DISTILL, incorporating each of the 10 processing procedures described above. The classification procedures derived in each case were applied to object sets of 54 LNRM and 54 ACLL cell images. The results of all tests are presented in Table 2.

The LNRM and ACLL cell images are well separated as evidenced by the extremely high likelihood ratios and low error rates. In fact, the descriptors RITE and PAVG(7) were individually capable of separating the training sets. For this reason, the use of DSELECT in procedure 2 caused the entire classification procedure to depend solely on RITE. Although the error rate in the object set increased slightly over that in the training sets, the sample divergences were quite consistent.

The dominance of RITE and PAVG(7) such that the ACLL cells consistently produce higher values of RITE but lower values of PAVG(7) indicates that the ACLL cells tend to have a smaller but denser nucleus.

Normal Endometrium vs Hyperplasia of the Endometrium

The second application of the programs determined a procedure to classify an unknown cell as either normal endometrium (NRME) or hyperplasia of the endometrium (HYPE). Feature vectors derived from 30 each NRME and HYPE cell images were processed by DISTILL using the first four

Table 2. Results of LNRM vs ACLL

	DSELECT-EXTRACT MVA		DSELECT-EXTRACT DSELECT-EXTRACT		EXTRACT MVA		EXTRACT EXTRACT		EXTRACT DSELECT	
	OPT	no OPT	OPT	no OPT	OPT	no OPT	OPT	no OPT	OPT	no OPT
Training Sets										
I(1:2)	54.8	66.7	17.7	17.7	149	25.2	26.2	11.0	-	-
I(2:1)	17.0	15.2	3.2	3.2	7.8	7.7	2.9	4.9	-	-
J(1,2)	35.9	41.0	10.4	10.4	78.0	16.4	14.5	7.9	-	-
PE12	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
PE21	.000	.000	.020	.020	.020	.000	.020	.000	.000	.000
PE	.000	.000	.010	.010	.010	.000	.010	.000	.000	.000
Object Sets										
I(1:2)	61.0	71.0	20.3	20.3	183.0	24.2	28.0	11.9	-	-
I(2:1)	15.1	15.6	2.6	2.6	6.0	6.3	2.5	4.3	-	-
J(1,2)	38.0	42.9	11.4	11.4	94.5	15.2	15.2	8.1	-	-
PE12	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
PE21	.040	.040	.120	.120	.080	.040	.060	.040	.060	.060
PE	.020	.020	.060	.060	.040	.020	.030	.020	.030	.030

I(j:k) is the mean information for discrimination of class j against class k.
PEjk is the probability of classifying a pattern in class j as class k.

procedures listed above with and without optimization in EXTRACT. The fifth procedure was not used because no object set was available. Since the fifth procedure uses DSELECT to classify patterns by decision boundaries, few errors are made in the training sets even when difficult problems are analyzed.

Unlike the first problem, the cell types considered here appear to be very similar since the likelihood ratios are relatively low. Distinguishing between these cell types is in fact an extremely difficult cytodiagnostic problem. Trained pathologists can rarely tell the difference on a cell for cell basis.

The results of the various tests have been tabulated in Table 3. Spurious information levels have been estimated using Figure 12. Although the spurious information was determined using 10 variables in each block as compared to an average of 14 in these problems, the variables are correlated to the extent that the spurious information produced by EXTRACT is significantly reduced. Since the spurious information induced by the linear combinations is not highly dependent on the dimensionality, the spurious information estimates are considered realistic.

The third procedure of applying EXTRACT to individual blocks and then using MVA on the composite descriptors appeared to give the best results. The sample divergence of

Table 3. Results of NRME vs HYPE and NRME vs WDAE

	DSELECT-EXTRACT / MVA		DSELECT-EXTRACT / DSELECT-EXTRACT		EXTRACT / MVA		EXTRACT / EXTRACT	
	OPT	no OPT	OPT	no OPT	OPT	no OPT	OPT	no OPT
				NRME vs HYPE				
I(1:2)	1.55	.83	.658	.502	2.14	1.42	.722	.741
I(2:1)	1.00	1.41	.514	.477	3.10	1.07	.450	.531
J(1,2)	1.27	1.12	.586	.489	2.62	1.24	.586	.636
PE12	.233	.133	.200	.200	.133	.133	.233	.233
PE21	.167	.400	.167	.300	.100	.200	.200	.200
PE	.200	.267	.183	.250	.117	.167	.217	.217
				NRME vs WDAE				
I(1:2)	2.81	1.87	.880	1.04	4.47	1.48	1.13	1.03
I(2:1)	4.09	2.93	3.14	1.82	4.91	2.27	3.72	1.85
J(1,2)	3.45	2.40	2.01	1.46	4.67	1.87	2.42	1.44
PE12	.000	.000	.100	.033	.000	.133	.067	.033
PE21	.033	.033	.033	.067	.033	.067	.000	.133
PE	.017	.017	.067	.050	.017	.100	.033	.083

I(j:k) is the mean information for discrimination of class j against class k.
PEjk is the probability of classifying a pattern in class j as class k.

2.62 is sufficiently higher than the spurious information level of 1.1 in comparable problems to indicate that a difference does exist. An examination of the tables in Appendix B indicates, however, that comparable spurious information samples are not impossible. The application of the classification procedure to an extensive set of known images is therefore the only way to be confident that useful results have indeed been obtained.

The difficulty in separating the training sets is also indicated by the fact that DSELECT chose as many descriptors in each block as allowed by the limit of six. The most useful descriptor in each of the four blocks as determined by the relative weights (defined in equation 2-10) was: HIST(2), RITE, ROWNF1(6,6), and PAVG(7).

The linear combination of composite descriptors derived in both procedures 2 and 4 produced relative weights that were about equal. Hence, the composite descriptors for this problem contain equal amounts of information.

Normal Endometrium vs Well Differentiated Adenocarcinoma of the Endometrium

The third investigation analyzed the program's application to the problem of distinguishing between normal endometrium (NRME) and well differentiated adenocarcinoma of the endometrium (WDAE). Training sets with 30 samples of each cell type were processed in the same manner as in the

previous problem. The results obtained by program DISTILL are tabulated in Table 3.

In this case the sample divergences obtained for the training sets are sufficiently high to indicate a definite difference in the populations. The best procedure appeared to be the first, in which DSELECT and EXTRACT are applied to each individual block of data before applying MVA to the resulting composite variables.

DSELECT was able to separate the training sets with only four descriptors in the HISTGM and POLAR blocks, and three descriptors in the ARRATIO and CELLID blocks. The most useful descriptor of each block (as determined by relative weight) was: HIST(1), RN, XTRANS, and PAVG(1).

The intervals used by CELLID were consecutively defined as in Table 1. Thus, every neighboring pair of points in the image defined a transition that caused XTRANS to be incremented. When this occurs XTRANS essentially represents the cell area. Hence, the concurrence of RN and XTRANS as useful descriptors is not surprising.

The fact that RN and XTRANS are highly correlated is reflected in the relative weight vector of the linear combination of composite variables derived by procedure 2. In this case, the relative weight of the second composite variable was approximately equal to the relative weight of the composite variable representing the third block. Since the major part of the information contained in these

composite variables was derived from RN and XTRANS respectively, equal weights should be expected.

Normal Lymphocytes vs Lymphocytes from Lymphocytic Leukemia

The fourth investigation examined differences between normal lymphocytes (NRML) and lymphocytes from lymphocytic leukemia (LLLE), a case similar to the one analyzed in the first application. All 10 procedures were applied to training sets of 30 NRML cells and 34 LLLE cells. An object set of 39 NRML cells was also processed to determine the consistency of the classification procedures. All results are presented in Table 4.

The sample divergences obtained for each procedure indicated an inherent difference in the populations had been found. In fact, when the procedures were applied to the object set of NRML cells, results were obtained which were quite consistent with those predicted by the training sets.

The most useful descriptors as determined by the relative weights in the linear combinations of their respective blocks were; HIST(7), RITE, ROWNF1(5,5), and PAVG(8). The final linear combination employed by procedure 2 indicated that the composite variables were equally useful with the exception of the first (HISTGM) whose relative weight was about one-third that of the other three.

Table 4. Results of NRML vs LLLE

	DSELECT-EXTRACT MVA		DSELECT-EXTRACT DSELECT-EXTRACT		EXTRACT-MVA		EXTRACT EXTRACT		EXTRACT DSELECT	
	OPT	no OPT	OPT	no OPT	OPT	no OPT	OPT	no OPT	OPT	no OPT
Training Sets										
I(1:2)	1.88	1.52	1.13	1.06	2.90	1.78	1.08	.99	-	-
I(2:1)	1.45	1.12	.93	.90	1.96	1.58	.92	.86	-	-
J(1,2)	1.67	1.32	.98	.98	2.44	1.68	1.00	.92	-	-
PE12	.154	.180	.180	.205	.180	.154	.154	.154	.029	.000
PE21	.118	.147	.147	.147	.118	.147	.147	.147	.000	.000
PE	.136	.630	.163	.176	.150	.150	.150	.150	.015	.000
NRML Object Set										
I(1:2)	1.44	1.80	1.43	1.25	3.89	2.03	1.43	1.27	-	-
PE12	.17	.13	.07	.13	.13	.13	.10	.13	.17	.13

I(j:k) is the mean information for discrimination of class j against class k.
PEjk is the probability of classifying a pattern in class j as class k.

CHAPTER 7

CONCLUSIONS

A system of programs capable of solving a general statistical pattern recognition problem has been developed. Supervised learning procedures embodied in program DISTILL incorporate a variety of techniques designed to determine the information necessary to discriminate between pattern classes represented by corresponding training sets. The training sets are assumed to represent pattern measurements which exhibit significant variability within classes. The learning procedures are not suited to a problem in which the discriminatory information is primarily contained in the occurrence in context of morphologic features.

The programs have been successfully applied to the recognition of human cell types. Using digitized representations of cell images, program REDUCE has been designed to extract certain features describing statistical properties of the digitized array values. Program DISTILL then determines which descriptors or combinations of descriptors are needed to discriminate between cell types.

The techniques described here have proven useful in assessing certain discriminatory cell properties that have not been assessed before. The statistical properties of

the values in an image array are not easily assimilated into a visual impression. These same properties have proven useful in determining a cell type or class. Although no clinical diagnosis is made on the basis of a single cell, the techniques described here can augment the available information upon which a diagnosis is made.

An automated cytodiagnostic system as described here provides an objective quantitative definition of a cell type. This is in contrast with the subjective qualitative impressions used by cytopathologists to assess a cell type. Thus, the automated approach more readily enables the establishment of universal standards and definitions of cell type.

An analysis has been made to determine when a classification procedure can be assumed to provide useful discrimination. The amount of success derived from training sets with no inherent differences has indicated to what degree the various procedures assimilate information derived solely from the variability of the measurements and not from the features themselves. The performance of the classification procedures has also been analyzed by applying the procedures to images representing cells whose types are known a priori.

No single procedure appears to work more effectively than the others when applied to a variety of problems. It

can be determined, however, that DSELECT should not be used in its present form to pre-select variables when the pattern classes are extremely well separated. When DSELECT was used in this manner on the LNRM vs ACLL problem presented in Chapter 6, only one variable was selected. Thus, much highly discriminating information was overlooked. DSELECT should only be used as a descriptor selector when it is necessary to reduce the dimensionality before other processing procedures can be attempted.

The results presented in Chapter 4 show that when MVA is directly applied to a pair of training sets, the number of samples must be much larger than the number of variables. On the other hand, linear discriminants are not likely to separate smaller training sets when the actual classes are inseparable. A single linear combination, however, is generally incapable of utilizing a high percentage of the available information. The composite discriminant function EXTRACT-MVA is therefore a compromise between the generation of meaningless procedures and the rejection of highly discriminatory procedures.

The optimization procedure of EXTRACT has been shown to increase greatly the average log of the likelihood ratio for a classification procedure. The optimization does, however, have some undesirable effects. When a training set has a single outlying point, the optimization procedure

tends to adjust the weight vector so that the outlier's likelihood ratio is increased greatly while reducing the likelihood ratios of all other samples. The net effect is to increase the average of the logs as designed, but in this case the number of errors may increase.

The above problem is particularly evident when small training sets with little separation are encountered. Since the tests for spurious information used training sets with little separation, the tables in Appendix B indicate the frequency of an outlier dominating the sample divergence. When the sample mean information $\hat{I}(1:2)$ for discrimination of c_1 against c_2 differed greatly from $\hat{I}(2:1)$, the difference was usually due to a few outlying points. Table 2 of Appendix B shows a much higher variation in the relative differences between $\hat{I}(1:2)$ and $\hat{I}(2:1)$ for sets of 20 samples.

The effects of outlying points can be reduced by limiting the contribution of any single point to the sample divergence. One method would be to use the integration indicated in (2-1) rather than the sample average defined in (2-12). When sample averages are used, an outlying point assumes the same probability as each point in the center of the training set. Thus, when the outlying point is accentuated by the optimization procedure, a true indication of the expected log is not obtained. Apart from the added complexity in computation, integration is undesirable

when the variables are not normally distributed. The likelihood ratio derived from normal models becomes a discriminant with decreasing quantitative significance as the true distribution differs from the normal. Since many of the descriptors used in the cytodiagnostic application are not normally distributed, the signs of the sample likelihood ratios are needed as an indication of the error probability.

Another similar problem with the optimization procedure can be seen in the results of the EXTRACT-EXTRACT procedure in the first problem of Chapter 6. The well separated pattern classes produced very large likelihood ratios even without the optimization procedure. When optimization was applied, the mean information for discrimination of the normal cells against the leukemic cells $\hat{I}(1:2)$ increased greatly while $\hat{I}(2:1)$ decreased slightly. The decrease in $\hat{I}(2:1)$ is significant, however, in that leukemic cells are more likely to be called normal than vice versa. Whether or not the shift in probabilities of error types is desirable in this case, such a shift should be under program control. Changing the objective function from the divergence to error probability or more generally the expected risk provides the necessary control. Such a function requires integration of an estimated distribution function and thus presents the problems mentioned above.

The results of Chapter 4 indicated that the optimization procedure does not always find the linear combination providing the maximum divergence as intended. Whenever the variables are highly correlated, the starting point supplied by the assumption of uncorrelated variables is a poor one. For this reason the use of a different starting point as derived from the general covariance matrix in equation (2-11) may be advisable. This added feature would only be advantageous when the pattern classes are very similar. The optimization procedure readily determines the optimum weight vector when the training sets are somewhat separated.

The various features have been examined to determine their relative merits in various recognition problems. Although only a limited number of tests have been made, the most useful descriptors appear to be RITE in the ARRATIO block and PAVG(7) (sometimes PAVG(8)) in the POLAR block.

Although many of the descriptors do not possess a variability indicative of a normal distribution, the normality approximation for composite variables appears justified. The chi-square statistic derived for all but RN, RITE, RIAE, and the PAVG descriptors indicates that the sample distributions are not well represented by a normal

distribution, Linear combinations of more than 2 of 3 of these variables, however, produces variables whose chi-square statistic suggests a normal approximation is justified.

APPENDIX A

THEOREMS AND DERIVATIONS

Derivation of Threshold for Minimum Risk Classification

A minimum risk (Bayes) classifier can be realized as an extension of a maximum likelihood classifier with a modified decision threshold.

The minimum risk (mr) classifier chooses class 1 when the expected cost of a decision is greater than the expected cost if class 2 is chosen instead. That is,

$$mr(\underline{x}) = c_1 \quad \text{when } r_{12}p(c_1)f_1(\underline{x}) > r_{21}p(c_2)f_2(\underline{x})$$

$$= c_2 \quad \text{otherwise}$$

where r_{12} is the cost of an error when class 2 is chosen and r_{21} is the cost of an error when class 1 is chosen.

$$mr(\underline{x}) = c_1 \quad \text{when} \quad lr = \frac{p(c_2)f_2(\underline{x})}{p(c_1)f_1(\underline{x})} < \frac{r_{12}}{r_{21}}$$

$$= c_2 \quad \text{otherwise}$$

Hence, the likelihood ratio lr can be used as a minimum risk discriminant when the threshold is changed from unity to r_{12}/r_{21}.

Derivation of Canonical Form

Any set of normally distributed variables with a nonsingular covariance matrix Σ can be represented as a set of normally distributed variables with covariance I transformed linearly by a matrix P.

Let R be a matrix whose columns are normalized eigenvectors of Σ.

then $R^T \Sigma R = \lambda \Rightarrow \Sigma = R\lambda R^T$

where λ is the diagonal matrix of eigenvalues of Σ.

Let $B_i = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \sqrt{\lambda_i} & \vdots \\ 0 & \cdots & \cdots & 1 \end{bmatrix}$

then $\lambda = B_1 B_2 \cdots B_n B_n \cdots B_2 B_1$

and $\Sigma = R B_1 B_2 \cdots B_n B_n \cdots B_2 B_1 R^T$

$= PIP^T$

where $P = R B_1 B_2 \cdots B_n$

The multivariate normal density function:

$$f(\underline{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}\left\{(\underline{x}-\underline{u})\Sigma^{-1}(\underline{x}-\underline{u})^T\right\}}$$

$$= \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}\left\{(\underline{x}-\underline{u})P^{T^{-1}}IP^{-1}(\underline{x}-\underline{u})^T\right\}}$$

$$= \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}\left\{(\underline{z}-\underline{v})I(\underline{z}-\underline{v})^T\right\}} \qquad \text{(A-1)}$$

where $\underline{z} = \underline{x}P^{T^{-1}} \Rightarrow \underline{x} = \underline{z}P^T$

$\underline{v} = \underline{u}P^{T^{-1}} \Rightarrow \underline{u} = \underline{v}P^T$

$$f(\underline{z}) = f(\underline{x})\ |P|$$

$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\left\{(\underline{z}-\underline{v})(\underline{z}-\underline{v})^T\right\}}$$

Derivation of Equation (3-1)

The log of the likelihood ratio derived from two multivariate normal distributions with covariance matrices Σ_1 and Σ_2 can be determined as:

$$\log_{10}(lr) = \frac{1}{4.605}(r_{12}+d_1^2-d_2^2)$$

where $r_{12} = \log_e\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)$ $\quad \underline{z}_k = \underline{x}P_k^{T^{-1}}$

$$\underline{v}_k = \underline{u}_k P_k^{T^{-1}} \qquad k = 1,2$$

$$d_k^2 = (\underline{z}_k-\underline{v}_k)(\underline{z}_k-\underline{v}_k)^T$$

From equation (A-1)

$$f_k(\underline{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma_k|^{\frac{1}{2}}} \; e^{-\frac{1}{2}\left\{(\underline{z}_k-\underline{v}_k)(\underline{z}_k-\underline{v}_k)^T\right\}} \qquad k = 1,2$$

$$\log_e(lr) = \tfrac{1}{2}\left\{\log_e\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + (\underline{z}_1-\underline{v}_1)(\underline{z}_1-\underline{v}_1)^T - (\underline{z}_2-\underline{v}_2)(\underline{z}_2-\underline{v}_2)^T\right\}$$

$$= \tfrac{1}{2}(r_{12}+d_1^2-d_2^2)$$

$$\log_{10}(lr) = \frac{1}{2.3025}\log_e(lr) = \frac{1}{4.605}(r_{12}+d_1^2-d_2^2)$$

Theorem 1

The divergence between two multivariate normal distributions is unchanged by a nonsingular linear transformation of variables.

Proof:

$$J(1,2) = \tfrac{1}{2}\left\{\int_{E^n} f_1(\underline{x})\log\left(\frac{f_1(\underline{x})}{f_2(\underline{x})}\right)d\underline{x} + \int_{E^n} f_2(\underline{x})\log\left(\frac{f_2(\underline{x})}{f_1(\underline{x})}\right)d\underline{x}\right\}$$

$$= \tfrac{1}{2}\left\{\int_{E^n}\left[f_1(\underline{x})-f_2(\underline{x})\right]\log\left(\frac{f_1(\underline{x})}{f_2(\underline{x})}\right)d\underline{x}\right\}$$

let $\bar{x} = \bar{z}P^T$

$$J(1,2) = \tfrac{1}{2}\int_E n\left[f_1(\underline{z}P^T) - f_2(\underline{z}P^T)\right] \log\left(\frac{f_1(\underline{z}P^T)}{f_2(\underline{z}P^T)}\right)|P|\, d\underline{z}$$

in X space

when class 1 and class 2 are distributed normally

$$f_k(\underline{z}P^T) = \frac{1}{(2\pi)^{n/2}|\Sigma_k|^{\frac{1}{2}}}\, e^{-\frac{1}{2}\left\{(\underline{z}-\underline{u})P^T\Sigma_k\, P(\underline{z}-\underline{u}')^T\right\}} \qquad k=1,2$$

in X space

$$f_k(\underline{z}) = \frac{|P|}{(2\pi)^{n/2}|\Sigma_k|^{\frac{1}{2}}}\, e^{-\frac{1}{2}\left\{(\underline{z}-\underline{u}')P^T\Sigma_k P(\underline{z}-\underline{u}')^T\right\}}$$

in Z space

$$= \frac{1}{(2\pi)^{n/2}|\Sigma_k'|^{\frac{1}{2}}}\, e^{-\frac{1}{2}\left\{(\underline{z}-\underline{u}')\Sigma_k'(\underline{z}-\underline{u}')^T\right\}}$$

where $\Sigma_k' = P^T\Sigma_k P$ $\qquad k=1,2$

Hence $J(1,2) = \tfrac{1}{2}\int_E n\left[f_1(\underline{z}) - f_2(\underline{z})\right] \log\left(\frac{f_1(\underline{z})}{f_2(\underline{z})}\right) d\underline{z}$

in Z space

Q.E.D.

Corollary

The distribution of spurious information derived by MVA from samples of identical normal distributions is independent of cross-correlation.

Proof:

As shown above any multivariate sample set can be obtained from a sample set of independent variables transformed by a matrix P. From Theorem 1 the divergence between two distributions (or sample sets) is unchanged by a linear transformation. Thus, a difference in sample sets of an arbitrary normal distribution can be represented as a difference in sample sets of an independent normal distribution.

Q.E.D.

Theorem 2

The maximum divergence of a linear combination of normally distributed variables is unchanged by a non-singular linear transformation.

Proof:

$$\text{let } J(1,2)_{max} = \tfrac{1}{2}\int \left[f_1(z)-f_2(z)\right] \log\left(\frac{f_1(z)}{f_2(z)}\right) dz$$

where $z = \underline{w} \cdot \underline{x}^T$ and $\underline{w}$ is the weight vector maximizing $J(1,2)$ over all possible weight vectors.

$$f_k(z) = \frac{1}{2\pi\sigma_k} e^{-\frac{1}{2}\left(\frac{z-u_k}{\sigma_k}\right)^2}$$

$$\text{where } u_k = \underline{w} \cdot \underline{u}_k^T$$

$$\sigma_k^2 = \underline{w}\Sigma_k\underline{w}^T \qquad k=1,2$$

when the $\underline{x}$ are transformed to $\underline{x}'$ by P

$$\underline{x}' = \underline{x}(P^T)^{-1}$$

$$\Rightarrow \quad \underline{x} = \underline{x}'P^T$$

then $\underline{u}_k' = \underline{u}_k(P^T)^{-1} \qquad \Sigma_k' = P^{-1}\Sigma_k(P^T)^{-1}$

let $\underline{w}' = \underline{w}P$

then $u_k' = \underline{w}'\underline{u}_k'^T \qquad (\sigma_k')^2 = \underline{w}'\Sigma_k'\underline{w}'^T$

$= \underline{w}P\left(\underline{u}_k(P^T)^{-1}\right)^T \qquad = \underline{w}PP^{-1}\Sigma_k(P^T)^{-1}(\underline{w}P)^T$

$= \underline{w}\underline{u}_k^T \qquad = \underline{w}\Sigma_k\underline{w}^T$

$= u_k \qquad = \sigma_k^2$

Hence the maximum divergence obtainable from the transformed variables is at least as large as the divergence obtainable from the original variables. A reverse argument implies that $\underline{w}'$ does in fact maximize the divergence of a linear combination of transformed variables.

Q.E.D.

Divergence Between Two Distributions with Equal Variances

The divergence between two univariate normal distributions with equal variances σ and means difference Δu is $\frac{1}{2}\Delta u^2/\sigma^2$.

$$J(1,2) = \tfrac{1}{2}I(1:2) + \tfrac{1}{2}I(2:1)$$

$$= \tfrac{1}{2}\int f_1(z) \ln\left(\frac{f_1(z)}{f_2(z)}\right) dz + \tfrac{1}{2}\int f_2(z) \ln\left(\frac{f_2(z)}{f_1(z)}\right) dz$$

without loss of generality let $u_1=0$ and $u_2=\Delta u$ so that

$$I(1:2) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2/\sigma^2} \frac{1}{2\sigma^2} \left(-2\Delta uz + \Delta u^2\right) dz$$

$$= \frac{1}{2(2\pi)^{\frac{1}{2}}\sigma^3} \left\{ \int_{-\infty}^{\infty} -2\Delta uz\, e^{-\frac{1}{2}z^2/\sigma^2} dz \right.$$

$$\left. + \int_{-\infty}^{\infty} u^2\, e^{-\frac{1}{2}z^2/\sigma^2} dz \right\}$$

$$= \frac{1}{2(2\pi)^{\frac{1}{2}}\sigma^3} \left\{ 0 + \Delta u^2 (2\pi)^{\frac{1}{2}}\sigma \right\}$$

$$= \frac{\Delta u^2}{2\sigma^2}$$

By a change of variables it is easy to show that $I(2:1) = I(1:2)$. Therefore,

$$J(1,2) = \frac{\Delta u^2}{2\sigma^2}$$

Two univariate normal distributions obtained as a linear combination of variables from two multivariate normal distributions with equal covariance matrices have identical variances (this is indicated in the proof of Theorem 2). The error probability associated with maximum likelihood classification on two univariate normal distributions with equal variances σ is dependent only on the relative means difference $r = (u_2-u_1)/\sigma$. Since the error probability decreases as r increases, the above derivation shows that maximizing the divergence between distributions of composite variables can minimize the probability of misclassification.

APPENDIX B

SPURIOUS INFORMATION SAMPLES

The following tables contain the results of the spurious information tests described in Chapter 4. The number of samples in each column heading refer to the size of each synthetic training set. $\hat{I}(1:2)$ is the sample mean information for discrimination of synthetic class 1 against synthetic classs 2. $\hat{I}(2:1)$ is the sample mean information for discrimination of synthetic class 2 against synthetic class 1.

Table 5. Spurious Information from 4 Variables Using MVA

20 Samples		40 Samples		60 Samples		80 Samples	
$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$
.35	.61	.16	.15	.11	.10	.10	.11
.42	.55	.30	.29	.06	.07	.05	.06
.24	.36	.19	.13	.12	.10	.14	.18
.38	.34	.12	.13	.07	.10	.07	.06
.13	.17	.27	.25	.12	.11	.06	.06
.50	.22	.33	.25	.12	.12	.07	.07
.37	.26	.42	.23	.08	.07	.06	.06
.45	.28	.11	.13	.15	.17	.15	.16
.25	.29	.17	.15	.13	.12	.10	.11
.47	.88	.21	.14	.14	.11	.07	.08
Average .38		.21		.11		.09	

Table 6. Spurious Information from 6 Variables Using MVA

20 Samples		40 Samples		60 Samples		80 Samples	
$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$
.56	.87	.29	.28	.19	.22	.22	.26
.52	.92	.46	.61	.27	.23	.16	.12
.47	.59	.31	.30	.14	.12	.16	.15
.73	.62	.44	.35	.18	.22	.18	.18
1.36	.76	.33	.30	.12	.15	.19	.18
1.30	1.36	.63	.55	.17	.19	.20	.14
1.05	.88	.43	.79	.33	.28	.22	.23
.41	.51	.33	.40	.19	.20	.18	.21
.81	.95	.29	.23	.24	.23	.16	.14
1.52	.76	.48	.46	.15	.23	.28	.20
Average	.85		.41		.19		.19

Table 7. Spurious Information from 8 Variables Using MVA

20 Samples		40 Samples		60 Samples		80 Samples	
$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$
1.64	1.66	.43	.57	.74	.54	.28	.21
1.00	2.57	.44	.55	.29	.31	.34	.31
3.24	1.39	.30	.43	.37	.27	.28	.33
1.28	1.97	.76	1.01	.41	.34	.34	.27
1.77	1.14	.56	.66	.32	.39	.24	.23
1.61	1.55	.69	.53	.57	.42	.29	.36
1.82	1.07	.62	.73	.55	.56	.36	.27
2.07	1.50	.66	.50	.37	.25	.35	.29
1.91	1.93	.49	.50	.57	.48	.36	.42
1.99	1.12	.62	.49	.48	.45	.27	.30
Avg.	1.71		.58		.44		.31

Table 8. Spurious Information from 10 Variables Using MVA

20 Samples		40 Samples		60 Samples		80 Samples		100 Samples	
$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$
4.27	2.51	1.55	1.00	.67	.53	.39	.53	.44	.43
3.96	8.01	1.17	1.05	.53	.37	.61	.46	.31	.33
2.81	3.01	1.10	1.24	.88	.78	.49	.44	.30	.28
4.02	2.66	1.01	1.14	.62	.63	.51	.63	.52	.42
7.77	4.07	.98	.97	.46	.40	.35	.38	.36	.30
5.20	3.52	.96	.67	.69	.58	.29	.32	.36	.34
3.24	5.25	1.03	.96	.52	.52	.64	.48	.27	.24
3.09	2.57			.54	.61	.35	.47	.31	.32
3.60	2.01			.52	.45	.33	.42	.37	.44
2.85	3.00			.47	.47	.36	.45	.35	.30
Avg.	3.90		1.06		.56		.44		.35

Table 9. Spurious Information from 10 Independent Variables Using EXTRACT with Optimization

20 Samples		40 Samples		60 Samples		80 Samples		100 Samples	
$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$
.36	.15	.36	.14	.06	.10	.06	.09	.09	.06
.36	1.89	.28	.12	.04	.04	.05	.07	.05	.03
.30	.77	.28	.15	.11	.12	.05	.07	.03	.05
1.12	.28	.19	.10	.15	.08	.14	.26	.06	.04
2.19	.39	.12	.21	.30	.13	.06	.08	.07	.08
.44	.16	.27	.12	.13	.08	.21	.11	.12	.08
.85	.41	.13	.23	.08	.10	.04	.07	.06	.05
.82	.25	.21	.14	.28	.13	.08	.08	.05	.07
.25	.11	.10	.21	.09	.12	.05	.09	.05	.09
.29	1.06	.20	.65	.05	.08	.08	.06	.04	.05
.73	.31	.18	.22	.05	.05	.06	.10	.05	.04
.68	.31	.20	.12	.06	.10	.10	.06	.08	.05
Avg.	1.07		.40		.11		.09		.06

Table 10. Spurious Information from 4 Blocks of 10 Independent Variables Using EXTRACT with Optimization on Each Block and on each of the Composite Variables.

20 Samples		40 Samples		60 Samples		80 Samples		100 Samples	
$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$
.37	1.94	.43	1.30	.31	.22	.28	.15	.15	.10
3.35	.50	.44	.60	.30	.25	.24	.15	.25	.25
1.87	.76	.45	1.09	.15	.29	.28	.15	.12	.08
1.08	.41			.18	.22	.24	.28		
.96	1.52			.26	.26				
Avg. 1.28		.72		.25		.22		.16	

Table 11. Spurious Information from 4 Blocks of 10 Independent Variables using EXTRACT with Optimization on Each Block and MVA on the Composite Variables.

20 Samples		40 Samples		60 Samples		80 Samples		100 Samples	
$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$	$\hat{I}(1:2)$	$\hat{I}(2:1)$
3.69	3.45	1.20	.50	.82	.48	.41	.35	.24	.19
5.13	1.49	.82	.90	.54	.47	.38	.45	.31	.31
2.38	1.93	.87	1.32	.35	.35			.24	.24
1.53	1.49	.82	.52						
1.66	3.47	1.68	.88						
Avg. 2.61		.95		.58		.40		.25	

REFERENCES

Amari, Shunichi, "A Theory of Adaptive Pattern Classifiers," IEEE Transactions of EC, Vol. 16, June 1967, pp. 299-307.

Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, New York, 1958.

Bartels, P. H., G. F. Bahr, J. C. Bellamy, M. Bibbo, D. L. Richards, and G. L. Wied, "A Self-Learning Computer Program for Cell Recognition," Acta Cytologica, Vol. 14, October, 1970, pp 486-494.

Fukunaga, Keinosuke and Thomas Krile, "Calculation of Bayes Recognition Error for Two Multivariate Gaussian Distributions," IEEE Transactions on Computers, Vol. C-18, March 1969, pp. 220-224.

Fukunaga, Keinosuke and Warren Koontz, "Application of the Karhunen-Loeve Expansion to Feature Selection and Ordering," IEEE Transactions on Electronic Computers, Vol. C-19, April 1970, pp. 311-318.

Henrichon, E. G. and King-sun Fu, "A Nonparametric Partitioning Procedure for Pattern Classification," IEEE Transactions on Computers, Vol. C-18, July 1969, pp. 614-624.

Huelsman, L. P., GOSPEL--A General Optimization Software Package for Electrical Network Design, Engineering Experiment Station, University of Arizona, Tucson, Arizona, September 1968.

Hughes, Gordon, "On the Mean Accuracy of Statistical Pattern Recognizers," IEEE Transactions on Information Theory, Vol. 14, January 1968, pp. 55-63.

Kullback, Soloman, Information Theory and Statistics, John Wiley and Sons, New York, 1959.

Nilsson, Nils J., Learning Machines, McGraw-Hill, New York, 1965.

Prabhu, Karkal Pulkeri Sheshagiri, "On Feature Reduction with Application to Electroencephalograms," Technical Report No. 615, Division of Engineering and Applied Research, Harvard University, September 1970.

Sammon, J. W., Jr., "An Optimal Discriminant Plane," IEEE Transactions on Computers, Vol. C-19, September 1970, pp. 826-830.

Sebestyen, G. S., "Pattern Recognition by an Adaptive Process of Sample Set Construction," IRE Trans. on Information Theory, Vol. II-8, September 1962, pp. 582-591.

Specht, Donald F., "Polynomial Discriminant Functions for Pattern Recognition," Pattern Recognition, Thompson, Washington, D.C., 1968, pp. 291-322.

Swonger, Claron W., "Property Learning in Pattern Recognition Systems Using Information Content Measurements," Pattern Recognition, L. V. Kanal, ed., Thompson, Washington, D.C., 1968, pp. 329-347.

Uhr, Leonard, "Feature Discovery and Pattern Description," Pattern Recognition, L. V. Kanal, ed., Thompson, Washington, D.C., 1968, pp. 159-181.

Wied, G. L., P. H. Bartels, G. F. Bahr and D. G. Oldfield, "Taxonomic Intra-Cellular Analytic System (TICAS) for Cell Identification," Acta Cytologica, Vol. 12, May-June, 1968, pp. 180-204.