To the Graduate Council:

I am submitting herewith a thesis written by Bernard Marcel Edmond Moret entitled "Interactive Feature Selection Using a Distance Criterion." I recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Electrical Engineering.

R. C. Gonzalez

R. C. Gonzalez

We have read this thesis and recommend its acceptance:

E R Hall

A. S. Jordan

Accepted for the Council:

L Evans Roth

Vice Chancellor
Graduate Studies and Research

cop. 2

INTERACTIVE FEATURE SELECTION USING A DISTANCE CRITERION

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Bernard Marcel Edmond Moret

August 1977

1325505

ACKNOWLEDGMENTS

The author wishes to thank Dr. R. C. Gonzalez for his guidance in research and assistance in proofreading this thesis. Appreciation is extended to Dr. E. L. Hall for his invaluable suggestions and to Dr. G. S. Jordan for serving on the thesis committee.

This work was supported by the Office of Naval Research, contract N00014-75-C-0545.

ABSTRACT

A survey of the problem of the selection of measurements is presented. The Mahalanobis distance is introduced as a selection criterion, and approached from an algebraic point of view. A new method for updating the inverse of a positive definite matrix is derived, and applied to the selection problem. Two interactive procedures are developed and used in experiments with the classic problem of Fisher, with high-resolution radar images, and with aircraft photographs. Moments are used as features in the last two experiments.

TABLE OF CONTENTS

CHAPTER		PAGE
1.	INTRODUCTION	1
2.	SELECTION OF MEASUREMENTS	2
2.1.	Desirability of a Selection	2
2.2.	Selection Criteria and Methods	3
2.3.	Mahalanobis Distance as a Criterion	8
3.	ALGEBRAIC APPROACH TO THE MAHALANOBIS DISTANCE	14
3.1.	Updating the Inverse of a Positive-Definite Matrix	14
3.2.	Geometric Interpretation	23
4.	INTERACTIVE SELECTION PROCEDURE	28
4.1.	Optimal and Suboptimal Searches	28
4.2.	Computational Problems	30
4.3.	Set-Up of the Procedures	31
5.	MOMENTS AS FEATURES	34
6.	EXPERIMENTAL RESULTS	38
6.1.	Iris Data	38
6.2.	Radar Returns of Aircraft	43
6.3.	Photographs of Aircraft	53
7.	CONCLUSIONS AND RECOMMENDATIONS	58
REFERENCES		59
APPENDICES		62
A.	CIRCULANT MATRICES AND THE MAHALANOBIS DISTANCE	63
B.	PROGRAM DESCRIPTIONS	65
VITA		76

LIST OF TABLES

TABLE		PAGE
1.	Iris Classification Matrix Using all Measurements	39
2.	Iris Classification Matrices (I) Using (a) Measurements 2, 3, 4; (b) Measurements 3, 4; (c) Measurement 3	40
3.	Error Rates Associated with Iris Measurement Subsets (I) .	41
4.	Iris Classification Matrices (II) Using (a) Measurements 1, 3, 4; (b) Measurements 3, 4; (c) Measurement 4	42
5.	Error Rates Associated with Iris Measurement Subsets (II) .	43
6.	Aircraft Used in the Experiments with Radar Returns	44
7.	Aircraft Classification Matrix Using All 33 Moments	49
8.	Aircraft Classification Matrix Using 11 Selected Moments .	50
9.	Classification Results for Various Subsets	51
10.	Fighter Aircraft Classification Matrices Using (a) Nine Selected Moments and (b) Eighteen Selected Moments	52
11.	Transport Aircraft Classification Matrices Using (a) Nine Selected Moments and (b) Eighteen Selected Moments	53

LIST OF FIGURES

FIGURE		PAGE
1.	Illustration of the Role of Spread in Distance Measurement .	9
2.	Mahalanobis Distance between Two Classes	10
3.	Illustration of the Mahalanobis Distance	25
4.	Illustration of the First Type of Dimensionality Reduction .	26
5.	Illustration of the Second Type of Dimensionality Reduction	26
6.	Typical Simulated Radar Image	45
7.	Sample Patterns of the FB-111A Class	46
8.	Sample Patterns of the B-737 Class	47
9.	Relationship between the Number of Selected Components and the Error Rate	51
10.	Sample Photographs of the Four Aircraft	54

NOTATION AND TERMS

$\underline{x}$ (lower case underlined letter)	:	column vector
$\underline{x}^T$	:	transpose of $\underline{x}$
$\|\|\underline{x}\|\|_2$ (= $\sqrt{\underline{x}^T \cdot \underline{x}}$)	:	natural 2-norm of $\underline{x}$
$\underline{m}_i$	:	mean vector of the i-th class
$\underline{A}$ (upper case underlined letter)	:	matrix
$\underline{A}^T$	:	transpose of $\underline{A}$
$\|\underline{A}\|$	:	determinant of $\underline{A}$
$\underline{A}^{-1}$	:	inverse of $\underline{A}$
$\underline{C}_i$	:	covariance matrix of the i-th class
$\underline{F}$	:	complex Fourier matrix
$\underline{F}^H$	:	Hermitean transpose of $\underline{F}$
ω_i	:	i-th class (a set of patterns)
$p(\underline{x})$	:	probability density function of $\underline{x}$
$p(\underline{x}/\omega_i)$	:	conditional probability density function of $\underline{x}$ given ω_i

Unless otherwise specified, all numbers, vectors, and matrices are real, and all matrices are square.

The terms "measurement" and "component" are used interchangeably; likewise, the terms "pattern" and "(measurement) vector" are substituted indifferently for one another.

As used in the text, a principal submatrix of a matrix $\underline{A}$ is a matrix of dimension inferior or equal to the dimension of $\underline{A}$, and the diagonal elements of which are diagonal elements of $\underline{A}$.

CHAPTER 1

INTRODUCTION

The recognition of patterns is a basic act of human life, indeed, of animal life. The duplication of this activity in a machine has been the object of numerous studies, and the number of publications dealing with this problem has been steadily increasing over the past decade.

Three stages can be distinguished in a typical pattern recognition process: feature extraction, measurement selection, and measurement classification. This study attempts to provide a new solution to the second problem, the selection of measurements. Since no stage can be examined independently of the others, consideration is given to the other two problems as well, but only in the measure where they influence decision processes in the selection problem. Various authors have extensively surveyed the field of pattern recognition ([14], [19], [23]). A detailed review of contributions to the subject can be found in an article by Das Gupta [9].

CHAPTER 2

SELECTION OF MEASUREMENTS

2.1. Desirability of a Selection

The question might arise as to why a selection of measurements is desirable. As answers to this question are of obvious importance in the design of a selection scheme, some are presented below.

First, not all the measurements are necessarily useful to classification. Indeed, if the distribution of a measurement is the same for all pattern classes, that measurement is useless for the purpose of discrimination. Though this can be avoided by a careful design of the feature extraction scheme, a priori information about the data to be classified may be insufficient, or the features may be too difficult to interpret (as in the case discussed in Chapter 5).

Secondly, an excess of measurements can cause difficulties during a training process. As only a limited set of patterns—the training set—is available during training, the number of measurements can reach or exceed the number of training samples. This results in a classifier being sensitive to the separation of sample patterns instead of to the distributions the latter should represent. In the particular case of a linear decision function, this corresponds to the fact that, if the patterns are well-distributed (i.e., in general position), there exists a linear dichotomy separating any two given distinct subsets of the training set [21, pp. 58-60]. As an enlargement of the training set is not always possible, it may be necessary to reduce the number of measurements.

A third answer is economy. Processing a large number of measurements is costly in terms of computation time, storage memory, and communication channels. Likewise, performing the measurements might require expensive sensors and exert a perturbing, if not disrupting, influence on the process to be measured. Thus it is desirable that the number of measurements be limited to the minimum necessary to insure proper classification. In fact, one is often forced to trade low error rates for reduced number of measurements.

2.2. Selection Criteria and Methods

In order to carry out a selection procedure, some criterion of choice is needed. In general, given a set of N measurements, a subset of N_1 ($N_1 < N$) measurements will be chosen such that no other subset yields better results, according to some definition of a "good result." This problem is often constrained by the condition that the chosen subset does not contain more than a given limit number of measurements.

The selection procedure depends on the definition of a "good subset" which, in turn, often depends on the type of classifier to be ultimately used. Consequently, there exists a variety of criteria, each resulting in a different selection procedure. Some of these are examined below.

The optimal strategy can be defined as that incurring the minimum loss. This is the basis of the Bayes strategy, widely used in pattern classification [21, Chapter 4]. When such a classifier is to be used, the selection criterion is naturally derived from it. Therefore, a subset of measurements will be said to be better than, and will consequently be preferred to, another subset if the average loss

associated with its choice is less than that associated with the choice of the other subset. The average loss incurred when classifying a pattern can be expressed as follows. Let M denote the number of classes, ω_i be the set of all patterns pertaining to class i, and $\underline{x}$ be an unknown measurement vector; finally, let L_{ij} denote the loss incurred when classifying a pattern of class i into class j. Then the average loss associated with the classification of a pattern into class j is given by:

$$r_j(\underline{x}) = \sum_{i=1}^{M} L_{ij}\, p(\omega_i/\underline{x}) \qquad (2.2\text{-}1)$$

and the classification strategy is to assign $\underline{x}$ to class j if $r_j(\underline{x}) < r_i(\underline{x})$, for all i different from j. The L_{ij} coefficients can usually be taken as:

$$L_{ij} = 1 - \delta_{ij} \qquad (2.2\text{-}2)$$

where δ_{ij} is the Kronecker function; (2.2-1) then reduces to:

$$r_j(\underline{x}) = \frac{1}{p(\underline{x})} \sum_{i=1}^{M} (1 - \delta_{ij}) p(\underline{x}/\omega_i) p(\omega_i)$$

$$= 1 - \frac{1}{p(\underline{x})}\, p(\underline{x}/\omega_j) p(\omega_j) \qquad (2.2\text{-}3)$$

where use has been made of Bayes' formula. When used for classification,

$\underline{r}_j(\underline{x})$ can be simplified by dropping the terms independent from the class index. Thus (2.2-3) finally becomes:

$$r_j'(\underline{x}) = -p(\underline{x}/\omega_j) \cdot p(\omega_j) \qquad (2.2\text{-}4)$$

It is noted, however, that the assumption of (2.2-2) cannot always be made. As an example, consider a medical application where the classifier attempts to distinguish healthy persons from unhealthy ones. Clearly, the loss associated with the classification of an ill subject as healthy is much higher than that associated with the reverse case. In such cases, equation (2.2-1) must be retained in spite of its higher complexity.

The problem encountered in implementing a Bayes classifier is the estimation of the probability density functions, $p(\omega_i)$ and $p(\underline{x}/\omega_i)$. This problem becomes even worse in a selection process as the measurement vector, $\underline{x}$, is then of variable dimensionality. Indeed, the problem is, in most cases, intractable. However, solutions exist in the case of some special distributions, in particular normal and Poisson distributions [1]. It will be recalled that a normal, or Gaussian, process is entirely characterized by its mean and covariance matrix (i.e., by statistical moments of the first and second order). Let $\underline{m}$ be the mean vector and $\underline{C}$ the covariance matrix; then the probability density function of a normal distribution is given by:

$$p(\underline{x}) = 2(\pi)^{-n/2} \, |\underline{C}|^{-1/2} \exp \left\{ -\frac{1}{2} (\underline{x} - \underline{m})^T \, \underline{C}^{-1} \, (\underline{x} - \underline{m}) \right\} \qquad (2.2\text{-}5)$$

where $|\underline{C}|$ denotes the determinant of $\underline{C}$. When this is substituted in (2.2-4), and after some manipulations, the following decision function is obtained [21, p. 121]:

$$d_j(\underline{x}) = \ln p(\omega_j) - \frac{1}{2} \ln |\underline{C}_j| - \frac{1}{2} (\underline{x} - \underline{m}_j)^T \underline{C}_j^{-1} (\underline{x} - \underline{m}_j) \qquad (2.2\text{-}6)$$

where the mean vector and covariance matrix have been indexed to indicate their relation to class j. Both the mean vector and the covariance matrix can be easily estimated from a training set; the class probability can be derived by observation of sample data. Thus the Bayes decision rule can be implemented for normal pattern distribution with a minimum of complexity (see [2], [9]). The probability of misclassification, however, remains a highly complex function, unless more restrictions are imposed [7].

Thus, most investigations in the field of measurement selection based on Bayes' probability of error have assumed normal distributions with equal covariance matrices—in which case the decision functions are linear [21, p. 121].

A second type of criterion is based upon the results of information theory and tends to minimize the correlation between components in order to minimize the redundancy of the measurements. This can be achieved in two ways. The first is not properly a selection procedure but rather a reduction of dimensionality and uses the Karhunen-Loève expansion to decorrelate the components of the measurement vector. The result is a rotation in the measurement space; the new components, linear combinations of the originals, can then be selected on an

individual basis. The second approach is sequential and, starting with an arbitrary choice, selects at each step the component least correlated with those already chosen.

The third major category includes distance criteria. Measurements are selected which maximize some type of distance between the classes of patterns. In order of increasing complexity, the following measures are commonly used: Euclidean distance, Mahalanobis distance, and Bhattacharyya distance. Given two patterns, $\underline{x}$ and $\underline{y}$, the Euclidean distance is given by:

$$d_E(\underline{x},\underline{y}) = \sqrt{(\underline{x} - \underline{y})^T(\underline{x} - \underline{y})} \qquad (2.2\text{-}7)$$

Given a class characterized by its mean vector, $\underline{m}_i$, and covariance matrix, $\underline{C}_i$, the Mahalanobis distance from the pattern $\underline{x}$ to the class i is given [17] by:

$$d_M(\underline{x},\omega_i) = \sqrt{(\underline{x} - \underline{m}_i)^T \underline{C}_i^{-1} (\underline{x} - \underline{m}_i)} \qquad (2.2\text{-}8)$$

This is recognized as the square root of the exponent of the normal probability density function of (2.2-5). Finally, given two classes given by their probability density functions, $p_1(\underline{x})$ and $p_2(\underline{x})$, the Bhattacharyya distance between the classes is defined as:

$$d_B(p_1, p_2) = -\ell n \int \sqrt{p_1(\underline{x}) \cdot p_2(\underline{x})}\ d\underline{x} \qquad (2.2\text{-}9)$$

This is a measure of divergence between the two distributions, p_1 and p_2; the integral can be interpreted as giving the cosine of the "angle" between the distributions [19].

These three measures have different properties and do not measure distances between the same objects. Each, however, is a special case of the next. The Bhattacharyya distance is the most efficient criterion of the three for selection. Unfortunately, it requires, as the Bayes' criterion, the estimation of the density functions. In the case of normal distributions, it reduces to the Mahalanobis distance, which is more tractable. Finally, when the covariance matrix is the identity matrix, the Mahalanobis distance degenerates into the Euclidean distance.

Other, more specialized criteria have been proposed to fulfill particular requirements. All of these criteria result, in practice, in selections that are better—in terms of classification performance—than trivial random selections.

This study proposes the use of the Mahalanobis distance as a selection criterion for several reasons that are noted in the following section.

2.3. Mahalanobis Distance as a Criterion

Considered as a distance measure, the Mahalanobis distance presents numerous advantages. First, it is a true metric in real space; that is, it is positive (or zero, which correctly implies that the two considered vectors are identical), and it obeys the triangle inequality. Moreover, this measure stays invariant under general nonsingular linear transformations [6], whereas the Euclidean distance, for instance, is invariant

only under unitary transformations (i.e., rotations). In particular, this makes scaling unnecessary, as it would not affect the measure. (Scaling might be necessary, however, for computational reasons.) Finally, as a consequence of the above, it is seen that the Mahalanobis distance measure has a "built-in" scaling which weighs the Euclidean distance by the set variance along the axis of measure. This is important in two respects. First, measures taken in disproportioned units are considered only as relative variations and thus rescaled. More significantly, the spread of the sets is taken into account, as illustrated in Figure 1a and 1b. In these figures, the three mean points are at fixed Euclidean distances of each other. Consequently, according to a Euclidean measure, the pattern P would be said to be closer to the mean M_1 in both cases. It is apparent, however, that a Mahalanobis measure would result in two different conclusions, in accordance with intuitive reasoning.

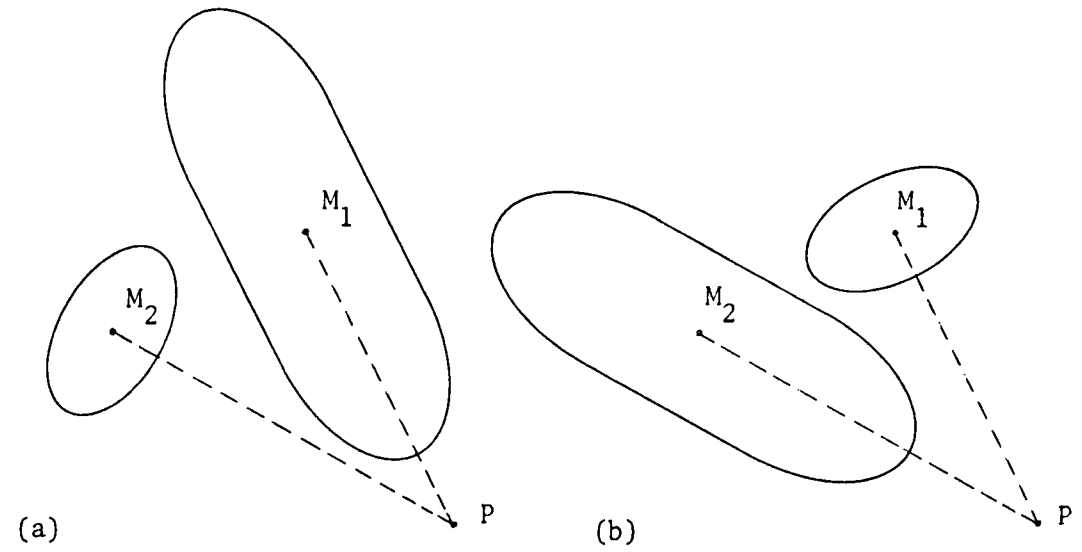


Figure 1. Illustration of the role of spread in distance measurement.

Thus the use of second statistical moments allows a refinement of the notion of distance. Though higher order statistical moments could be used (for non-Gaussian data), as is the case with the Bhattacharyya distance, the second-degree approximation of the Mahalanobis distance constitutes a satisfactory description of the situation in most cases.

Selection procedures usually involve several classes of patterns. The problem is how to characterize interclass distances with Mahalanobis distances. Let each class be described by its mean $\underline{m}_i$ and covariance matrix $\underline{C}_i$. Then the distance between classes i and j can be measured in two ways, in terms of their respective means and covariance matrices.

$$\begin{cases} d_{ij} = \sqrt{(\underline{m}_i - \underline{m}_j)^T \, \underline{C}_i^{-1} \, (\underline{m}_i - \underline{m}_j)} \\ d_{ji} = \sqrt{(\underline{m}_i - \underline{m}_j)^T \, \underline{C}_j^{-1} \, (\underline{m}_i - \underline{m}_j)} \end{cases} \tag{2.3-1}$$

Depending on the spread of the classes, these two measures can be considerably different, as illustrated in Figure 2. A better characterization of the interclass distance would be the average of the two measures, that is, the double of the distance from the median of the

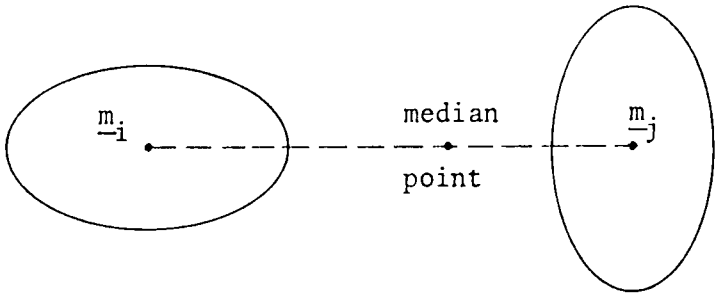


Figure 2. Mahalanobis distance between two classes.

segment $\overline{\underline{m}_i \, \underline{m}_j}$ (in terms of Mahalanobis distance) to either class. A rapid calculation shows this to be the harmonic mean of the two distances:

$$d_{ij} \text{ aver.} = 2\, d_{ji} d_{ij} / (d_{ji} + d_{ij}) \qquad (2.3\text{-}2)$$

Thus, as would be expected, the average of the two distances is closer to the smaller of the two.

For more than two classes, however, the case is more complex. The average interclass distance should possess a property similar to the harmonic mean of equation (2.3-2) so that excessively large distances do not completely override smaller ones. There are, for N classes, $N \cdot (N-1)/2$ interclass distances (or $N \cdot (N-1)$ if measured as in equation (2.3-1)). There is no point in using (2.3-2) anymore; rather, two possibilities are left. The first is almost simplistic and involves taking the arithmetic mean of the $N \cdot (N-1)$ distances computed as in (2.3-1). The second approach is a generalization of the Mahalanobis distance to the multiclass case, yielding a unique measure of average class separation, based upon scatter matrices introduced by Wilks (see [1], [6]).

Let ω_i, i=1, ..., M, denote M classes, each comprised of N_i vectors of measurements; then define:

$$\underline{m}_i = \frac{1}{N_i} \sum_{\underline{x} \epsilon \omega_i} \underline{x} \quad \text{mean vectors}$$

$$\underline{W}_i = \sum_{\underline{x} \epsilon \omega_i} (\underline{x} - \underline{m}_i)(\underline{x} - \underline{m}_i)^T \quad \text{within-class scatter matrices}$$

$$\underline{W} = \sum_{i=1}^{M} \underline{W}_i \quad \text{total within-class scatter matrix}$$

$$\underline{B} = \sum_{i=1}^{M} N_i \underline{m}_i \underline{m}_i^T \quad \text{between-classes scatter matrix} \tag{2.3-3}$$

$\underline{W}$ is also referred to as the pooled covariance matrix [10]. Finally, define the following measure:

$$d^2 = \text{tr}(\underline{W}^{-1} \cdot \underline{B}) \tag{2.3-4}$$

where tr denotes the trace of the matrix. This measure is the natural extension of the Mahalanobis distance to a multiclass case. This is more easily understood by noting that (2.3-4) can be rewritten as follows:

$$d^2 = \text{tr}\ (\underline{W}^{-1} \cdot \underline{B}) = \text{tr}(\underline{W}^{-1} \cdot \sum_{i=1}^{M} N_i \underline{m}_i \underline{m}_i^T)$$

$$= \text{tr}\ (\sum_{i=1}^{M} N_i \underline{W}^{-1} \underline{m}_i \underline{m}_i^T) = \sum_{i=1}^{M} N_i \text{tr}\ (\underline{W}^{-1} \cdot \underline{m}_i \cdot \underline{m}_i^T)$$

which finally yields:

$$d^2 = \sum_{i=1}^{M} N_i \underline{m}_i^T \underline{W}^{-1} \underline{m}_i \tag{2.3-5}$$

The analogy of (2.3-5) with the Mahalanobis distance defined in (2.2-8) is evident.

The selection of measurements according to the criterion of maximum interclass distance then requires the evaluation of either (2.3-1) or (2.3-5) for each subset of components; this in turn necessitates the updating of the inverse matrix $\underline{C}^{-1}$ (or $\underline{W}^{-1}$) as rows and columns are added to, or deleted from, $\underline{C}_i$ (or $\underline{W}$). Thus an algebraic study of the Mahalanobis distance seems appropriate.

CHAPTER 3

ALGEBRAIC APPROACH TO THE MAHALANOBIS DISTANCE

3.1. Updating the Inverse of a Positive-Definite Matrix

As mentioned in the previous chapter, the selection of measurements using the interclass Mahalanobis distance as a criterion involves the computation of the inverse of various submatrices of the covariance matrix. As the inversion of a matrix is a lengthy process, some kind of recursiveness would be desirable; this would have the further advantage of providing some insight into the selection problem.

The properties of the covariance matrix simplify the problem considerably. The sample covariance matrix may be computed by using the following definition [21, pp. 137-139]:

$$\underline{C} = \frac{1}{N} \sum_{i=1}^{N} (\underline{x}_i - \underline{m})(\underline{x}_i - \underline{m})^T \qquad (3.1\text{-}1)$$

where $\underline{m}$ denotes the mean of the set and $\underline{x}_i$, i=1, ..., N, are the patterns in the set. (This is the so-called maximum likelihood estimator.) Thus the sample covariance matrix is guaranteed to be real symmetric and positive semidefinite (since any matrix of the type $\underline{A} \cdot \underline{A}^T$ is positive semidefinite and since the sum of positive semidefinite matrices is a semidefinite matrix [5], [18]). The true covariance matrix is known to be positive definite [1]; the sample covariance matrix is guaranteed to be so under certain conditions [1], [8].

Positive definite matrices possess a number of interesting properties; those of interest to the present discussion are enumerated below:

-a- The inverse of a positive definite matrix exists and is itself positive definite

-b- Any principal submatrix of a positive definite matrix is itself positive definite; in particular all diagonal elements are real positive

-c- A positive-definite matrix possesses a Cholesky decomposition; i.e., there exists a lower triangular matrix $\underline{L}$, with real elements (positive on the diagonal) such that the product $\underline{L} \cdot \underline{L}^T$ is equal to the matrix in question.

(3.1-2)

Demonstrations of these various properties can be found in most books on linear algebra [5], [18], [22].

Property (3.1-2-b-) guarantees that submatrices of the covariance matrix will also give rise to positive definite quadratic forms. This proves the statement made in the previous chapter (section 2.3) about the metric property of the Mahalanobis distance.

Of particular interest is the fact that the inverse of a positive-definite matrix can be easily updated when either a rank-one matrix is added to the positive definite matrix [4] or a row and a column (or several rows and several columns) are added to the original positive definite matrix, as demonstrated below.

Let $\underline{C}$ and $\underline{C}_k$ be positive definite matrices of dimension n and k respectively, and such that $\underline{C}_k$ is a principal submatrix of $\underline{C}$. Without loss of generality, $\underline{C}_k$ can be taken as the upper left-hand principal submatrix. Let $\underline{C}^{-1}$ and $\underline{C}_k^{-1}$ denote the inverse of $\underline{C}$ and $\underline{C}_k$ respectively, and let $\underline{C}^{-1}$ be partitioned in a manner analogous to $\underline{C}$:

$$\underline{C} = \begin{bmatrix} \underline{C}_k & \underline{X} \\ \underline{X}^T & \underline{Y} \end{bmatrix} \quad , \quad \underline{C}^{-1} = \begin{bmatrix} \underline{U} & \underline{V} \\ \underline{V}^T & \underline{W} \end{bmatrix} \tag{3.1-3}$$

By property (3.12-a-b) $\underline{Y}$ and $\underline{W}$ are positive definite and possess inverses, $\underline{Y}^{-1}$ and $\underline{W}^{-1}$ respectively. $\underline{X}$ and $\underline{V}$ are $k \times (n-k)$ rectangular matrices. By definition, $\underline{C} \cdot \underline{C}^{-1} = \underline{I}_n$. This can be rewritten as follows:

$$\begin{bmatrix} \underline{C}_k & \underline{X} \\ \underline{X}^T & \underline{Y} \end{bmatrix} \cdot \begin{bmatrix} \underline{U} & \underline{V} \\ \underline{V}^T & \underline{W} \end{bmatrix} = \begin{bmatrix} \underline{I}_k & \underline{O} \\ \underline{O} & \underline{I}_{n-k} \end{bmatrix} \tag{3.1-4}$$

which yields three matrix equations in three unknowns, $\underline{U}$, $\underline{V}$, and $\underline{W}$ (a fourth equation is redundant).

$$\begin{cases} \underline{C}_k \cdot \underline{U} + \underline{X} \cdot \underline{V}^T = \underline{I}_k \\ \underline{C}_k \cdot \underline{V} + \underline{X} \cdot \underline{W} = \underline{O} \\ \underline{X}^T \cdot \underline{V} + \underline{Y} \cdot \underline{W} = \underline{I}_{n-k} \end{cases} \tag{3.1-5}$$

Multiplying the first two equations by $\underline{C}_k^{-1}$ on the left, and the third by $\underline{Y}^{-1}$ on the left yields:

$$\begin{cases} \underline{U} + \underline{C}_k^{-1} \cdot \underline{X} \cdot \underline{V}^T = \underline{C}_k^{-1} \\ \underline{V} + \underline{C}_k^{-1} \cdot \underline{X} \cdot \underline{W} = \underline{0} \\ \underline{W} + \underline{Y}^{-1} \cdot \underline{X}^T \cdot \underline{V} = \underline{Y}^{-1} \end{cases} \tag{3.1-6}$$

Finally, using the second equation in the third and in the first, and since $\underline{W}$ is known to exist and to be positive definite, the following solution is obtained:

$$\begin{cases} \underline{U} = (\underline{C}_k - \underline{X} \cdot \underline{Y}^{-1} \underline{X}^T)^{-1} = \underline{C}_k^{-1} + (\underline{C}_k^{-1} \cdot \underline{X}) \cdot \underline{W} \cdot (\underline{C}_k^{-1} \cdot \underline{X})^T \\ \underline{V} = -\underline{C}_k^{-1} \cdot \underline{X} \cdot \underline{W} \\ \underline{W} = (\underline{Y} - \underline{X}^T \cdot \underline{C}_k^{-1} \cdot \underline{X})^{-1} \end{cases} \tag{3.1-7}$$

In particular, if k = n-1, $\underline{X}$ and $\underline{V}$ reduce to the vectors $\underline{x}$ and $\underline{v}$ and $\underline{Y}$ and $\underline{W}$ reduce to the scalars α and β so that:

$$\underline{C} = \begin{bmatrix} \underline{C}_k & \underline{x} \\ \underline{x}^T & \alpha \end{bmatrix} \quad , \quad \underline{C}^{-1} = \begin{bmatrix} \underline{U} & \underline{v} \\ \underline{v}^T & \beta \end{bmatrix} \tag{3.1-8}$$

and the system (3.1-7) reduces to the following:

$$\begin{cases} \underline{U} = \underline{C}_k^{-1} + \beta \cdot (\underline{C}_k^{-1} \cdot \underline{x}) \cdot (\underline{C}_k^{-1} \cdot \underline{x})^T \\ \underline{v} = -\beta \cdot \underline{C}_k^{-1} \cdot \underline{x} \\ \beta = 1/(\alpha - \underline{x}^T \cdot \underline{C}_k^{-1} \cdot \underline{x}) \end{cases} \tag{3.1-9}$$

The system (3.1-9) provides a fast and efficient method for updating the inverse of a positive-definite matrix when the latter is augmented by one row and one column. Moreover, if no a priori knowledge can be assumed about the augmented matrix, the value of β provides all the necessary information: if β is positive, the matrix is positive definite; if β is infinite, the matrix is positive semidefinite singular; if β is negative, the matrix is indefinite.

Both (3.1-7) and (3.1-9) can easily be applied to quadratic forms. Let $\underline{m}$ be an $n \times 1$ vector and $\underline{m}_k$ a $k \times 1$ vector formed with the k first components (again this is assumed without loss of generality as the value of a quadratic form does not depend on the arrangement of the components). Thus:

$$\underline{m} = \begin{bmatrix} \underline{m}_k \\ \underline{m}_{n-k} \end{bmatrix} \tag{3.1-10}$$

Now the quadratic form $\underline{m}^T \cdot \underline{C}^{-1} \cdot \underline{m}$ can be rewritten:

$$\underline{m}^T \cdot \underline{C}^{-1} \cdot \underline{m} = \begin{bmatrix} \underline{m}_k^T & \underline{m}_{n-k}^T \end{bmatrix} \cdot \begin{bmatrix} \underline{U} & \underline{V} \\ \underline{V}^T & \underline{W} \end{bmatrix} \cdot \begin{bmatrix} \underline{m}_k \\ \underline{m}_{n-k} \end{bmatrix}$$

$$= \underline{m}_k^T \cdot \underline{U} \cdot \underline{m}_k + 2\, \underline{m}_k^T \cdot \underline{V} \cdot \underline{m}_{n-k} + \underline{m}_{n-k}^T \cdot \underline{W} \cdot \underline{m}_{n-k} \,. \qquad (3.1\text{-}11)$$

Making use of (3.1-7) and writing the Cholesky decomposition of $\underline{W}$: $\underline{W} = \underline{L} \cdot \underline{L}^T$, (3.1-11) becomes:

$$\underline{m}^T \cdot \underline{C}^{-1} \cdot \underline{m} = \underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{m}_k + \underline{m}_{n-k}^T \cdot \underline{L} \cdot \underline{L}^T \cdot \underline{m}_{n-k} + \underline{m}_k^T \cdot (\underline{C}_k^{-1} \cdot \underline{X}) \cdot \underline{L} \cdot \underline{L}^T \cdot (\underline{C}_k^{-1} \cdot \underline{X})^T \cdot \underline{m}_k$$

$$- 2 \cdot \underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{X} \cdot \underline{L} \cdot \underline{L}^T \cdot \underline{m}_{n-k}$$

which reduces to:

$$\underline{m}^T \cdot \underline{C}^{-1} \cdot \underline{m} = \underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{m}_k + \left|\left| \underline{L}^T \cdot (\underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{X} - \underline{m}_{n-k}^T)^T \right|\right|_2^2 \qquad (3.1\text{-}12)$$

Denoting by d_n^2 the quadratic form $\underline{m}^T \cdot \underline{C}^{-1} \cdot \underline{m}$ and by d_k^2 the quadratic form $\underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{m}_k$, (3.1-12) can be reformulated as:

$$d_n^2 = d_k^2 + \Delta_{n-k}^2 \qquad (3.1\text{-}13)$$

The reduction to the case k = n-1 is straightforward. Let μ be the last component of $\underline{m}$. Then, using the notation of (3.1-9), the equation (3.1-12) reduces to:

$$\underline{m}^T \cdot \underline{C}^{-1} \cdot \underline{m} = \underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{m}_k + \beta \cdot (\underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{x} - \mu)^2 \qquad (3.1\text{-}14)$$

To the knowledge of the author, equations (3.1-12) and (3.1-14) represent a new result, together with the systems (3.1-7) and (3.1-9). Equations (3.1-12) and (3.1-14) are particularly significant in relation with the selection problem: they show that positive definite quadratic forms, the Mahalanobis distance (squared) in particular, increase monotonously as new components are added (thus the process of selecting a subset of components can only decrease the separation between the classes). Moreover, these equations provide a way to compute the change in distance associated with the addition of components without actually computing the new inverse matrix, thus obviously saving time in a selection process. Finally, once a selection is accomplished, the inverse matrix can be rapidly updated.

The matrix equations (3.1-4) have been solved for an updating in the direction of increased dimension. Though this is likely to be the most commonly encountered situation, it is also desirable to derive solutions for an updating in the direction of decreased dimension. Thus $\underline{C}_k^{-1}$ must be expressed in terms of $\underline{U}$, $\underline{V}$, and $\underline{W}$. This is presented only for the case of k = n-1, as it is of more practical interest than the general case. Two of the three equations are rewritten for k = n-1:

$$\begin{cases} \underline{C}_k \cdot \underline{U} + \underline{x} \cdot \underline{v}^T = \underline{I} \\ \underline{C}_k \cdot \underline{v} + \beta\, \underline{x} = \underline{0} \end{cases}$$

Multiplying both on the left by $\underline{C}_k^{-1}$ yields:

$$\begin{cases} \underline{U} + \underline{C}_k^{-1} \cdot \underline{x} \cdot \underline{v}^T = \underline{C}_k^{-1} \\ \underline{v} + \beta \cdot \underline{C}_k^{-1} \cdot \underline{x} = \underline{0} \end{cases}$$

Substituting the second equation into the first, the following expression is obtained for $\underline{C}_k^{-1}$:

$$\underline{C}_k^{-1} = \underline{U} - 1/\beta \cdot \underline{v} \cdot \underline{v}^T \qquad (3.1\text{-}15)$$

This result can readily be applied to the evaluation of the quadratic form—where now $\underline{m}_k^T \underline{C}_k^{-1} \underline{m}_k$ is desired—as follows. From (3.1-14):

$$\underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{m}_k = \underline{m}^T \cdot \underline{C}^{-1} \cdot \underline{m} - \beta (\underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{x} - \mu)^2$$

Using the above results, this becomes:

$$\underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{m}_k = \underline{m}^T \cdot \underline{C}^{-1} \cdot \underline{m} - \beta (-\underline{m}_k^T \cdot \frac{1}{\beta} \cdot \underline{v} - \mu)^2$$

which can be rewritten as:

$$\underline{m}_k^T \cdot C_k^{-1} \cdot \underline{m}_k = \underline{m}^T \cdot \underline{C}^{-1} \cdot \underline{m} - \frac{1}{\beta} (\underline{m}_k^T \cdot \underline{v} + \beta\mu)^2 \qquad (3.1\text{-}16)$$

Equations (3.1-15) and (3.1-16) are the counterpart of system (3.1-9) and equation (3.1-14). All the remarks made above concerning the advantages of the latter also hold for the "downward" updating procedure. Thus these results allow for an overall simplification of the computations involved in a selection process.

As the incorporation of all measurements results in the maximum distance, a computation of this upper bound may be desirable. This requires the evaluation of $\underline{C}^{-1}$; as the value of the upper bound should be available at the beginning of the selection process, a stepwise computation of $\underline{C}^{-1}$ is not efficient. Several methods exist for the direct calculation of the inverse of a positive-definite matrix [6], [22]. Once this inverse is obtained, however, the computation of inverse submatrices might not be necessary, as a corresponding submatrix of $\underline{C}^{-1}$ could directly be used. This possibility, developed below, presents the particularity that it makes a distance computed with k measurements, $k < N$, dependent on all initially available N measurements, with the risk of distorting the distance measure by biasing it towards nonexistent measurements.

Using this measure is equivalent to setting $\underline{m}_{n-k} = \underline{0}$ in (3.1-12), thus the equation becomes:

$$d_k'^2 = d_k^2 + ||(\underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{X} \cdot \underline{L})^T||_2^2 \qquad (3.1\text{-}17)$$

and, in the case where $k = n-1$:

$$d_k'^2 = d_k^2 + \beta(\underline{m}_k^T \cdot \underline{C}_k^{-1} \cdot \underline{x})^2 \qquad (3.1\text{-}18)$$

These equations show the new measure, $d_k'^2$, to be always larger than the equivalent measure, d_k^2, which uses the true inverse of $\underline{C}_k$. Therefore, this new measure presents a double advantage: better separation and a single inverse matrix computation. It is noted, however, that this

computation may not always be possible, for reasons of computation time, storage memory, and rounding errors.

Unfortunately, this new measure does not increase monotonously as new components are added. Indeed, using the notation of (3.1-12), the relationship between the distances computed with n and with k components is as follows:

$$d'^2_n = d'^2_k + \underline{m}^T_{n-k} \cdot (2 \cdot \underline{V}^T \cdot \underline{m}_k + \underline{W} \cdot \underline{m}_{n-k}) \qquad (3.1\text{-}19)$$

and, in the case where k = n-1:

$$d'^2_n = d'^2_k + \mu \cdot (2 \cdot \underline{v}^T \cdot \underline{m}_k + \beta \cdot \mu) \qquad (3.1\text{-}20)$$

Thus added components may actually decrease the total distance and the maximum value is no longer necessarily reached by using all components. This measure then appears as less representative of the actual situation than the other; nevertheless, its computational advantages way outweigh this problem.

3.2. Geometric Interpretation

The results of the previous section can be better understood through the use of a geometric representation. In order to appreciate the full meaning of the latter, a few remarks are necessary.

It is known from matrix theory [18] that the quadric of equation:

$$\underline{x}^T \cdot \underline{A} \cdot \underline{x} = \text{constant} \qquad (3.2\text{-}1)$$

where $\underline{A}$ is positive definite, represents an ellipsoid. Moreover, the problem of minimizing $||\underline{x}||_2$ subject to the condition (3.2-1) is equivalent to that of maximizing $\underline{x}^T \cdot \underline{A} \cdot \underline{x}$ subject to the condition $||\underline{x}||_2$ = constant [18, p. 405]. Finally, it is noted that, if $\underline{A} = \underline{C}^{-1}$, the variance in the direction of $\underline{x}$ is given by:

$$1/\sigma^2(\underline{C},\underline{x}) = \underline{x}^T \cdot \underline{C}^{-1} \cdot \underline{x} / \underline{x}^T \cdot \underline{x} \qquad (3.2\text{-}2)$$

This ratio is known as the Rayleigh quotient. Thus the two problems mentioned above are equivalent to maximizing the Rayleigh quotient, or minimizing the variance, along all axes.

This allows a geometric arrangement for interpreting the Mahalanobis distance: the quadric (3.2-1) is drawn for an arbitrary value of the constant, and a line segment drawn from the unknown vector to the mean of the class (this is the support of the vector $\underline{x}$ used above). The Mahalanobis distance is then directly proportional to the ratio of the Euclidean distance between the unknown vector and the mean of the class (i.e., $\sqrt{\underline{x}^T \cdot \underline{x}}$) to the radius of the quadric in this direction. (If the quadric is drawn for a constant value of 1, this radius is the standard deviation of the set in this direction, and the Mahalanobis distance is directly equal to the ratio defined above.) This is illustrated in Figure 3. By using this representation, it becomes possible to illustrate geometrically the results of section 3.1.

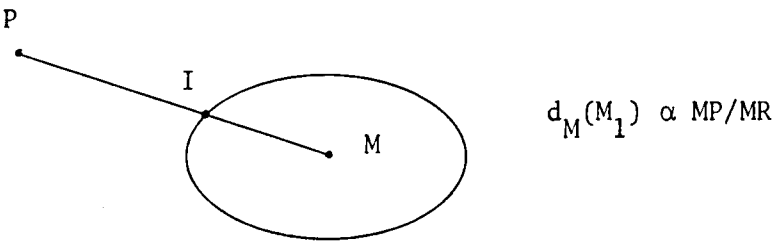


Figure 3. Illustration of the Mahalanobis distance.

First, consider Figure 4. The point M is the mean of the class, and the point P is the unknown pattern. The coordinates of these two points on both axes are denoted by a subscript x or y. The Mahalanobis distance using both x and y components is proportional to $d_E(M,P)/d_E(M,I)$; using only the x component, the distance is proportional to $d_E(M_x,P_x)/d_E(M_x,I_x)$, where I_x is not, in general, the x-coordinate of I. Since I_x is always situated between P_x and the projection of I on the x axis, elementary geometry insures that the measure using both components is always larger than that using only the x component. (Similar reasoning can be held for the y component.) Equality occurs when the principal axes of the ellipsoid are parallel to the coordinate axes and MP is along one of the axes (i.e., $\underline{C}_k$ is diagonal and $\underline{x}$ has only one nonzero component).

Then consider Figure 5. As the original covariance matrix is kept, the result is that the original quadric, not its projection, is used as a measure of spread. Clearly, the ratio $d_E(M,P_x)/d_E(M,I_x')$ is always larger than, or equal to, the ratio $d_E(M_x,P_x)/d_E(M_x,I_x)$; the same remark holds for measures along the y-axis. It is noted that the ratio $d_E(P_x,M)/d_E(M,I_x')$ is even larger than the ratio $d_E(P,M)/d_E(M,I)$,

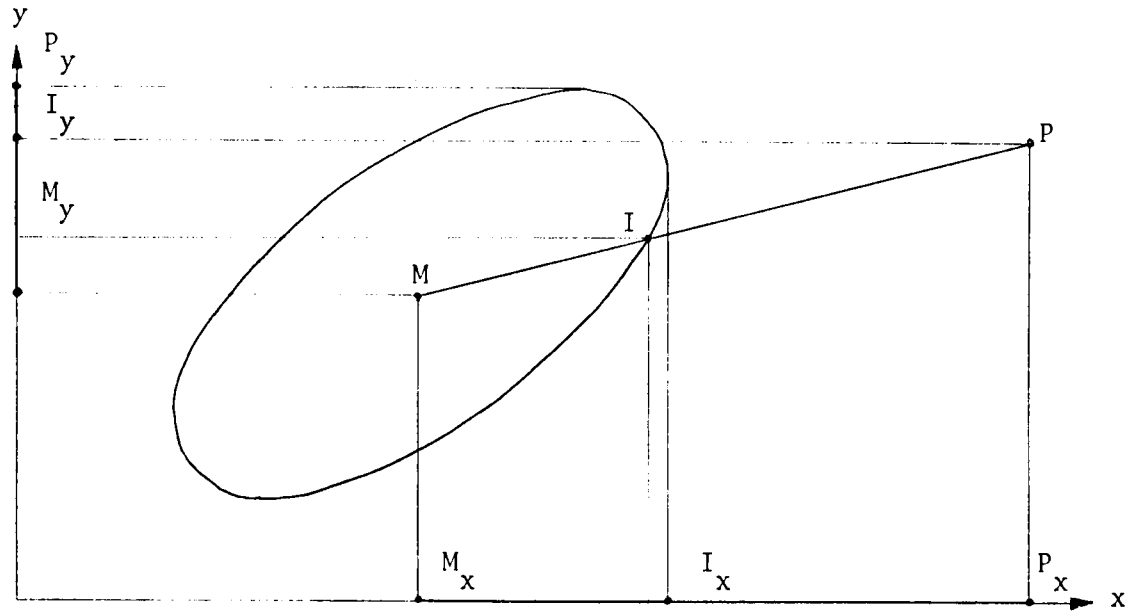


Figure 4. Illustration of the first type of dimensionality reduction.

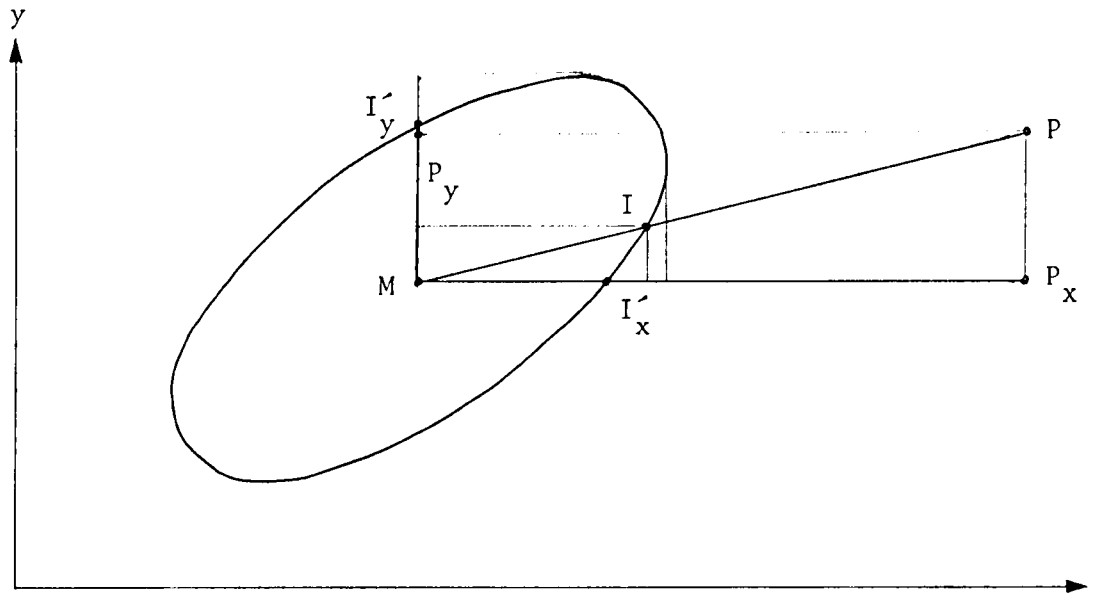


Figure 5. Illustration of the second type of dimensionality reduction.

meaning that the maximum of this second type of measure is not reached when using both components, but when using the x component alone.

However, in a multiclass case, an increased distance does not imply better classification performance, at least when using classifiers of the minimum distance type. This is due to the fact that only ratios change; the actual information available constantly decreases when components are removed.

CHAPTER 4

INTERACTIVE SELECTION PROCEDURE

4.1. Optimal and Suboptimal Searches

The problem of selecting the "best" subset of components is a search problem. That is to say, it requires the evaluation of each possible subset against the chosen criterion and then the choice of the most satisfactory subset. However, for a set of N measurements, there are 2^N-1 non empty subsets; thus the number of possible subsets gets rapidly out of hand. Even when the choice is limited to subsets of, say, k components, there still are $\binom{N}{k}$ possible subsets. To appreciate the size of these numbers, let N = 50 and k = 20; then 2^N-1 is of the order of 10^{15} and $\binom{N}{k}$ of the order of 10^{13}. These are clearly forbidding numbers; thus an exhaustive search is, for all practical purposes, impossible.

An attractive solution is the stepwise search, in which, at the k-th step, that of the remaining N-k components which, when added to the k already selected, yields the best subset of k+1, is selected. To obtain k components, this method will require $N + (N-1) + \ldots + (N-k) = N(k+1) - 1 - \frac{1}{2}k(k+3)$ searches. For the previous example, this yields 819 searches, which is well within feasible values. Unfortunately, such an approach does not, in general, yield the optimal solution. This is a typical feature of multivariate problems, where the interdependence of the various parameters requires that all of them be considered together. In the particular case where the variables are statistically

independent, a stepwise procedure will yield the optimal solution; otherwise, nothing can be said about the value of the selected subset. (This is true for the selection problem, but not for the feature ordering problem—as encountered in sequential pattern recognition—in which case some optimality properties can be demonstrated [13]). Nevertheless, stepwise procedures (also called sequential procedures) constitute the only tools available for deterministic selections.

A sequential selection procedure can start at both ends; that is, it can either add the next best, or delete the next poorest, measurement. These two methods are referred to as forward and backward sequential searches. Stopping criteria must be provided for automated search procedures to avoid ending with a trivial selection (all components or none at all). Such criteria are chosen in accordance with the selection strategy to determine the significance of prospective new measurements. Equipped with well-chosen criteria, sequential methods will find suboptimal solutions.

All selection procedures, however, are based on statistical strategies. The structural information which may be available is not taken into consideration. Other choice criteria, such as cost of measurement, sensitivity to noise, or reliability of individual measurements, are difficult to incorporate, despite some efforts in the field of dynamic programming. In general, the above-mentioned information, structural in particular, cannot very well be fed to computers. Human interpretation is necessary at this level; this is the rationale behind interactive procedures.

Therefore, it seems desirable to develop a procedure which provides a human user with all the facilities of automated search methods and accepts human input to direct the use of searches.

4.2. Computational Problems

As stressed before, the computation of the inverse matrix, $\underline{C}^{-1}$, is a delicate task when the dimension of $\underline{C}$ is large. This problem is further complicated by the requirement that $\underline{C}^{-1}$ be positive definite despite the rounding errors. Unfortunately, sample covariance matrices are often ill-conditioned, due to the lack of sample vectors. (In this respect, $\underline{W}$—as defined by (2.3-3)—is likely to present a better behavior.)

The best way to minimize rounding errors is to use the Cholesky decomposition; this decomposition is very stable from a numerical point of view. The quadratic form can then be rewritten:

$$d^2 = \underline{x}^T \cdot \underline{C}^{-1} \cdot \underline{x} = \underline{x}^T (\underline{L} \cdot \underline{L}^T) \bar{\underline{x}}^1 = ||(\underline{L}^T)^{-1} \cdot \underline{x}||_2^2 \qquad (4.2\text{-}1)$$

which can also be written:

$$d^2 = ||\underline{y}||_2^2 \quad , \quad \underline{L}^T \cdot \underline{y} = \underline{x} \qquad (4.2\text{-}2)$$

As $\underline{L}$ is lower triangular, $\underline{y}$ can be immediately found by back-substitution This method has the added advantage of insuring that the quadratic form will be nonnegative. Unfortunately, it is extremely difficult to introduce a recurrence (of the type developed in section 3.1) in these formulas.

It is noted, however, that the updating procedure summarized by equations (3.1-9) is very stable and presents the advantage of determining if the updated inverse is positive definite (thanks to the test on the value of β). Thus the selection process will not be more sensitive to the condition of the covariance matrices than direct inversion procedures.

Inverting the matrices directly might still be desirable; since the estimation of the distance requires more operations using (4.2-2) than by straight matrix multiplication using the inverse matrix, computation time might be reduced by using inverses. In such a case, it usually proves better to invert the triangular matrix, $\underline{L}^T$, than the original matrix ([6],[22]). (This can be related to the fact that the condition number of $\underline{C}$ is the square of that of $\underline{L}^T$.) Nevertheless, negative values might be obtained when computing distances. When using these distances for classification, however, the problem is lessened since the error introduced in the measure is approximately constant and thus, within reasonable limits, should not change the classification.

4.3. Set-Up of the Procedures

Several procedures have been programmed to implement interactive sequential selection, making use of the results of sections 2.3 and 3.1. Both measures of average interclass distance—as discussed in section 2.3—can be used; it was felt that the arithmetic average of all interclass distances could prove useful as, with the same set-up, selection criteria could be directed towards various subsets of these distances (an unrealizable option while using the generalized measure).

Programs were designed to use the updating procedures described in section 3.1. An additional procedure, using the arithmetic mean criterion and working with the large inverse matrices, was programmed, to be used with sets of reduced dimension. All procedures can perform forward and backward sequential searches—though the backward search can only start from the point reached by the forward search. Variable thresholds are used as stopping criteria, either in absolute or in relative value. When using the arithmetic mean criterion, weights can be assigned to specific distances in order to bias the search; moreover, the search can be directed to try to isolate a specific class or to use only the distance between a specified pair of classes.

The results of the selection are judged by attempting to classify the training patterns (with and without a jackknife) and patterns generated by slight modifications to the training set. It will be recalled that the jackknife procedure consists of leaving some patterns out of the training set, classifying them with the selection obtained by using the diminished training set, and repeating the procedure until all patterns have been used. In the absence of real data, the jackknife method allows a more realistic estimation of the performance of a recognition system; however, it is a time-consuming procedure.

The classifier used was chosen to be well suited to the selection criteria, simple (since the purpose of the experiments was selection and not classification), and still able to produce good results. An obvious choice was the minimum distance classifier (using the Mahalanobis distance), which assigns a pattern to the class—represented by its mean—closest to it.

The programs were written in FORTRAN V and run on the DEC-system 10 computer of the University of Tennessee, with an allocation of 65 k words of main memory.

CHAPTER 5

MOMENTS AS FEATURES

As mentioned in Chapter 1, feature extraction is the first stage in a pattern recognition process. Thus, in order to test the selection procedures developed in this study on real data, some method of feature extraction must be chosen. In the absence of any information concerning the data, global feature descriptors are more desirable than local ones; also, some properties of invariance to several types of perturbation should be found in any good feature descriptor.

Moments satisfy all these requirements. They have been extensively used in classical mechanics and statistics, and several researchers have used them for pattern recognition purposes ([11], [16]). It will be recalled that the moment of order p of a discrete function f(x) over a finite interval is defined by:

$$m_p = \sum_x x^p \cdot f(x) \qquad (5\text{-}1)$$

where the summation is taken over all x in the interval. This is easily extended to n-dimensional moments as follows. The $p_1 p_2 \ldots p_n$-th moment (of order $p_1 + p_2 + \ldots + p_n$) of the function $f(x_1, x_2, \ldots x_n)$ is defined as:

$$m_{p_1 p_2 \cdots p_n} = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} x_1^{p_1} \cdot x_2^{p_2} \cdots x_n^{p_n} \cdot f(x_1, x_2, \ldots x_n) \qquad (5\text{-}2)$$

M. K. Hu [16] has noted that two-dimensional moments of a continuous function—the summations are then replaced by integrals—will, under some finiteness assumptions, uniquely determine the function itself and, vice versa, the function uniquely determines the sequence of moments. In a discrete case, it is logical to assume that only a finite set of moments is necessary to represent a function over a finite interval. The following proof is due to Barrero [3] and restricted, for simplicity, to a one-dimensional case.

Let the discrete function f of a discrete variable x_i be defined over the interval i = 1, ..., N; this function can then be represented in vector form as follows:

$$\underline{f} = [f(x_1)\ f(x_2)\ \ldots\ f(x_n)]^T \tag{5-3}$$

Let the sequence of moments of order p, p = 0, ..., N - 1, also be represented by a vector:

$$\underline{m} = [m_0(F)\ m_1(F)\ \ldots\ m_{N-1}(F)] \tag{5-4}$$

Then a matrix $\underline{X}$ can be defined such that:

$$\underline{m} = \underline{X} \cdot \underline{f} \tag{5-5}$$

and this matrix reads as follows:

$$\underline{X} = \begin{bmatrix} x_1^0 & \cdots & x_i^0 & \cdots & x_N^0 \\ x_1^1 & \cdots & x_i^1 & \cdots & x_N^1 \\ \vdots & & \vdots & & \vdots \\ x_1^{N-1} & \cdots & x_i^{N-1} & \cdots & x_N^{N-1} \end{bmatrix} \tag{5-6}$$

If $x_i \neq x_j$, $i \neq j$, this matrix is nonsingular (its determinant is known in matrix algebra as a Vandermonde determinant [5], [18]). Thus it has an inverse, $\underline{X}^{-1}$, and the transformation can be reversed:

$$\underline{f} = \underline{X}^{-1} \cdot \underline{m} \tag{5-7}$$

which proves that N moments are sufficient to represent uniquely N samples of a discrete function. It is further noted that the use of x in the above equations implies only that x be a single-valued function of the index i. This allows the use of various functions in order, for instance, to reduce the computational problems arising with high order moments, or to pick out specific regions in the interval.

Invariant properties of moments are easily developed. The following moments, called central moments, are invariant under translation:

$$\mu_p = \sum_x (x - \bar{x})^p \cdot f(x), \quad x \text{ in the interval} \tag{5-8}$$

where $\bar{x}$ $(= m_1/m_0)$ corresponds to the center of mass of the function. In order to achieve invariance to size, these moments are normalized as follows:

$$\eta_p = \mu_p / m_0^{(p/2 + 1)} \tag{5-9}$$

In particular, $\eta_0 = 1$, $\eta_1 = 0$.

It can be further demonstrated [16] that the theory of algebraic invariants is applicable to moments, so that combinations of moments can be developed that are invariant under general linear (nonsingular) transformations.

Some moments are well known throughout the technical literature. The moment of order zero is the sum of all values of the function in the interval and thus represents the "mass" of the function. The noncentral moments of order one are related to the center of mass; the moments of order two are related to the principal components (central moments of order two express variances and covariances). Finally, moments of order three are associated with the skewness of the function and moments of order four with its kurtosis.

Thus moments appear as useful global feature descriptors; they are easily combined to form invariants; with the use of an appropriate index function, they can be made to extract local features. Finally, the computational problems associated with moments of large orders can be overcome by scaling methods or by choosing an index function with a reduced dynamic range (it is noted, however, that this last procedure is likely to diminish the discriminant power of the moments, and to make the matrix in (5-6) ill-conditioned).

CHAPTER 6

EXPERIMENTAL RESULTS

6.1. Iris Data

The first application example is based on the classic taxonomy problem presented by Fisher [12]. This problem involves three species of Iris, namely I. setosa, I. versicolor, and I. virginica, on which four measurements have been performed: sepal length, sepal width, petal length, and petal width. The class of I. setosa is known to be linearly separable from the two others, which present considerable overlap. This is obviously a rather simple case, which can be investigated by exhaustive search; this, and the fact that numerous researchers have studied the problem from various points of view, make the Iris data a useful basis of comparison.

Fisher's original data consist of fifty sets of measurements—fifty patterns—per class. Tests were run without jackknife procedure, using both criteria presented in section 2.3. The additional procedure using the complete inverse matrices was also used, since the small number of measurements would not allow severe distortions of the measure. For comparison purposes, exhaustive classification tests were run, using both the "true" inverse covariance matrix (i.e., the inverse of the reduced covariance matrix) and the inverse matrix subsets. First, a classification test was run using all four measurements to establish a basis for comparison. The resulting classification matrix is shown

in Table 1. It corresponds to a total misclassification rate of 2 percent. Using a similar classifier, the BMD program, P7M [10], presents the same error percentage, where a method based on density estimation [6] resulted in a perfect score.

Table 1. Iris classification matrix using all measurements.

from: to:	I. setosa	I. versicolor	I. virginica
I. setosa	50	0	0
I. versicolor	0	47	3
I. virginica	0	0	50

The selection program based on the large inverse matrices was then run using both forward and backward sequential procedures; these always converged to the same result. Thus subsets of three, two, and one measurements were determined, which resulted in misclassification rates of 2, 3-1/3, and 4-2/3 percent respectively. The classification matrices for these selections are shown in Table 2, where the measurements have been numbered for commodity (1: sepal length, 2: sepal width, 3: petal length, 4: petal width).

The exhaustive classification tests using the large inverse matrices showed all selections to be indeed optimal. A second subset of three measurements and a second subset of two measurements demonstrated the same error percentage as the selected subsets, but none better could be found. It is interesting to note that twelve of the fourteen possible

Table 2. Iris classification matrices (I) using (a) measurements 2, 3, 4; (b) measurements 3, 4; (c) measurement 3.

from: to:	I. setosa	I. versicolor	I. virginica
	(a) Measurements 2, 3, 4		
I. setosa	50	0	0
I. versicolor	0	48	2
I. virginica	0	1	49
	(b) Measurements 3, 4		
I. setosa	50	0	0
I. versicolor	0	47	3
I. virginica	0	2	48
	(c) Measurement 3		
I. setosa	50	0	0
I. versicolor	0	46	4
I. virginica	0	3	47

subsets correctly classified all patterns of the first class; only the first or the second measure taken alone resulted in a few errors for this class (while the same subsets completely confused I. versicolor and I. virginica). The classification results are summarized in Table 3.

Table 3. Error rates associated with Iris measurement subsets (I).

Subset	Error Rate	Subset	Error Rate
1,2	18	2,3	4-2/3
1,3	3-1/3	2,4	4-2/3
1,4	8	3,4	3-1/3
1,2,3	4	1	28
1,2,4	8	2	38
1,3,4	2	3	4-2/3
2,3,4	2	4	5-1/3

The same procedure was repeated with the other two procedures, which use the updating method to compute the "true" inverse at each step. The generalized distance criterion led to rather poor selections (with 4, 4, and 28 percent of error respectively), while the arithmetic mean criterion resulted in the optimal selection for each subset dimension. Classification matrices are shown in Table 4, and the exhaustive classification results summarized in Table 5. It is noted that a better classification was achieved by using three measurements than by using all four of them. The comparison between the two approaches (subsets of the large inverse matrix versus "true" inverse matrix) confirms what has been said in section 3.1: the classification results are better

Table 4. Iris classification matrices (II) using (a) measurements 1, 3, 4; (b) measurements 3, 4; (c) measurement 4.

from: to:	I. setosa	I. versicolor	I. virginica
	(a) Measurements 1, 3, 4		
I. setosa	50	0	0
I. versicolor	0	48	2
I. virginica	0	0	50
	(b) Measurements 3, 4		
I. setosa	50	0	0
I. versicolor	0	47	3
I. virginica	0	1	49
	(c) Measurement 4		
I. setosa	50	0	0
I. versicolor	0	48	2
I. virginica	0	4	46

Table 5. Error rates associated with Iris measurement subsets (II).

Subset	Error Rate	Subset	Error Rate
1,2	19-1/3	2,3	4
1,3	4	2,4	5-1/3
1,4	3-1/3	3,4	2-2/3
1,2,3	4	1	28
1,2,4	4-2/3	2	44-2/3
1,3,4	1-1/3	3	4-2/3
2,3,4	3-1/3	4	4

with the "true" inverse. Moreover, the selection operated by one method is not optimal when used with the other method. Both methods, however, achieved optimal solutions; while this cannot be expected to remain true for more complex problems, it is a proof that sequential procedures are able to choose better-than-average subsets.

6.2. Radar Returns of Aircraft

In this experiment, eight different aircraft were used, four American planes and four Russian planes. These aircraft had been selected for a previous investigation [15] because they form four more or less similar pairs: two large transports, two light bireactor transports, two large fighters, and two small fighters. A list of the eight planes with their wingspan is given in Table 6.

The radar returns were carefully simulated by M. Harris [15], using model airplanes. As used in the following experiments, a radar image consists of a set of reflection points, characterized by their coordinates on a 512 by 512 grid, around which electromagnetic scatter is simulated

Table 6. Aircraft used in the experiments with radar returns.

Name	Type	Nationality	Wingspan
Mikoyan MIG-21	Fighter	USSR	7.15 m
McDonnell-Douglas A4M	Fighter	USA	8.38 m
Mikoyan MIG-25	Fighter	USSR	12.19 m
General Dynamics FB-111A	Fighter	USA	19.20 m
Tupolev TU-22	Transport	USSR	27.70 m
Boeing 737	Transport	USA	28.35 m
Antonov AN-22	Transport	USSR	64.39 m
Lockheed C-5A	Transport	USA	67.88 m

by generating a cluster of points with a two-dimensional Gaussian random generator. Typical images have from five to fifteen reflection points, each with a cluster of twenty to fifty points; one such image is shown in Figure 6.

Sixty-four patterns were generated for each class, for a total of 512 different images. The patterns correspond to about thirty different positions of each aircraft, each pattern thus obtained being used twice with the random generator. Figures 7 and 8 illustrate some of the patterns thus generated for the FB-111A and the B-737.

Central moments were used to describe the images. Thirty-three moments were computed, namely all moments of order two to order seven, using a reduced index described by the function:

$$x = 1 + i/(512 \cdot 10^{-2}) \quad , \quad i = 1, \ldots, 512 \qquad (6.2\text{-}1)$$

The moments were all divided by m_{00}, the zero-th order moment, but were not normalized since invariance to size was not desired. In fact, the



Figure 6. Typical simulated radar image.

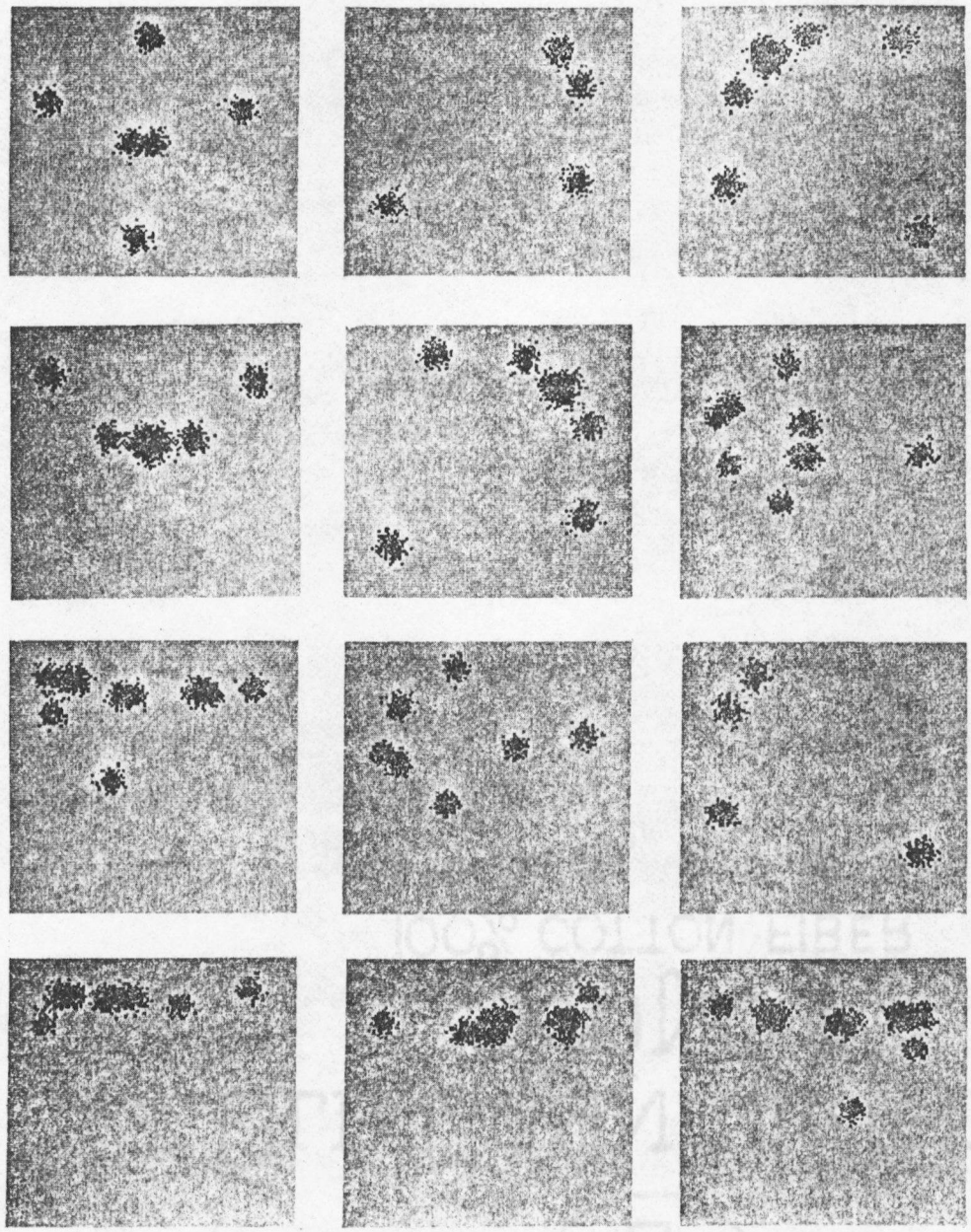


Figure 7. Sample patterns of the FB-111A class.

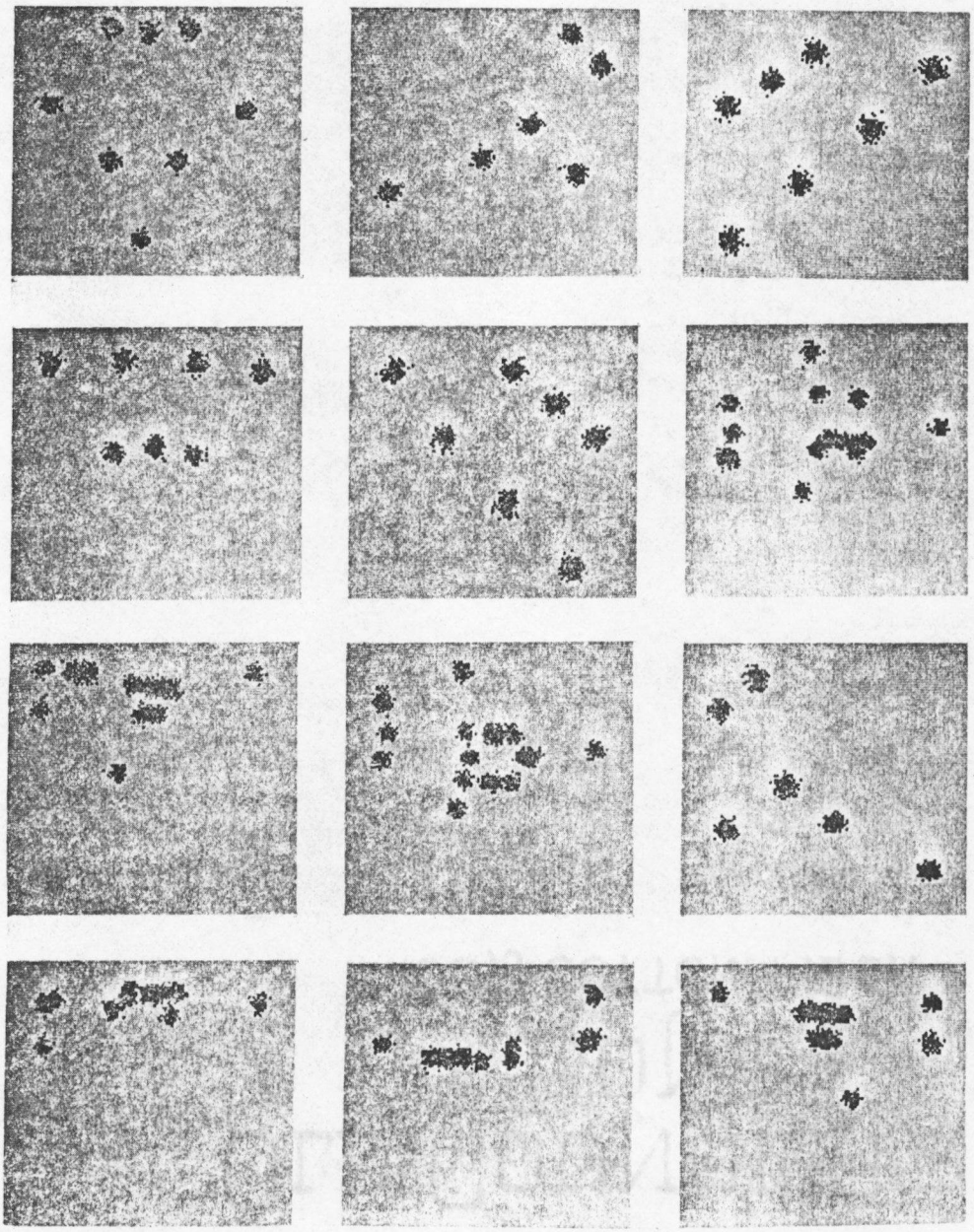


Figure 8. Sample patterns of the B-737 class.

size of the airplanes is a very important factor in this type of recognition problem. As radars usually provide range information, or even automatic scaling, the size can be taken into account irrespective of the distance at which the plane is flying. These experiments assume an automatic prescaling so that images are all of correct proportions. The scaling can also be accomplished with the moments themselves: the gyration radius of an image (i.e., $\sqrt{\mu_{02} + \mu_{20}}$) is known to be inversely proportional to the distance between the object and the imaging device and can thus be used together with the range information to scale the other moments [11].

Three series of experiments were conducted. In the first series, all eight airplanes were used, whereas the second and the third series dealt with the fighters and the transports respectively. As a preliminary step, classification tests were run using all thirty-three moments. The results were surprisingly good: only 6 out of the 512 patterns were misclassified, yielding an error rate of 1.17 percent. The classification matrix for this experiment is shown in Table 7. It is noted that there was not any confusion between fighters and transports.

Jackknife procedures resulted in an overall error rate of 2.9 percent (fifteen errors). Since the difference between both results was so small, and because jackknife classification required much more time than normal classification, the jackknife procedure was abandoned for the rest of the experiments. It is felt that the comparison between the performances of two subsets is as valid with normal evaluation methods as with jackknife procedures.

Table 7. Aircraft classification matrix using all 33 moments.

from: to:	MIG21	A4M	MIG25	FB111	TU22	B737	AN22	C5A
MIG21	63		1					
A4M	1	63						
MIG25			64					
FB111A			3	61				
TU22					64			
B737					1	63		
AN22							64	
C5A								64

The first test was designed to evaluate the usefulness of the large inverse matrix method as applied to high-dimensional data. A selection of 11 components was obtained using the arithmetic mean criterion, which yielded a catastrophic error rate of 80.2 percent. Directing the selection to use the smallest distance only as criterion, the result improved to 76.4 percent of error. Using the "true" inverse covariance matrix of each subset for classification diminished the error rates to 64.5 and 51.8 percent respectively. Most patterns were classified as C5A, i.e., attributed to the class which had the largest spread. Other selections confirmed these results, clearly showing the inapplicability of this procedure to high-dimensional data.

Random selections of 11 components were then performed and classification tests run to obtain a mean value of performance with which to compare the results of the first two selection procedures. For the classification among eight airplanes, an average error rate of 40 percent was observed, while the two-class fighter-transport

classification was correct 90 percent of the time. Both selection procedures were then used; both selected the same subset, for which classification tests showed an error rate of 17.2 percent for the eight-class problem and 2.7 percent for the two-class problem. The classification matrix for this experiment is given in Table 8.

Table 8. Aircraft classification matrix using 11 selected moments.

from: to:	MIG21	A4M	MIG25	FB111A	TU22	B737	AN22	C5A
MIG21	37	13	14					
A4M	3	50	11					
MIG25		2	53	1	7	1		
FB111A			12	46	1	5		
TU22					52	8		4
B737					5	58		1
AN22							64	
C5A								64

Various other subsets were selected using these two procedures; for each selected subset, several other subsets were randomly chosen (with the same number of components) to provide a basis for comparison. Some of these results are summarized in Table 9. It is noted that the selected subset always allowed much more accurate classifications. A second important feature is the relationship between the number of selected components and the error rate; Figure 9 illustrates this relationship. The curve presents two thresholds, between which the increase in performance is very slow; thus a practical pattern recognition system would use either a dozen measurements or all of them.

Table 9. Classification results for various subsets.

Number of Components	Selected Subset Error Rate		Random Subset Error Rate	
	8-class	2-class	8-class	2-class
1	62.6	31.2	81.	50.
3	38.3	16.7	70.	32.
5	37.1	15.2	54.	28.
9	16.1	6.5	45.	16.
11	17.2	2.7	40.	10.
18	18.1	3.	27.	7.
23	12.7	3.1	21.	6.

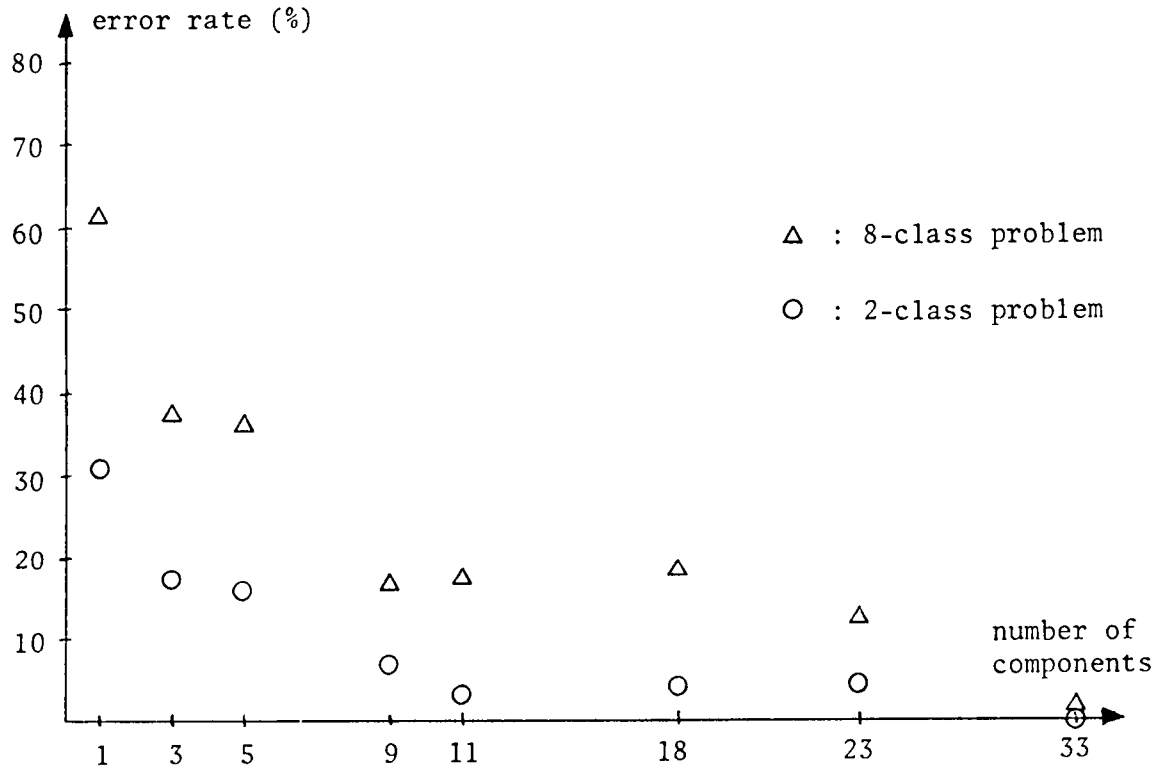


Figure 9. Relationship between the number of selected components and the error rate.

The second series of experiments used only the four fighters, since the distinction between fighters and transports seemed well-defined. Two selections were obtained, one of nine, and one of eighteen components. In agreement with the curve of Figure 9, the selection of nine measurements yielded slightly better results (9 percent versus 9.4 percent of error). Classification matrices are shown in Table 10. It is noted that the first seven moments selected were the same as those selected for the 8-class problem. (The same seven moments will also be the first selected in the next series of experiments.)

Table 10. Fighter aircraft classification matrices using (a) nine selected moments and (b) eighteen selected moments.

from: to:	MIG21	A4M	MIG25	FB111A	MIG21	A4M	MIG25	FB111A
MIG21	45	7	12		47	7	10	
A4M		63	1		1	61	2	
MIG25		1	63			1	63	
FB111A			2	62			3	61

The last series of tests was a repetition of the second, this time using the four transports. As expected, there were less misclassifications with these planes as the selected subsets resulted in error rates of 4.7 and 7.8 percent for nine and eighteen components respectively. The classification matrices are shown in Table 11.

These three series of experiments have shown both the discriminating power of moments and of the minimum Mahalanobis-distance classifier and the usefulness of sequential search procedures. Selected subsets

Table 11. Transport aircraft classification matrices using (a) nine selected moments and (b) eighteen selected moments.

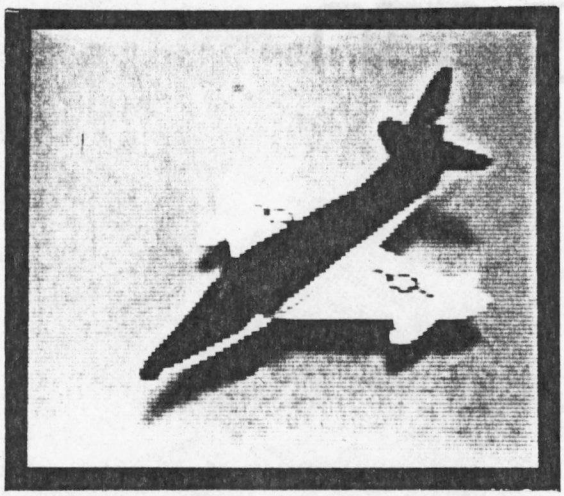
from: to:	TU22	B737	AN22	C5A	TU22	B737	AN22	C5A
TU22	55	8		1	57	7		
B737	1	62		1	13	51		
AN22			63	1			64	
C5A				64				64

resulted in two to three times less misclassifications than equivalent subsets chosen at random. Moreover, the selection allowed a reduction of a factor three in the number of measurements for a decrease in performance of about 10 percent only.

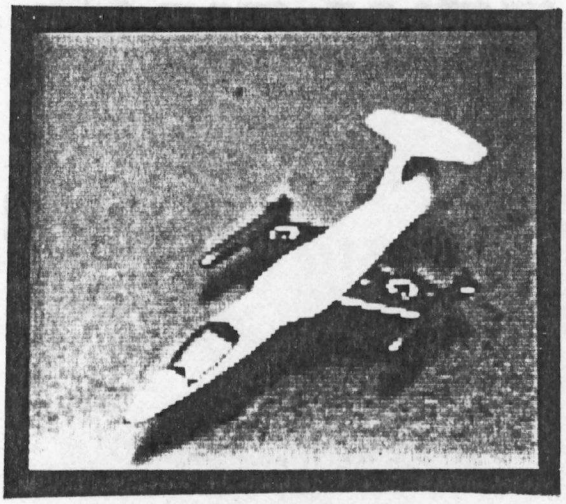
6.3. Photographs of Aircraft

In this experiment, a series of four pictures was taken for each of four different toy aircraft. All four planes were fighters: the A4M "Skyhawk," the F104, the A7D "Corsair," and the Mirage F1. Each photograph represents a plane in a different position. Typical photographs are shown in Figure 10. The pictures were digitized in a grid of 128 by 128 points, using 256 levels of gray. Each picture was then numerically rotated 90, 180, and 270 degrees, and symmetrized around both vertical and horizontal axes, to yield a total of twenty-four images per airplane.

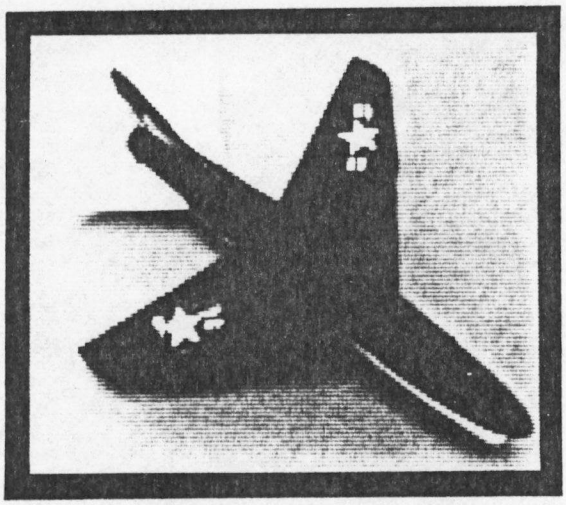
Six moment invariants were used to describe these pictures. (The choice of such a small number of features was based on the fact that photographs convey a large amount of information about aircraft—much



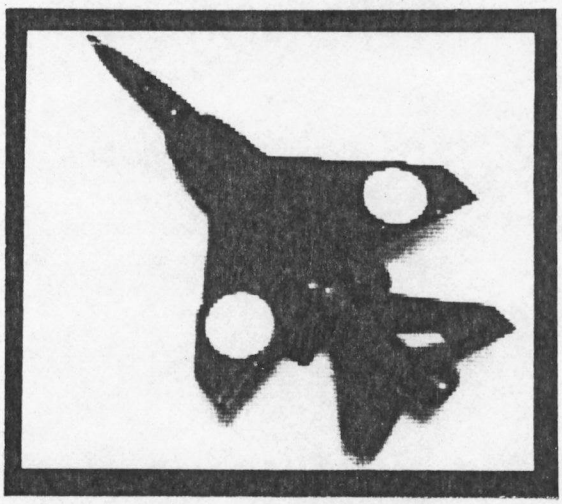
(a) A4M "Skyhawk"



(b) F104



(c) A7D "Corsair"



(d) Mirage F1

Figure 10. Sample photographs of the four aircraft.

more, for instance, than radar returns.) These invariants were derived from the central normalized moments of order two and three using formulas developed by M. K. Hu [16], and adopting a normalization by the gyration radius introduced by Dudani [11]. These invariants can be shown to remain unchanged under general linear transformations.

In terms of the central normalized moments, the invariants can be expressed as follows:

$$I_1 = [(\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2] \; / \; (\eta_{02} + \eta_{20})^2$$

$$I_2 = [(\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2] \; / \; (\eta_{02} + \eta_{20})^3$$

$$I_3 = [(\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2] \; / \; (\eta_{02} + \eta_{20})^3$$

$$I_4 = \{(\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$$
$$+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\} \; /$$
$$(\eta_{02} + \eta_{20})^6$$

$$I_5 = \{(\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$
$$+ 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\} \; / \; (\eta_{02} + \eta_{20})^4$$

$$I_6 = \{(3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$$
$$- (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\} \; /$$
$$(\eta_{02} + \eta_{20})^6$$

(6.3-1)

The last invariant, I_6, is taken in absolute value, since it changes sign under improper transformations.

The computed invariants differed only in the fourth significant digit; such accuracy is due to the absence of digitizing errors in the rotations and symmetries, and cannot be expected for more general transformations. This result shows, however, that moment invariants can be easily and accurately computed.

Preliminary classification tests were run using all six measurements: they yielded a perfect score of 96 correct classifications. Both procedures were then used to select subsets of one, two, three, and four components, which were compared against randomly selected subsets. All subsets of four components were found to result in perfect classification, thereby confirming the assertion made above concerning the informational content of a photograph.

Subsets of three components as selected by both procedures yielded error rates of 31 percent and 17 percent. Typical random subsets resulted in an error rate of 25 percent; one subset, however, gave rise to a perfect classification (first three components). This exemplifies the fact that sequential methods do not insure an optimal solution; indeed some selections might be worse than trivial random ones, as was the case with the above-mentioned subset of three components.

While searching for subsets of two components, it was noticed that the interclass distance measure fell sharply between three and two components. Thus, such subsets were not expected to result in satisfactory performances. Indeed, classification tests demonstrated an error

rate of 44 and 50 percent for both selected subsets, and 60 percent for random selections.

The usefulness of interactive procedures was demonstrated by the discovery of the optimal subset of three components: this happened while trying to keep the simplest invariants, and directing the searches in consequence. It is also noted that, in this experiment as in the other two, the selection procedure based on the arithmetic mean criterion outperformed that based on the Wilks' measure criterion, although it required more time and more attention from the user.

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

The development of two types of sequential feature selection procedures has been presented along with a new inverse matrix updating algorithm. These procedures have been applied to three different classification problems, two of which make use of moments as pattern features.

The first experiment demonstrated the ability of the procedures to find suboptimum—even optimum—measurement subsets. The second experiment verified this property by comparing the selected subsets with randomly chosen ones; it also evidenced the discriminatory power of moments. The third experiment demonstrated the invariant properties of moments and proved the usefulness of interactive procedures.

Further experiments should be performed using known structural information as features, in order to determine the extent to which the interactive capabilities of the procedures improve the selection. Research should also be carried on to reduce the computational errors, especially to establish bounds on the final resulting errors as related to the condition of the covariance matrix.

REFERENCES

REFERENCES

1. Anderson, T. W., Introduction to Multivariate Analysis, John Wiley & Sons, New York, 1958.

2. _____, and R. R. Bahadur, "Classification into Two Multivariate Normal Distributions with Different Covariance Matrices," Annals of Mathematical Statistics, vol. 33, pp. 420-431, 1962.

3. Barrero, A., "Existence of a Moment Transformation," Private Communication.

4. Bartlett, M. S., "An Inverse Matrix Adjustment Arising in Discriminant Analysis," Annals of Mathematical Statistics, vol. 22, pp. 107-111, 1951.

5. Bellman, R. E., Introduction to Matrix Analysis, McGraw-Hill, New York, 1970.

6. Bryant, T. K., "Classification and Clustering Using Density Estimation," Ph.D. Dissertation, University of Missouri, Columbia, 1971.

7. Chang, T. K., and S. T. Dwyer III, "New Approach to Measurement Selection in Pattern Recognition," Proceedings of the Second International Joint Conference on Pattern Recognition, pp. 35-40, 1974.

8. Das Gupta, S., "Non-Singularity of the Sample Covariance Matrix," Sankhya A, vol. 33, pp. 475-478, 1971.

9. _____, "Classification Procedures: A Review," in Discriminant Analysis and Applications, T. Cacoullos, ed., Academic Press, New York, 1973.

10. Dixon, W. J., ed., BMD: Biomedical Computer Programs, University of California Press, Berkeley, 1975.

11. Dudani, S. A., et al., "Aircraft Identification by Moment Invariants," IEEE Transactions on Computers, vol. C-26, no. 1, pp. 39-45, January 1977.

12. Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, vol. 8, pp. 179-188, 1936.

13. Fu, K. S., Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York, 1968.

14. Gonzalez, R. C., and P. A. Wintz, Digital Image Processing, Addison-Wesley, Reading, Mass., 1977.

15. Harris, J. M., "Feature Extraction and Recognition of Two-Dimensional Data by the Method of Moments," M.S. Thesis, University of Tennessee, Knoxville, 1976.

16. Hu, M. K., "Visual Pattern Recognition by Moment Invariants," IRE Transactions on Information Theory, vol. IT-8, pp. 179-187, 1962.

17. Mahalanobis, P. C., "Historical Note on the D^2-Statistic," Sankhyā, vol. 9, pp. 237-240, 1949.

18. Noble, B., Applied Linear Algebra, Prentice-Hall, Englewood Cliffs, N.J., 1969.

19. Patrick, E. A., Fundamentals of Pattern Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1972.

20. Rao, C. R., Linear Statistical Inference and Its Applications, John Wiley & Sons, New York, 1968.

21. Tou, J. T., and R. C. Gonzalez, Pattern Recognition Principles, Addison-Wesley, Reading, Mass., 1974.

22. Wilkinson, J. H., and C. Reinsch, Linear Algebra, Springer Verlag, New York, 1971.

23. Young, T. Y., and T. W. Calvert, Classification, Estimation and Pattern Recognition, American Elsevier, New York, 1974.

APPENDICES

APPENDIX A

CIRCULANT MATRICES AND THE MAHALANOBIS DISTANCE

Circulant matrices have been used in digital image processing [14] to reduce the computational work involved in image restoration. A circulant matrix is such that any of its rows is obtained by a circular shift of the previous row (considering the last row as being before the first):

$$\underline{C} = \begin{bmatrix} C_0 & C_1 & \cdots & C_{N-1} \\ C_{N-1} & C_0 & \cdots & C_{N-2} \\ C_{N-2} & C_{N-1} & \cdots & C_{N-3} \\ \vdots & \vdots & & \vdots \\ C_1 & C_2 & \cdots & C_0 \end{bmatrix} \tag{A-1}$$

These matrices have the remarkable property that they can be diagonalized by a two-dimensional Fourier transformation:

$$\underline{C} = \underline{F}^{-1} \cdot \underline{\Lambda} \cdot \underline{F} \tag{A-2}$$

where $\underline{F}$ denotes the forward Fourier transformation matrix. As Fourier matrices are unitary, $\underline{F}^{-1} = \underline{F}^{H}$ (where $\underline{F}^{H}$ denotes the Hermitean transpose of $\underline{F}$), and equation (A-2) becomes:

$$\underline{C} = \underline{F}^{H} \cdot \underline{\Lambda} \cdot \underline{F} \tag{A-3}$$

Thus the inverse of $\underline{C}$, $\underline{C}^{-1}$, can be written:

$$\underline{C}^{-1} = \underline{F}^{H} \cdot \underline{\Lambda}^{-1} \cdot \underline{F} \tag{A-4}$$

Introducing this in the expression of a quadratic form yields:

$$\underline{m}^{T} \cdot \underline{C}^{-1} \cdot \underline{m} = \underline{m}^{T} \cdot \underline{F}^{H} \cdot \underline{\Lambda}^{-1} \cdot \underline{F} \cdot \underline{m} = (\underline{F} \cdot \underline{m})^{H} \cdot \underline{\Lambda}^{-1} \cdot \underline{F} \cdot \underline{m} \tag{A-5}$$

It is noted that $\underline{F} \cdot \underline{m}$ is the two-dimensional forward Fourier transform of $\underline{m}$; moreover, each diagonal element of $\underline{\Lambda}$ is the one-dimensional backward transform of the corresponding row of $\underline{C}$ [5]. The computation of the quadratic form thus requires a series of Fourier transformations —which can be rapidly performed by an FFT algorithm—and a vector innerproduct, instead of a matrix inversion and a matrix multiplication followed by a vector innerproduct. As there exist dedicated hardware circuits to compute fast Fourier transforms, the gain in computation time over a conventional method can be quite substantial.

However, this approach is not suitable for recursive computations such as used in the procedures developed in this thesis. Moreover, it appears unlikely that covariance matrices present a circulant structure, as this would reduce the number of different elements of the matrix to N/2.

APPENDIX B

PROGRAM DESCRIPTIONS

There are four basic programs. Two of these perform the selection, one is used to invert the appropriate covariance submatrices, and the last classifies the training set using the inverses previously computed. Another program computes the covariance matrices and mean vectors from the training sets; still other programs are used in conjunction with each experiment to prepare training samples. Each program is described in some detail in the following pages. The order in which programs are surveyed corresponds to the order in which they are used in practice, from the preparation of sample patterns to the classification test. It must be kept in mind that these programs have been written for a 36-bit machine with a limited allocation of main memory (65k words). It is finally noted that the execution times given for each program are only indicative, since they are highly dependent on the amount of data to be processed.

B.1. Program PATGEN.FOR

This program is used to generate simulated radar patterns. It reads (from a specified disk file) a specified number of images; each image is a list of coordinate pairs, locating reflection points on a 512 by 512 grid, and ending with a terminator pair. Each image is then symmetrized around vertical and horizontal axes, and around the first bisector, to yield three more images. A random number generator is

used to generate scatter points around each reflection point; the number of scatter points is specified by the user. (The final number may be somewhat smaller, due to the elimination of identical pairs.) The resulting patterns are written to a specified disk file; each pattern is a list of coordinate pairs ending with a terminator. A typical application used 8 original images, each formed of about a dozen reflection points, to generate 64 patterns (as each image is used twice with the random number generator); each pattern consisted of about a thousand coordinate pairs; the CPU time—including the user dialogue—was around 40 seconds (the length is mainly due to the elimination of redundant coordinate pairs). The Gaussian generator is one-dimensional, giving rise to spherical clusters of scatter points around each reflection point.

B.2. Program MOMENT.FOR

This program is used to compute moments from simulated radar patterns. It processes several classes at once. Each class is represented by a disk file in which radar patterns of an aircraft are stored (as created by the program PATGEN.FOR). Central moments are computed, up to a specified order, and normalized by the zero-th order moment. The index function used in the computation is $X = 1 + I/(512 \cdot \alpha)$, where the constant α is specified by the user. When all moments have been computed for a given class, they are written to a specified disk file; each moment vector (i.e., the moments of a pattern) is written in turn, beginning with the low order moments, and varying the order in the

second coordinate faster than that in the first coordinate. A typical application involved 8 classes of 64 patterns each, using moments up to order 7 (i.e., 33 moments) and required a little less than 2 minutes of CPU time.

B.3. Program MOMINV.FOR

This program is used to compute six general moment invariants from digitized pictures. Each picture is a 128 by 128 array of integers in the range 0-255. Four integers are stored in each word in the disk files. The program reads a specified number of pictures, each from a different disk file. Pictures are read in column by column; after each new column has been read, a call is made to a routine, written in DEC-10 assembler code, to unpack the 4 bytes. Each picture is then rotated 90, 180, and 270 degrees, and symmetrized around both vertical and horizontal axes (all of these operations are performed by simple reordering of the picture elements), to yield 5 new pictures. Central normalized moments of order two and three are then computed for each picture, and combined to form the six invariants described in section 6.3. For reasons of storage space, each picture is processed individually and written to disk; this output file is unique and contains 6 moment vectors for each original picture. A typical application used 4 original pictures to yield 24 vectors of six invariants, and required a little more than four minutes of CPU time. (This stems from the necessity of processing $128^2 = 16384$ items of data for each picture.)

B.4. Program STAT.FOR

This program is used to compute the statistics of a training set, that is, the set mean vector and covariance matrix. Several sets can be processed at once. A set consists of a number of pattern vectors written without separation in a disk file. The program will read a set in main memory, compute—in double precision—its mean vector and covariance matrix, and write them in a specified disk file (one for each class). The mean vector is written first, followed by the covariance matrix, in full storage. The double precision is useful in reducing the rounding errors when inverting the covariance matrix. A typical application involved 8 classes, each with 64 patterns of 33 components, to yield 8 mean vectors and 33 by 33 covariance matrices, using less than a minute of CPU time (including user dialogue).

B.5. Program INVERT.FOR

This program is used to compute the inverse of selected covariance submatrices. Several classes can be processed at once; a class is represented by a disk file where class mean vector and covariance matrix are stored in double precision (as created by the program STAT.FOR). The user specifies the components to be used; only the corresponding submatrix is then inverted. The inversion routine uses a triangular decomposition with a column search procedure, as developed by J. K. Bryant [6], in order to minimize the rounding errors. When the submatrix of a class is inverted, the subset of the mean vector and the inverse submatrix are written—in single precision—to a specified file (one

per class). A message is issued if a matrix is found to be singular. No messages are printed if the inverse is detected as not positive-definite; the FORTRAN run-time library, however, will issue a warning stating an attempt to take the square root of a negative number. A typical application used 8 classes of 33 components and required less than 20 seconds of CPU time.

B.6. Program CLASS.FOR

This program is used to run classification tests. The user must specify a number of classes and give, for each class, the name of the disk file containing the training samples, the number of these samples, and the name of the disk file containing the inverse covariance (sub) matrix to be used. There exist two versions of the program: the first, CLASS1.FOR, requires that the number of components to be used be equal to the dimension of the inverse matrix, while the second, CLASS2.FOR, uses a submatrix of the inverse covariance matrix if the number of components to be used is less than the dimension of the inverse matrix. Both programs use only the components that the user specifies. The Mahalanobis distances from each training sample to each class are computed and stored in main memory. The computations, for reasons of storage space, are performed on a class per class basis: the mean vector and inverse covariance matrix of a class are read, and all training samples processed with this class. The programs print an extensive list, class per class, of all distances, with the result of the classification in the margin; in addition, the number of correctly

and incorrectly classified patterns is printed for each class. A typical application involved 8 classes of 64 training patterns each; using all 33 components resulted in a CPU time occupation of a little more than two minutes. The program CLASS1.FOR is intended to be used with the selection programs SELEC1.FOR and SELEC2.FOR, while the program CLASS2.FOR should be used in conjunction with the selection program SELEC3.FOR.

B.7. Program SELEC1.FOR

This program is used to select subsets of measurements using Wilks' matrices and the generalized Mahalanobis interclass distance measure as selection criterion. To initialize the program, the user must specify the number of classes, the total number of components, and, for each class, the name of the file where the mean vector and covariance matrix are stored (as created by the program STAT.FOR) and the number of patterns used to compute these statistics. The program reads, and stores in main memory, all mean vectors, and adds all covariance matrices (weighed by the number of patterns used to compute them) into a single matrix, the total within-class scatter matrix. Once the initialization is complete, the user has a choice of eleven commands at his disposition, namely:

1. Reject all components (partial initialization).
2. Add a specified component.
3. Reject a specified component.
4. Select the next best component(s) (unless its contribution is below the current threshold-to-select).

5. Reject the next poorest component(s) (unless its contribution is above the current threshold-to-reject).
6. Set the threshold-to-select to a specified value (in percent of the total interclass distance measure).
7. Set the threshold-to-reject to a specified value.
8. Print a separation line (to isolate particular sections of the printed output).
9. Type (to the user's terminal) the current status (i.e., number of selected components, value of the interclass distance measure, and both threshold values).
10. Print the status.
11. End of program (return to the monitor).

Each selection, or rejection, is indicated to the user, together with the contribution of the component in cause; this information is also printed. It is thus possible to request a complete feature ordering by starting with no components and requesting the maximum number of steps on the fourth command. Each selection, or rejection is immediately followed by an updating of the interclass distance measure and of the current inverse matrix.

Using 65k words of main memory, this program can deal with problems involving several dozens classes and up to eighty components. Apart from the initialization, which may require a few seconds of CPU time, no command takes more than a tenth of a second of CPU time.

B.8. Program SELEC2.FOR

This program is used to select subsets of measurements using the mean of all interclass distances as selection criterion. To initialize the program, the user must specify the number of classes, the total number of components, and, for each class, the name of the file where the mean vector and covariance matrix are stored (as created by the program STAT.FOR) and the name of the file to which the selected subset's mean vector and inverse covariance matrix are to be written. The program reads, and stores in main memory, all mean vectors and covariance matrices. Once the initialization is complete, the user has a choice of twenty-one commands at his disposition, namely:

1. Print a separation line (to isolate particular sections of the printed output).
2. Reject all components, reset all weights to one (partial initialization).
3. Add a specified component.
4. Reject a specified component.
5. Select the next best component(s), using the average of all weighted interclass distances as criterion (unless the resulting contribution is below the current threshold-to-select).
6. Reject the next poorest component(s), using the same criterion as under 5 (unless the resulting contribution is above the current threshold-to-reject).
7. Select the next best component(s), using the average of the weighted distances from a specified class to all others as

criterion (unless the resulting contribution is below the current threshold-to-select).

8. Reject the next poorest component(s), using the same criterion as under 7 (unless the resulting contribution is above the current threshold-to-reject).
9. Select the next best component(s), using the distance from one specified class to another specified class as criterion (unless the resulting contribution is below the current threshold-to-select).
10. Reject the next poorest component(s), using the same criterion as under 9 (unless the resulting contribution is above the current threshold-to-reject).
11. Type (to the user's terminal) the distances between two specified classes.
12. Type (to the user's terminal) the list of all selected components.
13. Type (to the user's terminal) the current short status (i.e., the number of selected components, the unweighted average interclass distance, its standard deviation, the minimum interclass distance, and both threshold values).
14. Print the current short status, plus the list of all selected components.
15. Print the current short status, plus the list of all selected components, plus the list of all interclass distances and their respective weights.

16. Set the threshold-to-select to a specified value (an absolute measure).
17. Set the threshold-to-reject to a specified value.
18. Assign a specified weight to the distance from a specified class to another specified class.
19. Type (to the user's terminal) the weight of the distance from a specified class to another specified class.
20. Type (to the user's terminal) the minimum interclass distance—in harmonic mean—and both class numbers.
21. Write (to the disk files specified during the initialization) the current inverse covariance matrices to disk and return to the monitor (end of program).

Each selection, or rejection, is indicated to the user, together with the weighted contribution of the component in cause; this information is also printed. Each selection, or rejection, is immediately followed by an updating of all interclass distances and all current inverse matrices.

Using 65k words of main memory, this program can deal with problems involving, for instance, eight classes of 33 components. Apart from the initialization, which may require a few seconds of CPU time, no command takes more than a tenth of a second of CPU time.

B.9. Program SELEC3.FOR

This program is used to select subsets of measurements using the mean of all interclass distances as selection criterion. It is basically identical to the program SELEC2.FOR, except that it uses subsets of

inverse matrices in the computations. In consequence, it requires as input one file per class, where the mean vector and inverse covariance matrix are stored (as created by the program INVERT.FOR). For more details, the listing of the program should be consulted.

VITA

Bernard Marcel Edmond Moret was born in Vevey, Switzerland, on September 13, 1953. He attended elementary schools in Vevey and Lausanne, Switzerland, and received his Baccalaureate in Latin, Greek, and Mathematics in July 1970. The following September, he entered the Swiss Federal Institute of Technology at Lausanne, which awarded him the Diploma of Electrical Engineering in January 1975.

In January 1975, Mr. Moret went to work with the Omega Company, Bienne, Switzerland, where he was in charge of the software development for sports timing and public address applications. He resigned after the Games of the XXI Olympiad and accepted a research assistantship at the University of Tennessee, Knoxville. He was conferred a Master of Science degree, with a major in Electrical Engineering, in August 1977.