AFCRL-70-0540

# ANALYTICAL AND INTERACTIVE TECHNIQUES FOR MULTIVARIATE DATA COMPRESSION AND CLASSIFICATION

Robert B. Roper

Systems Research Laboratories, Inc.
7001 Indian Ripple Road
Dayton, Ohio 45440

Contract No. F19628-67-C-0150
Project No. 5632
Task No. 563201
Unit No. 56320101

FINAL REPORT

Period Covered: 1 December 1966 - 31 May 1970

25 September 1970

Contract Monitor: Charlton M. Walter, LRM



Prepared
for
AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS 01730

AFCRL-70-0540

# ANALYTICAL AND INTERACTIVE TECHNIQUES FOR MULTIVARIATE DATA COMPRESSION AND CLASSIFICATION

Robert B. Roper

Systems Research Laboratories, Inc.
7001 Indian Ripple Road
Dayton, Ohio 45440

Contract No. F19628-67-C-0150
Project No. 5632
Task No. 563201
Unit No. 56320101

FINAL REPORT

Period Covered: 1 December 1966 - 31 May 1970

25 September 1970

Contract Monitor: Charlton M. Walter, LRM



Prepared
for
AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS 01730

ABSTRACT

This report demonstrates the applicability of classical statistical techniques to problems involving compression and classification of multivariate data. The theoretical foundations of two such techniques, intrinsic analysis and discriminant analysis, are treated in detail. Efficient digital computer implementation is discussed, including the combined application of intrinsic and discriminant analysis and a new algorithm for computing approximate intrinsic bases for very large problems. Experimental results are presented on the application of these techniques as feature extractors in a signal classification problem. Also included is a description of the interactive graphics-oriented system software which has been developed to facilitate the application of these techniques.

## FORWARD

This report was prepared by Systems Research Laboratories, Inc., 7001 Indian Ripple Road, Dayton, Ohio, under Contract No. F19628-67-C-0150, for the Air Force Cambridge Research Laboratories, Air Force Systems Command, United States Air Force, L. G. Hanscom Field, Bedford, Massachusetts, in support of Project 5632, Task 563201. The research was initiated 1 December 1966. This report covers the period through 31 May 1970. The technical monitor was Charlton M. Walter, Multisensor Signal Processing Branch, Data Sciences Laboratory (LRM).

The author wishes to acknowledge the major contributions of Edward N. Chase, Richard J. Daesen and Donald A. Stremsky of AFCRL toward implementation of the system software and applications programs described in Section VI.

Previous publications under this contract include: "Approximate Eigensystems of Large Covariance Matrices," R. Roper, 1 May 1968; Choice of Axes for Projection of Data on a CRT Using Multivariate Analysis of Variance," R. Colomb, 1 May 1968; "Implementation of Lie Group Transforms of Spatial Data on Parallel Logic Arrays," R. Colomb, 1 August 1968; "Effects of Computing Discriminants in a Truncated Intrinsic Basis," R. Colomb, 1 September 1968; "AMP Macros for Algebraic Computations: AMAC User's Manual," R. Roper, 1 February 1969; "The DX-1 Disk File System," R. Roper, 5 September 1969; and "Assembler Macros for Algebraic Computations: AMAC User's Manual," R. Roper, E. Chase and R. Daesen, 30 January 1970.

TABLE OF CONTENTS

## Section I

## INTRODUCTION

The research effort reported herein has been directed toward the implementation of an interactive, graphics oriented computer system for the representation, analysis and classification of multivariate data. Work has proceeded in two parallel areas: development of the appropriate analytical techniques for, and design of, a software system tailored to the requirements of such a facility. The ultimate purpose of the system is to provide a tool for the systems engineer or scientist dealing with large scale problems requiring reduction and/or classification of high dimensional data, by supplying a means of evaluating the effectiveness (or lack of effectiveness) of proposed approaches to his specific problems. The system can also be used to investigate the interrelationships among standard analytical techniques and to develop new data analysis methods.

The typical application deals with sets of data whose members are measurement vectors, for example, simultaneous outputs of a bank of sensors or discrete time samples of a continuous function. It is easy to display these vectors component by component, but this reveals little information about the overall statistical properties of the random processes from which they have been sampled. Therefore it is desirable to find two-dimensional representations for the measurement space in which the members of entire data sets appear as projected points. If the coordinates of this representation are chosen judiciously, the resulting projection may yield valuable insight into the statistical relationships among the data elements.

One of the major goals of this effort has been the development of analytical methods for selecting such coordinates. These include means for efficient representation of data sets with highly redundant measurements (intrinsic analysis, Section II) and for viewing the separability of several distinct data sets (discriminant analysis, Section III). Classification problems require automatic pattern recognition algorithms. The pattern classification problem and several specific methods are discussed in Section IV.

The primary consideration in the selection of all of these methods is that they have a sound mathematical formulation. This is to assure that the resulting system is sufficiently general to be applicable to a wide range of problems, and to facilitate the analysis of its performance. We, therefore, have avoided methods requiring interactive human supervision in their execution. Human interaction is generally restricted to the selection of the sequence of processes, with their parameters, to be applied to the data and control of the interactive display programs.

By applying sequences of elementary processes and observing the results at each stage, the user may develop compound procedures appropriate to his application. To emphasize the value of this building block approach, this report stresses the interrelationships among the analytical techniques. An example involving the application of information compression methods as feature extractors to improve the performance of a pattern classification algorithm is presented in Section V.

The computer software which has been designed to aid in the implementation of this system is described in Section VI. This includes a disk file system, an on-line monitor, an interactive vector projection display program and an extended macro assembly language, in addition to the mathematical routines.

The mathematical developments which follow are descriptive enough for the general reader with limited mathematical background to understand the underlying concepts, although familiarity with probability theory and matrix algebra is desirable. Extensive use is made of the concept of a random vector, which is a vector whose components are random variables. No notational convention is used to distinguish scalars, vectors and matrices; the distinctions should be clear from context. Vectors are always column vectors; transposes of vectors are always row vectors. Specific notations are defined as needed in the text.

## Section II

## INFORMATION COMPRESSION

A major problem in many data analysis and classification problems, as well as data transmission applications, is the high dimensionality of the data. Data vectors arising from sampled continuous signals or their Fourier transforms, for example, typically contain hundreds or thousands of sample points. All but the most straightforward data analysis and pattern recognition techniques tend to bog down in numerical computations or become ineffective when dealing with such large problems. One way to alleviate such problems is to find a more compact representation for the data which preserves as much as possible of its original information content. We shall refer to this process as information compression or dimension reduction.

The canonical data representation to be developed here is similar to Fourier analysis in that its components are inner products of the data with members of an orthogonal function set. However, the form of the orthogonal functions is not restricted to sinusoids. Thus it may be thought of as generalized harmonic analysis. It also has the desirable property that its components are uncorrelated.

Essentially the same technique has been developed by many authors in several disciplines. In multivariate statistics (Wilks[1] and Anderson[2]) it is referred to as principal components analysis. It was applied by Kramer and Mathews[3] to speech bandwidth compression by encoding the output of a channel vocoder. The term intrinsic analysis is due to Young and Huggins[4] and is also used by Walter[5] and Colomb.[6] In communication theory (Davenport and Root[7]) and probability theory (Loève[8]) it appears as Loève-Karhunen analysis. The technique is equally applicable to continuous (real or complex-valued) functions or vectors. For our purposes, the vector formulation is more convenient, and will be developed here. The extension to continuous functions is routine (see Colomb[6] or Watanabe[9]).

The approximate data representations obtained through intrinsic

analysis are optimal in the least mean square error sense. It develops that they are also optimal in the sense of minimizing an entropy function defined on the coefficients. This is shown by Watanabe,[9] who relates these two properties in the context of a pattern clustering and recognition problem. An instructive proof of the error minimization property is given by Anderson.[2] Mostly for heuristic purposes, we offer here a simplified version which relies more heavily on algebraic eigenvalue theory. We will also relate Watanabe's results on entropy minimization.

1. Intrinsic Analysis

Let $X$ be a random vector in the n-dimensional real vector space $V$ with probability density function $p(X)$. If $F(X)$ is a function of $X$, the expectation of $F(X)$ is defined by

$$E\,F(X) = \int_V F(X)\, p(X)\, dX$$

The mean vector $\mu$ of $X$ is

$$\mu = E\,X$$

and the autocorrelation matrix $A$ of $X$ is

$$A = E\,X\,X'$$

where prime denotes transpose. $A$ is symmetric ($A(i,j) = A(j,i)$) and the element $A(i,j) = E\,X(i)\,X(j)$ is the correlation of the i-th and j-th elements of $X$. The euclidian norm, or length, or a vector $v$ in $V$ is

$$||v|| = \left( \sum_{j=1}^{n} v\,(j)^2 \right)^{\frac{1}{2}}$$

We define the energy of $X$ as the expected value of the squared norm of $X$, namely

$$E(X) = E\,||X||^2 = E \sum_{j=1}^{n} X(j)^2$$

Note that the energy of $X$ is equal to the trace of $A$:

$$E(X) = \sum_{j=1}^{n} E(X(j))^2 = \sum_{j=1}^{n} A_{jj} = \text{tr } A$$

Our approach to the dimension reduction problem is to find, for any $k \leq n$, a k-dimensional subspace $V_k$ of $V$ which maximizes the energy of the approximation of $X$ by projection onto $V_k$. It is sufficient to find $k$ orthonormal vectors $\{\phi_i | i=1,\ldots,k\}$ which span $V_k$. By orthonormal, we mean that

$$\phi_i' \phi_j = 0 \quad \text{for } i \neq j$$

$$\phi_i' \phi_i = 1 \quad \text{for } i=1,\ldots,k$$

The coordinates of $X$ in $V_k$ are the projections of $X$ on the $\phi_i$

$$c_i = \phi_i' X$$

and the approximation of $X$ in the standard basis of $V$ is

$$\hat{X}_k = \sum_{i=1}^{k} c_i \phi_i$$

which permits reconstruction of the n-space representation of the approximation.

We define the relevance $o_i$ of $\phi_i$ in representing $X$ as the mean squared projection of $X$ on $\phi_j$

$$o_i = E(\phi_i' X)^2 = E(c_i^2)$$

Let $\Phi_k$ be the matrix whose columns are the $\phi_i$. Then the projection from $V$ onto the $\Phi_k$ basis of $V_k$ is given by the standard change of basis transformation

$$\hat{X}_k = \Phi_k' X$$

and the energy of the desired optimal approximation is

$$E(\hat{X}_k) = E\| \Phi_k' X \|^2 = E \sum_{i=1}^{k} (\phi_i' X)^2 = \sum_{i=1}^{k} \rho_i$$

So maximizing the energy of approximation in a subspace is equivalent to maximizing the sum of the relevances of the basis vectors.

When $k = 1$, with $\phi_1$ denoted by $\phi$, the problem is to maximize

$$\rho_1 = E(\phi'X)(X'\phi) = \phi' E(XX') \phi = \phi' A\phi$$

which is the quadratic form associated with $A$, the autocorrelation matrix of $X$. So the normal vector $\phi$ which maximizes $\rho$ must maximize the quadratic form of $A$ subject to the normality constraint $\phi' \phi - 1 = 0$. Using the Lagrange multiplier $\lambda$, define

$$\alpha = \phi' A\phi - \lambda (\phi' \phi - 1)$$

The vector of partial derivatives of $\alpha$ w.r.t. the components of $\phi$ is, according to Anderson,[2] p. 347,

$$\frac{\partial \alpha}{\partial \phi} = 2A\phi - 2\lambda\phi$$

If $\phi$ maximizes $\phi'A\phi$ with $\phi'\phi - 1 = 0$, $\partial\alpha/\partial\phi$ must be zero, which yields

$$A\phi = \lambda\phi$$

the algebraic eigenvalue equation. Since $A$ is an autocorrelation matrix, it is positive semidefinite, that is, all of its eigenvalues are nonnegative. Premultiplying by $\phi'$, we have $\phi'A\phi = \phi'\lambda\phi = \lambda$, which implies that $\rho_1 = \lambda$. So the maximum possible relevance to $X$ of a normal basis vector is $\lambda_1$, the largest eigenvalue of $A$ and that vector is the eigenvector $\phi_1$ of $A$ corresponding to $\lambda_1$.

We extend this result to arbitrary $k$ by induction. Having determined $\phi_1,\ldots,\phi_{i-1}$ and $E(\hat{X}_{i-1})$, we seek the optimal orthonormal basis for $V_i$ and its energy $E(\hat{X}_i)$. This basis must include $\phi_1,\ldots,\phi_{i-1}$, for otherwise $E(\hat{X}_i)$ determined by the new basis could be improved by substituting

the missing $\phi$'s for members of the new basis orthogonal to the $\phi$'s contained in $V_{i-1}$, since the old basis is optimal for $k = i-1$. So the problem is to maximize the additional energy $E(\hat{X}_i) - E(\hat{X}_{i-1})$ obtained by adding one more basis vector. As before, this is $\rho_i = \phi_i' A \phi_i$, which leads to the eigenvalue equation. Therefore, the maximum additional energy is obtained by projection on the eigenvector $\phi$ of $A$ corresponding to the largest remaining eigenvalue $\lambda_i$ of $A$, and

$$E(\hat{X}_i) - E(\hat{X}_{i-1}) = \rho_i = \lambda_i$$

We conclude that the $k$-dimensional subspace $V_k$ of $V$ which contains the greatest fraction of the energy of $X$ is spanned by the eigenvectors $\phi_1,\ldots,\phi_k$ of the autocorrelation matrix $A$ of $X$ corresponding to the $k$ largest eigenvalues $\lambda_1,\ldots,\lambda_k$. The $\phi_i$ are called the <u>intrinsic basis vectors</u> of $X$. The energy of the approximation in $V_k$ is

$$E(\hat{X}_k) = \sum_{j=1}^{k} \rho_i = \sum_{j=1}^{k} \lambda_k$$

where $\rho_j$ is the relevance of $\phi_j$ to $X$. A convenient measure of the performance of the intrinsic analysis is the fraction of the energy retained, which is

$$\frac{E(\hat{X}_k)}{E(X)} = \frac{\sum_{j=1}^{k} \lambda_i}{\mathrm{tr}A}$$

The above treatment is valid if $\lambda_i \neq \lambda_j$ whenever $i \neq j$ and $\lambda_i = 0$, $i = 1,\ldots,m$. Equal eigenvalues determine an eigensubspace of the corresponding dimension and within this subspace, the orthonormal intrinsic basis vectors may be chosen arbitrarily. Similarly, zero eigenvalues determine a subspace orthogonal to the rest of the eigenspace. For a detailed treatment of these special cases, see Anderson.[2]

Maximization of the energy in the intrinsic basis approximation is equivalent to minimization of the mean square error of the approximation. To see this, we observe that when $k = n$, the intrinsic basis approximation is

exact, so $||v_n|| = ||v||$. The mean square error (the average squared norm of the error vector) is then

$$\epsilon(\hat{X}_k) = E||\hat{X}_n - \hat{X}_k||^2$$

$$= E\sum_{j=k+1}^{n}(\phi_j'X)^2$$

$$= E\left(\sum_{j=1}^{n}(\phi_j'X)^2 - \sum_{j=1}^{k}(\phi_j'X)^2\right)$$

$$= E||\Phi_n'X|| - E||\Phi_k'X||^2$$

$$= E(X) - E(\hat{X}_k)$$

Expressing the approximation energy in terms of the eigenvalues, the error becomes

$$\epsilon(\hat{X}_k) = E(X) - \sum_{j=1}^{k}\lambda_j = trA - \sum_{j=1}^{k}\lambda_j$$

and as a fraction of total energy, the error is

$$\frac{\epsilon(\hat{X}_k)}{E(X)} = 1 - \frac{\sum_{j=1}^{k}\lambda_j}{tr\ A}$$

Intrinsic analysis corresponds very closely to a method of multivariate statistics called principal components analysis. The relationship between these two techniques is as follows: if the mean $u$ of $X$ is nonzero, the first intrinsic basis vector $\phi_1$ may tend to resemble it. If $X$ is symmetrically distributed about $u \neq 0$, then $\phi_1$ is a scalar multiple of $u$. On the other hand, it is noted by Colomb,[6] p. 13, that there exist distributions with zero mean having the same covariance matrix as a distribution with an arbitrary mean $u$. In practical situations, however, we may expect that if $u$ is large, $\phi_1$ will tend to have high correlations with most of the elements of $X$ and, therefore, have a strong similarity to $u$. In

problems where interest is focused on covariances among the elements of $X$ rather than on cross correlations, it is desirable to eliminate this effect of the mean vector by performing the intrinsic analysis on the variable $X-\mu$, which has zero mean. This is called principal components analysis. The autocorrelation matrix $\Sigma$ of $X-\mu$ is called the covariance matrix of $X$. A diagonal element of this matrix, $E(X(i) - \mu(i))^2$ is the variance of $X(i)$; a non-diagonal element, $E(X(i) - \mu(i))(X(j) - \mu(j))$ is the covariance of $X(i)$ and $X(j)$. With this change, references to energy in the development for intrinsic analysis may be read as variance, which is the energy of $X-\mu$. The eigenvectors of $\Sigma$ are called the principal components of $X$ because, properly ordered, they account for the "principal components" of the variance of $X$ about $\mu$.

At the level of machine calculations, these techniques are easily interchangeable because the autocorrelation and mean of $X$ determine its covariance by the following relation

$$\begin{aligned}\Sigma &= E(X-\mu)(X-\mu)' \\ &= E(XX' - X\mu' - \mu X' + \mu\mu') \\ &= E\,XX' - E\,X\mu' - \mu E X' + \mu\mu' \\ &= A - \mu\mu'\end{aligned}$$

Watanabe[9] has shown that intrinsic analysis satisfies another consideration in selecting a basis for information compression. This is that the relevance measures $c_i$ should be highly concentrated on a few of the basis vectors rather than spread out more evenly. If $\psi_1,\ldots,\psi_n$ is any orthonormal basis of $V$ and we normalize $X$ so that $E(X) = 1$, then $c_i$ is a probability measure on $\{\psi_i\}$ with

$$c_i \geq 0 \quad , \quad \sum_{i=1}^{n} c_i = 1$$

It is then possible to introduce the entropy function

$$H(\Psi) = -\sum_{i=1}^{n} c_i \log c_i$$

which is a measure of the concentration of the $\rho$'s. Watanabe demonstrates that the intrinsic basis minimizes the entropy function as well as the mean square error.

## 2. Estimates of the Intrinsic Basis

In practice, the mean $\mu$ and autocorrelation $A$ of the random vector $X$ are seldom available. It is then necessary to estimate them from a finite sample $\{x_1,\ldots,x_m\}$ of $X$. The estimates are simple averages involving the sample vectors

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

$$\hat{A} = \frac{1}{m} \sum_{i=1}^{m} x_i x_i'$$

Estimates of the intrinsic basis vectors $\hat{\phi}_1,\ldots,\hat{\phi}_n$ and their relevances $\hat{\lambda}_1,\ldots,\hat{\lambda}_n$ are given by the eigensystem of $\hat{A}$ (or of $\hat{\Sigma} = \hat{A} - \hat{\mu}\hat{\mu}'$ in the case of principal components analysis). Anderson,[2] p. 279, shows that if the distribution of $X$ is multivariate normal, then this process defines the maximum likelihood estimates of $\phi_1,\ldots,\phi_n$ and $\lambda_1,\ldots,\lambda_n$.

In real problems, the distributions involved are usually not multivariate normal and are frequently unknown. However, it is demonstrated by K. Miller in Ref. 6, p. 7, that $\hat{A}$ (or $\hat{\Sigma}$) is the minimum variance linear unbiased estimate of $A$ (or $\Sigma$). Since the eigenvalues and eigenvectors are highly nonlinear functions of $\hat{A}$, it is difficult to infer from this the variance of the errors in the estimated eigensystems. Here we will merely mention some conclusions of a detailed discussion of the problem by Colomb,[6] pp. 18-24. First, the eigenvalues are stable with respect to perturbations of $\hat{A}$ in that, under certain conditions, their variances are approximately equal to the variances of the diagonal elements of the error matrix $A-\hat{A}$. Second, the eigenvectors are very stable if their eigenvalues are well separated "and as the eigenvalues get close together . . . the stability decreases until, when the separation of the values is of the order of the perturbation, no certain information is attained about any individual eigenvector." In the specific problems studied thus far, the largest eigenvalues are most separated while eigenvalues belonging to low-energy subspaces tend

to cluster together. However, if the best eigenvectors collectively are used to represent X (which is the typical case) the errors in the eigenvectors resulting from confusion with other eigenvectors being used are not harmful. Thus, the accuracy of the resulting representation may be better than that indicated by consideration of the eigenvectors separately. Finally it is noted that the ambiguity in $\hat{A}$ usually decreases as the square root of the number of samples increases.

Errors in the intrinsic basis can also result from measurement and computational errors. For purposes of error analysis, these can be treated as part of the ambiguity of $\hat{A}$. While measurement errors (sensor additive noise and quantization noise) are unavoidable, they are likely to be less significant than the statistical sampling error already mentioned. In implementations using a general purpose digital computer, the computational error may be reduced to insignificance by using an accurate eigensystem algorithm such as Householder's Method (see Wilkinson,[10] pp. 290-335).

## 3. Approximations for Large Problems

In many information compression problems, the dimension n of the sample data is in the hundreds or thousands. Such high dimensions introduce large costs in terms of both computation time and computer memory. The time required for estimation of the autocorrelation matrix increases with $n^2$ and for calculation of its eigensystem roughly with $n^3$. More serious, perhaps, are the random access memory requirements of these processes. Unless the efficiency of the computations is drastically reduced, it is necessary to retain the entire matrix in memory simultaneously. Since it is symmetric, this means that at least $n(n+1)/2$ locations are required. Thus if there are S available storage locations, the maximum dimension which can be handled is less than $\sqrt{2S}$. For problems too large for existing hardware, it is sometimes possible to obtain useful approximations of the estimated intrinsic basis by the procedure described by Roper.[11] The rest of this section is devoted to a summary of this method.

The idea behind the approximation is to break a large eigensystem problem up into several small ones and use their solutions to reduce the size of the problem; this is equivalent to a piecewise dimension reduction by intrinsic analysis. We denote the n-dimensional symmetric matrix in question (in this case, the estimated autocorrelation matrix) by M, and its

true eigensystem by $(\Lambda,\Psi)$. We partition $M$ symmetrically into $p^2$ submatrices $M_{ij}$, and find the eigensystems $(\Lambda_i,\Phi_i)$ of the $p$ diagonal submatrices $M_{ii}$. Next we discard all but the $k_i$ most relevant eigenvectors from each $\Phi_i$ ($\sum_{i=1}^{p} k_i = k \leq n$), and form the $n \times k$ matrix $\Phi$ with the $\Phi_i$ as submatrices along the diagonal and zeros elsewhere. Note that $\Phi$ inherits the orthonormality of the $\Phi_i$. Therefore, $\Phi$ can be used to transform $M$ into a $k \times k$ dimensional approximation

$$\hat{M} = \Phi' M \Phi$$

whose eigensystem, given by $\hat{M}\,\Theta = \Theta\,\hat{\Lambda}$, satisfies

$$\Theta'\,\hat{M}\,\Theta = \hat{\Lambda}$$

Combining these, we have

$$\Theta'\,\Phi'\,M\,\Phi\,\Theta = \hat{\Lambda}$$

If we define

$$\hat{\Psi} = \Phi\,\Theta$$

the above relation becomes

$$\hat{\Psi}'\,M\,\hat{\Psi} = \hat{\Lambda}$$

and $(\hat{\Lambda},\ \hat{\Psi})$ is the approximation of the first $k$ members of $(\Lambda,\ \Psi)$. It is easy to show that the columns of $\hat{\Psi}$ are orthonormal, so if $M$ is an estimated autocorrelation matrix, the above expression implies that the relevances of the approximate eigenvectors (to the energy of the sample vectors) are equal to the corresponding approximate eigenvalues. This fact can help determine the value of such approximations in dimension reduction problems, but unfortunately it cannot tell us how far we are from the optimal representation. To minimize this deviation, it is advisable to make $k$ as large as possible even though fewer components may be used in the final representation.

One further refinement is possible (and necessary if M does not fit in memory). In addition to speeding up the eigensystem calculations, it allows us to reduce the time required for computation of M. This is done by computing directly only the $M_{11}$. The $\hat{M}_{1j}$ are then computed from data vectors in the reduced $\Phi$ basis. The savings realized are substantial, especially when the number of sample vectors is large. Details on implementation and further discussion of the reductions in computation time may be found in Roper.[11]

## Section III

## DISCRIMINATION AMONG SAMPLE CLASSES

In the previous section we were concerned with samples of a single random vector. Here we will consider several distinct classes of samples drawn from random vectors with different multivariate distribution functions.

The methods described here are motivated by the requirements of a graphics-oriented data analysis facility. In the typical data reduction and classification problem, the scientist or systems engineer is concerned with whether the available measurements (components of the sample vectors) are adequate to determine class membership, and with which class the measurements are most useful. For projection on a computer display, a two-dimensional subspace of the measurement space must be selected. Coordinate projections are not very useful since they ignore the information in all but two of the measurements. Intrinsic basis representations embody information from all or most of the measurements, but make no use of class membership information. The alternative suggested here is a low-dimensional representation for the sample vectors which tends to cluster together samples within each individual class while emphasizing the variations among all the classes. Such a projection gives the analyst the optional two-dimensional linear projection of his classification problem which can yield insights into its statistical characteristics, for example, the degree of linear separability of the sample classes.

One further motivation for using such representations lies in the implementation of automatic pattern classification algorithms. The performance of most algorithms is improved if the pattern vectors are first subjected to a transformation which clusters samples within the same class.

This section begins with a derivation of the discriminant analysis technique. There follows a discussion of estimation and computation of the discriminants. A serious computational problem arises when the number of sample vectors is small relative to their dimension. A way of circumventing this problem by prior application of intrinsic analysis is suggested.

1. Discriminant Analysis

Discriminant analysis, like principal components analysis, is a technique of multivariate analysis of variance. It is developed by Hotelling[12] and Wilks.[1] A detailed exposition, with applications to the data display problem, is offered by Colomb.[13] As in Section II, our exposition will begin by treating random vectors. Then, with the theoretical groundwork established, it will be expanded to include finite samples of them.

Let $X_1,\ldots,X_r$ be n-dimensional random vectors with mean vectors $\mu_1,\ldots,\mu_r$ and covariance matrices $W_1,\ldots,W_r$. Each random vector $X_i$ is identified with a sample class $C_i$. The within-class covariance matrix, which describes covariance about class means, is obtained by averaging the covariance matrices of the classes

$$W = \frac{1}{r} \sum_{i=1}^{r} W_i$$

the grand mean is the average of the class means

$$\mu = \frac{1}{r} \sum_{i=1}^{r} \mu_i$$

and the among-classes covariance matrix, which describes covariance of the class means about the grand mean, is

$$A = \frac{1}{r} \sum_{i=1}^{r} \mu_i \mu_i' - \mu\mu'$$

Principal components analysis produces vectors, or directions, in which the covariance of a random vector is maximized. Discriminant analysis finds vectors which maximize the among groups covariance while minimizing the within groups covariance. From the discussion of intrinsic analysis, we recall that the relevance of a vector $\phi$ to the covariance described by $\Sigma$ is $\phi'\Sigma\phi$. Thus we want to find a discriminant vector $d$ which maximizes

$$\frac{\sigma_A^2}{\sigma_W^2} = \frac{d'Ad}{d'Wd}$$

the ratio of the relevances, or variances in the direction of d. This can be accomplished by maximizing $\sigma_A^2$ with $\sigma_W^2$ held constant. Again using $\lambda$ to denote a Lagrange multiplier, a necessary condition for this maximization is

$$\frac{\partial}{\partial d}\left[d'Ad - \lambda(d'Wd - \text{constant})\right] = 0$$

$$2\,Ad - \lambda\,2\,Wd = 0$$

or

$$Ad = \lambda Wd$$

This is the generalized algebraic eigenvalue equation. The number of distinct, non-trivial solutions (discriminant vectors) is equal to the rank of A, which is r-1 (the number of classes minus one). The subspace spanned by $d_1,\ldots,d_{r-1}$ is called the discriminant space. (Since the discriminant vectors are not necessarily orthonormal, it is desirable, in practice, to orthogonalize them to obtain an orthonormal basis for the discriminant space.) Finally, we note that since we can pre-multiply the above by $d'$ to obtain $d'Ad = d'\lambda Wd$, we have

$$\lambda = \frac{d'Ad}{d'Wd} = \frac{\sigma_A^2}{\sigma_W^2}$$

Therefore, the eigenvalues indicate the ratio of the among-class to the within-class variances of projections on the corresponding discriminants.

## 2. Discriminant Computations

One of the advantages of intrinsic analysis is its ability to reduce a high dimensional set of data with redundant measurements to a lower dimensional representation. Provided that the number of samples available be sufficient to provide a reasonable estimate of the intrinsic basis. The situation is quite different with discriminant analysis. As we shall see, the discriminant computations are impossible if the number of samples is less

than their dimension.

Suppose the random vectors $X_i$, i=1,..,r are represented by sets of sample vectors

$$\{x_{ij} | j=1,\ldots,m_i\} \quad \text{were } m = \sum_{i=1}^{r} m_i$$

the means and covariance matrices of the sample classes are estimated by

$$\mu_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij} \; ; \qquad W_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij} x'_{ij} - \mu_i \mu_i'$$

The grand mean and among-groups covariance matrix are estimated by

$$\mu = \frac{1}{m} \sum_{i=1}^{r} m_i \mu_i \; ; \qquad A = \frac{1}{m} \sum_{i=1}^{r} m_i \mu_i \mu_i' - \mu\mu'$$

and the within-groups covariance matrix by

$$W = \frac{1}{m} \sum_{i=1}^{r} m_i W_i$$

Estimates of the discriminants are the solutions of generalized eigenvalue equation

$$Ad = \delta \, Wd$$

which may be reduced to the ordinary eigenvalue equation by a fast and accurate process given by Wilkinson,[10] pp. 337-40. For details of the computations and a discussion of the numerical errors involved, the reader is referred to Colomb,[13] pp. 18-20. The fact which concerns us here is that the reduction requires what amounts to an inversion of W. The rank of W is less than or equal to m-r. So if

$$m < n + r$$

where $n$ is the dimension of the samples, $W$ is singular and its inverse does not exist. Streeter and Raviv, [14] p. 16, have experimentally evaluated three different means of avoiding this singularity problem, including the Moore-Penrose generalized inverse and the 'H' inverse of T. Harley. Both of these approaches introduce artificial constraints on the solution and are cumbersome to implement. The third approach, suggested by Streeter and Raviv, gave the best results. Their idea is to use intrinsic analysis to reduce the dimension of the samples so that the discriminants may be computed.

It has been found convenient to use principal components analysis rather than intrinsic analysis. The principal components $\phi_1,\ldots,\phi_k$ are eigenvectors of the total covariance matrix of all the samples about the grand mean

$$T = \frac{1}{m} \sum_{i=1}^{r} \sum_{j=1}^{m_i} x_{ij} x_{ij}' - uu'$$

It can easily be shown that

$$T = A + W$$

that is, the total covariance equals the among-classes covariance plus the within-class covariance.

The samples are represented using the first $k$ principal components as

$$\hat{x}_{ij} = \phi_k' x_{ij}$$

where

$$\phi_k' T \phi_k = \Lambda_k$$

the approximate covariance matrices $\hat{A}$ and $\hat{W}$ are then computed as before using the $\hat{x}_{ij}$, and the new discriminant vectors are the solutions of

$$\hat{A} d = \hat{\lambda} \hat{W} d$$

Even when it is not mathematically necessary, this approximation may be valuable in reducing the storage requirements for discriminant analysis, which are roughly twice those of principal components analysis, or because of the time savings resulting from the lowered dimension of the discriminant problem. Those savings are detailed by Colomb,[13] pp. 20-23, who shows how to obtain further savings by taking advantage of the fact that

$$\hat{W} = \Lambda_k - \hat{A}$$

This follows because $\hat{A} + \hat{W} = \hat{T}$, which, in the principal components basis, is $\Lambda_k$. This makes it possible to avoid direct calculation of $\hat{W}$ and even the transformation of the sample vectors, since $\hat{A}$ can be obtained by transforming only the class means.

The principal components approximation can also be useful in improving the accuracy of the discriminant vectors. Even when $W$ is not singular, it may be very ill-conditioned. (The condition number of a symmetric matrix is the ratio of its largest and smallest eigenvalues.) An ill-conditioned $W$ may introduce instabilities into the inversion process and errors in the resulting discriminant vectors and values. The discriminant equation may be rewritten

$$Ad = \delta\,(T-A)d$$

or

$$Ad = \frac{\delta}{1-\delta}\,Td$$

Solution of this form requires inversion of $T$, but in the principal components basis, $\hat{T} = \Lambda_k$ is diagonal. Thus, by an appropriate choice of the reduced dimension $k$, we may ensure that $\hat{T}$ has any required condition number. The only remaining question is that of the optimal choice of $k$. This problem is treated in great detail by Colomb.[15]

Section IV

## METHODS OF DATA CLASSIFICATION

In many instances the ultimate goal of the information compression and discrimination techniques described above is the efficient implementation of a procedure for data classification or sorting. We are concerned here with classification of observations into one of several previously known categories. The pattern recognition problem has been formulated in several disciplines, including information theory, switching theory and control theory, and an informative survey of the field is offered by Nagy.[16] We will review here the usual formulation in terms of statistical decision theory. Works of general interest in this area include Sebestyen,[17] Highleyman[18] and Nilsson.[19] The treatment by Nilsson is most convenient and will be drawn on heavily here. We will discuss the implementation of pattern classifiers using discriminant functions, some optimal methods for normally distributed patterns, and the problem of estimating unknown multivariate probability density functions (called densities, for convenience) from finite training samples. We will review several approaches to this problem and discuss the most promising in some detail.

### 1. Bayes Optimal Decision Rules

Statistical decision theory provides a means of specifying rules for pattern classification which are optimal in the sense of minimizing average losses due to incorrect classification. A good treatment of this approach is found in Robbins.[20] We assume the existence of $r$ pattern categories $C_i$ with a priori probabilities of occurrence $p(i)$, for $i=1,\ldots,r$. We must also specify a loss function $\ell(i|j)$ which represents the loss resulting from assigning to category $i$ a pattern which actually belongs to category $j$. Using the loss function, the conditional average loss of assigning pattern $X$ to category $i$ is defined by

$$L_i(X) = \sum_{j=1}^{r} \ell(i|j)\, p(j|X)$$

where $p(j|X)$ is the probability that, given $X$, its category is actually $j$. The average loss is, therefore, minimized by assigning each $X$ to the category $i_o$ for which

$$L_{i_o}(X) \leq L_i(X) \text{ for } i = 1,\ldots,r$$

Such a decision rule is called a Bayes strategy. Using Bayes' rule we may write

$$p(j|X) = \frac{p(X|j)\ p(j)}{p(X)}$$

where $p(X|j)$ is the density function of category $j$ evaluated at $X$. The conditional average loss then becomes

$$L_i(X) = \frac{1}{p(X)} \sum_{j=1}^{r} \ell(i|j)\ p(X|j)\ p(j)$$

Since $P(X)$ is independent of $i$, it need not be evaluated in minimizing $L_i(X)$. It remains to evaluate the probability density functions $p(X|j)$ of each category at $X$. This is the central problem of pattern classification. In some instances the losses incurred by all misclassifications are equal. This situation is described by the <u>symmetric loss function</u>

$$\ell(i|j) = 1 - \delta_{ij}$$

which is zero for correct classifications and one otherwise. The problem is then reduced to minimizing

$$L_i(X) = 1 - \frac{p(X|i)\ p(i)}{p(X)}$$

which, if all categories are equally likely, is equivalent to maximizing $p(X|i)$. Such a rule is called a <u>maximum likelihood</u> decision rule.

The Bayes strategy may be explicitly implemented only if the $p(X|j)$ are already known. In practical situations, this is not the case and the densities must be estimated from samples of the categories. Their functional

form is sometimes known in advance (or more often assumed). In this case the problem is reduced to estimating the parameters of the density functions. This is called the "parametric" approach. For certain forms, notably the multivariate normal, convenient realizations of the optimal decision rules have been derived, several of which are described in this section. The weakness of this approach is that actual density functions do not usually conform to the assumed forms; this can result in badly suboptimal decision rules. An example of this is the case where the actual densities are multimodal.

The opposite "nonparametric" approach makes no assumptions about the density functions, except that they are reasonably smooth, and it approximates the entire function from sample patterns. These approximations become impractical as the dimension of the pattern vectors increases, because their domain is the pattern space and the number of points involved in a discrete approximation increases exponentially with the dimension. The problem is simplified if the components of the pattern vectors are statistically independent, for then

$$p(X|j) = \prod_{k=1}^{n} p(x_k|j)$$

where $n$ is the dimension. Here we need only approximate $n$ univariate densities for each category. If the components are not independent, but the categories have equal covariance matrices, then new, independent variables may be found by diagonalizing the covariance matrices (see Section II). However, there appears to be no general solution to the approximation problem. Most practical schemes approximate the densities indirectly through the use of discriminant functions (see below) which are equivalent, in terms of classification, to certain density function approximations. This problem, and various attempts to solve it, are discussed in greater detail under "Nonparametric Methods."

## 2. Discriminant Functions and Decision Surfaces

The theoretical foundation for the concept of discriminant functions and their role in pattern classification is summarized below. (The discriminant functions of decision theory should not be confused with the discriminant vectors of multivariate analysis of variance, treated in Section III.)

Geometrically, a pattern classification rule is equivalent to a

partition of the pattern space into disjoint regions corresponding to the categories. These regions are called decision regions and the surfaces separating them are called decision surfaces. The decision regions $R_1,\ldots,R_r$ of any r-category pattern classifier may be implicitly defined by a set of r discriminant functions $g_1(X),\ldots,g_r(X)$ which satisfy, for every $X$ in $R_i$

$$g_i(X) \geq g_j(X) \quad \text{for} \quad i,j=1,\ldots,r$$

Then the decision surface separating $R_i$ from $R_j$ is given by

$$g_i(X) - g_j(X) = 0$$

For example, the discriminant functions which determine the decision regions of the Bayes strategy are the negatives of the average loss functions: $-L_i(X)$.

Discriminant functions are widely used in pattern recognition applications because of their relative ease of implementation. However, as the probability densities of the pattern classes become more complex, so do the corresponding optimal discriminant functions. Therefore, much work in pattern recognition theory has been devoted to finding suboptimal discriminants of simple forms which closely approximate the performance of optimal functions. It is desirable to consider families of discriminant functions where members are determined by a modest number of parameters (which must be stored in any implementation of the resulting classifier). There is a useful class of function families whose members depend linearly on their weights. Such discriminant functions are referred to by Nilsson as $\Phi$ functions and may be written

$$\Phi(X) = w_0 + w_1 f_1(X) + \ldots + w_M f_M(X)$$

where the $f_i(X)$, $i=1,\ldots,M$, are linearly independent, real, single valued functions independent of the weights. The number of weights $M + 1$, which determine the $\Phi$ function, is called the number of degrees of freedom of the family. We will consider here only $\Phi$ function families whose members are polynomial functions of the components of $X$. The potential performance

of polynomial discriminant functions increases with the degree of the polynomials, but so does the number of weights necessary to implement them. In fact, increasing the maximum degree of the polynomials from d-1 to d adds $\binom{d+n-1}{n-1}$ possible degrees of freedom when $n \geq d$. It is, therefore, necessary to limit the maximum degree d of the polynomials.

When d = 1, we have

$$f_i(X) = x_i \text{ for } i=1,\ldots,n$$

and the $\Phi$ functions are <u>linear</u> in the components of X, with n+1 degrees of freedom. For d=2, the $f_i$ are of the form

$$f_i(X) = x_{k_1}^{p_1} x_{k_2}^{p_2} \qquad \text{for } k_1,k_2 = 1,\ldots,n$$

$$p_1,p_2 = 0,1$$

resulting in <u>quadric</u> $\Phi$ functions with (n+1) (n+2)/2 degrees of freedom. In general, the $f_i$ are d-th degrees polynomials

$$f_i(X) = x_{k_1}^{p_1} x_{k_2}^{p_2} \ldots x_{k_d}^{p_d} \qquad \text{for } k_1,\ldots,k_d = 1,\ldots,n$$

$$p_1,\ldots,p_d = 0, 1$$

and the $\Phi$ functions are general d-th degree polynomials in the components of X with $\binom{d+n}{n}$ degrees of freedom. Note that at least in theory, an arbitrary continuous discriminant function (for example, the likelihood functions $p(X|i)$) may be approximated to any desired degree of accuracy by appropriate choice of d. Later we will realize the usefulness of this fact.

It is instructive to consider the shape of the decision regions defined by polynomial discriminant functions. Since categories are assigned by finding the maximum $g_i(X)$, the decision surface separating regions $R_a$ and $R_b$ satisfies

$$g_a(X) - g_b(X) = 0$$

Since linear decision functions have the form $g(X) = w_0 + w_1x_1 + \ldots w_nx_n$, this is

$$(w_{a_1} - w_{b_1})\, x_1 + \ldots + (w_{a_n} - w_{b_n})\, x_n + (w_{a_0} - w_{b_0}) = 0$$

Thus each decision region of a linear pattern classifier is convex and is bonded by no more than $r - 1$ hyperplanes of dimension $n - 1$.

A quadric discriminant function has the form

$$g(X) = w_0 + \sum_{j=1}^{d} w_j x_j + \sum_{j=1}^{d} \sum_{k=j}^{d} q_{jk} x_j x_k$$

and is determined by its $1 + n + n(n-1)/2$ weights. This can be expressed in matrix form as

$$g(X) - w_0 + W'X + X'QX$$

where $W$ is the vector whose elements are the linear weights and $Q$ is the symmetric matrix whose elements are the weights of the quadratic terms. So the quadric decision surface separating $R_a$ and $R_b$ satisfies

$$X'(Q_a - Q_b)X + (W_a - W_b)'X + (w_{a_0} - w_{b_0}) = 0$$

The shape of the quadric surface depends upon the quadratic form $X'(Q_a - Q_b)X$. If $(Q_a - Q_b)$ is positive (or negative) definite, the surface is called a hyperellipsoid; if $(Q_a - Q_b)$ is not positive (or negative definite, the surface is called a hyperhyperboloid. In general, polynomial discriminant functions of degree $d$ result in d-th order decision surfaces in the pattern space.

The higher the order of the optimal polynomial decision surface, the better its ability to separate pattern categories with complex distributions. A set of categories which can be correctly identified by linear decision functions is called <u>linearly separable</u>. A family of $\Phi$ functions is determined by its component functions $f_i(X)$, $i=1,\ldots,M$. These component functions can be used to define a transformation

$$F(X) = (f_1(X),\ldots,f_M(X))$$

from the pattern space into an M-dimensional space called the <u>$\Phi$ space</u>.

Thus a decision surface in the pattern space implied by a given set of $\Phi$ functions has corresponding to it a hyperplane in the $\Phi$ space. If a set of categories is correctly identified by the $\Phi$ functions, it is linearly separable in the $\Phi$ space. In Section V we will see that the feature extraction problem may be viewed as the choice of an appropriate set of $\Phi$ functions.

3. Discriminant Functions for Normal Populations

If the probability density functions of the categories are multivariate normal, the optimal discriminant functions for a symmetric loss function may be derived explicitly (see Nilsson, [19] p. 55). They are

$$g_i(X) = w_0 + \left[(X - \mu_i)' \Sigma_i^{-1} (X - \mu_i)\right]$$

where $\mu_i$ and $\Sigma_i$ are the mean vector and covariance matrix of category $i$, and $w_0 = \ln p(i) - 1/2 \ln|\Sigma_i|$, for $i=1,\ldots,r$. So the optimum discriminant functions for normal patterns are quadric.

In the case of equal covariance matrices these can be reduced to linear discriminant functions. Furthermore, if the covariance matrices are the identity and the a priori probabilities are equal, the (maximum likelihood) discriminants are given by

$$g_i(X) = X' \mu_i - 1/2 \mu_i'\mu_i \quad \text{for} \quad i = 1,\ldots,r$$

Notice that equivalent classifications are obtained by minimizing the squared distance from $X$ to $\mu_i$, which is

$$\text{dist}^2 (X,\mu_i) = (X - \mu_i)' (X - \mu_i) = X'X - 2 X' \mu_i + \mu_i'\mu_i$$

because $X'X$ is constant over $i$, and so may be eliminated. These linear discriminants are widely used when little is known about the distributions of the patterns because they satisfy the intuitive notion that unknown patterns should be assigned to categories (represented by the means) to which they are close. Due to these considerations, we will use the minimum distance criterion to demonstrate empirically the value of intrinsic analysis and discriminant analysis for pattern classification in an experiment described in Section V.

## 4. Nonparametric Approaches

We now return to the problem of approximating arbitrary multivariate density functions from training samples. As we have seen, if the components of the patterns are statistically independent, these can be written as products of univariate densities. But the independence assumption is not usually justified, so we must, in effect, estimate the joint probabilities of the components of each possible pattern vector. For high dimensional pattern spaces this is very impractical and it is necessary to find some means of representing the densities indirectly. A productive technique for binary patterns is reported by Chow.[21] Here the pattern space B comprises the $2^n$ vertices of an n-dimensional cube; the density functions are expanded as linear combinstions of Walsh-Rademacher functions, which form a complete orthonormal basis for the space of real valued functions on B.

For continuous patterns the density function space becomes infinite dimensional. Various formal expansions for the continuous case have been proposed, for example, using Laguerre polynomials (Krishnomoorthy[22]), but they are quite impractical. Kanal,[23] pp. 4-20, reviews the problem of constructing orthonormal expansions and concludes: "In the multivariate case we are really faced with the curse of dimensionality and the prospect of constructing practical systems for adaptively approximating likelihood functions based on orthogonal expansions seems dim."

Another alternative is to give up direct estimation of densities and adopt a classification procedure which deals with the sample patterns directly and only implicitly involves the densities. Perhaps the most straightforward approach of this type is the "nearest neighbor decision rule" by which an unknown pattern is assigned to the category containing the training sample closest to it according to some metric defined on the pattern space. This is equivalent to the minimum distance criterion already described, with each sample point defining its own subcategory. The resulting decision surfaces are piecewise linear and will, in general, perform better than the optimal linear boundaries. It has been shown by Cover and Hart[24] that the error rate of this rule is at most twice that of the Bayes optimal classifier for an infinitely large training set. Of course, the problem is that as the number of training samples increases it becomes impractical to compute distances to all of them. One method which is frequently used to overcome this difficulty involves partitioning the samples into subcategories which tend

to cluster together. Modal points (typically means) of these subcategories are then used to implement the nearest neighbor rule. See, for example, Firschien and Fischler.[25] This "mode seeking" approach can be very productive but some care must be taken in the selection of clustering algorithms and their parameters for specific problems. In fact, according to Sammon,[26] p. 11, the performance of all known clustering algorithms is so sensitive to the settings of their parameters that "the proper setting usually can only be determined by a trial and error method."

Closely related to this is the attempt by Sebestyen[27] to estimate an arbitrary density function as the mean of a small number of normal densities approximated from subcategories. Besides ad hoc rules for "adaptive sample set construction," this approach involves division of the pattern space into cells and, therefore, runs into difficulty as the dimension increases. The fundamental assumption of both the mode seeking and adaptive sample set construction methods is that the densities in question can be well represented as the sum of symmetric normal densities. Thus they are particularly effective in handling multi-modal densities.

The idea of approximating density functions as means of normal densities is carried to its logical extreme in an elegant technique proposed by Specht,[28] who generates a symmetric density of normal form

$$g_1(X) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[\frac{-||X - S_1||^2}{2\sigma^2}\right]$$

about each sample pattern $S_1$. These "interpolation functions" are averaged over all patterns in the training set to obtain the approximation. It is shown, p. 31, that as the number of samples becomes infinite, and as the "smoothing parameter" $\sigma \to 0$, the approximation converges to the true density wherever it is continuous. In order to evaluate the approximation, the exponentials in the interpolation functions are written, using the series expansion, as polynomials in the components of $X$. The truncated expansions may then be used in $d$-th degree polynomial discriminant functions to implement a Bayes strategy. This is referred to as the "polynomial discriminant method."

It is interesting to note that as $\sigma \to \infty$, the resulting decision rule becomes the minimum distance classifier, and as $\sigma \to 0$ it becomes the nearest

neighbor rule; the corresponding decision surfaces range from strictly linear to highly nonlinear. In practice, the shape of the decision surfaces also depends on the degree d of the truncated polynomial approximation. (See Section IV-2 above.) Generally speaking, the higher the degree, the larger the value of $\sigma$ necessary to obtain an adequately smooth approximation.

This method, like all others, is subject to "the curse of dimensionality." Specht shows that the number of training samples required to obtain approximations of a given quality increases exponentially with the dimension of the pattern space. (It also increases as $\sigma$ is made smaller.) Nevertheless, the polynomial discriminant method seems to be applicable when only a small number of training samples is available. In fact, in an experiment involving separation of normally distributed categories, it actually outperformed the optimal (quadratic) classifier (see Section IV-3) based on estimates of the means and covariances. In the experiments eight samples were drawn from each category, but since their dimension was only five, it may not be valid to extrapolate the results to higher dimensions.

Finally, the polynomial discriminant method shares the practical advantages of matched filter methods over most other techniques. The coefficients of the polynomials are simple averages of the corresponding coefficients contributed by each training sample. Thus the classifier may be made adaptive simply by updating the discriminant functions as new samples are obtained. Also, unlike iterative techniques, only one look at each sample is required. The classifier can adapt to time varying statistics if exponential smoothing is used to update the coefficients. From both the practical and theoretical viewpoints, Specht's method is, in this author's opinion, the most promising nonparametric approach to Bayes optimal classification.

Section V

# INFORMATION COMPRESSION APPLIED TO DATA CLASSIFICATION PROBLEMS

In the previous Section we considered the pattern classification problem in isolation. The design of a system for pattern recognition generally includes two other stages: feature extraction, the problem of what measurements to use, and optimization of system parameters. Since the optimal parameters are dependent on the statistical properties of the data, they are usually estimated empirically; this problem will not be discussed further. This Section considers the feature extraction problem and the applicability of principal components analysis and discriminant analysis. Experimental results are described in which both methods were used as feature extractors for a minimum distance classifier.

## 1. The Feature Extraction Problem

Feature extraction is the process of selecting a relatively small number of measurements or combinations of measurements which tend to describe the characteristic features of the pattern classes. There are two basic goals: (a) minimizing the number of features and the resulting dimension of the classifier, and (b) finding features which determine a space in which the members of each pattern class will tend to cluster together, thus improving the performance of the classifier or making it possible to use a simpler algorithm. In some instances physical considerations will indicate an appropriate choice of measurements and feature extraction is primarily an engineering problem. Our consideration here is restricted to the situation in which a well defined set of sensor measurements already exists and the problem is to select features from these measurements. In this context feature extraction may be thought of as a mapping from the measurement space into a "feature space" which accomplishes either or both of the above goals.

In Section IV-2 we saw that the component functions of a $\phi$ function family determine a mapping from the measurement space into the $\phi$ space, in

which the decision regions are linear. The choice of good component functions may thus be regarded as feature selection if it improves the performance of a linear classifier in the $\phi$ space. Consider, for example, the polynomial discriminant function of Section IV-4. The terms of the polynomials are the component functions. Goal (b) of feature extraction is satisfied by these polynomial terms, because as the degree of the polynomials increases the Bayes optimal classifier is more nearly approximated. But goal (a) is not, because the number of terms increases rapidly with the degree of the polynomials so that the number of measurements is actually increased. Specht,[28] however, proposes methods for eliminating terms which are least useful in classification.

The feature extraction problem does not lend itself to a general solution. This is partly because the goodness of the features can ultimately be judged only on the performance of the recognition system, which depends also on the classification algorithm used. Another difficulty is the unbounded number of feature space transformations which are possible. If we consider only selection of measurements, for example, there are $\binom{n}{p}$ possible subsets of $p$ measurements chosen from a set of $n$. Some workers report success by simply choosing random subsets of redundant measurement sets. Another approach is to define some measure of the information content of each measurement relative to the classification of training categories. Measurements are then selected which have the largest information content. Since the above techniques treat measurements separately, they ignore the joint densities of the measurements. A nonparametric method for evaluating measurement subsets which does consider the joint densities is proposed by Fu.[29] This method employs direct estimation from multivariate density estimates of the error probability of a particular measurement subset; however, it offers no guidance for the choice of prospective subsets. For more detailed discussions of feature extraction and references, consult Fu or Nagy,[16] pp. 852-854. In the remainder of this Section, we consider the application of intrinsic analysis and discriminant analysis to the feature extraction problem.

## 2. Application of Principal Components Analysis and Discriminant Analysis

The linear dimension reduction and data discrimination techniques reviewed in Sections II and III find useful application in feature extraction. They produce linear transformations which may be applied to the measurement

space to obtain reduced dimension and improved performance of pattern classifiers, or at least linear classifiers, which we shall consider here. Highleyman,[18] p. 1505, shows that any pattern classifier using linear discriminant functions is invariant under nonsingular linear transformations of the measurement space. But an appropriate singular (dimension reducing) transformation can improve performance.

Principal components applied to the pooled sample sets (Section III-2) yields a linear transformation into the subspace of the measurement space spanned by the principal components. This transformation acts as a suboptimal feature selector by reducing the linear redundancy of the measurements. We have seen that the coefficients in the subspace are mutually uncorrelated over the ensemble of all the categories. This fact may tend to simplify the densities in the principal components basis. Also, by eliminating low variance components, the transformation could actually elliminate random noise present in the measurements. But its primary applicability is to goal (a) of feature extraction, dimension reduction. In an application to crop classification (Fu[29]) this approach was compared with the method of minimizing estimated error probabilities. The results were about equal if more than three features were allowed.

The above method does not consider class membership information and so could discard components related to it. The natural remedy to this danger is discriminant analysis, which maximizes the variance of class means relative to within class variance. The dimension reduction is extreme, since the number of discriminants is one less than the number of categories. Thus if there is a small number of categories, the representation in the discriminant space may not be adequate to represent complicated densities. We shall see, on the other hand, that it can be very effective for problems with any degree of linear separability.

### 3. Experimental Results

Principal components analysis and discriminant analysis were applied to a classification problem involving aircraft radar frequency signatures. Each sample pattern comprised measurements of 320 frequency components. Eight distinct categories were represented by a total of 281 samples. The samples of each category, an average of 35, were divided as evenly as possible into a training set and a testing set. Mean vectors of the categories and the grand mean were estimated from the training sets. Principal components of

the pooled training sets were estimated by the approximation method described in Section II-3. To reduce the dimension sufficiently to compute estimated discriminant vectors for the training sets, the samples were represented in terms of the first 70 principal components. The representation retained 97% of the variance of the training data, while achieving a dimension compression of more than four to one. Finally, the seven discriminant vectors were orthogonalized to form a basis for the discriminant space.

The approximate principal components and the orthonormalized discriminants both span subspaces of the pattern space with origin at $\mu$, the grand mean. Representations of the pattern vectors and the category mean vector estimates in these subspaces were obtained by the change of basis transformations

$$X_\Psi = \Psi' (X-\mu)$$

$$X_D = D' (X_\Psi) = D'\Psi' (X-\mu) = (\Psi D)' (X-\mu)$$

where $X$ is a vector in the pattern space and the columns of $\Psi$ and $D$ are the principal components and the discriminant vectors.

The minimum distance classification algorithm described in Section IV-3 was applied to the test patterns directly and in these two representations. The error rate in the original basis was 26.3%.

At best, the principal components representation improved this performance only negligibly, to 25.5%. On the other hand, it did at least as well even after components accounting for 15% of the variance of the training data (all but 17) had been discarded, a dimension reduction of twenty to one. (Due to statistical errors in the estimation of the principal components, it is likely that more than 15% of the variance of the test data was ignored in this representation.) As even more of the principal components were discarded, the error rate increased gradually to 30% for five vectors (43% of variance ignored) and rose sharply thereafter. There were only two instances in which the error rate actually decreased with increased loss of variance, at dimension 22 and 7; in both cases, the decrease was slight. These results are reasonable since the principal components representation preserves optimally the (squared) lengths of the patterns and the classifier compares distances to category means, which are lengths of

difference vectors. Since no significant improvement in classification was achieved, the primary value of principal components analysis was the reduction of the dimension so that discriminant analysis could be applied.

As expected, the performance of the minimum distance classifier improved substantially in the discriminant basis, with an error rate of 7.3%. This error is attributable to statistical error in estimation of the discriminants and mean vectors, because the test patterns are linearly separable. This was demonstrated by computing optimal discriminants directly from the test samples; that representation reduced the error rate to zero. For purposes of comparison, the mean vectors and principal components were also computed directly from the test data classification of the untransformed test data with the exact means resulting in an error rate of 19.7%; the difference between this and the "honest" error rate of 26.3% can be attributed to errors in estimation of the means from the training data. The percent error rates are summarized in the box below. It should be emphasized that the results in the second row cannot be achieved in practice and are included only to point out the estimation problem.

| Means and basis estimated from | Representation used | | |
|---|---|---|---|
| | Original data | First 20 principal components | Discriminant vectors |
| Training samples | 26.3 | 25.5 | 7.3 |
| Test samples | 19.7 | 19.7 | 0.0 |

## Section VI

## IMPLEMENTATION

Most of the above methods for data compression and classification have been implemented on the Dynamic Experimental Processor (DX-1) at the Multisensor Signal Processing Branch, Air Force Cambridge Research Laboratories. The hardware configuration includes two Digital Equipment Corporation PDP-1 central processors, an IBM 2311 disk storage unit, several CRT display consoles including a DEC color display, and a core-buffered, line-generating display unit called the Experimental Display Processor (XDP), which drives two of the consoles. In order to create a suitable environment for the development and operation of the computer programs involved, it has been necessary to design an operating system which provides the appropriate interactive data management and program execution capabilities.

A fundamental requirement is the ability to symbolically identify files of vectors on-line for random access storage and retrieval. This is accomplished by a disk based, fixed record length file system. Variable length information, including programs and relocatable subroutines is stored in partitioned files.

Programs are named, stored, loaded and executed on-line by the system monitor. Data files may also be manipulated through the monitor. Each user or problem is assigned a code which assures unique identification of his partitioned files and data files. Each user has read/write access to his own files and read-only access to all others. Thus all programs and data in the system may be shared by all users.

A convenient means of visually evaluating the results of data representation algorithms is provided by an interactive vector display program for the color CRT. Vectors may be displayed as graphs of their components or as projected points on a hyperplane determined by an arbitrary pair of vectors, or both, under user control. Commands are also supplied for scaling the projected images, saving them in random access storage, and

rotating the projection plane, in real time, to a plane determined by two new axes. The program is particularly flexible because of the on-line access to all data vectors by file name, which is provided by the vector file system.

To facilitate the programming of the data analysis algorithms, a special purpose language, called AMAC (Assembler Macros for Algebraic Computations), has been designed. Important features of AMAC include a run-time storage allocation capability, vector and matrix manipulation instructions and a comprehensive set of input/output macros.

Numerical algorithms for the system have been programmed to be as modular as possible, to allow flexibility in choosing the processes to be carried out. Therefore many of the techniques described in earlier sections require execution of several of the program modules described in this section. To simplify the execution process, it would be useful for the system to "remember" sequences of program calls which could then be activated by a single monitor command. Steps toward this goal are discussed at the end of this section.

1. Data Management

Random access storage for the DX-1 system is an IBM 2311 magnetic disk storage unit. The portion of the operating system which controls storage and retrieval of information on the disk is called the Disk File System. It stores information in fixed record length files, which is a particularly convenient form for vector data. An ID table is maintained which contains a unique six character name for each file and its size and location on the disk. (The first character of each file name is used by the system to designate the user or problem to which the file belongs, leaving five characters to be supplied by the user.) When a file is created, its name must be specified, along with its record length (number of data elements in each record) and element length (number of 18-bit words in each element). These parameters are fixed and are stored in the file system ID table along with the current file length and the physical location of records on the disk. The file length is never specified explicitly and may be increased at any time simply by writing more records. Records may be rewritten at any time.

The basic Disk File System commands are available to the user through the on-line monitor and to programs as standard I/O macros. They are the following:

assign - specify file parameters and enter file name in ID table
rename - change name of a file already assigned
delete - erase file and its ID table entry
lookup - retrieve file parameters (including file length) from ID table
write - write one or more contiguous records
read - read one or more contiguous records

For a thorough discussion of the implementation, characteristics and maintenance of the Disk File System, see Ref. 30.

The operating system also includes routines which handle partitioned files for variable length information. Each partitioned file occupies two ordinary files, a table of contents file, and a file for the actual information. Members of partitioned files are given six character alphanumeric names which are unique within the file. The file names are determined by the user identification and a type code. Partitioned file types used and anticipated include programs, relocatable subroutines, source programs, documentation and procedure definitions. Thus all programs belonging to a given user, for example, are stored in one partitioned file. By combining several separate elements of information into a single file, partitioned files increase disk space utilization and reduce average access time.

## 2. The On-Line Monitor

The monitor controls user identification, program storage, loading and execution, as well as on-line operations on data files. When using the system, each individual ordinarily supplies his identification character, which is added to his data file names and partitioned file names. One character, 1, is reserved to designate library files. This identification code must be used in order to update library files, and is assumed until the user provides his own code. The user is allowed to assign, rename, delete and write only files whose names are prefixed with his code. He may read or look up files prefixed with either his own or the library code, or other files by supplying an overriding prefix code.

Monitor commands are issued on the console typewriter or Soroban display keyboard, using only lower case characters. The format is a command symbol followed possibly by argument symbols separated by break characters. Legal symbols may contain alphanumeric characters, period or minus. All other characters, including comma, slash, space, tab, and carriage return,

are break characters which terminate symbols. Two special break characters are used to erase previous input. Backspaces erase previous characters and middle dot erases the entire command. Monitor commands and their descriptions are listed on page 39. Optional information is enclosed in brackets. The single character prefix, separated by a slash, overrides the current user identification, and the three character prefix designates a partitioned file type. Each of the commands may be abbreviated to three characters. The mask feature should be explained further. The parentheses may enclose up to five characters, any of which may be blank. All files are selected whose names correspond in the non-blank characters. If no characters are contained within the parentheses, all files are selected.

3. Dynamic Color Display of Vector Data

Graphic manipulations of the vector data files are carried out by a separate display program which is accessed through the monitor. Its primary function is vector projection of one or more files of vectors on a plane determined by any pair of non-colinear vectors. The program assigns a different color to each file to allow easy identification of the projected points. Such a projection may then be scaled up or down or transformed continuously into a projection on a new pair of coordinate axes. These functions, together with the data compression and discrimination algorithms already described, aid the user in discerning statistical relationships among several sets of data vectors. Properties of particular vectors may be presented by selecting one of the projected points with the light pen. This causes a graph of the coordinates of the projected vector to be displayed along with the projection. Graphs of the corresponding vector in different basis representations (and therefore different disk files) may also be requested. The sequence of display manipulations is determined by issuing commands on the display console keyboard. The effects of these commands are described in detail below.

Newdata: The program initializes the display, requests a list of names of files of vectors to be displayed, checks its validity, and divides all the data by the norm of the longest vector to assure that all projections will fit on the screen.

Project: The program asks for horizontal and vertical axes, which are specified by vector file name and logical record index within each file. These vectors are normalized to unit length so that only their

# DX-1 MONITOR COMMAND FORMATS

## Program Commands

| | |
|---|---|
| user | Examine current user identification |
| user u | Supply user identification letter |
| store pgname,ial,fal,...,sa | Store program on disk - ia is initial address and fa is final address of each block, sa is start address |
| load [u/] pgname[,m] | Load program into module m |
| start | Start loaded program |
| call [u/]pgname | Load and start program |
| list [u/][pgname] | Type names of programs of current or indicated user; type program addresses if pgname is specified |
| newname oldnam,newnam | Change program name |
| remove pgname | Delete program |

## Data File Commands

| | |
|---|---|
| assign fname,eltlen,reclen | Assign file parameters - fname is file name, eltlen is element length, reclen is record length |
| write fname,ia,index,nrecs | Write into file - ia is octal location of first record, index is position of record in file, nrecs is number of records transferred |
| read [u/]fname,is,index,nrecs | Read from file of current or indicated user |
| rename [pft/]oldnam,newnam | Change file or partition name |
| delete [pft/]fname | Delete file or partition |
| delete (mask) | Delete matching files |
| lookup [u/][pft/][fname] | Type file or partition names of current or indicated user; type file parameters if fname is specified |
| lookup [u/](mask) | Type matching file names of current or indicated user |

directions are used. In general the axes are not orthogonal and covariant projection is used, whereby a projected point is displayed at the intersection of the normals to the axes. Projection axes typically used include principal components, discriminant vectors, category mean vectors and standard basis vectors (for coordinate projection).

Scale: Due to the normalization of the data, the projections frequently do not fill the displayable area of the color CRT. The scale command allows the picture to be expanded (or contracted) in increments specified by the user.

Rotate: This command transforms the current projection into a projection of the same vectors onto a plane determined by a new pair of axes. Selection and normalization of the axes are carried out as in "project." The program in effect generates a whole sequence of projection planes, whose axes are located at equal angular intervals between the original axes and the new axes. The projections of the data vectors are computed directly only on the new axes. The projections on all of the intermediate axes are computed, in real time, using formulae involving sines and cosines of sums of angles. These calculations are so efficient that a great number of intermediate projections may be generated in a short time, even for hundreds of data vectors. The effect produced is an apparently continuous rotation of the plane of projection The user controls the speed (and smoothness) of this rotation by his choice of the number of intermediate projections. The rotation may be interrupted midway or reversed by sense switch control.

This feature has several applications. It facilitates the comparison of projection planes by allowing the user to follow the movement of individual points between them. It makes possible visual evaluation of the "stability" of a projection with respect to perturbations of its axes. Finally, the user may "explore" the vector space to discover projection planes (on intermediate axes) which may appear more desirable than those which are directly available.

Graph: Two modes are available for selection of vectors to be displayed in their component representations. Any vector already projected may be pointed out on the display screen with the light pen. Alternatively, any vector stored in a disk file may be indicated by file name and record index. Up to five graphs are displayed beside the current projection. Their colors are selected by the user to aid in distinguishing the graphs from one another.

## 4. Program Development

The DX-1 is an experimental system which undergoes frequent hardware modification. So programs are coded in the DX-1 MIDAS assembly language, which can be easily modified to accommodate new instructions and I/O operations. Coding of complicated mathematical algorithms, however, is difficult and tedious at the level of machine instructions. Therefore, a set of MIDAS macro instructions, called AMAC, has been written which provides some of the power of an algorithmic language without sacrificing the flexibility of assembly language.

AMAC includes a limited form of arithmetic statement with subscripted variables which may also contain bit manipulation and logical operations. Among other FORTRAN-like features are instructions for looping, conditional execution and subroutine linkage and a library of arithmetic function subroutines. Unlike FORTRAN, AMAC allows run-time storage allocation. Especially useful in this effort have been macros which call subroutines performing matrix/vector operations typically involved in statistical applications, such as inner products, sums, differences, and matrix products.

AMAC contains an integrated set of character-oriented I/O macros for the on-line typewriter, display console keyboard, paper tape reader, and punch and CRT display. Specific devices and formats are specified as arguments of the macros; thus the effective device may be a run-time variable. Disk I/O is performed by the macros described in Section VI-1. A thorough description of AMAC may be found in Ref. 31.

The program modules used in the experiments of Section V are executed by the monitor commands described below. Arguments are currently requested individually by the programs, but for clarity they are indicated here as lists. Parentheses contain arguments which may be lists; brackets indicate arguments which may be omitted.

call intan((files), dim, rdim,[gmean],[switch],eigsys)

Performs intrinsic analysis on the (pooled) vectors contained in files. The first dim elements of each vector are used; rdim eigenvalues and eigenvectors are computed. The values are typed out on-line and the vectors are written into the file to be named eigsys. The grand mean is written in file gmean if the name is supplied. Intrinsic analysis is performed if switch is nonzero; principal components (data centered about the mean) if it is null.

call chbas((ifiles),dim,[vector],[basis],(ofiles))

Performs the translation and change of basis transformation

output = basis' (input-vector)

on the first dim components of the vectors in ifiles. The dimension of the output vectors written in ofiles is the number of vectors in basis (its file length). If vector is null, only the matrix multiplication is performed; if basis is null, only the vector difference. This program is typically used to transform data into an intrinsic or principal components basis.

call intanp((ifiles),dim,p,rdim,[gmean],switch,prteig)

call chbasp((ifiles,dim,p,[gmean],prteig,(ofiles))

These programs are used together to perform the preliminary reduction for the approximate intrinsic analysis described in Section II-3. The vectors in ifiles are partitioned into p equal segments to produce segment eigensystems which are stored in prteig. The programs and their arguments ifiles, dim, rdim, mean, switch and ofiles are analogous to those described above. The output in ofiles is processed by intan and chbas to complete the approximation process.

call means((files),means,[gmean])

Computes mean vectors for all inputfiles and writes them in means; writes the grand mean in gmean if specified.

call discrm((files),dim,[gmean],mean,dvecs)

Performs linear discriminant analysis on the first dim components of the vectors in files. Each file represents one pattern category. Discriminant vectors are written in file dvecs; eigenvalues are typed on-line. Category means are written in means; the grand mean is written in gmean if indicated.

call orthog(ifile,ofile)

The vectors in ifile are orthogonalized and written in ofile.

call mindis((files),dim,means,[basis])

Applies the minimum distance classification rule to the first dim components of the vectors in files. The vectors in means are used as prototype patterns. If basis is not null, both data and means are first transformed into it. If the number of files matches the number of means, error percentages are printed in addition to a confusion matrix.

## 5. Extensions of the System

One of the goals of this effort has been to provide a facility for experimentally discovering or verifying sequences of analytical methods useful in data reduction and classification problems. Due to the modular nature of the algorithms developed for the system, the typical procedure may require calling several separate programs. The user must remember the sequence of programs to be performed along with the argument lists of each, which, as the above examples indicate, are often highly redundant. To ease the burden, a "cataloged procedure" facility has been designed. It will add three new commands to the monitor: define, termin and exec. The procedure name is supplied to define along with a list of dummy arguments. Any sequence of legitimate monitor commands using constant or dummy arguments follows. The definition is ended by the termin command. A procedure thus defined may then be invoked by the exec command with the procedure name and a list of actual arguments.

For example

```
define      intrep(ilist/old,dim/dec,rdim/dec,eigsys/new,olist/new)
call intan  (ilist,dim,rdim,,,eigsys)
call chbas  (ilist,dim,,eigsys,olist)
delete      eigsys
termin
```

will make possible the command

```
exec intrep ((file1,file2),120,30,,ps1,(file11,file21))
```

which produces the intrinsic basis representation of the vectors in file1 and file2 and then deletes the eigensystem from the disk.

A problem which arises immediately is argument screening. If arguments are to be supplied all at once in lists, there is a strong possibility of out-of-order arguments which would cause execution errors. Therefore, an argument processing routine will be added to the operating system which will screen arguments requested by programs for proper type and format. So that procedure arguments may be screened before the programs are called, their types are indicated in the definition by characters attached to the dummy names. The argument types currently recognized are the following:

| Code | Format of Argument |
|---|---|
| old | previously assigned file id |
| new | id of file assigned by this procedure |
| oct | octal integer |
| dec | decimal integer |
| flo | floating point number |
| tex | arbitrary text enclosed in brackets |

Finally, as the library of programs and procedures grows, it will be increasingly desirable to provide on-line graphic system documentation. This can be supplied by monitor commands to list the names of available procedures and display short writeups and argument descriptions for specific programs and procedures, which could be stored conveniently in partitioned files.

## Section VII

## SUMMARY AND CONCLUSIONS

This research has extended the theoretical foundations and developed the practical techniques necessary for implementation of an interactive multivariate data analysis facility. The analytical problems treated can be broken down roughly into three areas: efficient representation of high-dimensional information, representation of multicategory information for graphic display and determination of optimal or satisfactory decision procedures for data classification. Classical multivariate statistical methods have proven valuable in these applications. Principal components analysis (or intrinsic analysis) effects compression of vector data with minimum mean square error. Linear discriminant analysis produces axes which maximize the variance of multicategory data relative to the variance within categories.

These methods have been implemented on the Experimental Dynamic Processor (DX-1) at AFCRL, using state-of-the-art algebraic eigensystem algorithms. Also a new eigensystem approximation technique has been developed which allows approximate intrinsic analysis to be applied to very high dimensional problems which would otherwise be intractable due to computer time and storage requirements. The dimension of data to which discriminant analysis can be applied is limited by storage requirements and by the number of samples available. Both of these limitations have been overcome by first representing the data in a truncated intrinsic basis. This has also reduced the computation time required for the discriminant analysis.

In an interactive computer data analysis system, these methods are valuable in displaying information in a form which elucidates its statistical characteristics. This can help the system engineer or scientist determine the structure and degree of complexity of his problem. Since the ultimate goal of most data analysis systems is usually to improve a real world decision process, the applicability of these methods to pattern classification problems was considered. The minimum risk Bayes strategy for pattern

recognition was reviewed, as well as the concept of discriminant functions and their equivalent decision surfaces in the pattern space. It was also noted that when the measurements which make up the patterns are very numerous and/or highly interrelated, many pattern classification algorithms cannot be used effectively. It is then necessary to transform the original measurements into fewer and/or better measurements. This process is called feature extraction.

To test the applicability of intrinsic analysis and discriminant analysis to feature extraction, a commonly used, matched filter-type classifier was chosen: the minimum distance rule, which classifies an unknown pattern in the category to whose mean it is nearest. (This strategy is implemented by linear discriminant functions which are negatives of squared distances from category means, and is Bayes optimal when the categories have symmetric normal probability densities.) The minimum distance rule was applied to an aircraft radar frequency signature classification problem of high dimension with (a) no feature extraction, (b) feature extraction by principal components analysis, and (c) feature extraction by discriminant analysis. Principal components analysis greatly reduced the number of measurements but did not significantly improve the performance of the classifier. Discriminant analysis reduced the dimension even further and reduced the error rate substantially. These results are to be expected, since principal components are vectors which preserve, as well as possible, squared distances of all patterns from the origin in the pattern space; whereas discriminant vectors emphasize variance among categories.

All of the transformations described above for data reduction and classification are linear. They performed reasonably well because the category densities were essentially unimodal. Other experiments have shown that their performance is much worse when more complex, multimodal densities are involved. Nonlinear methods are then needed, which can estimate the densities directly in order to implement optimal decision rules. To this end most researchers have advocated either histograms or orthogonal expansions. For pattern dimensions greater than two, the former are too cumbersome, requiring a great deal of manual supervision, and the latter are hopelessly complicated. D. Specht[28] has successfully employed a far more promising approach involving multinomial expansions of multivariate density function estimates. It appears that future work along these lines should be directed toward development and refinement of his technique.

Implementation and effective utilization of an interactive data analysis facility using the methods described here has required the design of an operating system tailored to its needs. Its features include a disk-based vector file system for convenient manipulation of vector data, an on-line system monitor, a dynamic, color vector projection program, and a special purpose programming language. The need for such special purpose software and the high demands in computation time of many of the algorithms involved indicate that such systems can best be implemented on small or medium-scale dedicated machines rather than on large-scale, time-shared configurations.

The color CRT provides easy identification of projected points by category, which is valuable in classification problems. We have seen that discriminant analysis can determine good projection planes for multicategory data. It should be noted, however, that some intrinsically complex problems may not be sufficiently well represented by any two dimensional projection. Therefore, such displays should be used to augment the intuition but not to draw hard conclusions about the data structures involved.

The implementation of the algorithms has been as modular as possible to allow the greatest flexibility in their application. The vector projection program allows the results of intermediate results to be visually evaluated. Once a useful sequence of operations has been established, it is desirable, for simplicity of operation, to define it as a single procedure. For this purpose, future additions to the system will include a cataloged procedure facility. Also needed is a provision for on-line graphic documentation of programs and procedures.

REFERENCES


1. Wilks, S. Mathematical Statistics. New York: Wiley, 1962.

2. Anderson, T. An Introduction to Multivariate Statistical Analysis. New York: Wiley, 1958.

3. Kramer, H. and M. Mathews. "A Linear Coding for Transmitting a Set of Correlated Signals," IRE Trans. on Information Theory. IT-2, pp. 41-44, 1956.

4. Young, T. and W. Huggins. Representation of Electrocardiogram by Orthogonalized Exponentials. Report No. AFCRL-187, 1961. (AD 256 364).

5. Walter, C. "Intrinsic Analysis vs. Fourier Analysis in the Representation of Signal Data Structures," DECUS Proceedings, pp. 101-108, 1964.

6. Colomb, R. Techniques in Intrinsic Analysis. Report No. AFCRL-67-0153, Systems Research Labs., 1966. (AD 651 071).

7. Davenport, W. and W. Root. An Introduction to the Theory of Random Signals and Noise. New York: McGraw-Hill, 1958.

8. Loève, M. Probability Theory. Princeton: 3rd Ed., Van Nostrand, 1963.

9. Watanabe, S. The Loève-Karhunen Expansion as a Means of Information Compression for Classification of Continuous Signals. Report No. AMRL TR-65-114, IBM Corp., 1965. (AD 628 684).

10. Wilkinson, J. The Algebraic Eigenvalue Problem. Oxford: Clarendon, 1965.

11. Roper, R. Approximate Eigensystems of Large Covariance Matrices. Report No. AFCRL-68-0392, Systems Research Labs., Inc., 1968. (AD 678 172).

12. Hotelling, H. "A Generalized T Test and Measure of Multivariate Dispersion," Second Berkeley Symposium on Mathematical Statistics and Probability, pp. 23-41. Los Angeles: U. of Cal. Press, 1951.

13. Colomb, R. Choice of Axes for Projection of Data on a CRT Using Multivariate Analysis of Variance. Report No. AFCRL-68-0393, Systems Research Labs., Inc., 1968. (AD 678 105).

14. Streeter, D. and J. Raviv. Research on Advanced Computer Methods for Biological Data Processing. Report No. AMRL TR-66-24, IBM Corp., 1966. (AD 637 452).

15. Colomb, R. Effects of Computing Discriminants in a Truncated Intrinsic Basis. Interim Tech. Report, Contract F19628-67-C-0150, Systems Research Labs., Inc., 1968.

16. Nagy, G. "State of the Art in Pattern Recognition," Proc. IEEE, Vol. 56, No. 5, pp. 836-862, May 1968.

17. Sebestyen, G. Decision-Making Processes in Pattern Recognition ACM Monograph. New York: Macmillan, 1962.

18. Highleyman, W. "Linear Decision Functions, with Application to Pattern Recognition," Proc. IRE, Vol. 150, pp. 1501-1514, June 1962.

19. Nilsson, N. Learning Machines. New York: McGraw-Hill, 1965.

20. Robbins, H. "The Empirical Bayes Approach to Statistical Decision Problems," Annals of Mathematical Statistics, Vol. 35, No. 1, March 1964.

21. Chow, C. "A Class of Nonlinear Recognition Procedures," IEEE Trans. on Systems Science, Vol. SSC-2, No. 2, pp. 101-109, December 1966.

22. Krishnamoorthy, A. and M. Parthasarathy. "A Multivariate Gamma Type Distribution," Annals of Math. Statistics, Vol. 22, pp. 549-557, 1951.

23. Kanal, L. and K. Abend. Adaptive Modelling of Likelihood Classification - I, Report No. RADC-TR-66-190, Philco Corp., 1966. (AD 636 519).

24. Cover, T. and P. Hart. "Nearest Neighbor Pattern Classification," IEEE Trans. on Information Theory, IT-13, pp. 21-27, January 1967.

25. Firschien, O. and M. Fischler. "Automatic Subclass Determination for Pattern Recognition Applications," IEEE Trans. on Electronic Computers, EC-12, No. 2, pp. 137-141, April 1963.

26. Sammon, J. On-Line Pattern Analysis and Recognition System (OLPARS). Report No. RADC-TR-68-263, Rome Air Development Center, 1968. (AD 675 212).

27. Sebestyen, G. and J. Edie. "Pattern Recognition Research." Report No. AFCRL-64-821, Litton Systems, Inc., Waltham, Mass., 1964.

28. Specht, D. Generation of Polynomial Discriminant Functions for Pattern Recognition. Report No. SEL-66-029, Stanford University, Stanford, Cal., 1966. (AD 487 537).

29. Fu, K. S., et al. "Feature Selection in Pattern Recognition," IEEE Trans. on Systems Science, Vol. SCC-6, No. 1, January 1970.

30. Roper, R. The DX-1 Disk File System. Interim Tech. Report, Contract F19628-67-C-0150, Systems Research Labs., Inc., 1969.

31. Roper, R., N. Chase and R. Daesen. Assembler Macros for Algebraic Computations: AMAC User's Manual. Interim Tech. Report, Contract F19628-67-C-0150, Systems Research Labs., Inc., 1970.

Unclassified
Security Classification

**DOCUMENT CONTROL DATA - R&D**

*(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)*

1. ORIGINATING ACTIVITY *(Corporate author)*
SYSTEMS RESEARCH LABORATORIES, INC.
7001 Indian Ripple Road
Dayton, Ohio 45440

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
ANALYTICAL AND INTERACTIVE TECHNIQUES FOR MULTIVARIATE DATA ANALYSIS AND CLASSIFICATION

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Final Scientific Report, 1 December 1966 through 31 May 1970, Approved 4 Dec. 1970

5. AUTHOR(S) *(First name, middle initial, last name)*
Robert B. Roper

| 6. REPORT DATE | 7a. TOTAL NO. OF PAGES | 7b. NO. OF REFS |
|---|---|---|
| 25 September 1970 | 55 | 31 |

| 8a. CONTRACT OR GRANT NO. | 9a. ORIGINATOR'S REPORT NUMBER(S) |
|---|---|
| F19628-67-C-0150 | 60606 Final Report |
| b. PROJECT, TASK, WORK UNIT NOS. 5632, 563201, 56320101 | |
| c. DOD ELEMENT 61102F | 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* |
| d. DOD SUBELEMENT 681305 | |

10. DISTRIBUTION STATEMENT
1 - Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

| 11. SUPPLEMENTARY NOTES | 12. SPONSORING MILITARY ACTIVITY |
|---|---|
| Tech, Other | Air Force Cambridge Research Laboratories (LRM)<br>L. G. Hanscom Field<br>Bedford, Massachusetts 01730 |

13. ABSTRACT

This report demonstrates the applicability of classical statistical techniques to problems involving compression and classification of multivariate data. The theoretical foundations of two such techniques, intrinsic analysis and discriminant analysis, are treated in detail. Efficient digital computer implementation is discussed, including the combined application of intrinsic and discriminant analysis and a new algorithm for computing approximate intrinsic bases for very large problems. Experimental results are presented on the application of these techniques as feature extractors in a signal classification problem. Also included is a description of the interactive graphics-oriented system software which has been developed to facilitate the application of these techniques.

DD FORM 1 NOV 65 1473

Unclassified
Security Classification

| 14. KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|---|---|---|---|---|---|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| Information Compression | | | | | | |
| Pattern Classification | | | | | | |
| Feature Extraction | | | | | | |
| Multivariate Statistical Analysis | | | | | | |
| Principal Components Analysis | | | | | | |
| Discriminant Analysis | | | | | | |
| Interactive Graphics | | | | | | |