An Optimal Set of Discriminant Vectors

DONALD H. FOLEY, MEMBER, IEEE, AND JOHN W. SAMMON, JR., MEMBER, IEEE

***Abstract*—A new method for the extraction of features in a two-class pattern recognition problem is derived. The main advantage is that the method for selecting features is based entirely upon discrimination or separability as opposed to the more common approach of fitting. The classical example of fitting is the use of the eigenvectors of the lumped covariance matrix corresponding to the largest eigenvalues. In an analogous manner, the new technique selects discriminant vectors (or features) corresponding to the largest "discrim-values." The new method is compared to some of the more popular alternative techniques via both data-dependent and mathematical examples. In addition, a recursive method for obtaining the discriminant vectors is given.**

***Index Terms*—Dimensionality reduction, discriminants, eigenvectors, feature extraction, feature ranking, feature selection, Karhunen-Loeve expansions, multivariate data-analysis, pattern classification, pattern recognition.**

I. INTRODUCTION

IN MANY PATTERN recognition applications it is convenient to divide the problem into two parts, feature extraction or selection and pattern classification. Although numerous papers exist in the literature on designing classification logic given a set of features, it is only recently that significant attention has been given to the area of feature extraction. An indication of this emphasis on feature extraction is the special issue of the IEEE TRANSACTIONS ON COMPUTERS devoted entirely to this subject area [1].

Usually, an initial set of raw measurements is taken from examples of each class of interest. For example, the raw measurements might consist of digitized waveforms or images. The number of raw measurements is frequently quite large. Simply stated, the goal of feature extraction is to find a transformation which maps the raw measurements into a smaller set of features which hopefully contain all the relevant or discriminatory information needed to solve the overall pattern recognition problem.

Most of the feature extraction literature has centered around finding linear transformations. The most popular transform is the Karhunen-Loeve or eigenvector orthonormal expansion [2], [3]. Since each eigenvector can be ranked by its corresponding eigenvalue, a subset of the "best" eigenvectors can be chosen as the most "relevant" features. Unfortunately, the subset chosen provides the best fitting subspace. Recognizing that representational accuracy is not the ultimate objective of pattern recognition, Fukunaga and Koontz [4] suggest a modification of the eigenvector technique. A preliminary transformation for a two-class problem is found such that the eigenvectors which best fit class 1 are the poorest for representing class 2.

This paper suggests and derives an algorithm for extracting a set of features for a two-class problem in which the criteria for selecting each feature is based directly on its discriminatory potential. These features are based on an optimal set of discriminant vectors which are an extension of the discriminant plane derived by Sammon [5]. For the remainder of the paper, the term "discriminant vectors" will be used for the new technique.

The following sections of the paper give a quick review and counterexample for both the eigenvector and the Fukunaga-Koontz transforms, a derivation of the discriminant vectors, and a comparison of these different techniques.

Besides the obvious use of discriminant vectors for feature extraction, the discriminant vectors can be used in interactive pattern recognition systems such as the on-line pattern analysis and recognition system (OLPARS) [6]. One objective of OLPARS is to find projections of high dimensional data into one-, two-, or three-spaces so an on-line analyst can both observe the inherent structure of the data and design piecewise linear classification logic. The discriminant vectors provide promising candidates for the directions on which the data could be projected.

II. REVIEW OF KARHUNEN-LOEVE TRANSFORM

Most of the attention in the literature concerning feature selection has centered around the discrete Karhunen-Loeve or eigenvector expansion [2], [3]. In this procedure, the original L-dimensional data vectors are transformed by multiplying by the eigenvectors $\boldsymbol{e}_j$ of the estimated lumped covariance matrix $\sum$. The eigenvectors satisfy the equation

$$\sum \boldsymbol{e}_j = \lambda_j \boldsymbol{e}_j$$

where λ_j is the eigenvalue corresponding to the jth eigenvector. For real data, $\sum$ is a real symmetric matrix and all the eigenvectors are orthogonal, i.e., $\boldsymbol{e}_i{}^t \boldsymbol{e}_j = 0$ for $i \neq j$, and the eigenvalues are all greater than or equal to zero. Hence the eigenvalues can be ordered such that $\lambda_1 \geq$

Manuscript received September 30, 1973; revised August 20, 1974.
The authors are with the Pattern Analysis and Recognition Corporation, Rome, N. Y. 13440.

$\lambda_2 \geq \cdots \geq \lambda_L > 0$. If the eigenvectors are normalized such that $\| e_j \| = e_j{}^t e_j = 1$, then

$$\Phi = \begin{bmatrix} e_1{}^t \\ \cdot \\ \cdot \\ \cdot \\ e_L{}^t \end{bmatrix}$$

is an orthonormal transformation such that $y = \Phi x$ and each $y_j = e_j{}^t x$ is the jth feature in the new space for the input vector x. The mean square error obtained by projecting all the data into a subspace spanned by only a subset A of $K(K < L)$ eigenvectors is given by

$$\overline{\epsilon^2} = \sum_{j \notin A} \lambda_j$$

where A is the set containing the chosen K eigenvectors.

For any K, this error can be minimized by choosing the eigenvectors corresponding to the K largest eigenvalues. In fact, it is known (e.g., see Fukunaga [7]) that the mean square error in representing a set of random L-dimensional vectors with only K elements of *any* orthonormal transformation (i.e., any K-dimensional subspace) is minimized by selecting the first K elements of the discrete Karhunen–Loeve expansion. Although this transformation is optimal with respect to fitting the data, it is not necessarily optimal with respect to discriminating the data.

Fig. 1 shows a two-class two-dimensional pattern recognition problem. The eigenvector transform would rank direction e_1 as the best fitting direction or feature. However, it would obviously result in a poor recognition rate.

III. FUKUNAGA–KOONTZ TRANSFORM

To avoid the conflict of goals just mentioned, Fukunaga and Koontz [4] suggest in an earlier paper a preliminary transformation for a two-class problem such that the eigenvectors which best fit class 1 are the poorest for representing class 2. In order to briefly describe this method, let R_1 and R_2 be the weighted correlation matrices of class 1 and 2, respectively, i.e.,

$$R_i = P_i(\textstyle\sum_i + \mu_i \mu_i{}^t), \qquad i = 1,2$$

where P_i, $\sum_i$, and μ_i are the *a priori* probability, the covariance, and the mean of class i. Let T be a preliminary transform such that

$$T(R_1 + R_2)T^t = I.$$

Fukunaga shows that the eigenvectors of TR_1T^t are the same as the eigenvectors of TR_2T^t and all the eigenvalues are bounded by 0 and 1. Let the eigenvalues for class 1 be ordered such that

$$1 \geq \lambda_1{}^{(1)} \geq \lambda_2{}^{(1)} \geq \cdots \geq \lambda_L{}^{(1)} \geq 0.$$

Fukunaga proves next that $\lambda_i{}^{(2)} = 1 - \lambda_i{}^{(1)}$. Consequently, the best fitting eigenvectors for class 1 (e.g., $\lambda_i{}^{(1)} \cong 1$) are the poorest for class 2 (e.g., $\lambda_i{}^{(2)} = 1 - \lambda_i{}^{(1)} \cong 0$). Fukunaga recommends ordering the eigenvectors to be used as features by picking the largest of $| \lambda_j{}^{(1)} - 0.5 |$.

This implies that an eigenvalue of 0.5 is an extremely poor feature. This may not always be true. The following is a counterexample. Let

$$\mu_1{}^t = [0 \quad 5 \quad 0]; \quad \mu_2{}^t = [0 \quad -5 \quad 0]; \quad P_1 = P_2 = 1/2$$

$$\textstyle\sum_1 = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 8 \end{bmatrix}; \qquad \textstyle\sum_2 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Note that dimensions one and three have zero means for both classes, while there is considerable separation along dimension two. For example, if the marginal distribution along dimension two was normal, a Bayes error rate of 0.0003 percent would result. Along either dimensions one or three a Bayes error rate of 35 percent would result if the marginal distributions were normal. Also note that a double threshold is required. If a researcher used features which had zero means and different variances for the classes, the classification logic should not be as simple as any of the commonly used linear discriminant functions.

Proceeding with the Fukunaga–Koontz transform,

$$R_1 = 1/2 \begin{bmatrix} 8 & 0 & 0 \\ 0 & 26 & 0 \\ 0 & 0 & 8 \end{bmatrix}; \qquad R_2 = 1/2 \begin{bmatrix} 2 & 0 & 0 \\ 0 & 26 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

The matrix[1] T and the matrix $TR_1 T^t$ are given by

$$T = \begin{bmatrix} 1/(5)^{1/2} & 0 & 0 \\ 0 & 1/(26)^{1/2} & 0 \\ 0 & 0 & 1/(5)^{1/2} \end{bmatrix}$$

$$TR_1T^t = \begin{bmatrix} 0.8 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.8 \end{bmatrix}.$$

Using the Fukunaga–Koontz ranking (i.e., $| \lambda_j{}^{(1)} - 0.5 |$), dimensions one and three are tied with the highest rating (0.3), while dimension two received the lowest rating possible, i.e., 0.0.

In effect, two features with poor discriminatory information have been chosen over the one feature with almost perfect discriminatory capabilities. In addition, features which prevent the use of a simple linear discriminant have been chosen.

[1] T consists of each eigenvector of R divided by the square root of its corresponding eigenvalue.

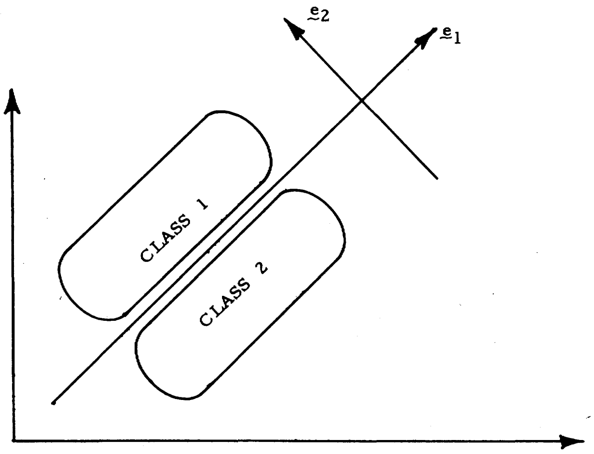


Fig. 1. Illustration of fitting deficiency. Underlined symbols indicate boldfaced symbols in text.

IV. DISCRIMINANT VECTORS

Before proceeding with a derivation of the discriminant vectors, it is worthwhile to place the roles of discrimination and fitting in perspective. Usually if the original set of features is carefully chosen, it is possible to design logic based on the statistics of a design set which achieves a very high recognition rate. In order to cut the error rate even further, the authors have found it practical to use fitting routines in order to create reject strategies. The fitting routines are applied after a class has been tentatively identified by the discrimination logic. Our experience has been to base the routines on the relationships and ranges of certain features. These fitting routines are usually designed from physical knowledge of what must be present if a particular class occurs. If the fitting criteria are not met, a reject decision is output. Other fitting routines based on criteria derived from eigenvector algorithms are also possible.

The objective of the discriminant vectors is to aid in solving the discrimination portion of the task.

In order to find the best vectors for discriminating between two classes, it is necessary to select some optimality criterion. As our measure of discrimination achieved by a vector $\boldsymbol{d}$, let us choose either 1) the Fisher ratio, that is, the ratio of the projected class differences to the sum of the within-class scatter along $\boldsymbol{d}$, i.e.,

$$R_1(\boldsymbol{d}) = \frac{(\boldsymbol{d}^t\boldsymbol{\Delta})^2}{\boldsymbol{d}^t W \boldsymbol{d}}$$

where

$\boldsymbol{d}$ = L-dimensional column vector on which the data are projected;

$\boldsymbol{d}^t$ = transpose of $\boldsymbol{d}$;

$\boldsymbol{x}_{ij}^t$ = $(x_{ij1} \cdots x_{ijL})$; jth sample vector for class i;

N_i = number of samples in class i;

$\hat{\boldsymbol{\mu}}_i$ = estimated mean of class i, $\hat{\boldsymbol{\mu}}_i = (1/N_i) \sum_{j=1}^{N_i} \boldsymbol{x}_{ij}$;

$\boldsymbol{\Delta}$ = difference in the estimated means, $\boldsymbol{\Delta} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$;

W_i = within-class scatter for class i, $W_i = \sum_{j=1}^{N_i} (\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_i)^t$;

W = sum of the within-class scatter, $W = W_1 + W_2$,

or 2) the modified Fisher ratio, that is, the ratio of the projected class differences to the sum of the within-class covariance along $\boldsymbol{d}$, i.e.,

$$R_2(\boldsymbol{d}) = (\boldsymbol{d}^t\boldsymbol{\Delta})^2 / \boldsymbol{d}^t[\hat{\Sigma}_1 + \hat{\Sigma}_2]\boldsymbol{d}$$

where

$$\hat{\Sigma}_i = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_i)^t = (N_i - 1)^{-1} W_i$$

is the estimated covariance of class i.

This criterion devotes equal attention to minimizing the covariance of each class, while the former devotes more attention to the class with more samples. Since both criteria have the same general form, Anderson and Bahadur [8] have suggested using

$$R(\boldsymbol{d}) = \frac{(\boldsymbol{d}^t\boldsymbol{\Delta})^2}{\boldsymbol{d}^t A \boldsymbol{d}}$$

where $A = cW_1 + (1 - c)W_2$ and $0 \leq c \leq 1$. $R(\boldsymbol{d})$ will be called the generalized Fisher criterion and will be used in the derivation of the discriminant vector. For $c = 0.5$, maximizing $R(\boldsymbol{d})$ with respect to $\boldsymbol{d}$ is equivalent to maximizing $R_1(\boldsymbol{d})$ with respect to $\boldsymbol{d}$, while for $c = (N_2 - 1)/(N_1 + N_2 - 2)$, maximizing $R(\boldsymbol{d})$ with respect to $\boldsymbol{d}$ is equivalent to maximizing $R_2(\boldsymbol{d})$ with respect to $\boldsymbol{d}$.

Caution should be exercised in using any pattern recognition technique, such as the Fisher criterion, which inherently makes use of the class means. This is due to the modality problem, and it might be necessary to partition each class into its modes before performing the discriminant vector analysis. If more than two classes are involved, a pairwise voting logic can always be used [5].

With this background, a derivation of the discriminant vectors can begin. Note that the generalized ratio R is independent of the magnitude of the vector $\boldsymbol{d}$ (since

$R(\boldsymbol{d}) = R(\alpha \boldsymbol{d})$). Consequently, only the direction of $\boldsymbol{d}$ is important and $\boldsymbol{d}$ will be normalized such that $\boldsymbol{d}^t\boldsymbol{d} = 1$. Sammon [5] gives an easily accessible derivation of the first two directions which maximize the Fisher criterion R. The first direction $\boldsymbol{d}_1$ is

$$\boldsymbol{d}_1 = \alpha_1 A^{-1}\boldsymbol{\Delta}$$

where α_1 is chosen such that $\boldsymbol{d}_1{}^t\boldsymbol{d}_1 = 1$ and

$$\alpha_1{}^2 = (\boldsymbol{\Delta}^t[A^{-1}]^2\boldsymbol{\Delta})^{-1}.$$

The second best direction for maximizing R subject to the constraint that $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ are orthogonal (i.e., $\boldsymbol{d}_2{}^t\boldsymbol{d}_1 = 0$ is

$$\boldsymbol{d}_2 = \alpha_2\{A^{-1} - b_1[A^{-1}]^2\}\boldsymbol{\Delta}$$

where the constant b_1 is

$$b_1 = \frac{\boldsymbol{\Delta}^t(A^{-1})^2\boldsymbol{\Delta}}{\boldsymbol{\Delta}^t(A^{-1})^3\boldsymbol{\Delta}}.$$

Now consider the nth discriminant direction $\boldsymbol{d}_n$. Using the method of Lagrange multipliers, we wish to maximize $R(\boldsymbol{d}_n)$ subject to the constraints that $\boldsymbol{d}_i{}^t\boldsymbol{d}_n = 0$ for $i = 1,2,\cdots,n-1$. Let C be

$$C = \frac{(\boldsymbol{d}_n{}^t\boldsymbol{\Delta})^2}{\boldsymbol{d}_n{}^tA\boldsymbol{d}_n} - \lambda_1\boldsymbol{d}_n{}^t\boldsymbol{d}_1 - \cdots - \lambda_{n-1}\boldsymbol{d}_n{}^t\boldsymbol{d}_{n-1}.$$

Setting the partial of C with respect to $\boldsymbol{d}_n$ equal to zero,

$$\frac{\partial C}{\partial \boldsymbol{d}_n} = 0 \Rightarrow 2K_n\boldsymbol{\Delta} - K_n{}^2 2A\boldsymbol{d}_n - \lambda_1\boldsymbol{d}_1 - \cdots - \lambda_{n-1}\boldsymbol{d}_{n-1} = \boldsymbol{0}$$

where

$$K_n = \frac{\boldsymbol{\Delta}^t\boldsymbol{d}_n}{\boldsymbol{d}_n{}^tA\boldsymbol{d}_n} = \frac{\boldsymbol{d}_n{}^t\boldsymbol{\Delta}}{\boldsymbol{d}_n{}^tA\boldsymbol{d}_n}.$$

Therefore,

$$\boldsymbol{d}_n = \frac{1}{K_n}A^{-1}\left[\boldsymbol{\Delta} - \frac{\lambda_1}{2K_n}\boldsymbol{d}_1 - \cdots - \frac{\lambda_{n-1}}{2K_n}\boldsymbol{d}_{n-1}\right]. \qquad (1)$$

Applying the constraints, and letting $\beta_i = \lambda_i/2K_n$ and $s_{ij} = \boldsymbol{d}_iA^{-1}\boldsymbol{d}_j$, we obtain

$$\begin{array}{l}
 s_{11}\beta_1 + s_{12}\beta_2 + \cdots + s_{1(n-1)}\beta_{n-1} = 1/\alpha_1 \\
 \qquad\qquad\cdot \\
\left.\begin{array}{r}\boldsymbol{d}_1{}^t\boldsymbol{d}_n = 0\\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{d}_{n-1}{}^t\boldsymbol{d}_n = 0\end{array}\right\} \Rightarrow s_{i1}\beta_1 + \cdots + s_{i(n-1)}\beta_{n-1} = 0 \\
 \qquad\qquad\cdot \\
 s_{(n-1)1}\beta_1 + \cdots + s_{(n-1)(n-1)}\beta_{(n-1)} = 0.
\end{array} \qquad (2)$$

Matrix notation can be used to simplify (1) and (2). For (1), let

$$\boldsymbol{d}_n = \alpha_nA^{-1}\{\boldsymbol{\Delta} - [\boldsymbol{d}_1\cdots\boldsymbol{d}_{n-1}]\boldsymbol{\beta}\} \qquad (3)$$

where $\boldsymbol{\beta}^t = [\beta_1\cdots\beta_{n-1}]$ and the constant $(1/K_n)$ has been replaced by the normalizing constant α_n, i.e., α_n is chosen such that $\boldsymbol{d}_n{}^t\boldsymbol{d}_n = 1$. For (2), let

$$S_{n-1}\boldsymbol{\beta} = \begin{bmatrix}1/\alpha_1\\ 0\\ \cdot\\ \cdot\\ \cdot\\ 0\end{bmatrix}\Bigg\updownarrow n-1 \quad\text{or}\quad \boldsymbol{\beta} = S_{n-1}{}^{-1}\begin{bmatrix}1/\alpha_1\\ 0\\ \cdot\\ \cdot\\ \cdot\\ 0\end{bmatrix} \qquad (4)$$

where

$$S_{n-1} = \begin{array}{c}\longleftarrow n-1 \longrightarrow\\ \begin{bmatrix} s_{11} & \cdots & s_{1(n-1)}\\ \cdot & & \cdot\\ \cdot & & \cdot\\ \cdot & & \cdot\\ s_{j1} & \cdots & s_{j(n-1)}\\ \cdot & & \cdot\\ \cdot & & \cdot\\ \cdot & & \cdot\\ s_{(n-1)1} & \cdots & s_{(n-1)(n-1)}\end{bmatrix}\end{array}\Bigg\updownarrow n-1.$$

Combining (3) and (4), a recursive definition for the nth discriminant vector can be obtained:

$$\boldsymbol{d}_n = \alpha_nA^{-1}\left\{\boldsymbol{\Delta} - [\boldsymbol{d}_1\cdots\boldsymbol{d}_{n-1}]S_{n-1}{}^{-1}\begin{bmatrix}1/\alpha_1\\ 0\\ \cdot\\ \cdot\\ \cdot\\ 0\end{bmatrix}\right\}. \qquad (5)$$

The computations involve the inversion of A and S_n. A recursive definition for finding the inverse of S_{n-1} is given in the following equations. A useful identity (e.g., see Rao [9]) for finding the inverse of a symmetric $n \times n$ matrix in terms of an $(n-1) \times (n-1)$ submatrix is given as follows.

Let us partition S_n in the following manner:

$$S_n = \left[\begin{array}{c:c} & s_{1n}\\ & \cdot\\ S_{n-1} & \cdot\\ & \cdot\\ & s_{(n-1)(n)}\\ \hline s_{(n)1}\cdots s_{(n)(n-1)} & s_{nn}\end{array}\right] = \left[\begin{array}{c:c} S_{n-1} & \boldsymbol{y}_n\\ \hline \boldsymbol{y}_n{}^t & s_{nn}\end{array}\right]$$

where $\boldsymbol{y}_n{}^t = [s_{(n)1}\cdots s_{(n)(n-1)}]$ and $n \geq 1$. Then

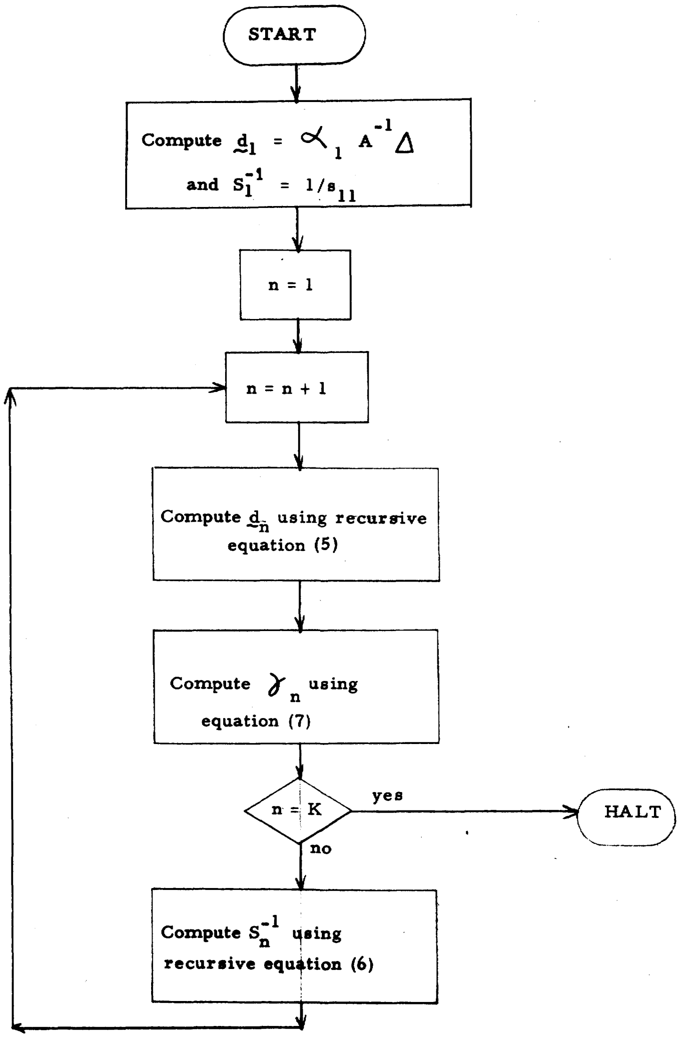

START
Compute $\underline{d}_1 = \alpha_1 A^{-1}\Delta$
and $S_1^{-1} = 1/s_{11}$
n = 1
n = n + 1
Compute $\underline{d}_n$ using recursive equation (5)
Compute γ_n using equation (7)
n = K
yes
no
HALT
Compute S_n^{-1} using recursive equation (6)

Fig. 2. Recursive procedure for computing K best discriminant vectors and discrim-values. Underlined symbols indicate bold-faced symbols in text.

$$S_n^{-1} = \frac{1}{c_n}\left[\begin{array}{c|c} c_nS_{n-1}^{-1} + S_{n-1}^{-1}\boldsymbol{y}_n\boldsymbol{y}_n^tS_{n-1}^{-1} & -S_{n-1}^{-1}\boldsymbol{y}_n \\ \hline -\boldsymbol{y}_n^tS_{n-1}^{-1} & 1 \end{array}\right] \tag{6}$$

where the scalar $c_n = s_{nn} - \boldsymbol{y}_n^tS_{n-1}^{-1}\boldsymbol{y}_n$.

Note that for each discriminant vector $\boldsymbol{d}_n$ there corresponds a "discrim-value" γ_n, given by

$$\gamma_n = \frac{(\boldsymbol{d}_n^t\boldsymbol{\Delta})^2}{\boldsymbol{d}_n^tA\boldsymbol{d}_n}. \tag{7}$$

Each γ_n represents the value of the discriminatory criterion $R(\boldsymbol{d}_n)$ for the corresponding discriminant vector $\boldsymbol{d}_n$. Noting that the discriminant vectors can be ordered according to their respective discrim-values such that

$$\gamma_1 \geq \gamma_2 \geq \gamma_3 \geq \cdots \geq \gamma_L \geq 0$$

the analogy between the eigenvectors and the discriminant vectors is complete. The former describes the best fitting subspace while the latter describes the best discriminating subspace.[2] Since the principal purpose of pattern recognition is discrimination, the discriminant vector subspace offers considerable potential as a feature extraction transformation.

Fig. 2 shows the recursive algorithm for computing the first K discriminant vectors and discrim-values.

V. COMPARISON OF FEATURE SELECTION TECHNIQUES

Comparing data-dependent transformations is a difficult task. One method is to assume a distribution for the data. For multivariate class-conditional normal distributions with a common covariance matrix, it is well known that projecting the data on the first discriminant vector yields the optimal Bayes results. Certainly for this case, the discriminant vector technique offers the best results.

[2] Of course, "best-discriminating subspace" is only optimal with respect to the criteria chosen.

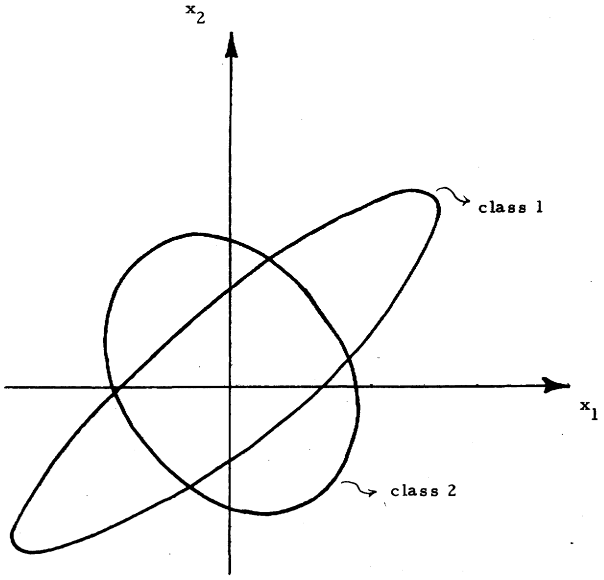


Fig. 3. Illustration of class covariance differences.

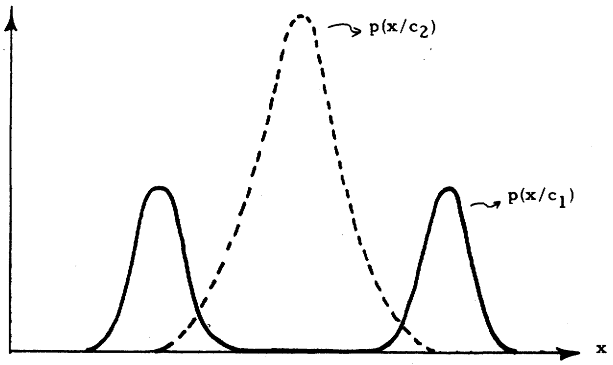


Fig. 4. Illustration of modality problem.

A second method of comparison is via a mathematical example. For the example given in Section III, the Fukunaga–Koontz transform fails to rank the features according to their discriminatory usefulness. However, the first discriminant vector $\boldsymbol{d} = k[\Sigma_1 + \Sigma_2]^{-1}\boldsymbol{\Delta}'$ lies along the second dimension, which is the most significant feature.

The common factor in the two previous methods of comparison is that the majority of the discriminatory information is contained in differences in the mean vectors as opposed to differences in class covariance. For these cases the discriminant vector technique appears superior.

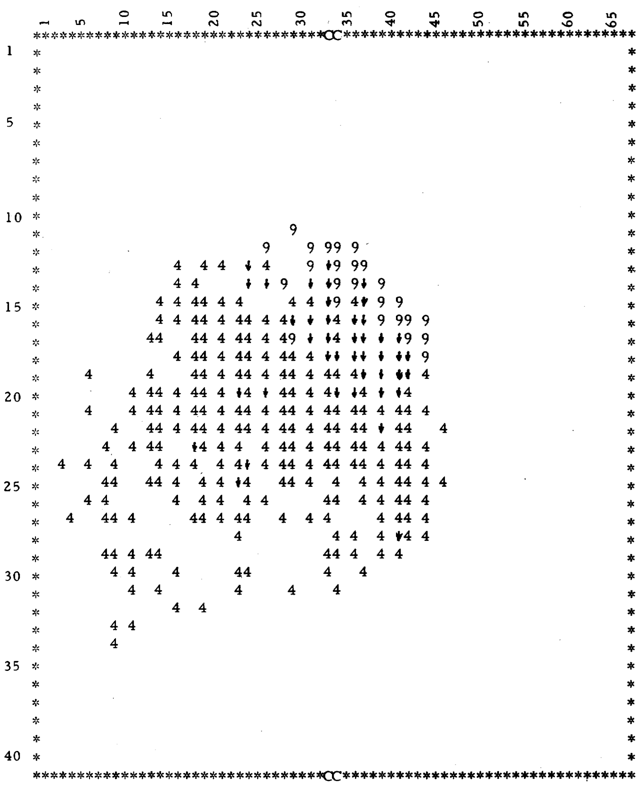
For those cases in which there is no difference in mean vectors and some difference in class covariance, the Fukunaga–Koontz transform is more effective since the direction chosen by the discriminant vector technique would be based only on sample-size "noise." Fig. 3 shows a typical illustration of this case.

Unfortunately, when these cases occur, the probability of error is usually significant. It has been the authors' experience in practical problems to avoid these cases by designing features whose discriminatory information is reflected in mean vector differences.

However, caution should be exercised in using any pattern recognition technique such as the Fisher criterion which inherently employs mean vector differences. This is caused by the modality problem discussed by Sammon [6] and illustrated here in Fig. 4. In this case the means of classes 1 and 2 are identical, yet significant discriminatory information exists along feature x.

To avoid missing this information, the authors have found it necessary to perform a modal analysis and partition each class into its respective modes before performing any logic design.

Another method of comparison is via a data-dependent example. One difficult data base readily available to the authors and also discussed in the open literature (see Sammon [6]) consisted of hand-printed characters for the class of fours and nines. The data base contained 175 fours and 908 nines. Each vector consisted of 40 features. Fig. 5 shows a cluster plot of the data projected on the plane defined by the eigenvectors of the lumped covariance matrix corresponding to the two maximum eigenvalues. A cluster plot is obtained by placing a mesh over the area in the plane spanned by the data. In this figure, each cell in the mesh contains a 4 if only sample vectors from the class of fours are present in the cell, a 9 if only nines are present, and an arrow (↓) if samples from both classes



```
SCALE= 4.06542          E 1/E 2        CLUSTER
X PLOT CENTER= -8.70167          Y PLOT CENTER= -4.54016
```

MULTIPLE POINT DISTRIBUTIONS

Fig. 5. Eigenvector plot.

TABLE I
Multiple Overprints for Eigenvector Plot

Column	Row	Class	# of Samples of Class in Cell	Class	# of Samples of Class in Cell
24	13	4	1	9	1
33	13	4	1	9	1
24	14	4	1	9	1
26	14	4	1	9	1
31	14	4	3	9	1
33	14	4	2	9	3
37	14	4	1	9	1
33	15	4	5	9	3
37	15	4	1	9	2
29	16	4	5	9	1
31	16	4	6	9	2
33	16	4	12	9	1
36	16	4	3	9	2
37	16	4	4	9	4
31	17	4	3	9	1
33	17	4	12	9	1
36	17	4	5	9	3
37	17	4	2	9	2
39	17	4	4	9	14
41	17	4	2	9	9
33	18	4	4	9	1
34	18	4	7	9	1
36	18	4	5	9	1
37	18	4	5	9	3
39	18	4	4	9	7
41	18	4	4	9	14
42	18	4	1	9	5
37	19	4	5	9	6
39	19	4	11	9	6
41	19	4	7	9	4
42	19	4	2	9	1
23	20	4	5	9	1
26	20	4	3	9	1
34	20	4	3	9	1
36	20	4	5	9	1
39	20	4	15	9	1
41	20	4	9	9	1
39	22	4	12	9	1
18	23	4	3	9	1
24	24	4	1	9	1
23	25	4	1	9	1
41	28	4	1	9	1

Note: Only cells containing both classes are shown.

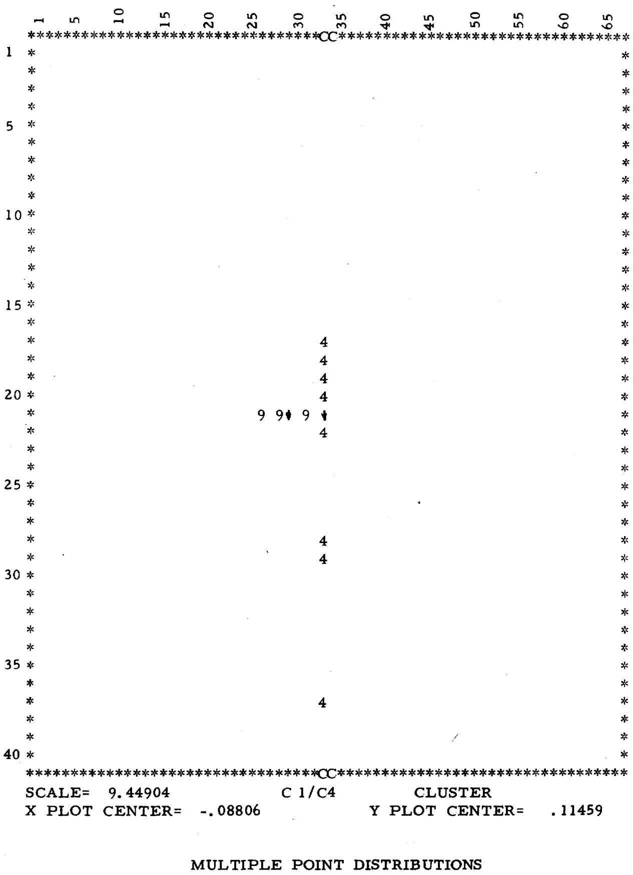


Fig. 6. Fukunaga–Koontz transform.

TABLE II
Multiple Overprints for Fukunaga-Koontz Transform

Column	Row	Class	# of Samples of Class in Cell	Class	# of Samples of Class in Cell
29	21	4	1	9	28
33	21	4	765	9	120

are present. Table I shows a list of cells containing both classes. One logic design procedure which results in a complicated boundary would assign a sample which fell in a cell to the class with the largest number of samples in that cell. Even with this technique, 77 errors would result. (Although this strategy could be used, the authors feel that is an example of overdesigning on the training set.)

Fig. 6 shows a plot of the Fukunaga–Koontz transform using the eigenvectors corresponding to the largest and smallest eigenvalues (essentially 1 and 0). Table II lists the multiple points in each cell. Since this plot may be misleading, it should be noted that the range for the class of nines in the vertical direction is from -0.000007061 to $+0.000004905$. Taking into account the accuracy of the computer, this range is essentially zero. Within this range are 290 fours and, of course, all 175 nines. Using the cell type logic previously mentioned, 121 errors would result.

Fig. 7 shows the data projected on the plane defined by the first and second discriminant vectors, while Table III lists the multiple points in each cell. The vertical boundary shown in the figure would result in only 17 errors. Quite a striking improvement over the two previous methods!

The questions of the type of logic which can be designed and the information contained in higher order discriminant vectors naturally arise. As previously mentioned in

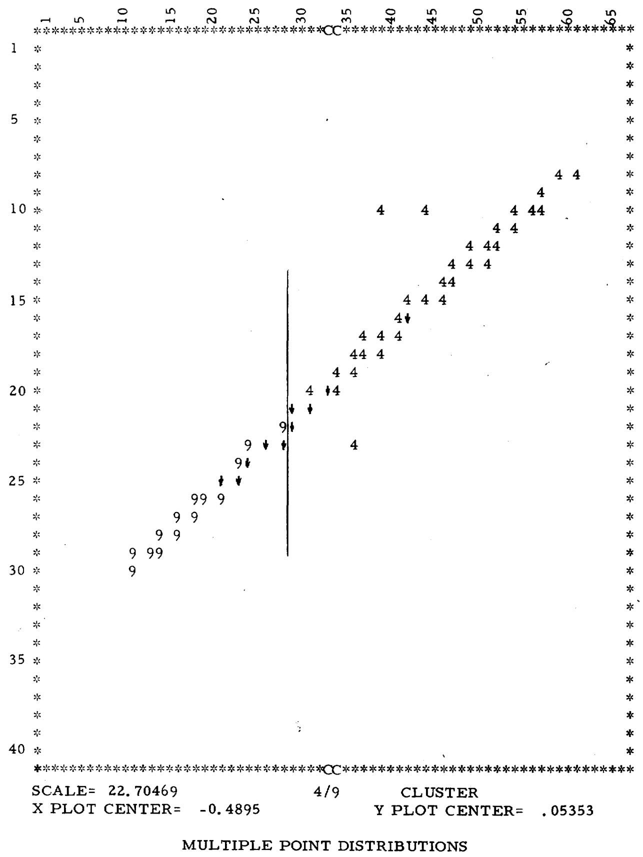


Fig. 7. Projection on first and second discriminant vectors.

TABLE III
MULTIPLE OVERPRINTS FOR FIRST AND SECOND DISCRIMINANT VECTORS

Column	Row	Class	# of Samples of Class in Cell	Class	# of Samples of Class in Cell
42	16	4	89	9	1
33	20	4	19	9	3
29	21	4	2	9	1
31	21	4	16	9	3
29	22	4	7	9	3
26	23	4	2	9	6
28	23	4	1	9	2
24	24	4	1	9	15
21	25	4	1	9	31
23	25	4	1	9	7

the paper, the higher order discriminant vectors can be used to define potentially useful projection planes in such man–machine interactive pattern recognition systems as OLPARS [6]. Fig. 8 contains a projection of this previous data base on the plane defined by the fifth and sixth discriminant vectors. Table IV contains the multiple overprints.

Since the objective of the discriminant vector transformation is to significantly reduce the dimensionality while retaining the discriminatory information, it should be possible to employ sophisticated pattern recognition techniques in the new space that were either computationally impractical or statistically insignificant[3] in the original high dimensional space.

A final justification for selecting the discriminant vectors is the intuitive notion that features based on discrimination should be better than features based on fitting or representing the data.

ACKNOWLEDGMENT

The authors wish to thank A. H. Proctor, Captain

[3] For a discussion of sample size as a function of dimensionality, see Foley [10].

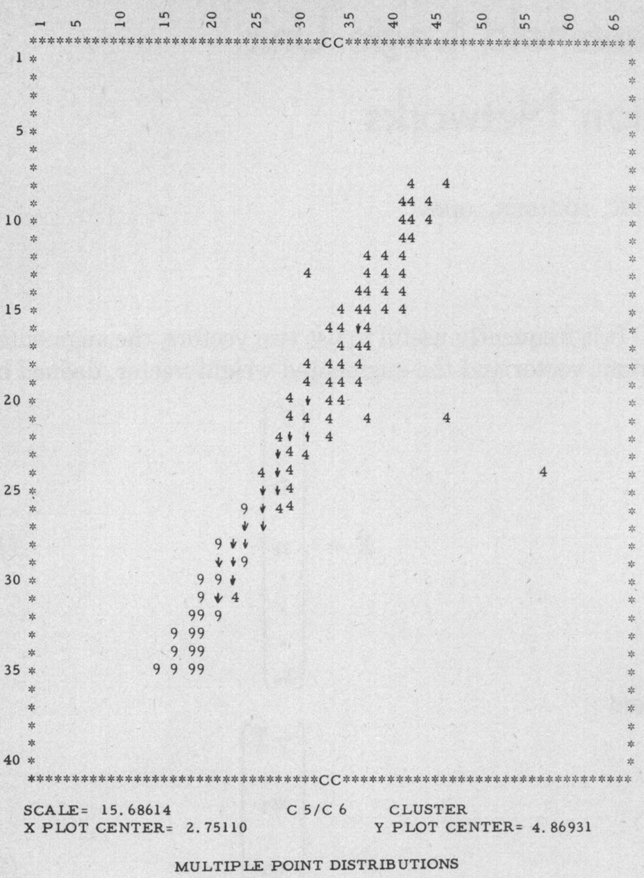


Fig. 8. Projection on fifth and sixth discriminant vectors.

TABLE IV
MULTIPLE OVERPRINTS FOR FIFTH AND SIXTH DISCRIMINANT VECTORS

Column	Row	Class	# of Samples of Class in Cell	Class	# of Samples of Class in Cell
36	16	4	31	9	1
31	20	4	18	9	1
29	22	4	22	9	1
31	22	4	47	9	1
28	23	4	2	9	1
28	24	4	18	9	4
26	25	4	8	9	3
28	25	4	15	9	1
26	26	4	14	9	1
24	27	4	2	9	9
26	27	4	2	9	1
23	28	4	1	9	1
24	28	4	5	9	8
21	29	4	1	9	11
23	29	4	1	9	17
23	30	4	1	9	6
21	31	4	1	9	13

R. D. Bouvier, and G. T. Stevens for their suggestions and programming assistance.

REFERENCES

[1] *IEEE Trans. Comput.* (*Special Issue on Feature Extraction and Selection in Pattern Recognition*), vol. C-20, pp. 967–1117, Sept. 1971.

[2] S. Watanabe, "Karhunen–Loeve expansion and factor analysis," in *Trans. 4th Prague Conf. Information Theory*, 1965.

[3] Y. T. Chien and K. S. Fu, "On the generalized Karhunen–Loève expansion," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-13, pp. 518–520, July 1967.

[4] K. Fukunaga and W. L. G. Koontz, "Application of the Karhunen-Loève expansion to feature selection and ordering," *IEEE Trans. Comput.*, vol. C-19, pp. 311–318, Apr. 1970.

[5] J. W. Sammon, Jr., "An optimal discriminant plane," *IEEE Trans. Comput.* (Short Notes), vol. C-19, pp. 826–829, Sept. 1970.

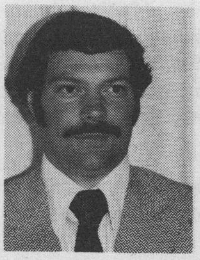
[6] ——, "Interactive pattern analysis and classification," *IEEE Trans. Comput.*, vol. C-19, pp. 594–616, July 1970.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic Press, 1972, ch. 8.

[8] T. W. Anderson and X. X. Bahadur, "Classification into two multivariate normal distributions with different covariance matrices," *Ann. Math. Stat.*, vol. 33, pp. 420–431, 1962.

[9] C. R. Rao, *Linear Statistical Inference and Its Applications*. New York: Wiley, 1965, p. 29.

[10] D. H. Foley, "Considerations of sample and feature size," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 618–626, Sept. 1972.



Donald H. Foley (S'66–M'66) was born in Hartford, Conn., on July 22, 1944. He received the B.S.E.E. and M.S.E.E. degrees from the Worcester Polytechnic Institute, Worcester, Mass., in 1966 and 1968, respectively, and the Ph.D. degree in electrical engineering from Syracuse University, Syracuse, N. T., in 1971.

In 1972 he joined Pattern Analysis and Recognition Corporation, Rome, N. Y., where he is now Vice President. In 1972 he also was appointed an Adjunct Professor in the Department of Systems and Information Sciences, Syracuse University, Syracuse, N. Y. Previous experience includes Rome Air Development Center, Bell Telephone Laboratories, and Sprague Electric. His area of specialization is in the field of pattern recognition, and he has published papers on sample size problems and interactive systems for the extraction and evaluation of features in signal classification problems.

Dr. Foley is a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi. He is a Reviewer for *Computing Reviews*, the IEEE TRANSACTIONS ON COMPUTERS, and the IEEE TRANSACTIONS ON INFORMATION THEORY.

John W. Sammon, Jr. (M'66) was born in Buffalo, N. Y., on February 19, 1939. He received the B.S. degree from the U. S. Naval Academy, Annapolis, Md., the S.M. degree in aeronautics and astronautics from the Massachusetts Institute of Technology, Cambridge, Mass., and the Ph.D. degree in electrical engineering from Syracuse University, Syracuse, N. Y., in 1960, 1962, and 1966, respectively.

From 1962 to 1969 he was Chief of the Computer Applications Section of Rome Air Development Center, Rome, N. Y., where he worked primarily in the areas of graphic system design, adaptive systems, and pattern recognition. He was responsible for the design, implementation, and application of an on-line graphics-oriented system (OLPARS) for solving both pattern recognition and pattern analysis problems. He is the founder and President of Pattern Analysis and Recognition Corporation, Rome, N. Y. Currently he is a member of the visiting staff of Marcy State Hospital and an Adjunct Professor of both Electrical Engineering and Systems and Information Sciences, Syracuse University.

Dr. Sammon is a member of Sigma Gamma Tau and Sigma Tau.