V. Conclusions

In this paper we have presented a smoothing technique which has some interesting scale preserving properties. If ξ and τ are two polygons with τ as a scaled version of ξ, their smoothed versions ξ_ϵ and τ_ϵ are exactly scaled versions of each other and the scale factor is the same provided ϵ obeys a simple inequality constraint. The smoothed version known as the linear minimum perimeter polygon is a method of approximating ξ by ξ_ϵ where the latter has a number of edges less than or equal to the former. It also gives us a single parameter ϵ to control the degree of approximation.

A consequence of this is that we can represent a polygon ξ approximately by a string of real number pairs and this string is invariant to the scale and the coordinate system of ξ.

The use of the LMPP to smooth maps and characters has been demonstrated.

References

[1] H. Freeman, "On the encoding of arbitrary geometric configurations," *IRE Trans. Electron Comput.*, vol. EC-10, pp. 260–268, 1961.

[2] R. L. Kashyap and B. J. Oommen, "A geometrical approach to polygonal dissimilarity and the classification of closed boundaries," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-4, pp. 649–654, 1982.

[3] U. Montanari, "A note on minimal length polygonal approximation to a digitized contour," *Commun. Ass. Comput. Mach.*, vol. 13, pp. 41–47, 1970.

[4] *Nat. Geographic Magazine*, maps of the Great Lakes Region, drawn Dec. 1953.

[5] T. Pavlidis, *Structural Pattern Recognition*. New York: Springer-Verlag, 1977.

[6] J. Sklansky, R. L. Chazin, and B. J. Hansen, "Minimum-perimeter polygons of digitized silhouettes," *IEEE Trans. Comput.*, vol. C-21, pp. 260–268, 1972.

[7] J. Sklansky and D. F. Kibler, "A theory of nonuniformly digitized binary pictures," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, pp. 637–647, 1976.

[8] A. Ramer, "An iterative procedure for the polygonal approximation of plane curves," *Comput. Graphics Image Processing*, pp. 244–256, 1972.

Nonparametric Discriminant Analysis

K. FUKUNAGA AND J. M. MANTOCK

Abstract—**A nonparametric method of discriminant analysis is proposed. It is based on nonparametric extensions of commonly used scatter matrices. Two advantages result from the use of the proposed nonparametric scatter matrices. First, they are generally of full rank. This provides the ability to specify the number of extracted features desired. This is in contrast to parametric discriminant analysis, which for an L class problem typically can determine at most $L-1$ features. Second, the nonparametric nature of the scatter matrices allows the procedure to work well even for non-Gaussian data sets.**

Using the same basic framework, a procedure is proposed to test the structural similarity of two distributions. The procedure works in high-dimensional space. It specifies a linear decomposition of the original data space in which a relative indication of dissimilarity along each new basis vector is provided.

The nonparametric scatter matrices are also used to derive a clustering procedure, which is recognized as a k-nearest neighbor version of the nonparametric valley seeking algorithm. The form which results provides a unified view of the parametric nearest mean reclassification algorithm and the nonparametric valley seeking algorithm.

Index Terms—**Clustering, dimensionality reduction, discriminant analysis, distributional tests, linear mapping, nonparametric feature extraction, scatter matrices.**

I. Introduction

To extract features either linear or nonlinear mappings can be used. In this paper we propose a linear feature extraction procedure. The determination of a linear mapping can be thought of as a problem of finding a rotation and multiplicative scaling of the original data space, followed by the selection of a subspace in which to perform all subsequent work.

Thus, two problems must be solved. First, a procedure to determine a rotation and scaling must be specified. Second, a procedure to rank the rotated and scaled features must be specified to facilitate the selection process. One method that accomplishes these goals simultaneously was proposed in a classic paper by Fisher [1]. Commonly known as discriminant analysis, it is a frequently used feature extraction tool. Since the procedure proposed in this paper has the same structure as is used in discriminant analysis, a brief review is appropriate. Discriminant analysis is based on the use of a criterion J, which has the following form:

$$J = \operatorname{tr} S_2^{-1} S_1 \tag{1}$$

where S_1 and S_2 are square matrices and $\operatorname{tr}(\cdot)$ is the trace operation. The matrices S_1 and S_2 generally indicate the scatter of sample vectors about given mean vectors.

Typically, S_2 represents a matrix that contains either within-class scatter information or class independent scatter information. The within-class scatter matrix is denoted as S_w and is usually computed as

$$S_w = \sum_{i=1}^{L} P(\omega_i)\, \hat{\Sigma}_i \tag{2}$$

where $P(\omega_i)$ is the *a priori* probability of class ω_i, $\hat{\Sigma}_i$ is the sample covariance matrix for class ω_i, and L is the number of classes. The scatter matrix for the entire data set, referred to as the mixture scatter matrix, is denoted by S_m and is usually computed as

$$S_m = \sum_{i=1}^{L} P(\omega_i)\, [\hat{\Sigma}_i + (\hat{M}_i - \hat{M}_0)(\hat{M}_i - \hat{M}_0)^T] \tag{3}$$

where $\hat{M}_i$ is the sample mean vector of class ω_i, and $\hat{M}_0$ is the global sample mean vector. Equation (3) is recognized as the sample covariance matrix of the entire data set.

Typically, S_1 represents between-class scatter, which is denoted as S_b. Many forms have been proposed for S_b [2]. One of the most common multiclass forms is defined as

$$S_b = \sum_{i=1}^{L} P(\omega_i)(\hat{M}_i - \hat{M}_0)(\hat{M}_i - \hat{M}_0)^T. \tag{4}$$

If S_w is used for S_2 and S_b for S_1, a value for J can easily be computed in the original n-dimensional space. Observe that for classes that are tightly grouped about their mean vector

Manuscript received March 11, 1982; revised December 27, 1982. This work was supported in part by the National Science Foundation under Grant ECS-80-05482.

K. Fukunaga is with the Department of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

J. M. Mantock was with the Department of Electrical Engineering, Purdue University, West Lafayette, IN 47907 and the Aerospace Corporation, Los Angeles, CA 90009. He is now with Texas Instruments, Inc., Lewisville, TX 75067.

0162-8828/83/1100-0671$01.00 © 1983 IEEE

and are well separated from each other, J will be large. Since this situation is ideal for classification, we would like to maximize the J that results after mapping down from n-dimensional space to m-dimensional space. It can be shown that this problem is readily solved by finding the eigenvalues and eigenvectors of $S_2^{-1}S_1$ [2]. To maximize J all eigenvectors corresponding to nonzero eigenvalues are chosen. The relative computational simplicity of this procedure is largely responsible for its popularity. However, if we refer to (4) it is seen that S_b can have at most $L - 1$ nonzero eigenvalues. Thus, for the two class case, $S_2^{-1}S_1$ will have at most one nonzero eigenvalue. This will result in only a single feature being extracted.

The fact that the algorithm produces only $L - 1$ extracted features does not guarantee acceptable performance of the extracted features. If error estimates establish that more features need to be extracted, some method must be devised to augment the feature extraction process. One possibility is to artificially increase the number of classes. In this way we can increase the rank of S_b. This could be accomplished by dividing each class into a number of clusters. For those cases where multimodal behavior is present, and a clustering algorithm can be found that "properly" identifies the clusters, this may work well. As a second possibility, after determining the L-1 features, one could remove them leaving a subspace orthogonal to the extracted features. A similar procedure could then be applied to the subspace to extract additional features. Foley and Sammon [3] followed this approach for the two class case.

A more fundamental problem is the parametric nature of (2), (3), and (4). If the distributions are significantly non-Gaussian, the use of such forms cannot be expected to accurately indicate which features should be extracted to preserve any complex structure that might be needed for classification.

In this paper a nonparametric form of discriminant analysis is presented which overcomes both of the previously mentioned problems. The basis of the extension is a nonparametric S_b. It measures between-class scatter on a local basis, using k-nearest neighbor (k-NN) techniques, and is generally of full rank. As a result, neither artificial class generation, nor sequential methods are necessary.

The nonparametric discriminant analysis and experimental results are detailed in Section II. Use of the algorithm for linear classifier design is presented in Section III. In Section IV we propose a method to compare two distributions. Based on a simple modification of the functional form of our nonparametric J, the procedure allows one to test the structural similarity of two distributions. The solution also indicates along which axes the greatest differences in distribution occur. In Section V the nonparametric form for J is used to derive the nonparametric valley-seeking algorithm for clustering proposed by Fukunaga and Koontz [4]. It is shown that the nearest mean clustering algorithm (used in ISODATA [5]) and the valley seeking algorithm are very closely related. Section VI contains a summary.

II. The Nonparametric Feature Extraction Algorithm

In this and the following section only the two class problem will be discussed for simplicity.

We begin by examining a slightly different parametric form for S_b

$$S_b' = \frac{P(\omega_1)}{N_1}\sum_{l=1}^{N_1}(X_l - \hat{M}_2)(X_l - \hat{M}_2)^T + \frac{P(\omega_2)}{N_2}\sum_{l=N_1+1}^{N_1+N_2}(X_l - \hat{M}_1)(X_l - \hat{M}_1)^T \quad (5)$$

where N_1 and N_2 are the number of samples in classes ω_1 and ω_2, respectively, and the measurements X_l are ordered so that $X_l \in \omega_1$ for $l = 1, \cdots, N_1$ and $X_l \in \omega_2$ for $l = N_1 + 1, \cdots, N_1 + N_2$. Equation (5) measures the scatter of the class ω_1 measurements about the class ω_2 sample mean and the class ω_2 measurements about the class ω_1 mean. Note that (5) can be simplified to

$$S_b' = P(\omega_1)\hat{\Sigma}_1 + P(\omega_2)\hat{\Sigma}_2 + (\hat{M}_1 - \hat{M}_2)(\hat{M}_1 - \hat{M}_2)^T. \quad (6)$$

It is shown in the Appendix that S_b' and the S_b defined in (4) result in the same eigenvectors when solving (1).

Returning to (5), observe that mean vectors are being used represent global information about each class. Instead of using $\hat{M}_i$ we propose to use $\boldsymbol{M}_i^k(X_l)$, defined as

$$\boldsymbol{M}_i^k(X_l) = \frac{1}{k}\sum_{j=1}^{k} XNN_{il}^j \quad (7)$$

where XNN_{il}^j is the jth NN from class ω_i to sample X_l. That is, $\boldsymbol{M}_i^k(X_l)$ is the mean vector of the k-NN's from class ω_i to sample X_l. The resulting nonparametric form, denoted as $\boldsymbol{S}_{bk}$, is

$$\boldsymbol{S}_{bk} = \frac{P(\omega_1)}{N_1}\sum_{l=1}^{N_1}(X_l - \boldsymbol{M}_2^k(X_l))(X_l - \boldsymbol{M}_2^k(X_l))^T + \frac{P(\omega_2)}{N_2}\sum_{l=N_1+1}^{N_1+N_2}(X_l - \boldsymbol{M}_1^k(X_l))(X_l - \boldsymbol{M}_1^k(X_l))^T. \quad (8)$$

Note that as k is increased to N_i, $\boldsymbol{M}_i^k(X_l)$ converges to $\hat{M}_i$. Thus, (8) is a generalization of (5). Our primary concern, however, in this paper is the use of $\boldsymbol{S}_{bk}$ with small values of k. Further understanding of $\boldsymbol{S}_{bk}$ is obtained by examining the vector $(X_l - \boldsymbol{M}_i^k(X_l))$. Fukunaga and Hostetler [6] have shown that it points in the direction of the local gradient of the class ω_i density function at X_l. This is an encouraging result, because Short and Fukunaga [7] have shown that the optimal metric for k-NN classification is obtained by projecting samples down onto a local gradient. This indicates that the vectors we are using to compute $\boldsymbol{S}_{bk}$ are precisely the vectors necessary to preserve classification structure.

This viewpoint suggests that careful attention should be paid to the magnitude of each gradient direction term, $(X_l - \boldsymbol{M}_i^k(X_l))$. Samples which are far away from the decision boundary tend to have gradient direction terms with large magnitudes. These large magnitudes can exert a considerable influence on the resulting eigenvalue and eigenvector determination. Since we are primarily concerned with preserving classification structure, some method of deemphasizing samples far from the classification boundary seems appropriate. To accomplish this we propose to use a weighting function for each $(X_l - M_i^k(X_l))(X_l - M_i^k(X_l))^T$. The value of the weighting function, denoted as w_l, is defined as

$$w_l = \frac{\min\{d^\alpha(X_l, XNN_{1l}^k), d^\alpha(X_l, XNN_{2l}^k)\}}{d(X_l, XNN_{1l}^k)^\alpha + d(X_l, XNN_{2l}^k)^\alpha} \quad (9)$$

where α is a control parameter between zero and infinity, and $d(X_l, XNN_{il}^k)$ is the Euclidean distance from X_l to its k-NN from class ω_i. Observe that if α is selected as n, w_l corresponds to the k-NN risk estimate of Fukunaga and Hostetler [8].

This weighting function has the property that near the classification boundary it takes on values close to 0.5 and drops off to 0.0 as we move away from the classification boundary. The control parameter α adjusts how rapidly w_l falls to 0.0 as we move away.

The final form for S_{bk} is as follows:

$$S_{bk} = \frac{1}{N}\sum_{l=1}^{N_1} w_l(X_l - M_2^k(X_l))(X_l - M_2^k(X_l))^T + \frac{1}{N}\sum_{l=N_1+1}^{N_1+N_2} w_l(X_l - M_1^k(X_l))(X_l - M_1^k(X_l))^T \quad (10)$$

where $N = N_1 + N_2$, and the $P(\omega_i)$ are replaced by N_i/N.

We now turn to the choice of S_2. Recall that S_2 typically represents within-class or mixture scatter properties. As such, the term S_2^{-1} in (1) can be viewed as a "global" normalization. This is more readily seen if the maximization of (1) is performed as a two step process (we assume either S_w or S_m is chosen for S_2 and it is of full rank). First, find a linear transformation that maps S_2 to the identity matrix in the transformed space. This is really done using the whitening algorithm presented in [2]. Compute S_b in the normalized space. Second, find the eigenvalues and eigenvectors of transformed space S_b. Order the eigenvectors so that their associated eigenvalues are in decreasing order. The first m eigenvectors are selected as a basis for computing the m extracted features. Subsequent samples are transformed using the linear normalization transform and then projected onto each of the m eigenvectors to form the new m-dimensional feature vector. Observe that in the first step the data are transformed so that the weighted sum of the sample class covariance matrices is the identity matrix for $S_2 = S_w$, or the sample global covariance matrix is the identity matrix for $S_2 = S_m$. Hence, we view S_2^{-1} as a "global" normalization. Viewing the feature extraction process as a two step procedure with normalization is particularly appealing when one recalls that the k-NN determination for S_{bk} is done using the Euclidean metric. Intuitively, one would probably prefer to apply the Euclidean metric to data whose covariance matrix was the identity matrix. However, transforming two data sets simultaneously so that both have $\Sigma_i = I$ is generally not possible. As a compromise, we propose to transform the data so that $S_w = I$. S_{bk} is then computed in the transformed space.

We now present the algorithm in its entirety.

1) Linearly transform the data so that S_w is the identity matrix in the transformed space. This is accomplished using the $n \times n$ matrix

$$A = \Lambda^{-1/2}\Phi^T$$

where Λ is a $n \times n$ diagonal matrix whose diagonal elements are the eigenvalues of S_w and $\Phi = [\phi_1 \cdots \phi_n]$ where ϕ_i is the ith eigenvector of S_w corresponding to the ith eigenvalue of S_w [2].

2) Select k and α. Compute S_{bk} in (10) in the normalized space using w_l for weighting.

3) Find the eigenvalues and eigenvectors of S_{bk}.

4) Order the eigenvectors so that the corresponding eigenvalues are arranged from largest to smallest.

5) Select the desired number of eigenvectors based either on classification requirements or the magnitudes of the eigenvalues. Denote the selected eigenvectors as ψ_i $i = 1, \cdots, m$ and their eigenvector matrix as $\Psi = [\psi_1 \cdots \psi_m]$.

6) The m extracted features in the transformed space are formed by projecting onto the m selected eigenvectors. Thus, the full transformation from the original n-dimensional space to the m extracted features is $\Psi^T\Lambda^{-1/2}\Phi^T$.

Note that the eigenvectors are computed in the normalized space. In some situations, particularly for testing of the algorithm, it is preferable to express the eigenvectors in the unnormalized data space. This is easily accomplished by observing that the jth extracted feature is formed by projecting onto $\psi_j^T\Lambda^{-1/2}\Phi^T$ in step 6). Thus, the eigenvector expressed in unnormalized space is $\eta_j = \Phi\Lambda^{-1/2}\psi_j$.

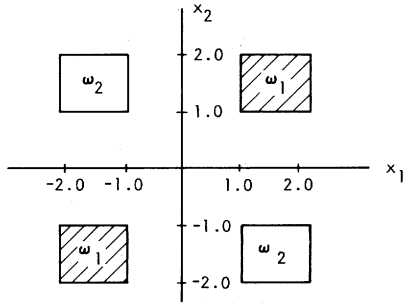

Fig. 1. Distribution of the two classes along x_1 and x_2 in the first feature extraction experiment.

The experiments presented in the remainder of this section were performed with $k = 3$ and $\alpha = 2$.

A. Experiment 1

For the first experiment two groups of three-dimensional data were generated. The first two measurements were generated using random numbers distributed uniformly as shown in Fig. 1. A Gaussian distribution with zero mean and unit variance was used for both classes to generate the third measurement. Each class was comprised of 100 vectors. After applying the nonparametric feature extraction procedure the following results were obtained:

$$\lambda_1 = 0.56, \eta_1 = \begin{bmatrix} -0.59 \\ -0.29 \\ -0.03 \end{bmatrix}, \lambda_2 = 0.40, \eta_2 = \begin{bmatrix} 0.28 \\ -0.58 \\ -0.05 \end{bmatrix},$$

and

$$\lambda_3 = 0.04, \eta_3 = \begin{bmatrix} 0.03 \\ -0.07 \\ 1.00 \end{bmatrix} \quad (11)$$

where the eigenvalues have been normalized so that $\lambda_1 + \lambda_2 + \lambda_3 = 1.0$ and the η_i have been normalized so that $\|\eta_i\| = 1.0$. The eigenvalues clearly indicate only two features are needed. Furthermore, it is observed that both η_1 and η_2 effectively exclude the third measurement, which possessed no structure that would assist in classification.

B. Experiment 2

The second experiment used data sets that were obtained by time sampling waveforms which had random parameters. The time signals were represented in eight-dimensional space by sampling the waveforms at eight uniformly spaced times. The total time interval used for both classes was

$$0.0 \leqslant t \leqslant 1.05 \quad (12)$$

with 250 samples being generated for each class. The first class was obtained by sampling a Gaussian pulse given by

$$x(t) = a_g \exp\{-(t - m_g)^2/2\sigma_g^2\} \quad (13)$$

where the random parameters a_g, m_g, and σ_g^2 were distributed uniformly as

$$0.7 \leqslant a_g \leqslant 1.3$$
$$0.3 \leqslant m_g \leqslant 0.7$$
$$0.2 \leqslant \sigma_g \leqslant 0.4. \quad (14)$$

The second class was obtained using a double exponential pulse given by

$$x(t) = a_e \exp\{-|t - m_e|/\tau_e\} \quad (15)$$

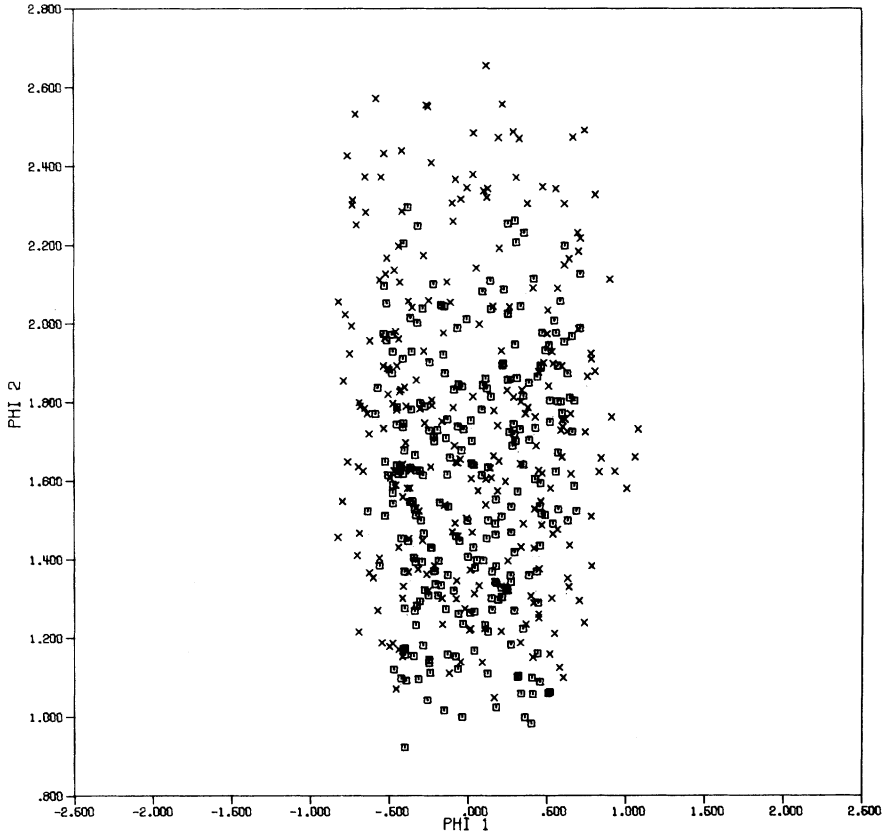

Fig. 2. Plot resulting from projection onto two dominant eigenvectors of the Karhunen–Loeve expansion.

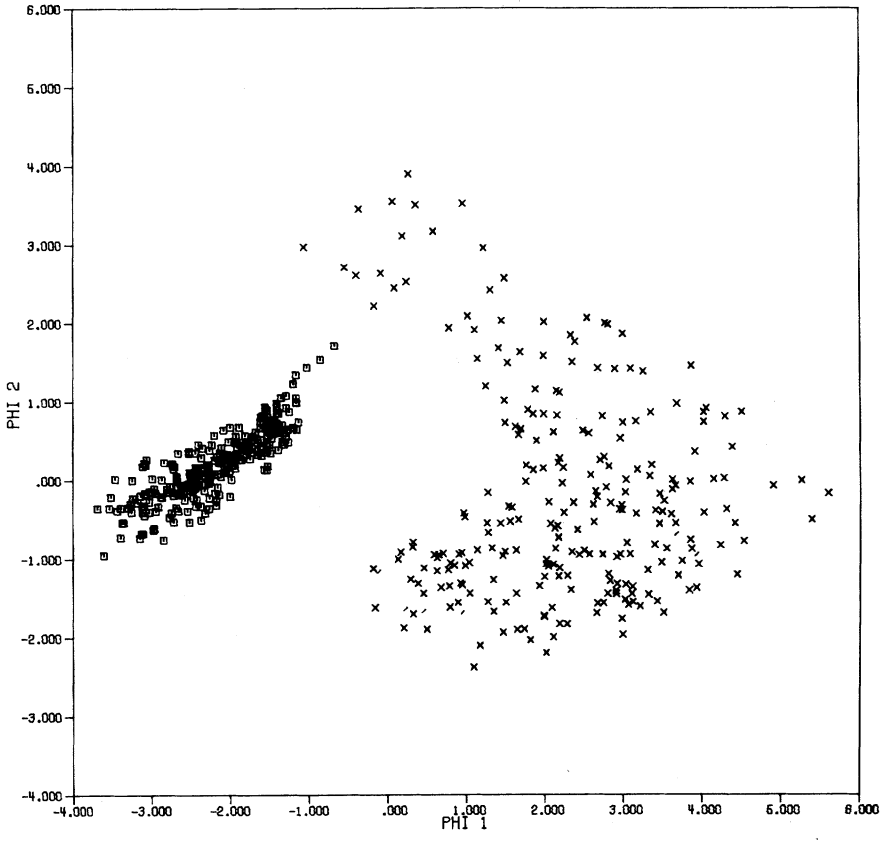

Fig. 3. Plot resulting from projection onto two dominant eigenvectors of the nonparametric discriminant analysis procedure.

where the random parameters a_e, m_e and τ_e were distributed uniformly as

$$0.7 \leqslant a_e \leqslant 1.3$$

$$0.3 \leqslant m_e \leqslant 0.7$$

$$0.3 \leqslant \tau_e \leqslant 0.6. \tag{16}$$

Our experience with these data sets has indicated that the samples in each class are distributed as a warped three-dimensional surface in eight-dimensional space. To provide a display of the data we grouped both classes together, computed the Karhunen–Loeve expansion [2], and projected the data onto the two eigenvectors corresponding to the two largest eigenvalues. This result is presented in Fig. 2. Obviously, this linear mapping does a poor job of feature extraction for classification. In Fig. 3 the result of the nonparametric feature extraction is presented. The normalized eigenvalues were 0.45, 0.18, 0.11, 0.08, 0.06, 0.05, 0.04, and 0.0. Fig. 3 clearly shows that the first extracted feature embodies the majority of classification structure, as its 0.45 eigenvalue indicates. The second extracted feature, while not effective if used alone (as the eigenvalue of 0.18 suggests), establishes that the two classes are separable by a quadratic form.

III. Linear Classifier Design

Linear classifier design is a special case of feature extraction, involving the selection of a linear mapping to reduce the dimensionality to one. Classification is then performed by specifying a single threshold.

When $L = 2$, parametric discriminant analysis always results in the extraction of a single feature. Thus, parametric discriminant analysis for the two class case is essentially linear classifier design.

For linear classifier design using the nonparametric procedure we select the eigenvector corresponding to the largest eigenvalue as our linear transformation.

A. Experiment 3

The third experiment was two-dimensional with the classes distributed uniformly in the areas shown in Fig. 4. It was assumed both classes were equally likely and incorrect decisions were of equal cost. Clearly, the minimum error rate of 0.0 percent is obtained by selecting a mapping that retains only the first feature x_1. To provide a comparison with nonparametric discriminant analysis two other mapping algorithms were tried, the Peterson–Mattson procedure [9] and parametric discriminant analysis.

The Peterson–Mattson procedure finds the optimal linear classifier in the minimum cost sense, assuming that the data are Gaussianly distributed. The population parameters of the distributions were used as input, providing theoretical values for the mapping. Given the mapping, along with our knowledge of the true distribution of the data, a theoretical minimum error of 7.4 percent was achievable using a single threshold.

The population parameters were also used in the parametric discriminant analysis procedure. The resultant linear mapping produced a theoretical minimum error of 7.6 percent.

The nonparametric discriminant analysis procedure was then run with $k = 3$ and $\alpha = 2.0$. Since theoretical values of the nonparametric scatter matrix are not easily determined, a Monte Carlo approach was used. Fifty trials were run using 100 samples per class. The mean value of 50 error estimates was 5.1 percent with a standard deviation of 0.8 percent. Note that each error estimate was determined by mapping the true distribution onto the estimated feature vector and then computing the minimum error.

IV. Distributional Testing

Another application of the proposed nonparametric scatter matrices is in the testing of structural similarity of two distributions. The ability to compare two distributions has numerous applications. One can gather data sets at different times or under different conditions. Distributional testing can then be used to determine what invariances exist. For those cases where only a single feature is used, distributional testing can be solved in a fairly straightforward manner. However, multiple features considerably complicate the testing.

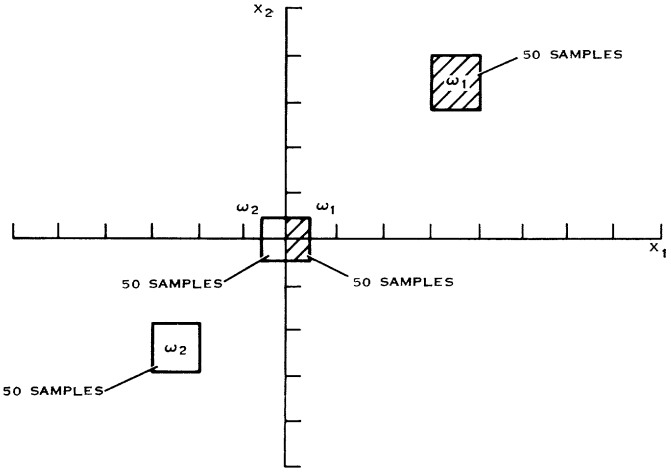

Fig. 4. Distributions of the two classes in the linear classifier design experiment.

One method of testing the similarity of two distributions in a high-dimensional space is to compare the mean vectors and covariance matrices of the distributions. Unfortunately, this method is only necessarily meaningful when both distributions are jointly normal.

A second alternative is to estimate the degree of overlap of the two distributions. This could be estimated using the NN error count, Bhattacharyya distance, divergence, etc. This method fails to indicate the subspace in which the differences are most prominent, or what type of differences exist.

Our proposed test requires no distributional assumptions, and produces an eigenvalue and eigenvector decomposition that is ranked by distributional similarity.

To develop the test we first separate S_{bk} into two parts:

$$S_{bk1} = \frac{1}{N_1} \sum_{l=1}^{N_1} w_l (X_l - M_2^k(X_l))(X_l - M_2^k(X_l))^T$$

and

$$S_{bk2} = \frac{1}{N_2} \sum_{l=N_1+1}^{N_1+N_2} w_l (X_l - M_1^k(X_l))(X_l - M_1^k(X_l))^T. \tag{17}$$

We can interpret S_{bki} as a nonparametric between-class scatter matrix, computed with respect to (w.r.t.) class ω_i. In addition, we will define two nonparametric within-class scatter matrices, denoted by S_{wk1} and S_{wk2}, as

$$S_{wk1} = \frac{1}{N_1} \sum_{l=1}^{N_1} w_l (X_l - M_1^k(X_l))(X_l - M_1^k(X_l))^T$$

and

$$S_{wk2} = \frac{1}{N_2} \sum_{l=N_1+1}^{N_1+N_2} w_l (X_l - M_2^k(X_l))(X_l - M_2^k(X_l))^T. \tag{18}$$

The only difference between S_{wki} and S_{bki} is the class of the local mean used in the computation. If the two distributions are identical, it is expected that $S_{wk1} \simeq S_{bk1}$ and $S_{wk2} \simeq S_{bk2}$. This suggests that the matrix products $S_{wk1}^{-1} S_{bk1}$ and $S_{wk2}^{-1} S_{bk2}$ should be close to I. To reduce the number of comparisons from n^2 to n for each matrix product we can diagonalize

TABLE I
EIGENVALUE RESULTS (FIFTY TRIALS)

	ω_1–Gaussian ω_2–Gaussian				ω_1–Gaussian ω_2–Uniform			
	$S_{wk1}^{-1} S_{bk1}$ λ_1	λ_2	$S_{wk2}^{-1} S_{bk2}$ λ_1	λ_2	$S_{wk1}^{-1} S_{bk1}$ λ_1	λ_2	$S_{wk2}^{-1} S_{bk2}$ λ_1	λ_2
Mean	0.91	0.65	0.94	0.67	1.31	0.99	1.12	1.02
Standard Deviation	0.10	0.12	0.18	0.09	0.41	0.38	0.46	0.40

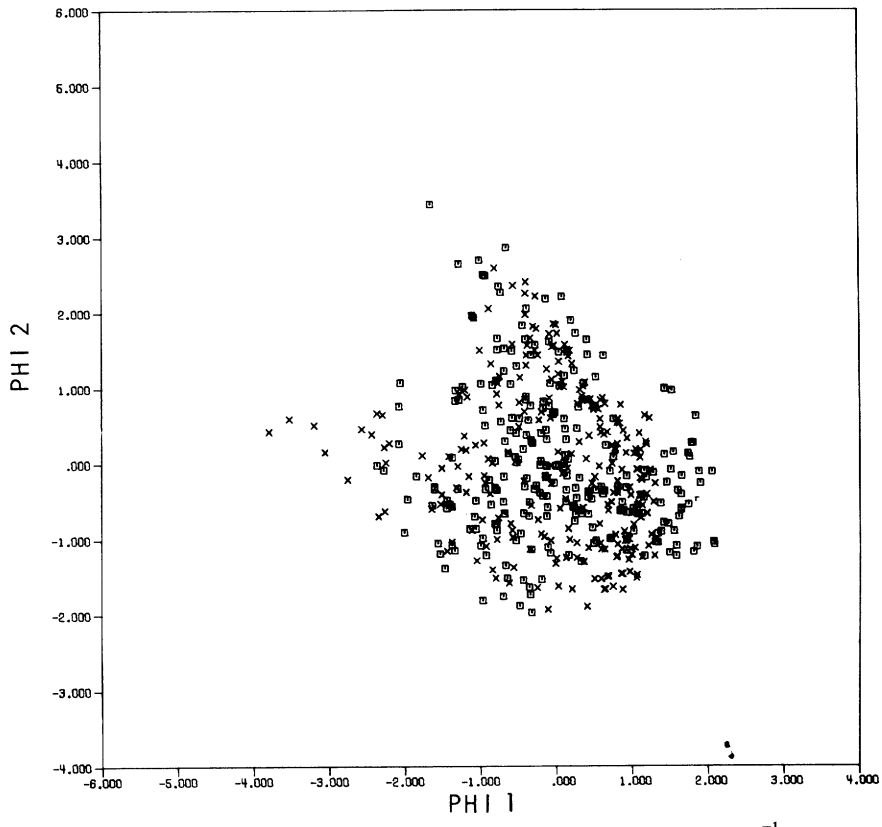

Fig. 5. Plot resulting from projection onto two dominant eigenvectors of $S_{wk1}^{-1} S_{bk1}$.

$S_{wk1}^{-1} S_{bk1}$ and $S_{wk2}^{-1} S_{bk2}$. The n diagonal elements of each matrix can then be compared to 1.0.

Before presenting the experimental results a final note about S_{wki}. For the sample X_l, when the local mean of its own class is computed, we do not consider X_l as its own NN.

A. Experiment 4

The fourth experiment consisted of two parts. First, two Gaussian distributions with mean vectors equal to the zero vector and covariance matrices equal to the identity matrix were compared. The distributions were two-dimensional. One hundred samples per class were generated, k was chosen as three, and α was chosen as zero, i.e., $w_l = 1.0$ or no weighting. $S_{wk1}^{-1} S_{bk1}$ and $S_{wk2}^{-1} S_{bk2}$ were computed, and their eigenvalues determined using 50 independent trials. The mean value and standard deviation of the eigenvalues were then determined. The results are summarized in Table I. The fact that the eigenvalues are less than one is not particularly surprising. Recall that when we computed S_{wki} it was necessary to exclude the sample X_l from our k-NN determination. As such, this introduces a bias into the resultant distance to the local mean, causing it to be larger than one actually expects. This increased magnitude, when inverted, will reduce the measured eigenvalues.

To complete the experiment a second comparison was performed. A Gaussian distribution was compared with a uniform distribution, both two-dimensional with mean vector equal to the zero vector and covariance matrix equal to the identity matrix. As before 100 samples per class were used, k was chosen as three, and α was chosen as zero. The mean value and standard deviation results of the eigenvalue calculations are presented in Table I. When compared to the Gaussian versus Gaussian results, a distributional difference is clearly evident.

B. Experiment 5

In the fifth experiment the time sampled Gaussian pulse was compared with the time sampled double exponential pulse. Refer to Section II for additional information about these data sets.

To provide a reference two eight-dimensional Gaussian distributions (not time sampled Gaussian pulses) were generated, both with mean vector equal to the zero vector and covariance matrix equal to the identity matrix. The resulting eigenvalues were 1.36, 1.21, 1.11, 1.06, 0.89, 0.88, 0.79, and 0.57 for $S_{wk1}^{-1} S_{bk1}$ and 1.29, 1.19, 1.11, 1.08, 0.93, 0.81, 0.70, and 0.57 for $S_{wk2}^{-1} S_{bk2}$ for a single trial.

To assure that we would be testing for structural differences, both time sampled data sets were independently whitened, i.e., sample mean vector transformed to the zero vector and sample covariance matrix transformed to the identity matrix. When

the whitened time sampled data sets were compared the eigenvalues were 34.4, 17.3, 14.3, 10.6, 6.6, 9, 2.7, and 1.4 for $S_{wk1}^{-1}S_{bk1}$ and 0.87, 0.75, 0.67, 0.52, 0.41, 0.32, 0.23, and 0.14 for $S_{wk2}^{-1}S_{bk2}$. These results clearly indicate that significant distributional differences exist. In addition they indicate why the $S_{wki}^{-1}S_{bki}$ should not be combined. It is possible that if they are combined the eigenvectors of the result may not exhibit the same level of discrimination. This is due to the fact that the eigenvalues are averaged in some fashion.

As well as having the ability to test distributional differences, the axes of major difference can be plotted, if the eigenvectors are computed. This is shown in Fig. 5 where we have projected the data down on the two eigenvectors of $S_{wk1}^{-1}S_{bk1}$ with corresponding eigenvalues, 34.4 and 17. The plot shows differences in the structure of the distributions.

V. Nonparametric Clustering

It is also possible to use the proposed nonparametric scatter matrices in clustering. In fact we will show in this section that the parametric nearest mean reclassification algorithm [2] and the nonparametric valley seeking algorithm [4] can be derived from the same basic form for discriminant analysis, the only difference being the choice of k.

We begin by describing the basis of a conventional parametric clustering technique, known as the nearest mean reclassification algorithm. With the number of classes specified *a priori* this algorithm seeks a class assignment of available samples that minimizes the criterion

$$J = \text{tr } S_m^{-1}S_w. \tag{19}$$

That is, we seek to achieve the "smallest" within-class scatter matrix, normalized by the mixture scatter matrix (which is class membership independent).

The solution of this minimization problem is fairly straightforward, if iterative techniques are used. The result is, that at each iteration, every sample is assigned to the class of the nearest mean vector. For additional details refer to [2].

To derive a nonparametric counterpart we will use the criterion

$$J = \text{tr } S_m^{-1}S_{wk} \tag{20}$$

where the multiclass version of S_{wk} is defined as

$$S_{wk} = \sum_{i=1}^{L} \frac{P(\omega_i)}{N_i} \sum_{X_l \in \omega_i} (X_l - M_i^k(X_l))(X_l - M_i^k(X_l))^T. \tag{21}$$

Observe that we do not include any weighting coefficients. Since the w_l defined earlier require knowledge of the true class membership of samples, the use of weighting coefficients is deemed inappropriate for clustering. If one compares S_{wk} with S_w, it is seen that the only difference is the use of $M_i^k(X_l)$ in S_{wk} and $\hat{M}_i$ in S_w. One might surmise then, that the only differences between the nonparametric clustering procedure and the nearest mean reclassification algorithm is the choice of k and the use of local mean vectors instead of global class mean vectors. In fact this conjecture is entirely correct. The details of the solution exactly follow the parametric clustering counterpart.

The nonparametric clustering algorithm that results is as follows.

1) Whiten the data set so that $S_m = I$ in the transformed data space (as discussed in Section II).

2) Select the number of clusters L, and choose k. Set the iteration counter l to zero, and choose an initial classification, denoted as $\Omega(0)$.

3) For every sample compute its class ω_i local mean vector for $i = 1, \cdots, L$. Find the closest local mean vector. If the class of the sample differs from the class of the closest local mean vector, flag the sample for transfer to the closest local mean vector's class.

4) Check all of the samples for flags and change the class assignments of those with flags. This forms $\Omega(l+1)$. If no flags are found, stop. Otherwise, set $l = l + 1$ and go to 3).

If $k \ll N_i$, the nonparametric clustering algorithm proposed is recognized as a k-NN version of the nonparametric valley seeking algorithm of Fukunaga and Koontz [4]. Note that, as for the feature extraction problem, by increasing k the nonparametric procedure is transformed into a parametric procedure. That is, when $k = N_i$, the above procedure becomes the parametric nearest mean reclassification algorithm.

Unfortunately, it is not clear what advantages are obtained by selecting intermediate values of k. Furthermore, if one were to use intermediate values of k, it is not clear what role the ordering of samples should play. Our choice of ordering by NN distance represents only one possible ordering scheme.

VI. Summary

A method was proposed to compute nonparametric scatter matrices. It was shown that they offer advantages over the parametric versions used in discriminant analysis. Specifically, the resultant linear mapping is designed to preserve classification structure, the degree of dimensionality reduction is easily controlled by the user, and each feature has a scalar associated with it that can be used as a relative indicator of separability.

An extension of the nonparametric feature extraction procedure was presented. It allows the user to test the dissimilarity of two distributions. In addition to providing an indication of dissimilarity, it also computes basis vectors that are ranked by dissimilarity.

Finally, a nonparametric clustering technique was presented, based on a derivation used in determining the parametric nearest mean clustering algorithm. The resultant nonparametric clustering technique was recognized as a valley-seeking clustering algorithm. By varying one of the control parameters, a connection between parametric and nonparametric procedures was made evident. This allows both the parametric and nonparametric procedure to be viewed in a more unified framework.

Appendix

Given that $S_2 = S_w$, we seek to prove that the features extracted, when using the criterion $J = \text{tr } S_2^{-1}S_1$, are identical for

$$S_1 = S_b = P(\omega_1)(\hat{M}_1 - \hat{M}_0)(\hat{M}_1 - \hat{M}_0)^T + P(\omega_2)(\hat{M}_2 - \hat{M}_0)(\hat{M}_2 - \hat{M}_0)^T \tag{A1}$$

and

$$S_1 = S_b' = P(\omega_1)\hat{\Sigma}_1 + P(\omega_2)\hat{\Sigma}_2 + (\hat{M}_1 - \hat{M}_2)(\hat{M}_1 - \hat{M}_2)^T \tag{A2}$$

where $\hat{M}_0$, $\hat{M}_1$, and $\hat{M}_2$ are the global, class ω_1 and class ω_2 sample mean vectors, respectively, and $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are the sample covariance matrices for classes ω_1 and ω_2, respectively.

Proof: Observe that both S_b and S_b' are independent of the location of the global sample mean vectors $\hat{M}_0$. As such, we will assume, without loss of generality, that $\hat{M}_0 = 0$. Thus $P(\omega_1)\hat{M}_1 + P(\omega_2)\hat{M}_2 = 0$ and

$$\hat{M}_2 = -\frac{P(\omega_1)}{P(\omega_2)}\hat{M}_1. \tag{A3}$$

Substituting (A3) into (A1) we get

$$S_b = \frac{P(\omega_1)}{P(\omega_2)}\hat{M}_1\hat{M}_1^T. \tag{A4}$$

The resulting form for the criterion using $S_1 = S_b$ is

$$J = \frac{P(\omega_1)}{P(\omega_2)} \text{ tr } S_w^{-1}\hat{M}_1\hat{M}_1^T. \tag{A5}$$

Since constant multipliers have no effect on the *feature extraction* process, we can equally well express the criterion as

$$J = \text{tr } S_w^{-1}\hat{M}_1\hat{M}_1^T. \tag{A6}$$

Observe that (A2) can also be expressed as

$$S_b' = S_w + (\hat{M}_1 - \hat{M}_2)(\hat{M}_1 - \hat{M}_2)^T. \tag{A7}$$

Substituting (A3) into (A7) we obtain

$$S_b' = S_w + \frac{1}{P(\omega_2)^2}\hat{M}_1\hat{M}_1^T. \tag{A8}$$

Premultiplying by S_w^{-1} in (A8) we get

$$S_w^{-1}S_b' = I + \frac{1}{P(\omega_2)^2} S_w^{-1}\hat{M}_1\hat{M}_1^T. \tag{A9}$$

The resulting criterion using $S_1 = S_b'$ is

$$J' = \text{tr } S_w^{-1}S_b' = n + \frac{1}{P(\omega_2)^2} \text{ tr } S_w^{-1}\hat{M}_1\hat{M}_1^T. \tag{A10}$$

Since constant multipliers and additive constants have no effect on the *feature extraction* process, we can equally well express the criterion as

$$J' = \text{tr } S_w^{-1}\hat{M}_1\hat{M}_1^T. \tag{A11}$$

The reduced forms for J and J' are identical. Therefore, the features extracted using J or J' will be identical. Q.E.D.

In a similar fashion it can be shown that (A1) and (A2) produce identical extracted features, if $S_2 = S_m$.

References

[1] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, part II, pp. 179–188, 1936.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic, 1972.

[3] D. H. Foley and J. W. Sammon, "An optimal set of discriminant vectors," *IEEE Trans. Comput.*, vol. C-24, pp. 281–289, Mar. 1975.

[4] W. L. G. Koontz and K. Fukunaga, "A nonparametric valley seeking technique for cluster analysis," *IEEE Trans. Comput.*, vol. C-21, pp. 171–178, Feb. 1972.

[5] G. H. Ball and D. J. Hall, "ISODATA: A novel method of data analysis and pattern classification," Stanford Res. Inst. Tech. Rep. AD-699616, 1965.

[6] K. Fukunaga and L. D. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 32–40, Jan. 1975.

[7] R. D. Short and K. Fukunaga, "The optimal distance measure for nearest neighbor classification," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 622–627, Sept. 1981.

[8] K. Fukunaga and R. D. Hostetler, "*k*-nearest-neighbor Bayes-risk estimation," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 285–293, May 1975.

[9] D. W. Peterson and R. C. Mattson, "A method of finding linear discriminant functions for a class of performance criteria," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 380–387, July 1966.

The Asymptotic Optimal Frequency Domain Filter for Edge Detection

W. H. H. J. LUNSCHER

Abstract—**In an earlier paper by Shanmugam, Dickey, and Green, an edge detection filter was derived which maximized the energy within a specified interval about an edge feature. The initial expression of this filter involved a prolate spheroidal wave function. However, a careful analysis of the application of an asymptotic approximation to this function uncovered a major dimensional error. The corrected derivation of the asymptotic optimal filter forms the subject of this paper. To verify the results, the filter found is compared to a similar filter developed independently by Marr and Hildreth.**

Index Terms—**Bandpass image filters, edge detection, image analysis, optimal edge enhancement filters.**

Through a rigorous analysis utilizing the special properties of prolate spheroidal wave functions, Shanmugam *et al.* [1] succeeded in deriving an optimal filter for the detection of step edges. However, a dimensional analysis of the subsequent application of an asymptotic approximation to this filter reveals the presence of errors in the results. These errors prevent proper scaling of the filter to match observed image structure, and make comparison to other edge filters (e.g., [2]) awkward. This correspondence will briefly review the properties of the optimal filter as derived by Shanmugam *et al.* The corrected application of the asymptotic approximation to the filter will follow. Finally, the results will be compared to a similar filter derived through a different method for verification.

The optimum edge filter, as defined by Shanmugam *et al.*, produces the maximum energy in the vicinity of the location of the edge in image space. Let $g(x)$ and $h(x)$ represent the filtered image and filter line spread function with Fourier transforms $G(\omega)$ and $H(\omega)$. For an edge feature centered at $x = 0$ the optimal filter maximizes

$$\gamma = \frac{\int_{-I/2}^{I/2} |g(x)|^2\, dx}{\int_{-\infty}^{\infty} |g(x)|^2\, dx}, \tag{1}$$

i.e., the proportion of image energy within the resolution interval $|x| \leqslant I/2$.

The response of the filter was constrained to be even and bandpass:

$$H(\omega) = 0 \quad \text{for} \quad |\omega| > \Omega \tag{2a}$$

$$H(0) = 0 \tag{2b}$$

$$H(\omega) = H(-\omega). \tag{2c}$$

On adopting the unit step function as the ideal edge model, Shanmugam *et al.* were able to show that the optimal filter transfer function is

$$H_{\text{step}}(\omega) = \begin{cases} K_1\omega\psi_1(c, \omega I/2\Omega), & |\omega| < \Omega \\ 0 & \text{elsewhere} \end{cases} \tag{3}$$

Manuscript received July 20, 1982; revised April 11, 1983.

The author is with the Department of Electrical Engineering, University of British Columbia, Vancouver, B.C., Canada V6T 1Z2.

0162-8828/83/1100-0678$01.00 © 1983 IEEE