

# Linear Discriminant Analysis for Two Classes via Removal of Classification Structure

Mayer Aladjem

**Abstract**—A new method for two-class linear discriminant analysis, called “removal of classification structure,” is proposed. Its novelty lies in the transformation of the data along an identified discriminant direction into data without discriminant information and iteration to obtain the next discriminant direction. It is free to search for discriminant directions oblique to each other and ensures that the informative directions already found will not be chosen again at a later stage. The efficacy of the method is examined for two discriminant criteria. Studies with a wide spectrum of synthetic data sets and a real data set indicate that the discrimination quality of these criteria can be improved by the proposed method.

**Index Terms**—Exploratory data analysis, dimension reduction, linear discriminant analysis, discriminant plots, structure removal.

## 1 INTRODUCTION

To obtain a visual representation of high dimensional data, we must reduce the dimensionality, and for data visualization two-dimensional representations (scatter plots) are most useful [5], [13]. In this paper we discuss *discriminant analysis for two classes* [12]. Scatter plots intended for discriminant analysis are called *discriminant plots*. Our consideration is limited to plots obtained by the linear mapping  $\mathbf{y} = [\mathbf{r}_1, \mathbf{r}_2]^T \mathbf{x}$ ,  $\mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{y} \in \mathbb{R}^2$ ,  $n > 2$  with  $\mathbf{x}$  an arbitrary  $n$ -dimensional observation, and  $\mathbf{r}_1, \mathbf{r}_2$  direction vectors (each having unit length). Vectors  $\mathbf{r}_1, \mathbf{r}_2$  which optimize a discriminant criterion are called *discriminant vectors*.

We discuss two classes of discriminant criteria, namely, the extended Fisher criterion previously proposed by us [1], [2] and the nonparametric criterion proposed by Fukunaga [8]. Our goal is to assess the discrimination qualities of the vectors  $\mathbf{r}_1$  and  $\mathbf{r}_2$  obtained by successive optimization of these criteria. In the past, two methods for successive optimization were applied. The first method uses orthogonal constraints on the discriminant vectors. It is called ORTH in this paper. The second method does not so constrain the discriminant vectors. It is called FREE. Our experience shows that method FREE does not always create new discriminant information along the second discriminant vector  $\mathbf{r}_2$ . In order to include new information in  $\mathbf{r}_2$ , method ORTH is used [9], [13], but it is known [5], [8], [13] that orthogonal directions do not always suffice; directions containing information for class separation may be oblique to each other.

In this paper, we propose a new method which is better than FREE and ORTH. It is free to search for discriminant directions oblique to each other and ensures that informative directions already found will not be chosen again at a later stage. Some preliminary results of our work were presented in [3]. This paper contains a more thorough analysis and more complete results.

• M. Aladjem is with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, P.O.B. 653, 84105 Beer-Sheva, Israel.  
E-mail: aladjem@bgu.ac.il

Manuscript received Oct. 9, 1995; revised Aug. 26, 1996. Recommended for acceptance by D.M. Titterton.

For information on obtaining reprints of this article, please send e-mail to: [transpami@computer.org](mailto:transpami@computer.org), and reference IEEECS Log Number P96093.

In Section 2 we describe a normalization of the data which simplifies expression of the criteria. Section 3 presents the criteria and their successive optimization by the methods FREE and ORTH. The new method, called REM, is presented in Section 4. Section 5 contains the results and analyses of comparative studies of the methods.

## 2 NORMALIZED DATA

Suppose we are given an original set of  $N$  labeled observations  $(\mathbf{z}_1, l_1), (\mathbf{z}_2, l_2), \dots, (\mathbf{z}_N, l_N)$  in  $n$ -dimensional sample space:  $\mathbf{z}_j \in \mathbb{R}^n$ ,  $(n > 2)$ ,  $j = 1, 2, \dots, N$ . The label  $l_j \in \{\omega_1, \omega_2\}$  shows that  $\mathbf{z}_j$  belongs to one of the classes  $\omega_1$  or  $\omega_2$ . These labels imply a decomposition of the set  $Z = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N\}$  into two subsets corresponding to the unique classes. Let the decomposition be  $Z = Z_1 \cup Z_2$ , where  $Z_i$  contains  $N_i$  observations in the class labeled by  $\omega_i$ , for  $i = 1, 2$ . Our aim is to assess *discrimination qualities* of discriminant plots. For this purpose we divide each  $Z_i$  into design ( $Z_{di}$ ) and test ( $Z_{ti}$ ) sets (see Section 5). Using  $Z_{di}$  we obtain direction vectors  $\mathbf{r}_1$  and  $\mathbf{r}_2$  defining the discriminant plots. We then project  $Z_{ti}$  on to the plots and compute the error rates of a *nearest neighbor* (NN) allocation rule applied to the test data.

Let  $N_{di}$  and  $N_{ti}$  denote the number of design and test observations for class  $\omega_i$ ;  $N_d = N_{d1} + N_{d2}$  and  $N_t = N_{t1} + N_{t2}$  are the total numbers of design and test observations. We arrange that the design and test sets preserve the data-based relative frequencies, i.e.,  $N_{di}/N_d \cong N_{ti}/N_t \cong N_i/N$ . To normalize ([8], p.470; [12], pp.193-194), we perform an eigenvalue-eigenvector decomposition  $S_z = \mathbf{RDR}^T$  of the pooled within-class sample covariance matrix  $S_z = (N_{d1}/N_d)S_{z1} + (N_{d2}/N_d)S_{z2}$  with  $S_{zi}$  the class-conditional sample covariance matrices estimated over the design data  $Z_{di}$  for  $i = 1, 2$ , and  $\mathbf{R}$  and  $\mathbf{D}$   $n \times n$  matrices;  $\mathbf{R}$  is orthonormal and  $\mathbf{D}$  diagonal. We then define the normalization matrix  $\mathbf{A} = \mathbf{D}^{-1/2}\mathbf{R}^T$ . The matrix  $S_z$  is assumed to be nonsingular, otherwise a preliminary dimensional reduction of the original data must be carried out or an augmented covariance matrix from  $S_z$  must be computed and used instead [10].

In the remainder of the paper, all operations are performed on the *normalized design data*  $\mathcal{X}_{di} = \{\mathbf{x}: \mathbf{x} = \mathbf{A}\mathbf{z}, \mathbf{z} \in Z_{di}\}$  and *normalized test data*  $\mathcal{X}_{ti} = \{\mathbf{x}: \mathbf{x} = \mathbf{A}\mathbf{z}, \mathbf{z} \in Z_{ti}\}$  for  $i = 1, 2$ . The pooled within-class sample covariance matrix estimated over  $\mathcal{X}_{di}$  becomes the identity matrix  $\mathbf{A}S_z\mathbf{A}^T = \mathbf{I}$ . This implies that, for any unit direction vector  $\mathbf{r}$ , the projections  $\tau = \mathbf{r}^T\mathbf{x}$  of the normalized design observations  $\mathbf{x} \in \{\mathcal{X}_{d1} \cup \mathcal{X}_{d2}\}$  have unit pooled within-class sample variance.

## 3 DISCRIMINANT PLOTS OBTAINED VIA THE METHODS FREE AND ORTH

### 3.1 Extended Fisher Discriminant Plots (ExF\_FREE and ExF\_ORTH)

The *extended Fisher (ExF) criterion* [1], [2] is a generalization of Malina's discriminant criterion [11], i.e.,

$$G_p(\mathbf{r}, \beta) = (1 - \beta)\mathbf{r}^T\mathbf{B}\mathbf{r} + \beta|\mathbf{r}^T\mathbf{S}^{(-)}\mathbf{r}|, 0 \leq \beta \leq 1 \quad (1)$$

with direction vector  $\mathbf{r}$ , control parameter  $\beta$  and

$$\mathbf{B} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T, \quad (2)$$

$$\mathbf{S}^{(-)} = \mathbf{S}_1 - \mathbf{S}_2 \text{ or } \mathbf{S}_2 - \mathbf{S}_1. \quad (3)$$

Here,  $\mathbf{B}$  is the sample between-class scatter matrix,  $\mathbf{m}_i$  are the class-conditional sample mean vectors and  $\mathbf{S}_i$  are the class-conditional sample covariance matrices, computed for the normalized design

data sets  $\mathcal{X}_{di}$ ,  $i = 1, 2$ . In (1), the symbol  $\mathbf{S}^{(-)}$  implies two forms of the criterion  $G_p(\mathbf{r}, \beta)$ .

The ExF discriminant vector maximizes  $G_p(\mathbf{r}, \beta)$ . It is the eigenvector corresponding to the largest eigenvalue of the matrices  $(1 - \beta)\mathbf{B} + \beta(\mathbf{S}_1 - \mathbf{S}_2)$  and  $(1 - \beta)\mathbf{B} + \beta(\mathbf{S}_2 - \mathbf{S}_1)$ . We obtain the sequence of ExF discriminant vectors  $\mathbf{r}_1$ ,  $\mathbf{r}_{2(\text{orth})}$  and  $\mathbf{r}_{2(\text{free}\perp)}$  by successive optimization of  $G_p(\mathbf{r}, \beta)$  for three specific values  $\beta_1$ ,  $\beta_{2(\text{orth})}$  and  $\beta_{2(\text{free})}$  of the control parameter  $\beta$ . The vector  $\mathbf{r}_1$  is obtained by optimization of  $G_p(\mathbf{r}, \beta_1)$ ;  $\mathbf{r}_{2(\text{orth})}$  is computed by method ORTH:  $\mathbf{r}_{2(\text{orth})}$  maximizes  $G_p(\mathbf{r}, \beta_{2(\text{orth})})$  with  $\mathbf{r}_{2(\text{orth})}$  constrained to be orthogonal to  $\mathbf{r}_1$ ; and  $\mathbf{r}_{2(\text{free}\perp)}$  is obtained by method FREE: first we calculate  $\mathbf{r}_{2(\text{free})}$  that maximizes  $G_p(\mathbf{r}, \beta_{2(\text{free})})$  then we obtain the vector  $\mathbf{r}_{2(\text{free}\perp)}$  orthogonal to  $\mathbf{r}_1$  and lying in the plane spanned by the vectors  $\mathbf{r}_1$  and  $\mathbf{r}_{2(\text{free})}$  (external orthogonalization [13]). The plot with discriminant vectors  $\mathbf{r}_1$ ,  $\mathbf{r}_{2(\text{orth})}$  is named ExF\_ORTH, and the plot defined by  $\mathbf{r}_1$ ,  $\mathbf{r}_{2(\text{free}\perp)}$  is named ExF\_FREE.

Appropriate values for  $\beta_1$ ,  $\beta_{2(\text{orth})}$  and  $\beta_{2(\text{free})}$  are not known in advance. We search for them using a trial and error procedure. Our approach to *model selection* is to choose the values of  $\beta$  that minimize the error rates of a NN allocation rule applied to the projections of the design data on to the discriminant plots. First we choose the value  $\beta_1$  which minimizes the error rate of the projections of the design data on to the straight line with direction  $\mathbf{r}_1$ . After that we obtain  $\beta_{2(\text{orth})}$  and  $\beta_{2(\text{free})}$  to minimize the error rates of the projections of the design observations on to the plots ExF\_ORTH and ExF\_FREE respectively. This gives rise to three problems of parameter optimization. Our strategy for solving them is to choose a grid of values in the interval  $(0 \leq \beta \leq 1)$ , to calculate the error rates for each value and then to choose the values with the smallest error rates as  $\beta_1$ ,  $\beta_{2(\text{orth})}$  and  $\beta_{2(\text{free})}$ . Our experience is that a uniform grid with 21 values (step-size 0.05) is suitable. A possible refinement of our method would be to assess the error rates by cross-validation (see Friedman [6] and Krzanowski et al. [10]).

We calculate error rates using the *2NN error-counting procedure* [7]. Let  $\mathbf{y}$ ,  $\mathbf{y}_1$ ,  $\mathbf{y}_2$  be one- or two-dimensional projections of three observations of the examined data where  $\mathbf{y}_1$ ,  $\mathbf{y}_2$  are 2NNs of  $\mathbf{y}$  among all projections. The 2NN procedure misclassifies  $\mathbf{y}$  when  $\mathbf{y} \in \omega_1$  but  $\mathbf{y}_1, \mathbf{y}_2 \in \omega_2$ , or  $\mathbf{y} \in \omega_2$  but  $\mathbf{y}_1, \mathbf{y}_2 \in \omega_1$ . No error is counted when the two NNs belong to different classes. We use Euclidean distance to define the NNs and calculate the *2NN error rate* as the percent of the misclassified observations among the examined data.

### 3.2 Fukunaga's Nonparametric Discriminant Plot (Fuku\_ORTH)

Fukunaga's (Fuku) nonparametric discriminant criterion ([8], pp.466-476) is

$$G_n(\mathbf{r}, \alpha, k) = \mathbf{r}^T \mathbf{B}_{k\alpha} \mathbf{r}, \quad (4)$$

where  $\mathbf{B}_{k\alpha}$  is a nonparametric between-class scatter matrix which expresses the local data structure along the class separation boundary. To be precise,

$$\begin{aligned} \mathbf{B}_{k\alpha} = & \frac{1}{N_d} \sum_{\mathbf{x} \in \mathcal{X}_{d1}} u(\mathbf{x}) [\mathbf{x} - \mathbf{m}_{k2}(\mathbf{x})] [\mathbf{x} - \mathbf{m}_{k2}(\mathbf{x})]^T \\ & + \frac{1}{N_d} \sum_{\mathbf{x} \in \mathcal{X}_{d2}} u(\mathbf{x}) [\mathbf{x} - \mathbf{m}_{k1}(\mathbf{x})] [\mathbf{x} - \mathbf{m}_{k1}(\mathbf{x})]^T \end{aligned} \quad (5)$$

where  $\mathbf{m}_{ki}(\mathbf{x})$  is the mean of the  $k$ NNs from class  $\omega_i$  to point  $\mathbf{x}$  for  $i = 1, 2$ , and  $u(\mathbf{x})$  is a weight function which is a kNN estimate of the Bayes risk of the dichotomy  $\omega_1$  and  $\omega_2$ . The function  $u(\mathbf{x})$  deemphasizes points far away from the class separation boundary. A parameter in  $u(\mathbf{x})$ , denoted by  $\alpha$ , controls the width of the band along the class separation boundary in which points with large weights  $u(\mathbf{x})$  are located. The parameter  $k$  controls the location of the pivotal point  $\mathbf{m}_{ki}(\mathbf{x})$ .

Following Fukunaga [8], we compute the discriminant vectors  $\mathbf{r}_1$  and  $\mathbf{r}_{2(\text{orth})}$  as the eigenvectors corresponding to the two largest eigenvalues of  $\mathbf{B}_{k\alpha}$ . The plot defined by these vectors is called Fuku\_ORTH. Our strategy for *model selection* is similar to that used previously. From experience, we define the optimization grid of  $(\alpha, k)$  values by the outer product of  $\alpha = (1, 2, 3, 4, 5)$  and  $k = (1, 2, 3, 4, 5, 6, 7, 8)$ , and we choose the pair of  $(\alpha, k)$ -values which jointly minimize the 2NN error rate of the projections of the design data on to the plot Fuku\_ORTH.

## 4 DISCRIMINANT ANALYSIS VIA REMOVAL OF CLASSIFICATION STRUCTURE

The new method, called REM, "*removal of classification structure*," consists of obtaining a discriminant vector  $\mathbf{r}_1$ , transforming the data along it into data without classification structure, and obtaining a discriminant vector  $\mathbf{r}_2$ . The proposed method is stimulated by an idea of Friedman [5] called "*structure removal*" which is oriented to cluster analysis (analysis of data that has not been previously grouped into classes). The aim of this paper is to employ "*structure removal*" in discriminant analysis. Following Friedman [5], we describe the method in its abstract version based on probability distributions. The application to observed data is obtained by substituting an estimate of the distributions over the design sets  $\mathcal{X}_{d1}$  and  $\mathcal{X}_{d2}$ .

### 4.1 Zero Informative Direction Vector

We start by discussing the properties of a directional vector  $\mathbf{a}$  which has no classification structure in terms of its density function. Such a vector  $\mathbf{a}$  is called a *zero informative direction vector*. In discriminant analysis  $\mathbf{a}$  is zero informative if by observing the projection  $\zeta = \mathbf{a}^T \mathbf{x}$  of any realization  $\mathbf{x}$  of a random vector  $\mathbf{X}$  we cannot gain any information about the class to which  $\mathbf{x}$  belongs (see Devijver and Kittler [4], pp.198-199). Thus,  $p_a(\zeta | \omega_i) = p_a(\zeta)$ , for  $i = 1, 2$ , where  $p_a(\zeta | \omega_i)$  is the class-conditional density of  $\mathbf{a}^T \mathbf{X}$  and  $p_a(\zeta)$  is the unconditional (mixture) density of  $\mathbf{a}^T \mathbf{X}$ . It is known [5] that, for most high-dimensional data, most low-dimensional projections are approximately normal. Therefore it seems reasonable to approximate  $p_a(\zeta)$  by a normal density function. Note also that in order to preserve the properties of the normalized data the variance of  $\mathbf{a}^T \mathbf{X}$  must be one. Without loss of generality, the mean of  $\mathbf{a}^T \mathbf{X}$  is assumed to be zero. Taking into account these observations, we conclude that class-conditional densities  $p_a(\zeta | \omega_i)$  along the zero informative direction vector  $\mathbf{a}$  can be approximated by the standard normal density  $N(0, 1)$ .

### 4.2 Removal of Classification Structure

In this section, we describe an algorithm for removal of the classification structure along a direction in the  $n$ -dimensional sample space. The idea is to transform the projected class-conditional densities to  $N(0, 1)$ . Let  $\mathbf{r}_1$  be a direction vector. Assume that  $\mathbf{U}$  is an orthonormal  $n \times n$  matrix with  $\mathbf{r}_1$  as the first row. Then applying the linear transformation  $\mathbf{t} = \mathbf{U}\mathbf{x}$  results in a rotation such that the new first coordinate of an observation  $\mathbf{x}$  is  $\tau_1 = \mathbf{r}_1^T \mathbf{x}$ . We denote the other coordinates by  $\tau_2, \tau_3, \dots, \tau_n$  ( $\mathbf{t} = [\tau_1, \tau_2, \dots, \tau_n]^T$ ). In order to remove the classification structure from the direction defined by  $\mathbf{r}_1$  we require a transformation that takes the class-conditional densities along  $\mathbf{r}_1$  to  $N(0, 1)$ , but leaves all other coordinates  $\tau_2, \tau_3, \dots, \tau_n$

unchanged. Let  $\mathbf{q}$  be a vector function with components  $q_1, q_2, \dots, q_n$  that carries out this transformation:  $\tau_1 = q_1(\tau_1)$  with  $q_1(\mathbf{r}_1^T \mathbf{X})$  having class-conditional  $N(0, 1)$  and  $\tau_i = q_i(\tau_i)$ ,  $i = 2, 3, \dots, n$  each given by the identity transformation. The function  $q_1$  is obtained by the percentile transformation method:

$$\text{for } \mathbf{x} \text{ from class } \omega_1: q_1(\tau_1) = \Phi^{-1}(F_{r_1}(\tau_1|\omega_1)); \quad (6)$$

$$\text{for } \mathbf{x} \text{ from class } \omega_2: q_1(\tau_1) = \Phi^{-1}(F_{r_1}(\tau_1|\omega_2)); \quad (7)$$

where  $(F_{r_1}(\tau_1|\omega_i))$  is the class-conditional (cumulative) distribution function along  $r_1$  for  $i = 1, 2$  and  $\Phi^{-1}$  is the inverse of the standard normal distribution function  $\Phi$ . Finally,

$$\mathbf{x}' = \mathbf{U}^T \mathbf{q}(\mathbf{U}\mathbf{x}) \quad (8)$$

transforms  $r_1$  to be a zero informative direction vector leaving all orthogonal directions unchanged.

### 4.3 Computational Procedure

The procedure for obtaining a discriminant plot via removal of the classification structure is as follows:

- 1) Find  $r_1$  which optimizes a discriminant criterion. The criterion is calculated for the *normalized design data* (Section 2).
- 2) Remove the classification structure from  $r_1$ : Obtain an estimate of  $F_{r_1}(\tau_1|\omega_i)$  over the projections  $\tau_1 = r_1^T \mathbf{x}$  of the design observations  $\mathbf{x} \in \mathcal{X}_{di}$ , for  $i = 1, 2$  ([5], p.254). Substitute the estimate of  $F_{r_1}(\tau_1|\omega_i)$  for  $i = 1, 2$  into (6) and (7). Transform  $\mathbf{x} \in \{\mathcal{X}_{d1} \cup \mathcal{X}_{d2}\}$  using (8) and obtain new design sets  $\mathcal{X}'_{d1}$  and  $\mathcal{X}'_{d2}$  that are totally overlapped along  $r_1$ .
- 3) Compute the discriminant criterion for the new design sets  $\mathcal{X}'_{d1}$  and  $\mathcal{X}'_{d2}$ . Obtain  $r_{2(rem)}$  which optimizes this criterion.
- 4) Obtain an orthonormal basis  $r_1, r_{2(rem)}$  in the plane spanned by the vectors  $r_1$  and  $r_{2(rem)}$  (external orthogonalization of  $r_1$  and  $r_{2(rem)}$ ). We name the plot defined by  $r_1, r_{2(rem)}$  ExF\_REM for the ExF discriminant criterion and Fuku\_REM for the Fuku criterion.

Note that the proposed procedure can be applied to any discriminant criterion that determines a single linear one-dimensional subspace of the sample space. The procedure is restricted to computing the criterion for the *normalized design data* (Section 2). This does not decrease its generality because most of the discriminant criteria are invariant under any nonsingular linear transformation [4], [8]. For such criteria the data normalization does not influence the optimal discriminant plot. Moreover, discriminant criteria defined by kNN techniques must be applied to normalized data if Euclidean distance is used to determine the kNNs ([8], p. 470).

## 5 COMPARATIVE EXPERIMENTS

In this section, we compare the discrimination qualities of the plots described above. We study a wide spectrum of situations in terms of the size  $N_{di}$  of the design sample drawn from synthetic and real data.

### 5.1 Simulation Studies

Except for the last set of synthetic data, we set  $N_{di} = 10, 15, 20, 30, 50, 70$  for  $i = 1, 2$ . An experiment for a given combination of particular setting, class-conditional distributions and size  $N_{di}$  of the design sample, consisted of 50 replications of the following procedure. In each class, we generated design data of size  $N_{di}$  from the appropriate class distribution. Using this data we obtained the

discriminant plots (their direction vectors), then projected  $N_t = 3,000$  (extra) test observations on to the plots and computed the test error rates. Finally, we calculated the mean and the standard errors of the error rates over the 50 replications. We compared the discrimination qualities of the plots by the resulting differences of the averaged test errors (see McLachlan [12], pp. 373-375) and evaluated the significance of the differences on the basis of the standard errors (t-test,  $p = 0.05$ ). The existence of significant deviations of the mean errors of the REM from those of the FREE or the ORTH are indicated by stars (\*\*\*) in the figures, where averaged rates and standard errors are plotted (Figs.1-3). Note that this is in fact a pessimistic evaluation of the significance of the differences, since we estimated the error rates using the same samples for all methods and therefore the estimates of the errors are not independent (when a favorable or an unfavorable design sample is realized the resulting test error rates of the methods tend to be small or large together).

We computed the test errors of each plot by the following procedure: first we divided the test data into 30 subsets with 100 observations (50 observations per class) and then computed the 2NN error rate [7] (see Section 3.1) of the projection of each subset on to the plot; we then averaged the errors over the 30 test subsets. We used 30 replications of the error computation because the 2NN method is suitable for small sample sizes.

In all experiments with a particular population, we used the same value of  $\beta_1$  for the ExF criterion and the same values of  $\alpha, k$  for the Fuku criterion. We obtained them by averaging the results of the model selection over 50 replications of the design sample with sample size  $N_{di} = 50$ . The value of  $\beta$  for the second ExF discriminant vector was obtained by applying the model selection procedure to each run separately. We did not search separately for the appropriate  $(\alpha, k)$ -value for  $r_1$  and  $r_{2(rem)}$  in order not to favor method REM over Fukunaga's method by expanding the number of parameters to be optimized.

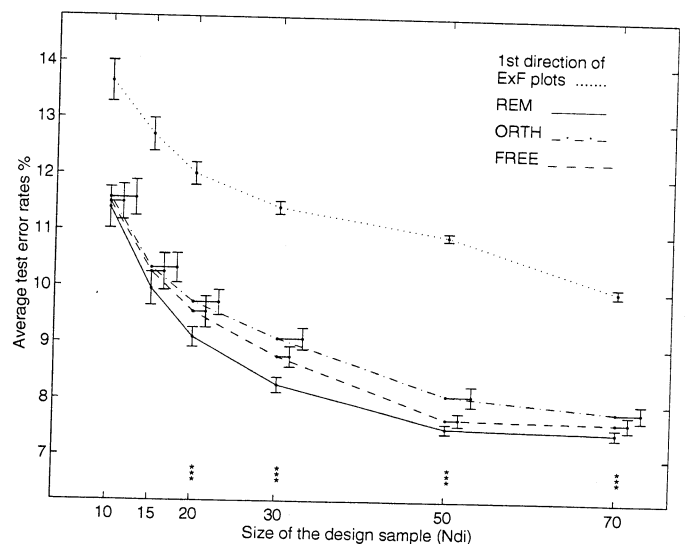


Fig.1. Normal class-conditional distributions. Test error rates in the ExF plots ( $\beta_1 = 0$ ). Bars represent the standard errors of the error rates. Stars (\*\*\*) indicate the existence of significant differences between the averaged errors of REM and ORTH (t-test,  $p = 0.05$ ).

#### 5.1.1 Normal Class-Conditional Distributions

The samples for two classes were drawn from six dimensional normal class-conditional distributions. The means and the covariance matrices were  $\mathbf{m}_1 = [0, 0, 0, 0, 0, 0]^T$ ,  $\mathbf{S}_1 = \text{diag}(0.18, 1.82, 0.67, 0.67, 0.67, 0.67)$  for class  $\omega_1$  and  $\mathbf{m}_2 = [1.5, 1.5, 0, 0, 0, 0]^T$ ,  $\mathbf{S}_2 = \text{diag}(1.82, 0.18, 1.33, 1.33, 1.33, 1.33)$  for class  $\omega_2$ . Here we

compared the methods FREE, ORTH and REM applied to the ExF criterion. Fig.1 presents the averaged test error rates. The dotted path shows the errors in the one-dimensional space defined by the vector  $r_1$ . The other paths show the errors in the plots obtained by methods ORTH, FREE and REM. The reduction of the errors in the plots compared with those along  $r_1$  (dotted path) shows the discrimination effectiveness of these methods. It is in the range 2-2.5%. REM outperforms the other methods by 0.25-1%, and REM deviates significantly from the ORTH for  $N_{di} = 20, 30, 50, 70$ .

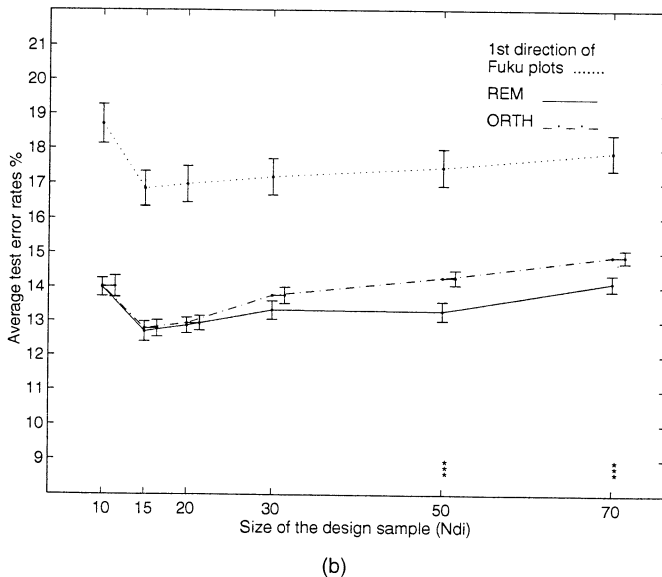
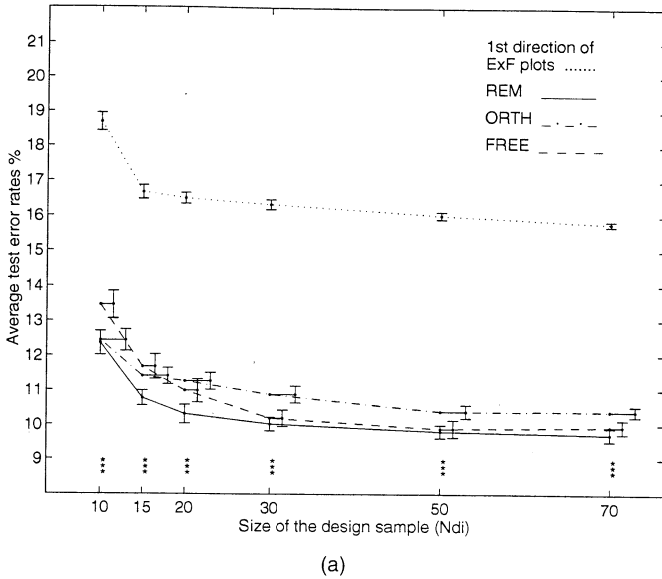


Fig.2. Nonnormal unimodal class-conditional distributions. Test error rates in: (a) ExF plots ( $\beta_1 = 0$ ). (b) Fuku plots ( $\alpha = 1, k = 8$ ).

### 5.1.2 Nonnormal Unimodal Class-Conditional Distributions.

Eight dimensional samples for the two classes were obtained by sampling two random signals  $X_1(t) = A \exp\{-(t - B)^2/2C^2\}$  and  $X_2(t) = A \exp\{-|t-B|^\delta/2C^2\}$  at eight uniformly spaced times in the interval  $(0 \leq t \leq 1.05)$ . Here  $\delta = 1.3$ , and  $A, B, C$  are random parameters with uniform distributions on the intervals  $(0.7, 1.3), (0.3, 0.7), (0.2, 0.4)$ .  $X_1(t)$  and  $X_2(t)$  are similar to those used by Fukunaga ([8], p.472). Originally, Fukunaga used  $\delta = 1$ , which implies en-

tirely separated classes. We chose  $\delta = 1.3$ , which results in some class overlap. In order to eliminate the small pooled (within-class) variations of the data we projected the eight-dimensional observations on to subspaces spanned by the five eigenvectors corresponding to the largest eigenvalues of the pooled within-class sample covariance matrix (modified canonical analysis—see Krzanowski et al. [10]). For this five-dimensional data set we ran experiments with the ExF and Fuku criteria. Fig. 2a shows the results for the ExF plots. REM significantly outperforms ORTH for  $N_{di} = 20, 30, 50, 70$  and FREE for  $N_{di} = 10, 15$ . We found that the test error rates in the ExF plots (Fig. 2a) are lower than those in Fuku plots (Fig. 2b). From this we conclude that the ExF criterion is more suitable than the Fuku criterion for this population. This would explain the surprising increase in the test error rates for the Fuku plots (Fig. 2b) for  $N_{di} = 20, 30, 50, 70$ . This unusual phenomenon is weaker for REM.

### 5.1.3 Multimodal Class-Conditional Distributions.

Eight dimensional samples were generated for each class  $\omega_1$  and  $\omega_2$ . The first two coordinates of the observations were drawn from normal mixtures:

$$\begin{aligned} p(x_1, x_2 | \omega_1) = & 1/3N([-3, 0]^T, 0.01I) \\ & + 1/3N([0.5, 3]^T, 0.01I) \\ & + 1/3N([-0.5, -3]^T, 0.01I) \end{aligned} \quad (9)$$

$$\begin{aligned} p(x_1, x_2 | \omega_2) = & 1/3N([-0.5, 3]^T, 0.01I) \\ & + 1/3N([3, 0]^T, 0.01I) \\ & + 1/3N([0.5, -3]^T, 0.01I) \end{aligned} \quad (10)$$

for  $\omega_1$  and  $\omega_2$ , respectively. Here,  $N([\mu_1, \mu_2]^T, 0.01I)$  denotes the bivariate normal density with a mean vector  $[\mu_1, \mu_2]^T$  and a diagonal covariance matrix. The other six coordinates were independent  $N(0, 1)$  for both classes. We ran simulations with a larger set of sample sizes  $N_{di} = 10, 15, 20, 30, 50, 60, 70, 80, 90, 100$ . REM outperforms FREE in the ExF plots (Fig. 3a). It succeeds in significantly decreasing the test error rates for  $N_{di} \geq 50$  in the Fuku plots (Fig. 3b). It is no surprise that the ExF plots (Fig. 3a) perform better than the original method of Fukunaga (ORTH in Fig. 3b). In this experiment we chose a data structure which is highly unfavorable to the latter method. Fig. 4, showing the projections of a test sample on to the plot Fuku\_REM, illustrates that REM succeeds in representing the classification structure of the data.

## 5.2 Experiments With a Real Data Set

A real data set concerning the medical diagnosis of the neurological disease cerebrovascular accident (CVA) contains pathologically-anatomically verified CVA cases: 200 cases with hemorrhages and 200 cases with infarction due to ischaemia. Twenty numerical results from a neurological examination were recorded for each CVA case [1]. We ran 50 replications of the following procedure. We randomly divided the data for each disease into a design set with 150 cases and a test set with 50 cases. Using the design data we obtained discriminant plots. After that we projected the test data on to the plots and computed the 2NN test error rates for each plot. Finally, we averaged the test errors over the 50 replications. Here, we studied the Fuku criterion for various numbers of the principal components used in the preliminary principal component reduction of the dimensionality of the observations (see Krzanowski et al. [10]). Fig.5 shows the results. REM significantly outperforms the original method of Fukunaga (Fuku\_ORTH) for 10 and 11 principal components ( $t$ -test,  $p = 0.05$ ).

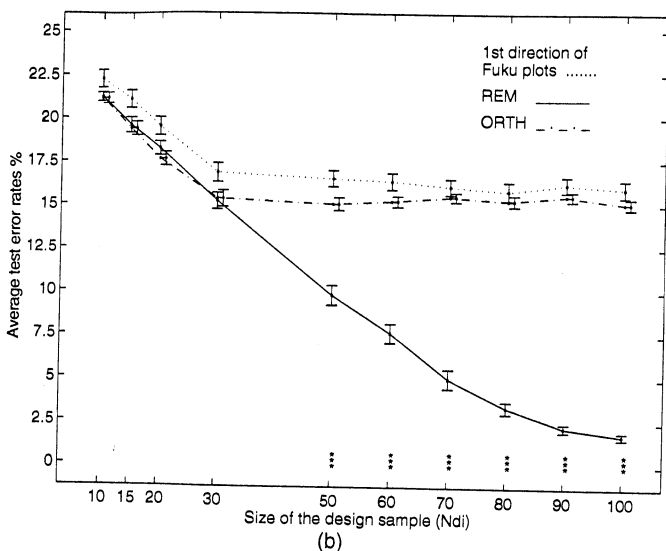
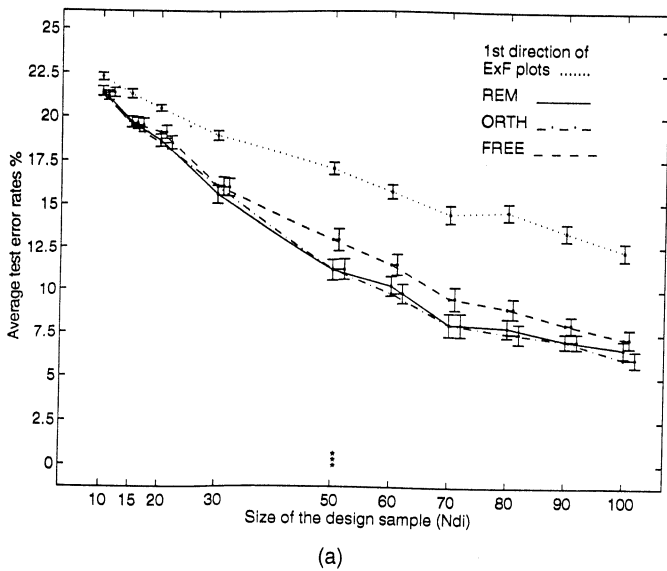


Fig. 3. Multimodal class-conditional distributions. Test error rates in: (a) ExF plots ( $\beta_1 = 0.5$ ), (b) Fuku plots ( $\alpha = 2, k = 3$ ).

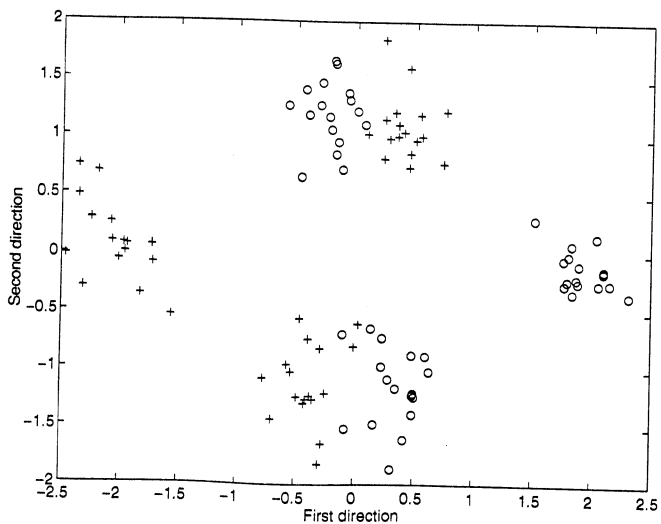


Fig. 4. Multimodal class-conditional distributions. Projection of 50 test observations per class on to the plot Fuku\_REM ("+" for class  $\omega_1$ , "o" for class  $\omega_2$ ,  $N_{d_i} = 100$ , 2NN test error rate 2%).

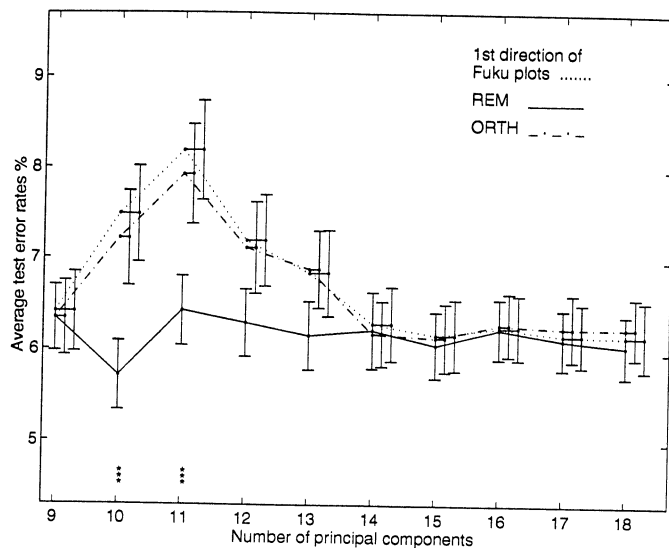


Fig. 5. CVA data. Test error rates in the Fuku plots ( $\alpha = 4, k = 5$ ).

### 6 CONCLUSIONS

The simulation studies and the real data experiment indicate that the proposed method "removal of classification structure" (REM) increases the quality of the discriminant plots obtained by optimization of two criteria, namely the extended Fisher criterion (ExF) [1], [2] and the nonparametric criterion of Fukunaga (Fuku) [8]. There appears to be no loss in applying REM instead of both methods ORTH, with an orthogonality constraint on the discriminant vectors, and FREE, without constraints on these vectors. It seems that REM provides a more effective constraint on the discriminant vectors than does the orthogonality constraint usually applied in statistical pattern recognition. Finally, we summarize the main features of the new method:

- 1) REM is free to search for discriminant vectors which are oblique to each other;
- 2) REM ensures that the informative directions already found will not be found again at a later stage;
- 3) REM can be applied to any discriminant criterion which determines a single linear one-dimensional subspace of the sample space.

### ACKNOWLEDGMENTS

The author wishes to thank associate editor Prof. D.M.Titterton and the reviewers for their critical reading of the manuscript and helpful comments. This work was supported in part by the Israeli Ministry of Science and Technology under contract number 3528, and in part by the Paul Ivanier Center for Robotics and Production Management, Ben Gurion University of the Negev, Israel.

### REFERENCES

- [1] M.E. Aladjem, "PNM: A Program for Parametric and Non-parametric Mapping of Multidimensional Data," *Computers in Biology and Medicine*, vol. 21, pp. 321-343, 1991.
- [2] M.E. Aladjem, "Multiclass Discriminant Mappings," *Signal Processing*, vol. 35, pp. 1-18, 1994.
- [3] M.E. Aladjem, "Discriminant Plots Obtained via Removal of Classification Structure," *Proc. 12th Int'l Conf. Pattern Recognition*, vol. 2, pp. 67-71, 1994.
- [4] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. London: Prentice-Hall, 1982.
- [5] J.H. Friedman, "Exploratory Projection Pursuit," *J. American Statistical Association*, vol. 82, pp. 249-266, 1987.

- [6] J.H. Friedman, "Regularized Discriminant Analysis," *J. American Statistical Association*, vol. 84, pp. 165-175, 1989.
- [7] K. Fukunaga and T.E. Flick, "The 2-NN Rule for More Accurate NN Risk Estimate," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 7, pp. 107-111, 1985.
- [8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second ed. New York: Academic Press, 1990.
- [9] Y. Hamamoto, Y. Matsuura, T. Kanaoka, and S. Tomita, "A Note on the Orthonormal Discriminant Vector Method for Feature Extraction," *Pattern Recognition*, vol. 24, pp. 681-684, 1991.
- [10] W.J. Krzanowski, P. Jonathan, W.V. McCarthy, and M.R. Thomas, "Discriminant Analysis With Singular Covariance Matrices: Methods and Applications to Spectroscopic Data," *Applied Statistics*, vol. 44, pp. 101-115, 1995.
- [11] W. Malina, "On an Extended Fisher Criterion for Feature Selection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 3, pp. 611-614, 1981.
- [12] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley & Sons, Inc., 1992.
- [13] W. Siedlecki, K. Siedlecka, and J. Sklansky, "An Overview of Mapping Techniques for Exploratory Pattern Analysis," *Pattern Recognition*, vol. 21, pp. 411-429, 1988.