

## Nonparametric Discriminant Analysis Via Recursive Optimization of Patrick-Fisher Distance

Mayer E. Aladjem

**Abstract**—A method for the linear discrimination of two classes is presented. It searches for the discriminant direction which maximizes the Patrick-Fisher (PF) distance between the projected class-conditional densities. It is a nonparametric method, in the sense that the densities are estimated from the data. Since the PF distance is a highly nonlinear function, we propose a recursive optimization procedure for searching the directions corresponding to several large local maxima of the PF distance. Its novelty lies in the transformation of the data along a found direction into data with deflated maxima of the PF distance and iteration to obtain the next direction. A simulation study and a medical data analysis indicate the potential of the method to find the sequence of directions with significant class separations.

### I. INTRODUCTION

We discuss discriminant analysis of two classes which is carried out by the linear mapping  $\tau = \mathbf{r}^T \mathbf{x}$ ,  $\mathbf{x} \in R^n$ ,  $\tau \in R^1$ ,  $n \geq 2$ , with  $\mathbf{x}$  an arbitrary  $n$ -dimensional observation, and  $\mathbf{r}$  a direction vector (having unit length). The vector  $\mathbf{r}$  maximizes the *Patrick-Fisher (PF) distance* [6] which measures the overlap of the class-conditional densities along  $\mathbf{r}$ . Unfortunately, the PF distance is not a unimodal function with respect to  $\mathbf{r}$  and has more than one maximum. In most applications [6], [12], [18], the optimal solution called the *PF discriminant vector* is searched for along the gradient of the PF distance, hoping that with a good starting point the procedure will converge to the global maximum or at least to a practical one. Some known techniques such as principal component and Fisher discriminant analysis [9], may be used for choosing a starting point for the optimization procedure. In this paper we use a technique which combines them. It is based on an *extended Fisher (ExF) discriminant criterion* previously proposed by us [1], [2]. The ExF criterion includes a control parameter for adjusting the criterion to the classification structure of the specific application. Nevertheless, the observed maximum of the PF distance can be merely a local maximum, which is far away from the global one in some data structures. In this paper we propose a recursive method which searches for several large local maxima of the PF distance. We are stimulated in this research by an idea of Friedman, called "*structure removal*" [8], which has great potential, like simulated annealing in the field of optimization [20, p. 78]. Some preliminary results of our work were presented in [3] and [4]. This paper contains a more thorough analysis and more complete results.

Section II presents a normalization of the data, called *sphering* [8], which is required by the recursive method. In Sections III and IV we describe the PF distance and the ExF criterion and the computation of the discriminant vectors related to them. The new method for recursive optimization of the PF distance is presented in Section V. Sections VI and VII contain the results and analyzes of a simulation study and an application to medical data.

Manuscript received March 16, 1996; revised November 25, 1996. This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University of the Negev, Israel.

The author is with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, 84105 Beer-Sheva, Israel (e-mail: aladjem@bgu.ac.il).

Publisher Item Identifier S 1083-4419(98)00216-7.

### II. SPHERED DATA

Suppose we are given a design (training) set of  $N_d$  labeled observations  $(\mathbf{z}_1, l_1), (\mathbf{z}_2, l_2), \dots, (\mathbf{z}_{N_d}, l_{N_d})$  in  $n$ -dimensional sample space,  $\mathbf{z}_j \in R^n$ , ( $n \geq 2$ ),  $j = 1, 2, \dots, N_d$ . We discuss the two class problem and 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 decomposition of the set  $Z_d = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{N_d}\}$  into two subsets corresponding to the unique classes. Let the decomposition be  $Z_d = Z_{d1} \cup Z_{d2}$ , where  $Z_{di}$  contains  $N_{di}$  observations in class labeled by  $\omega_i$ , for  $i = 1, 2$ . For the aim of data sphering [8, p. 251], we perform an eigenvalue-eigenvector decomposition  $\mathbf{S}_z = \mathbf{R}\mathbf{D}\mathbf{R}^T$  of the pooled sample covariance matrix  $\mathbf{S}_z$  with  $\mathbf{R}$  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  $\mathbf{S}_z$  is assumed to be nonsingular, otherwise a dimensional reduction must be done using only the eigenvectors corresponding to the nonzero eigenvalues [8]. In the remainder of the paper, all operations are performed on the *sphered design data*  $X_{di} = \{\mathbf{x}: \mathbf{x} = \mathbf{A}(\mathbf{z} - \mathbf{m}_z), \mathbf{z} \in Z_{di}\}$ , for  $i = 1, 2$  and *sphered new (arbitrary or test data)*  $\mathbf{x} = \mathbf{A}(\mathbf{z} - \mathbf{m}_z)$ ,  $\mathbf{z} \notin Z_d$  with  $\mathbf{m}_z$  the sample mean vector estimated over  $Z_d$ . For the sphered data the pooled sample covariance matrix becomes the identity matrix  $\mathbf{A}\mathbf{S}_z\mathbf{A}^T = \mathbf{I}$ . This implies that, for any unit direction vector  $\mathbf{r}$  the projections  $\tau = \mathbf{r}^T \mathbf{x}$  of the observations  $\mathbf{x} \in \{X_{d1} \cup X_{d2}\}$  have unit pooled sample variance.

### III. PATRICK-FISHER DISTANCE

The Patrick-Fisher (PF) distance is [6, pp. 277–280]

$$G_{PF}(\mathbf{r}, h) = \left\{ \int_{R^n} \left[ \frac{N_{d1}}{N_d} \hat{p}(\mathbf{r}^T \mathbf{x} | \omega_1) - \frac{N_{d2}}{N_d} \hat{p}(\mathbf{r}^T \mathbf{x} | \omega_2) \right]^2 d\mathbf{x} \right\}^{1/2} \quad (1)$$

with

$$\hat{p}(\mathbf{r}^T \mathbf{x} | \omega_i) = \frac{1}{h\sqrt{2\pi}N_{di}} \sum_{\mathbf{x}_{di} \in X_{di}} \exp \left\{ \frac{-1}{2h^2} [\mathbf{r}^T (\mathbf{x} - \mathbf{x}_{di})]^2 \right\}, \quad i = 1, 2 \quad (2)$$

the Parzen estimators with Gaussian kernels of the class-conditional densities of the projections  $\mathbf{r}^T \mathbf{x}$ . Here,  $\mathbf{x}$  is an arbitrary observation ( $\mathbf{x} \in R^n$ ),  $\mathbf{x}_{di} \in X_{di}$  are  $\omega_i$ -design observations, and  $h$  is a smoothing parameter. The effective performance of the Parzen estimator (2) is crucially dependent on the value of  $h$ . In our experiments (Sections VI and VII) we chose  $h$  subjectively [19]. For routine problems, an automatic choice of the smoothing parameter can be carried out [16, pp. 300–308], but this is beyond the scope of this work.

The theoretical motivation of the PF distance is its resultant upper bound on Bayes error along  $\mathbf{r}$ . It is known that the PF distance induces an upper bound which is larger than those of other probabilistic class separability measures [14]. Nevertheless,  $G_{PF}(\mathbf{r}, h)$  is more practical, because of its analytical simplification. The simplified forms of  $G_{PF}(\mathbf{r}, h)$  and its gradient  $\nabla_{\mathbf{r}} G_{PF}(\mathbf{r}, h)$  are [6, p. 279]

$$G(\mathbf{r}, h) = \left\{ \frac{1}{2h\sqrt{\pi}N_d^2} \sum_{i=1}^2 \sum_{j=1}^2 \sum_{\mathbf{x}_{di} \in X_{di}} \sum_{\mathbf{x}_{dj} \in X_{dj}} (-1)^{(i+j)} \times \exp \left\{ \frac{-1}{4h^2} [\mathbf{r}^T (\mathbf{x}_{di} - \mathbf{x}_{dj})]^2 \right\} \right\}^{1/2} \quad (3)$$

$$\begin{aligned} \nabla_{\mathbf{r}} G(\mathbf{r}, h) &= \frac{1}{4G(\mathbf{r}, h)h\sqrt{\pi}N_d^2} \sum_{i=1}^2 \sum_{j=1}^2 \sum_{\mathbf{x}_{d_i} \in X_{d_i}} \sum_{\mathbf{x}_{d_j} \in X_{d_j}} (-1)^{(i+j)} \\ &\times \exp \left\{ \frac{-1}{4h^2} \left[ \mathbf{r}^T (\mathbf{x}_{d_i} - \mathbf{x}_{d_j}) \right]^2 \right\} \\ &\times \left[ \frac{-1}{2h^2} (\mathbf{x}_{d_i} - \mathbf{x}_{d_j}) (\mathbf{x}_{d_i} - \mathbf{x}_{d_j})^T \mathbf{r} \right]. \end{aligned} \quad (4)$$

Using (3) and (4) we maximize  $G_{\text{PF}}(\mathbf{r}, h)$  with respect to  $\mathbf{r}$  by a sequential quadratic programming method [10] available as a routine E04UCF in the NAG Mathematical Library (mark 16, or latest version). It is considered one of the most efficient algorithms among the existing local optimizers [17]. In order to search among unit direction vectors we apply maximization of  $G_{\text{PF}}(\mathbf{r}, h)$  to nonlinear constraint  $\mathbf{r}^T \mathbf{r} = 1$ . The primary goal is to find the global maximum of  $G_{\text{PF}}(\mathbf{r}, h)$ . By a naive use of the optimization algorithm, the computed value for the observed  $\max\{G_{\text{PF}}(\mathbf{r}, h)\}$  can be merely a local maximum. The solution depends strongly on the starting point (vector) of the local optimizer. On the other hand, in some data structures more than one direction with significant (interesting) class separations exist. In this paper, we use an extended Fisher discriminant vector as a starting point because of its adaptation to the data structure under variations of a control parameter (Section IV). In order to search for several large local maxima we propose a method for recursive optimization of  $G_{\text{PF}}(\mathbf{r}, h)$  (Section V).

#### IV. EXTENDED FISHER CRITERION

The extended Fisher (ExF) criterion [1], [2] is a generalization of Malina's discriminant criterion [15], i.e.

$$G_{\text{ExF}}(\mathbf{r}, \beta) = [(1 - \beta)\mathbf{r}^T \mathbf{B} \mathbf{r} + \beta |\mathbf{r}^T \mathbf{S}^{(-)} \mathbf{r}|] [\mathbf{r}^T \mathbf{S}_W \mathbf{r}]^{-1}, \quad 0 \leq \beta \leq 1. \quad (5)$$

Here,  $\mathbf{r}$  is the direction vector,  $\beta$  ( $0 \leq \beta \leq 1$ ) is the control parameter,  $\mathbf{B} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$  is the sample between-class scatter matrix with  $\mathbf{m}_i$  the class-conditional sample mean vectors,  $\mathbf{S}^{(-)} = \mathbf{S}_1 - \mathbf{S}_2$  or  $\mathbf{S}_2 - \mathbf{S}_1$  with  $\mathbf{S}_i$  the class-conditional sample covariance matrices for  $i = 1, 2$  and  $\mathbf{S}_W = (N_{d1}/N_d)\mathbf{S}_1 + (N_{d2}/N_d)\mathbf{S}_2$ , the pooled within-class sample covariance matrix. All matrices are computed for the sphered design data sets  $X_{d_i}$ ,  $i = 1, 2$ . In (5) the symbol  $\mathbf{S}^{(-)}$  implies two forms of the criterion  $G_{\text{ExF}}(\mathbf{r}, \beta)$ . The ExF discriminant vector, which maximizes  $G_{\text{ExF}}(\mathbf{r}, \beta)$ , is the eigenvector corresponding to the largest eigenvalue of the matrices  $\mathbf{S}_W^{-1}[(1 - \beta)\mathbf{B} + \beta(\mathbf{S}_1 - \mathbf{S}_2)]$  and  $\mathbf{S}_W^{-1}[(1 - \beta)\mathbf{B} + \beta(\mathbf{S}_2 - \mathbf{S}_1)]$ . The criterion  $G_{\text{ExF}}(\mathbf{r}, \beta)$  is an extension of the most widely used criteria based on the scatter matrices which are obtained by introducing special values of  $\beta$  [1], [2]. This "universality" of the ExF criterion is our reason for expecting that the ExF discriminant vector can serve as a "good" starting point of the local optimizer of PF distance.

An appropriate value of the control parameter  $\beta$  is not known in advance. We search for it using a trial and error procedure. Our approach to *model selection* is to choose a suitable starting point of the local optimizer of the PF distance and to search for the value of  $\beta$  which maximizes the PF distance (1) along the ExF discriminant vector. This gives rise to a problem of parameter optimization. Our strategy for solving it is to choose the grid of values in the interval ( $0 \leq \beta \leq 1$ ), to calculate the PF distance for each value and then to choose the value with the largest PF distance as the  $\beta$ -value. Our experience is that a uniform grid with 11 values (step-size 0.1) is suitable.

#### V. RECURSIVE OPTIMIZATION OF THE PF DISTANCE

The proposed recursive method consists of obtaining a PF discriminant vector, transforming the data along it into data with

greater overlap of the classes (smaller PF distance), and obtaining a new PF discriminant vector. Like Friedman [8] we transform the densities along the direction with observed maxima into normal densities. The justification of the normal density assumption was made precise by Diaconis and Freedman [7]. They proved that most one-dimensional projections of high-dimensional data are approximately normal. Consequently the normal density assumption is more appropriate for data with higher dimension. Following Huber [13] and Friedman [8] we describe the method in its abstract version based on probability distributions. This makes some of the notation simpler. The application to observed data is obtained by substituting an estimate of the distributions over the design sets  $X_{d1}$  and  $X_{d2}$ .

#### A. Reduction of the Class Separation Along the PF Vector

In this section we describe an algorithm for reduction of the class separation along a direction with a local maximum of the PF distance. The idea is to transform the projected class conditional densities to normal densities in order to deflate that maximum. Let  $\mathbf{r}$  be a vector which defines a direction with a maximum of the PF distance. Assume that  $\mathbf{U}$  is an orthonormal ( $n \times n$ ) matrix with  $\mathbf{r}$  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  $\tau_1 = \mathbf{r}^T \mathbf{x}$ . We denote other coordinates as  $\tau_2, \tau_3, \dots, \tau_n$  ( $\mathbf{t} = [\tau_1 \tau_2 \dots \tau_n]^T$ ). Let  $p_{\mathbf{r}}(\tau_1 | \omega_i)$ ,  $i = 1, 2$  be the class-conditional densities along  $\mathbf{r}$  and  $m_{\mathbf{r}|\omega_i}$ ,  $\sigma_{\mathbf{r}|\omega_i}^2$  their means and variances. In order to reduce the class separation along  $\mathbf{r}$  we require a transformation that takes the class-conditional densities along  $\mathbf{r}$  to normal densities, 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  $\tau'_1$  having normal class-conditional distributions and  $\tau_i = q_i(\tau_i)$ ,  $i = 2, 3, \dots, n$  each given by the identity transformations. The function  $q_1$  is obtained by the percentile transformation method:

—for observations  $\mathbf{x}$  from class  $\omega_1$ :

$$q_1(\tau_1) = [\Phi^{-1}(F_{\mathbf{r}}(\tau_1 | \omega_1))] (\sigma_{\mathbf{r}|\omega_1}^2 \pm \Delta\sigma^2)^{1/2} + (m_{\mathbf{r}|\omega_1} - \Delta m_1); \quad (6)$$

—for observations  $\mathbf{x}$  from class  $\omega_2$ :

$$q_1(\tau_1) = [\Phi^{-1}(F_{\mathbf{r}}(\tau_1 | \omega_2))] (\sigma_{\mathbf{r}|\omega_2}^2 \pm \Delta\sigma^2)^{1/2} + (m_{\mathbf{r}|\omega_2} - \Delta m_2); \quad (7)$$

where  $\Delta\sigma^2$ ,  $\Delta m_1$ ,  $\Delta m_2$  are user-supplied parameters,  $F_{\mathbf{r}}(\tau_1 | \omega_i)$  is the class-conditional (cumulative) distribution functions along  $\mathbf{r}$  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 the class-conditional densities along  $\mathbf{r}$  to be normal densities  $p_{\mathbf{r}}(\tau_1 | \omega_i) = N(m_{\mathbf{r}|\omega_i} - \Delta m_i, \sigma_{\mathbf{r}|\omega_i}^2 \pm \Delta\sigma^2)$  leaving all orthogonal directions unchanged.

Now we are confronted with the problem of defining the values of the user-supplied parameters  $\Delta\sigma^2$ ,  $\Delta m_1$  and  $\Delta m_2$ . If we set  $\Delta\sigma^2 = 0$  and  $\Delta m_i = 0$ ,  $i = 1, 2$  we make minimal changes of the data in the sense of the minimal relative entropy distance measure between the original and transformed class-conditional distributions [8, p. 254], [13, Lemma 12.4, p. 456]. If  $\sigma_{\mathbf{r}|\omega_i}^2 \pm \Delta\sigma^2 = 1$  and  $m_{\mathbf{r}|\omega_i} - \Delta m_i = 0$ ,  $i = 1, 2$  we apply our previous method for successive optimization of discriminant criteria [5] which transforms the class-conditional densities along  $\mathbf{r}$  to  $N(0, 1)$  and results in full overlap of the classes along  $\mathbf{r}$ . This certainly eliminates the local maximum of the PF distance along  $\mathbf{r}$ , but it causes large changes of

the distributions of the transformed data  $\mathbf{x}'$  (8) in some applications. We refine our previous proposal [5] while keeping the data structure of  $\mathbf{x}'$  as close as possible to the original one. For this purpose (in our algorithm, Section V-B) we search for the smallest values of the parameters  $\Delta\sigma^2$ ,  $\Delta m_1$ ,  $\Delta m_2$  that result in a deflated PF distance along  $\mathbf{r}$ . We start our search with  $\Delta\sigma^2 = 0$  and  $\Delta m_i = 0$ ,  $i = 1, 2$  (minimal changes of the data) and then we make trials increasing the values of  $\Delta\sigma^2$  in the interval  $(0 \leq \Delta\sigma^2 \leq 1)$ . We choose the sign (+ or -) of the change ( $\pm\Delta\sigma^2$ ) in order to approach the value of  $\sigma_{\mathbf{r}|\omega_i}^2 \pm \Delta\sigma^2$  to 1. We assign the latter value to 1 if it crosses 1. For each  $\Delta\sigma^2$  we compute the values of  $\Delta m_1$  and  $\Delta m_2$  using the sphering conditions:

—zero unconditional mean

$$P(\omega_1)(m_{\mathbf{r}|\omega_1} - \Delta m_1) + P(\omega_2)(m_{\mathbf{r}|\omega_2} - \Delta m_2) = 0; \quad (9)$$

—unconditional variance equal to one

$$P(\omega_1)[(\sigma_{\mathbf{r}|\omega_1}^2 \pm \Delta\sigma^2) + (m_{\mathbf{r}|\omega_1} - \Delta m_1)^2] + P(\omega_2)[(\sigma_{\mathbf{r}|\omega_2}^2 \pm \Delta\sigma^2) + (m_{\mathbf{r}|\omega_2} - \Delta m_2)^2] = 1; \quad (10)$$

with  $P(\omega_i)$  the *a priori* probabilities of the classes  $\omega_i$ , for  $i = 1, 2$ .

### B. Recursive Optimization Procedure

The computation procedure of the sequence of PF discriminant vectors is as follows:

#### Initialization:

$\Delta\sigma^2 = 0$ ;  $X_1 = X_{d1}$ ,  $X_2 = X_{d2}$  where  $X_{d1}$ ,  $X_{d2}$  are the sphered design samples.

#### Step 1: Reduction of the class separation

- 1.1. Using the sample  $\{X_1 \cup X_2\}$ , compute the ExF vector with the largest PF distance (see Section IV).
- 1.2. Starting from the ExF vector, search to a convergence point by using a local maximizer (NAG routine E04UCF) of PF distance. The direction vector after convergence of the maximizer is a current PF vector denoted by  $\mathbf{r}$ . Save it.
- 1.3. Reduce the class separation along  $\mathbf{r}$ : Obtain an estimate of  $F_{\mathbf{r}}(\tau_1 | \omega_i)$  over the projections  $\tau_1 = \mathbf{r}^T \mathbf{x}$  of design observations  $\mathbf{x} \in X_i$  for  $i = 1, 2$  [8, p. 254]. Substitute the estimate of  $F_{\mathbf{r}}(\tau_1 | \omega_i)$  for  $i = 1, 2$  into (6) and (7). Transform  $\mathbf{x} \in \{X_1 \cup X_2\}$  using (8) and obtain new design sets  $X'_1$  and  $X'_2$  with reduced PF distance along  $\mathbf{r}$ . Assign the new sets to be the current sample sets, i.e.  $X_1 = X'_1$  and  $X_2 = X'_2$ .
- 1.4. If only one PF vector has been computed repeat above steps 1.1–1.3.

#### Step 2: Adjust (reoptimize) the PF vectors

Starting from PF vectors obtained in Step 1.2, search to the convergence points of the local optimizer of the PF distance for the original sphered data  $X_{d1}$ ,  $X_{d2}$ . The direction vectors after convergence of the algorithm are the *adjusted PF vectors*. Save them.

#### Step 3: Update the value of $\Delta\sigma^2$

Compare the class separation along the last two adjusted PF vectors, i.e. compare the PF distances and the class-conditional densities along them.

—If the class separation along the last two adjusted PF vectors is approximately equal, increase  $\Delta\sigma^2$ :  $\Delta\sigma^2 \leftarrow \Delta\sigma^2 + \varepsilon$ ,  $\varepsilon > 0$  (Here,  $\Delta\sigma^2 \leq 1$  and we recommend  $\varepsilon = 0.05 \div 0.1$ ).

—Otherwise assign  $\Delta\sigma^2 = 0$ .

#### Repeat Steps 1–3.

We stop the iterations if several adjusted PF vectors with different class separation along them are obtained. We regard the vectors

with the largest PF distances among all the adjusted PF vectors as “interesting” solutions (see Section VII—Experiments with real data set).

## VI. SIMULATION STUDIES

### A. Comparative Study of Friedman’s “Structure Removal” and Our Procedure for Reduction of the Class Separation

The main difference of our proposal for reduction of class separation (Section V-A) from Friedman’s [8] “structure removal” consists in the choice of the density which is transformed. Friedman’s algorithm is oriented to cluster analysis of unlabeled (unclassified) data. Applied to the labeled observations from two classes, it transforms the unconditional (mixture) density function  $p_{\mathbf{r}}(\zeta) = P(\omega_1)p_{\mathbf{r}}(\zeta | \omega_1) + P(\omega_2)p_{\mathbf{r}}(\zeta | \omega_2)$  along a direction vector  $\mathbf{r}$  ( $\zeta = \mathbf{r}^T \mathbf{x}$ ) to the standard normal density leaving all orthogonal directions unchanged. Our procedure separately transforms the projected class-conditional densities  $p_{\mathbf{r}}(\zeta | \omega_i)$ ,  $i = 1, 2$  to normal densities leaving all orthogonal directions to  $\mathbf{r}$  unchanged. We carried out an experiment in order to illustrate the difference between Friedman’s algorithm [8] and our procedure. The samples for two classes of the sample sizes  $N_{d1} = N_{d2} = 150$  were drawn from two-dimensional normal mixtures:

for class  $\omega_1$ :

$$p(x_1, x_2 | \omega_1) = 1/3N([0 \ 1]^T, \mathbf{I}) + 1/3N([5 \ 3]^T, \mathbf{I}) + 1/3N([0 \ 6]^T, \mathbf{I}), \quad (11)$$

for class  $\omega_2$ :

$$p(x_1, x_2 | \omega_2) = 1/3N([0 \ 3]^T, \mathbf{I}) + 1/3N([5 \ 6]^T, \mathbf{I}) + 1/3N([-5 \ 6]^T, \mathbf{I}). \quad (12)$$

Here,  $N([\mu_1 \ \mu_2]^T, \mathbf{I})$  denotes bivariate normal density with a mean vector  $[\mu_1 \ \mu_2]^T$  and an unit covariance matrix. Fig. 2(a) presents the sphered data (see Section II). We computed the PF distances for 91 equally angled directions into the  $(x_1, x_2)$ -plane. The result is shown in Fig. 1. The dotted path “---” presents the PF distances for the vectors  $\mathbf{r}$  directed under the angles  $0^\circ, 2^\circ, 4^\circ, \dots, 180^\circ$  with respect to  $x_1$ -axis. The angles  $50^\circ$  and  $142^\circ$  imply the local maxima of the PF distances. We applied Friedman’s and our procedures along the direction under  $50^\circ$ . Then we computed the PF distances for the transformed data sets for the same 91 directions as previously. The dashed path “- - -” of Fig. 1 presents the PF distances after Friedman’s “structure removal” and the solid path “—” the PF distances after our procedure for reduction of the class separation using  $\Delta\sigma^2 = 0$  (see Section V-A). Friedman’s algorithm preserved the classification structure of the data because the PF distances were approximately the same for the original (“---”) and the transformed (“- - -”) data. It didn’t deflate the local maximum along the direction under  $50^\circ$ . Our procedure eliminated the maximum at  $50^\circ$ . It caused destruction of the data which implied a unimodal shape of the PF distance (path “—” of Fig. 1). The unique maximum of the transformed PF distance (“—”) was close to the global maximum of the original PF distance (“---”). Our procedure exactly preserves the data in the subspace orthogonal to the direction which is the object of the reduction of the class separation (see Section V-A). In the case discussed here, we have not changed the data along the direction under  $140^\circ$ . In fact, we obtained approximately the same shapes of the paths “—” and “- - -” in the range  $140^\circ \pm 65^\circ$ . Therefore our procedure preserved the original classification structure in the latter range. In this experiment we eliminated the local maximum at  $50^\circ$  using the optimal normal approximation of the class-conditional densities ( $\Delta\sigma^2 = 0$ —see Section V-A). For other data a local maximum can be deflated using

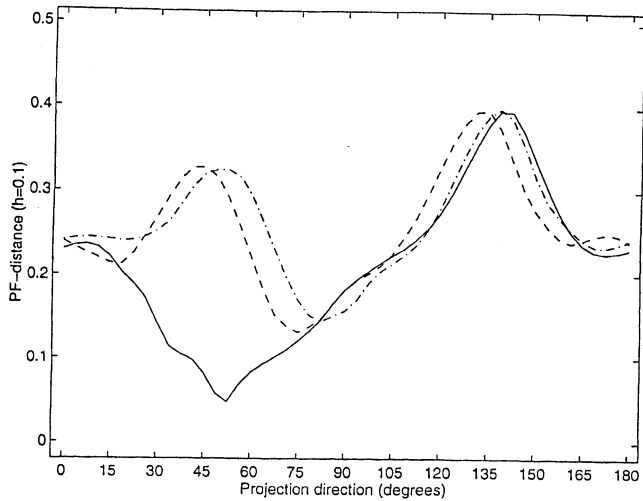


Fig. 1. PF distance for various directions into  $(x_1, x_2)$ -plane: ----- original sphered data; ---- transformed data by Friedman's "structure removal"; — transformed data by the procedure for reduction of the class separation.

a departure from optimality ( $\Delta\sigma^2 > 0$ ). This may cause stronger data destructuring and the range of the preserved data structure will be decreased. The proposed procedure in Section V-B preserves the data as much as possible after reduction of the class separation. For this purpose it searches for the smallest value of  $\Delta\sigma^2$  which directs the local optimizer to a new maximum of the PF distance (Step 3 of Section V-B). It seems reasonable to search for other (non-normal) density transformations which deflate the PF distance and cause less destructuring of the data. This is the object of our current research and it is beyond the scope of this work.

#### B. Recursive Optimization Applied to the Class-Conditional Distributions with Heavy Tails

In this experiment we studied our procedure for data which is highly unfavorable to it. We used data with significantly non-normal class-conditional distributions. Such distributions differ from the normal, mainly in the tails. We carried out an experiment with four-dimensional observations  $\mathbf{x} = [x_1 \ x_2 \ x_3 \ x_4]^T$  drawn from distributions  $p(\mathbf{x} | \omega_i) = p(x_1, x_2 | \omega_i)p(x_3, x_4 | \omega_i)$ ,  $i = 1, 2$ . Here the densities  $p(x_1, x_2 | \omega_i)$ ,  $i = 1, 2$  were the same as in the previous experiment [see (11) and (12)] and the densities  $p(x_3, x_4 | \omega_i)$ ,  $i = 1, 2$  were with heavy tails. The latter were constructed with the following mixtures of normal distributions:

for class  $\omega_1$ :

$$\begin{aligned} p(x_3, x_4 | \omega_1) &= 1/3N([-3 \ 0]^T, 0.01\mathbf{I}) + 1/3N([0.5 \ 3]^T, 0.01\mathbf{I}) \\ &\quad + 1/3N([-0.5 \ -3]^T, 0.01\mathbf{I}); \end{aligned}$$

for class  $\omega_2$ :

$$\begin{aligned} p(x_3, x_4 | \omega_2) &= 1/3N([-0.5 \ 3]^T, 0.01\mathbf{I}) + 1/3N([3 \ 0]^T, 0.01\mathbf{I}) \\ &\quad + 1/3N([0.5 \ -3]^T, 0.01\mathbf{I}). \end{aligned}$$

We generated 150 points per class ( $N_{d1} = N_{d2} = 150$ ). The classes were totally separated along a vector lying in the  $(x_3, x_4)$ -plane and directed under an angle of  $11^\circ$  with respect to  $x_3$ -axis. Fig. 2 presents the sphered data in the coordinate system spanned on the original  $x_i$ -axes. Because the sphering is not an orthonormal transformation (does not preserve the original interpoint distances), the best direction for

class separation is no longer in the plot shown in Fig. 2(b). Following the procedure of Section V-B, we computed the ExF discriminant vector for  $\beta = 0.5$ , which implied the maximal PF distance 0.3143. Fig. 3(a) shows the class-conditional densities along the latter vector. Starting from it we ran the local optimizer of the PF distance, which converged to a PF vector shown in Fig. 3(b). Inspecting the class-conditional densities along the latter vector [Fig. 3(b)], we concluded that it gains some of the class separation with respect to the starting ExF vector [Fig. 3(a)]. We iterated by a sequence of three reductions of the class separation with  $\Delta\sigma^2 = 0.2$  in order to illustrate the effect of a "large" value of the user-supplied parameter  $\Delta\sigma^2$ . Fig. 4 shows the data after the transformations were performed [see (8)]. Comparing the transformed data (Fig. 4) with the original (Fig. 2), we observed the following destructuring of the data:

- A significant class overlap into  $(x_1, x_2)$ -plane [Fig. 4(a)] was gained by this sequence of the reductions of the class separation. This was a desired result because our goal was to deflate the local maximum of the PF distance in the  $(x_1, x_2)$ -plane in order to direct the searching procedure to the  $(x_3, x_4)$ -plane with larger maxima.
- Some moving away of the encircled clusters in the  $(x_3, x_4)$ -plane [Fig. 4(b)] was caused by the reduction of the class separation. This is not a desired effect. It is a consequence of the "large" value of  $\Delta\sigma^2$  ( $\Delta\sigma^2 = 0.2$ ) which implied some shift ( $\Delta m_i$ ) of the classes. This result shows that a careful selection of the user-supplied parameter  $\Delta\sigma^2$  is crucial to the success of the proposed procedure.

Fig. 5(a) shows the final solution along the adjusted (reoptimized for the original data) PF vector after the third reduction of the class separation. We found a direction with large PF distance 0.6503, but we missed the global maximum of value 0.7353 [see Fig. 5(b)]. Consequently we do not view our procedure as a global optimization method, but as a tool which detects some directions with several large local maxima. We ran the procedure for recursive optimization with  $\Delta\sigma^2 = 0.0$ . After three reductions of the class separation we detected the global maximum of the PF distance [Fig. 5(b)].

## VII. EXPERIMENTS WITH A REAL DATA SET

The data concerning the medical diagnosis of the neurological disease cerebrovascular accident (CVA) contains pathologo-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]. In order to eliminate the small pooled variances of the data we used eight largest eigenvalues in the decomposition of the pooled sample covariance matrix (see Section II). Subsequently we reduced the dimensionality of the sphered data to eight.

#### A. Recursive Optimization

Following the recursive optimization procedure (Section V-B), we computed the ExF discriminant vector for  $\beta = 0$ , which implied a PF distance of maximal value 0.5046. Starting from it we ran the local optimizer of the PF distance, which converged to a PF discriminant vector with PF distance 0.8013. The class-conditional densities along the obtained discriminant vectors are shown in Fig. 7. The class separation was increased significantly along the PF vector [Fig. 7(b)]. Actually, this was the best result obtained in our study. In order to monitor the progress of our procedure we specified the coordinates of the sphered data which implied the class discrimination along the PF vector. They are the eighth, third, fourth, second, and seventh coordinates corresponding to the components of the PF vector with dominant values [see Fig. 7(b)]. We decided to monitor the

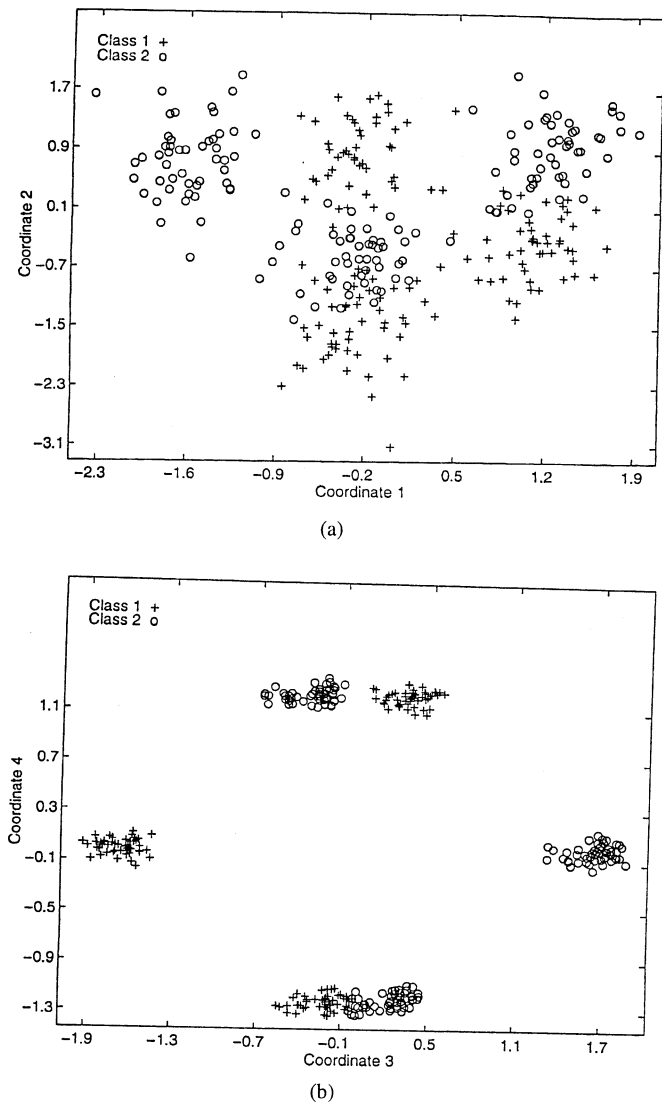


Fig. 2. Sphered data: (a)  $(x_1, x_2)$ -plane and (b)  $(x_3, x_4)$ -plane.

results using plots spanned on the third and fourth, and seventh and eighth coordinate-axes. Fig. 6 presents the data in these plots. We iterated by a sequence of reductions of the class separation in order to search for other directions with discriminant information. Following the proposed procedure we started trials with small values of  $\Delta\sigma^2$ . We iterated with  $\Delta\sigma^2 = 0.0, 0.0, 0.1, 0.2$  in the first to the fourth reductions, respectively. The adjusted PF vectors for this sequence converged to the result of the first trial [Fig. 7(b)]. Consequently these trials didn't direct the procedure to a new local maximum of the PF distance. We decided to continue with stronger reductions (larger values of  $\Delta\sigma^2$ ). We iterated with  $\Delta\sigma^2 = 0.3, 0.4, 0.5, 0.6$  and observed the adjusted PF distances 0.4475, 0.3498, 0.4129, 0.5069 in the fifth to the eighth trials. Fig. 8 shows the transformed data after the fifth reduction of the class separation. The presented destructuring of the data (with respect to the original data Fig. 6) directed the optimization procedure to maxima different from the initial solution. The eighth iteration implied an "interesting" result. We analyzed the discriminant information gained by the PF vector at this trial with respect to the best PF vector [Fig. 7(b)]. For this purpose we projected the original data on to the plot spanned on the latter vectors (Fig. 9). We detected a cluster (the encircled area in Fig. 9) of large (nearly full) overlap of the classes. The other

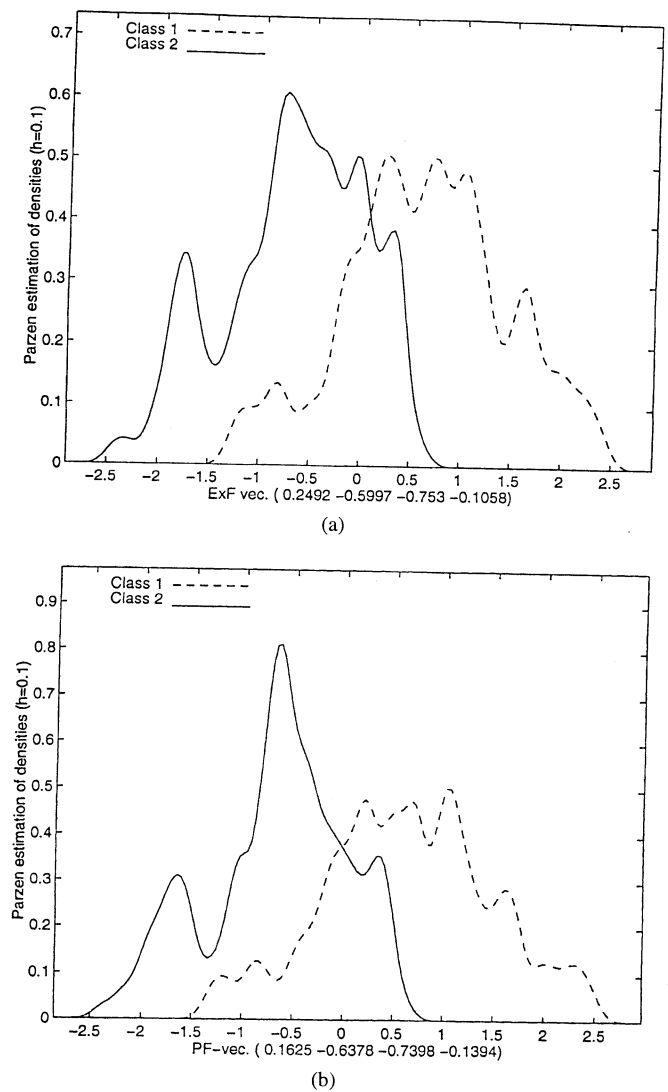


Fig. 3. Maximal class separation without reduction of the class separation: (a) along ExF vector ( $\beta = 0.5$ , PF distance 0.3143) and (b) along PF vector (PF distance 0.3226).

clusters define areas with a dominant number of the cases from one of the classes. We found that the two-dimensional presentation (Fig. 9) gains less class overlap compared with the one-dimensional projection shown in Fig. 7(b). We concluded that the PF vector at the eighth iteration adds "new" discriminant information to the best solution and consequently the obtained two-dimensional mapping may be a suitable choice for allocation of the CVA cases.

We assessed the discriminant performance by the design samples (samples used for computation of the PF vectors). Subsequently we carried out a resubstitution valuation of this performance. It is known that this approach leads to an optimistic result. In order to assess the actual allocation accuracy one has to estimate the probabilities of misclassification and rejection using "extra" (test) samples (see holdout and bootstrap methods [6], [9], [16]). Moreover, one could make trials with various values of the smoothing parameter  $h$  in the Parzen estimator (2) and could carry out a larger number of iterations with various values of  $\Delta\sigma^2$ . The latter detailed analysis of the CVA data is outside the scope of this paper. Our goal was to demonstrate that our procedure detected "interesting" directions which are worth examining for allocation of the CVA cases. We discussed a sequential procedure for selection the discriminant vectors. It seems more

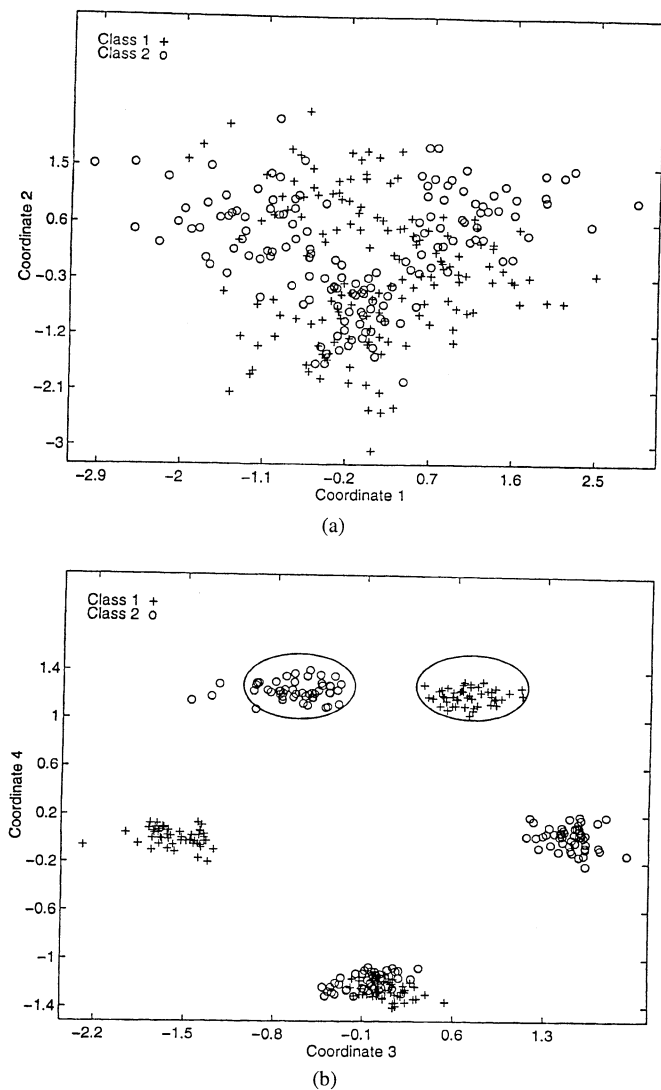


Fig. 4. Transformed data after three reductions of the class separation ( $\Delta\sigma^2 = 0.2$ ): (a)  $(x_1, x_2)$ -plane and (b)  $(x_3, x_4)$ -plane.

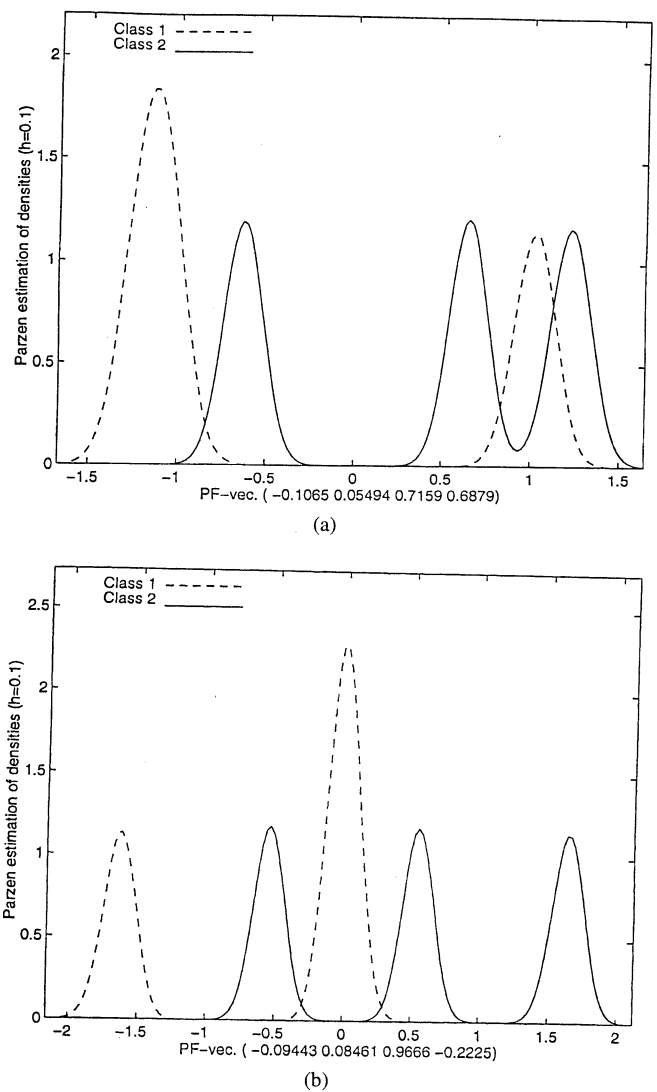


Fig. 5. Maximal class separation of the original data along the adjusted PF vectors after three reductions of the class separation: (a)  $\Delta\sigma^2 = 0.2$ , PF distance 0.6503 and (b)  $\Delta\sigma^2 = 0$ , PF distance 0.7353.

reasonable to make a one-shot search for  $m$  discriminant vectors ( $1 < m < n$ ,  $n$  is the dimension of the data) which maximize the PF distance into the subspace spanned on these vectors [6, pp. 277–280]. In the latter case we have to maximize over  $m \times n$  instead of  $n$  variables in the sequential approach. It is known [17, p. 591] that the computational complexity of the local optimizers is very high for more than 200 variables. Therefore the sequential approach, discussed in the paper, is an attractive choice for  $n$ -dimensional data with  $n > 100$ .

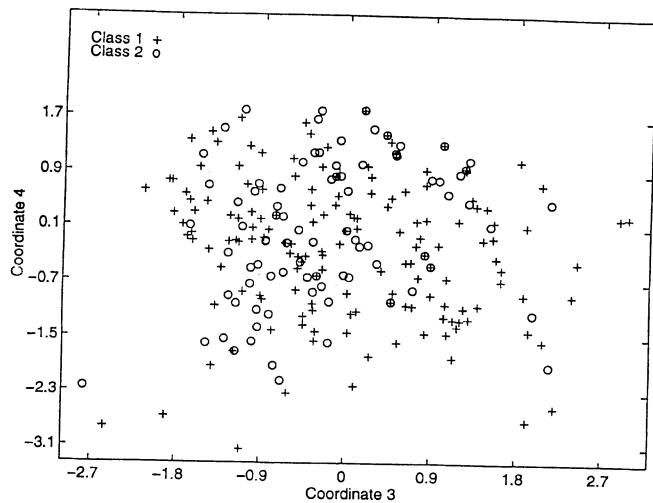
### B. Optimization Without Reduction of the Class Separation

The goal of these experiments was to examine the usefulness of the reduction of the class separation. For this purpose we compared the result of the previous section with the results of other trials for optimization of the PF distance of the CVA data.

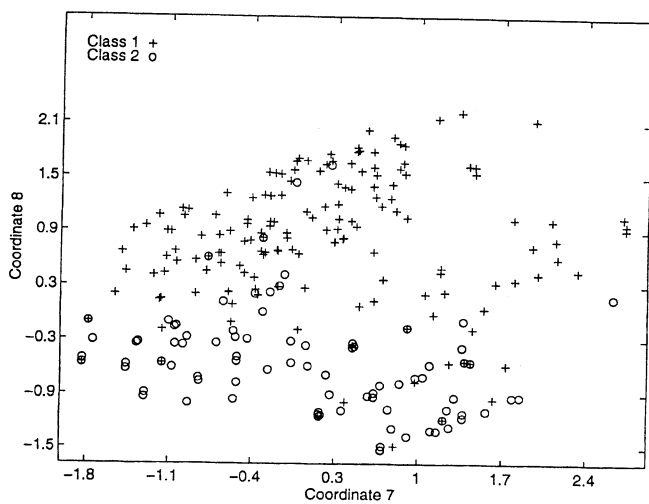
1) *ExF Initialization of the Local Optimizer*: We maximized (with no reductions of the class separation) the PF distance of the CVA data starting the local optimizer from the ExF vectors used in the previous experiment (ExF vectors for  $\beta = 0.0, 0.5, 0.7, 1.0$ ). After convergence of the local search we observed the PF distances 0.8013 for  $\beta = 0.0, 0.5, 0.7$  and 0.4436 for  $\beta = 1.0$ . The values  $\beta = 0.0, 0.5, 0.7$  implied the best solution obtained in Section VII-A

[Fig. 7(b)]. This was not a surprise because our procedure found this solution at the iteration without reduction of the class separation. The value of  $\beta = 1.0$  implied the local maximum of value 0.4436, which was less than the value of 0.5069 found by our procedure at the eighth iteration.

2) *Principal Component Initialization of the Local Optimizer*: We maximized (with no reductions of the class separation) the PF distance starting the optimizer from the principal component directions of the CVA data. Since we are working with the sphered data, its coordinate axes are in fact the principal component directions when referenced to the original data [8, p. 256]. In our experiment with the CVA data we used eight principal components—eight eigenvectors corresponding to the largest eigenvalues of the pooled sample covariance matrix. They were ordered in increasing eigenvalues, i.e. the eighth axis corresponds to the largest eigenvalue. We observed the PF distances 0.2169, 0.2055, 0.3291, 0.3100, 0.4468, 0.3968, 0.3699 and 0.8013 at the convergence points of the local optimizer starting from the first to the eighth axes of the sphered data. The eighth axis (the principal component corresponding to the largest eigenvalue) implied the best solution of our procedure [Fig. 7(b)]. This is not surprising, seeing that the PF vector of this solution has a dominant value 0.8139 for



(a)



(b)

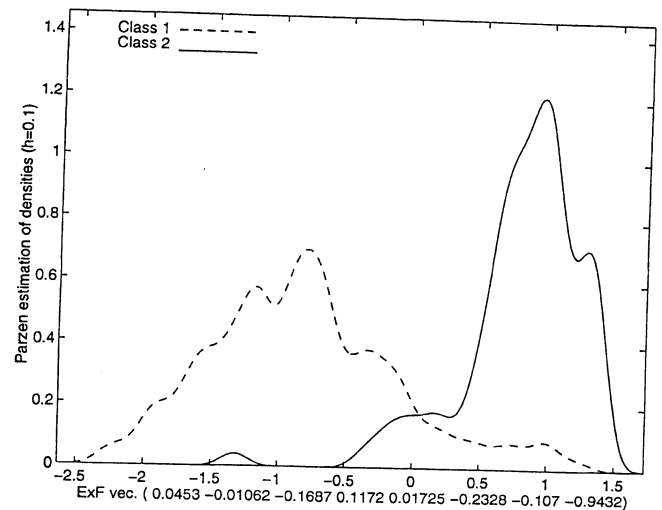
Fig. 6. Sphered CVA data: (a) along third and fourth coordinate axes and (b) along seventh and eighth coordinate axes.

its eighth component. Other axes implied values of the local maxima which were less than the PF distance 0.5069 observed at the eighth iteration of our procedure.

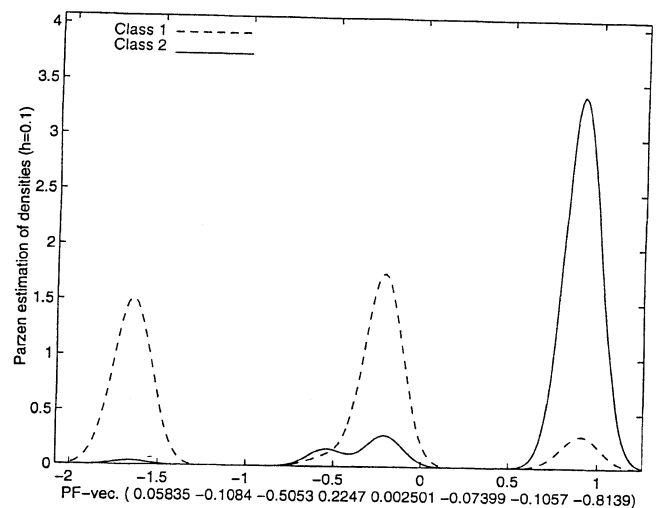
Finally, our procedure for recursive optimization managed to find, at the eighth iteration, the direction with a large PF distance of the CVA data, which was not achieved by the ExF and the principal component initializations of the local optimizer.

### VIII. SUMMARY AND CONCLUSION

We have presented a method for the linear discrimination of two classes based on the Patrick-Fisher (PF) distance. Since the PF distance is a highly nonlinear function, its optimization was carried out using a recursive method. Just like any other projection pursuit procedure [8], [13], our method searches for the discriminant directions corresponding to several large local maxima of the PF distance. The proposed method succeeded in finding the sequence of directions with significant discriminant information for a simulation study and a medical application considered in Sections VI and VII. Our recursive optimization procedure was more effective than the algorithms with the ExF and the principal component initializations of the local optimizer (Section VII-B).



(a)



(b)

Fig. 7. Maximal class-separation without reduction of the class separation (CVA data): (a) along ExF vector ( $\beta = 0$ , PF distance 0.5046) and (b) along PF vector (PF distance 0.8013).

Our method implements Friedman's [8] procedure for recursive optimization, called "structure removal." We summarize the main features of this implementation.

- As in Friedman's procedure we transform the densities along the direction with an observed maximum into normal densities. We said in Section VI-A that, depending on the context, other (non-normal) density transformation may serve better for the recursive optimization. Nevertheless, the normal transformation implied successful results for the experiments presented in the Sections VI and VII. They were carried out, purposely, for data highly unfavorable for the normal assumption- low dimensional data with highly structured classes (four- and eight-dimensional data with heavy tails of the class-conditional densities).
- The main difference of our procedure for reduction of the class separation from Friedman's [8] "structure removal" consists in the choice of the density which is transformed. Friedman's algorithm transforms the mixture (unconditional) density to the normal density while our algorithm processes the class-conditional densities separately. The latter is caused by the goal of our method. It is a tool for discriminant analysis while Friedman's procedure is oriented to cluster analysis of

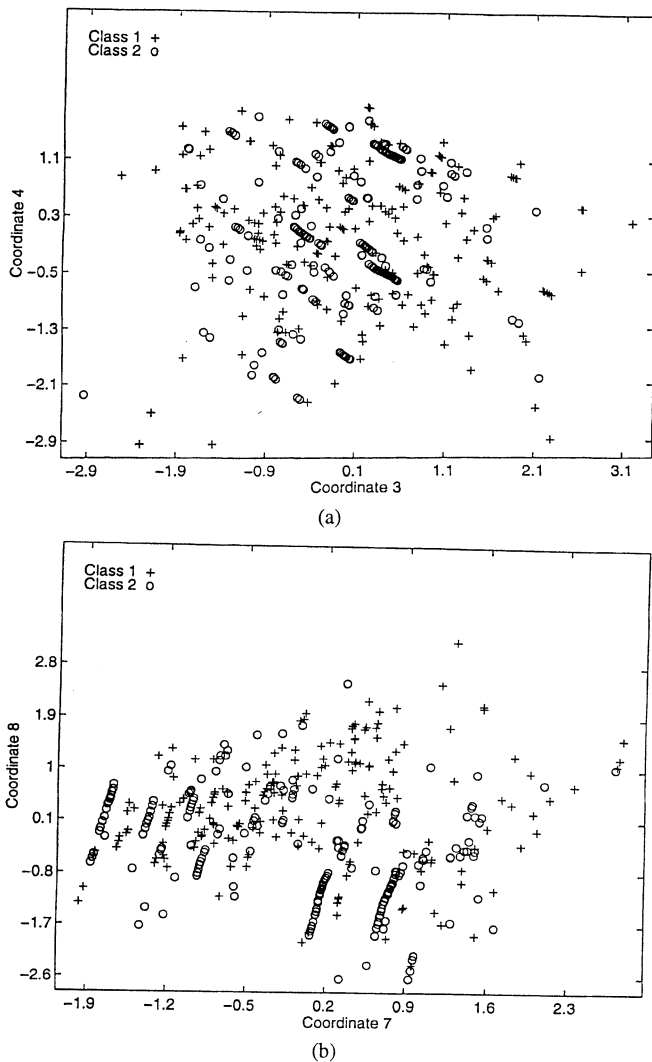


Fig. 8. Transformed CVA data, after the fifth reduction of the class separation: (a) along third and fourth coordinate axes, (b) along seventh and eighth coordinate axes.

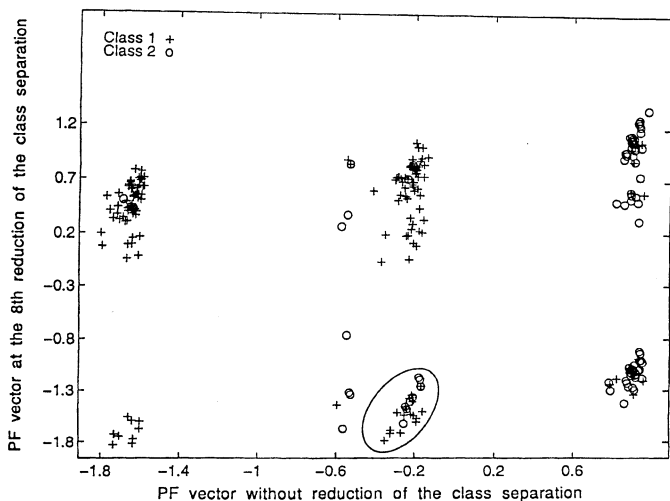


Fig. 9. Projection of the CVA data on to the plot spanned by the discriminant vectors with largest PF distances.

unclassified samples. We demonstrated the difference between Friedman's and our algorithms in Section VI-A.

- Our procedure, as is the case with the stochastic smoothing algorithms [11], may miss some largest local maxima including the global one. This was demonstrated in Section VI-B for the computer simulation with  $\Delta\sigma^2 = 0.2$ . We view this as desirable and do not think of our procedure as a global optimization method, but rather as a simple tool which detects directions with "interesting" discriminant information of  $n$ -dimensional data. If the goal is the global maximization of the PF distance then more complicated algorithms, like simulated annealing and other global optimizers, may be used. Unfortunately they require many evaluations of the objective function which leads to long run time. The sequential quadratic programming routine E04UCF in NAG, used by us as the local optimizer, was viewed by many users as the best one when the number  $n$  of the variables of the optimized function is less than two hundred [17, p. 591]. For larger problems ( $n > 200$ ) a refined version of this local optimizer [17] may be used.

#### REFERENCES

- [1] M. E. Aladjem, "PNM: A program for parametric and nonparametric mapping of multidimensional data," *Comput. Biol. Med.*, vol. 21, pp. 321–343, 1991.
- [2] —, "Multiclass discriminant mappings," *Signal Processing*, vol. 35, pp. 1–18, 1994.
- [3] —, "Two class pattern discrimination via recursive optimization of Patrick-Fisher distance," in *Proc. 13th Int. Conf. Pattern Recognition*, 1996, vol. 2, pp. 60–64.
- [4] —, "Nonparametric discriminant analysis applied to medical diagnosis," in *Proc. 19th Convention Electrical Electronics Engineers Israel*, 1996, pp. 422–425.
- [5] M. E. Aladjem, "Linear discriminant analysis for two classes via removal of classification structure," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 187–192, 1997.
- [6] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. London, U.K.: Prentice-Hall, 1982.
- [7] P. Diaconis and D. Freedman, "Asymptotics of graphical projection pursuit," *Ann. Statist.*, vol. 12, pp. 793–815, 1984.
- [8] J. H. Friedman, "Exploratory projection pursuit," *J. Amer. Statist. Assoc.*, vol. 82, pp. 249–266, 1987.
- [9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.
- [10] P. E. Gill, W. Murray, A. S. Michael, and M. H. Wright, "User's guide for NPSOL (version 4.0), Tech. Rep., Syst. Optim. Lab., Stanford Univ., Stanford, CA, 1986.
- [11] P. Gilmore and C. T. Kelley, "An implicit filtering algorithm for optimization of functions with many local minima," *SIAM J. Optim.*, vol. 5, pp. 269–285, 1995.
- [12] A. Hillion, P. Masson, and C. Roux, "A nonparametric approach to linear feature extraction; Application to classification of binary synthetic textures," in *Proc. 9th Int. Conf. Pattern Recognition*, 1988, pp. 1036–1039.
- [13] P. J. Huber, "Projection pursuit," including discussions, *Ann. Statist.*, vol. 13, pp. 435–525, 1985.
- [14] T. Lissack and K. Fu, "Error estimation in pattern recognition via  $L^\alpha$ -distance between posterior density functions," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 34–45, 1976.
- [15] W. Malina, "On an extended Fisher criterion for feature selection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 3, pp. 611–614, 1981.
- [16] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley, 1992.
- [17] W. Murray and F. J. Prieto, "A sequential quadratic programming algorithm using an incomplete solution of the subproblem," *SIAM J. Optim.*, vol. 5, pp. 590–640, 1995.
- [18] E. A. Patrick and F. P. Fisher II, "Nonparametric feature selection," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 577–584, 1969.
- [19] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Charman & Hall, 1986.
- [20] J. Sun, "Some practical aspects of exploratory projection pursuit," *SIAM J. Sci. Comput.*, vol. 14, pp. 68–80, 1993.