

Nonparametric Linear Discriminant Analysis by Recursive Optimization with Random Initialization ^{*}

Mayer Aladjem

Department of Electrical and Computer Engineering,
Ben-Gurion University of the Negev, P.O.B. 653,
84105 Beer-Sheva, Israel
aladjem@ee.bgu.ac.il

WWW home page: http://www.ee.bgu.ac.il/faculty/m_a.html

Abstract. A method for the linear discrimination of two classes has been proposed by us in [3]. It searches for the discriminant direction which maximizes the 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 distance between the projected densities is a highly nonlinear function with respect to the projected direction we maximize the objective function by an iterative optimization algorithm. The solution of this algorithm depends strongly on the starting point of the optimizer and the observed maximum can be merely a local maximum. In [3] we proposed a procedure for recursive optimization which searches for several local maxima of the objective function ensuring that a maximum already found will not be chosen again at a later stage. In this paper we refine this method. We propose a procedure which provides a batch mode optimization instead an interactive optimization employed in [3]. By means of a simulation we compare our procedure and the conventional optimization starting optimizers at random. The results obtained confirm the efficacy of our method.

1 Introduction

We discuss discriminant analysis which searches for a discriminant direction by maximizing the distance between the projected class-conditional densities. Unfortunately this distance is a highly nonlinear function with respect to the projected directions, and has more than one maximum. In most applications the optimal solution is searched for along the gradient of the objective function, hoping that with a good starting point the optimization procedure will converge to the global maximum or at least to a practical one. Some known techniques such as principal component analysis, Fisher discriminant analysis and their combination [1],[2] may be used for choosing a starting point for the optimization

^{*} This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University of the Negev, Israel.

procedure. Nevertheless, the observed maximum of the objective function can be merely a local maximum, which is far away from the global one in some data structures. In [3] we proposed a method for recursive optimization which searches for several large local maxima of the objective function. In this paper we refine this method. We propose a procedure for recursive optimization which ensures a batch mode optimization. Optimizing in this mode we replicate the recursive optimization using different starting points of the optimizer and then choose the best solutions from the trials done.

Section 2 describes our method for discriminant analysis [3], Section 3 presents our new proposal, and Sections 4 and 5 contain the results and analyses of the comparison based on the synthetic data sets.

2 Discriminant Analysis by Recursive Optimization [3]

Suppose we are given training data $(\mathbf{z}_1, \mathbf{c}_1), (\mathbf{z}_2, \mathbf{c}_2), \dots, (\mathbf{z}_{N_t}, \mathbf{c}_{N_t})$ comprising a set $Z_t = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{N_t}\}$ of N_t training observations in n -dimensional sample space ($\mathbf{z}_j \in \mathbb{R}^n, n \geq 2$) and their associated class-indicator vectors $\mathbf{c}_j, j = 1, 2, \dots, N_t$. We discuss a two class problem and we require that \mathbf{c}_j is a two-dimensional vector $\mathbf{c}_j = (c_{1j}, c_{2j})^T$ which shows that \mathbf{z}_j belongs to one of the classes ω_1 or ω_2 . The components c_{1j}, c_{2j} are defined to be one or zero according to the class-membership of \mathbf{z}_j , i.e. $c_{1j} = 1, c_{2j} = 0$ for $\mathbf{z}_j \in \omega_1$ and $c_{1j} = 0, c_{2j} = 1$ for $\mathbf{z}_j \in \omega_2$. The class-indicator vectors \mathbf{c}_j imply decomposition of the set Z_t into two subsets corresponding to the unique classes. We denote by N_{t_i} the number of the training observations in class ω_i , for $i = 1, 2$.

Our method requires a normalization of the data, called sphering [6]. To achieve data sphering 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 estimated over training set Z_t . Here \mathbf{R} and \mathbf{D} are $n \times n$ 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 non-singular, otherwise only the eigenvectors corresponding to the non-zero eigenvalues must be used in the decomposition [6]. In the remainder of the paper, all operations are performed on the *sphered training data* $X_t = \{\mathbf{x}_j : \mathbf{x}_j = \mathbf{A}(\mathbf{z}_j - \mathbf{m}_z), \mathbf{z}_j \in Z_t, j = 1, 2, \dots, N_t\}$ with \mathbf{m}_z the sample mean vector estimated over Z_t . For the sphered training data X_t the pooled sample covariance matrix becomes the identity matrix $\mathbf{A}\mathbf{S}_z\mathbf{A}^T = \mathbf{I}$.

We discuss discriminant analysis carried out by a linear mapping $y = \mathbf{w}^T \mathbf{x}$, $\mathbf{x} \in \mathbb{R}^n, y \in \mathbb{R}^1, n \geq 2$, with \mathbf{x} an arbitrary n -dimensional observation, and \mathbf{w} a direction vector. We require \mathbf{w} to have unit length, and $y = \mathbf{w}^T \mathbf{x}$ can be interpreted geometrically as the projection of the observation \mathbf{x} onto vector \mathbf{w} in \mathbf{x} -space (Fig.1).

We search for the discriminant vector \mathbf{w}^*

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \{PF(\mathbf{w})\}$$

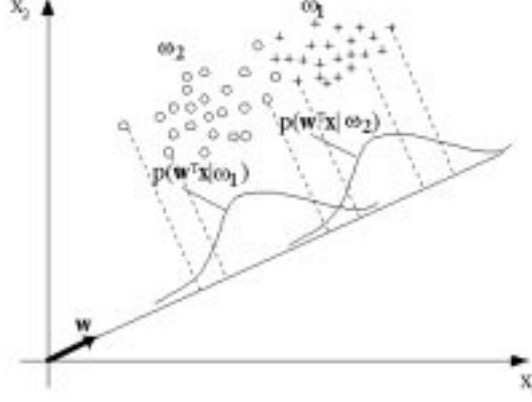


Fig. 1. Linear mapping ($y = \mathbf{w}^T \mathbf{x}$) in a two-dimensional \mathbf{x} -space. Class-conditional densities $p(\mathbf{w}^T \mathbf{x}|\omega_1)$ and $p(\mathbf{w}^T \mathbf{x}|\omega_2)$ along the vector \mathbf{w} .

which maximizes the *Patrick-Fisher (PF) distance* [5] between the class-conditional densities along it. $PF(\mathbf{w})$ denotes the PF distance along an arbitrary vector \mathbf{w}

$$PF(\mathbf{w}) = \left\{ \int_{\mathbb{R}^n} \left[\frac{N_{t1}}{N_t} \hat{p}(\mathbf{w}^T \mathbf{x}|\omega_1) - \frac{N_{t2}}{N_t} \hat{p}(\mathbf{w}^T \mathbf{x}|\omega_2) \right]^2 d\mathbf{x} \right\}^{1/2} \quad (1)$$

with

$$\hat{p}(\mathbf{w}^T \mathbf{x}|\omega_i) = \frac{1}{h\sqrt{2\pi}N_{ti}} \sum_{j=1}^{N_{ti}} c_{ij} \exp \left\{ \frac{-1}{2h^2} [\mathbf{w}^T (\mathbf{x} - \mathbf{x}_j)]^2 \right\}, i = 1, 2 \quad (2)$$

the Parzen estimators with Gaussian kernels of the class-conditional densities of the projections $y = \mathbf{w}^T \mathbf{x}$. Here \mathbf{x} is an arbitrary observation ($\mathbf{x} \in \mathbb{R}^n$), c_{ij} is the class-indicator which constrains the summation in (2) on the ω_i -training observations (\mathbf{x}_j corresponding to $c_{ij} = 1$), and h is a smoothing parameter.

$PF(\mathbf{w})$ is a nonlinear function with respect to \mathbf{w} . In order to search for several large local maxima of $PF(\mathbf{w})$ we have proposed a method for recursive maximization of $PF(\mathbf{w})$ [3]. We obtain a discriminant vector \mathbf{w}^* related to a local maximum $PF(\mathbf{w}^*)$ and then we transform the data along \mathbf{w}^* into data with greater overlap of the class-conditional densities (deflated maximum of $PF(\mathbf{w})$ at the solution \mathbf{w}^*), and iterate to obtain a new discriminant vector.

In our method we use the PF distance (1) because of the existence of an analytical expression of its gradient [5] used in the iterative optimization. Actually our method is not restricted to the PF distance only. It can be applied to any other discriminant criterion, which has several local maxima with respect to \mathbf{w} . In the case that an analytical expression of the gradient of the criterion can not be obtained we must estimate the gradient numerically.

The main point of the method is the procedure for deflating the local maximum of $PF(\mathbf{w})$ called *Reduction of the Class Separation* (RCS). In order to deflate $PF(\mathbf{w})$ at \mathbf{w}^* (to increase class overlap along \mathbf{w}^*), we transform class-conditional densities along \mathbf{w}^* to normal densities. For this purpose, we rotate the data applying the linear transformation

$$\mathbf{r} = \mathbf{U}\mathbf{x} \quad (3)$$

with \mathbf{U} an orthonormal ($n \times n$) matrix. We denote the new coordinates as r_1, r_2, \dots, r_n ($\mathbf{r} = (r_1, r_2, \dots, r_n)^T$). We require that the first row of \mathbf{U} is \mathbf{w}^* , which results in a rotation such that the new first coordinate of an observation \mathbf{x} is $r_1 = y = (\mathbf{w}^*)^T \mathbf{x}$. Assume that $p(y|\omega_i), i = 1, 2$ are the class-conditional densities of $y = (\mathbf{w}^*)^T \mathbf{x}$ and $\mathbf{m}_{y|\omega_i}, \sigma_{y|\omega_i}^2$ their means and variances. We transform $p(y|\omega_i)$ to normal densities and leave the coordinates r_2, r_3, \dots, r_n unchanged. Let \mathbf{q} be a vector function with components q_1, q_2, \dots, q_n that carries out this transformation: $r_1' = q_1(y)$ with r_1' having normal class-conditional distributions and $r_i' = q_i(r_i), i = 2, 3, \dots, n$ each given by the identity transformations. The function q_1 is obtained by the percentile transformation method:

- for observations \mathbf{x} from class ω_1 :

$$q_1(y) = [\Phi^{-1}(F(y|\omega_1))] (\sigma_{y|\omega_1}^2 \pm \Delta\sigma^2)^{1/2} + (\mathbf{m}_{y|\omega_1} - \Delta\mathbf{m}_1); \quad (4)$$

- for observations \mathbf{x} from class ω_2 :

$$q_1(y) = [\Phi^{-1}(F(y|\omega_2))] (\sigma_{y|\omega_2}^2 \pm \Delta\sigma^2)^{1/2} + (\mathbf{m}_{y|\omega_2} - \Delta\mathbf{m}_2). \quad (5)$$

Here, $\Delta\sigma^2 (0 \leq \Delta\sigma^2 \leq 1)$, $\Delta\mathbf{m}_1, \Delta\mathbf{m}_2$ are user-supplied parameters, $F(y|\omega_i)$ is the class-conditional (cumulative) distribution function of $y = (\mathbf{w}^*)^T \mathbf{x}$ for $i = 1, 2$ and Φ^{-1} is the inverse of the standard normal distribution function Φ . Finally,

$$\mathbf{x}' = U^T \mathbf{q}(\mathbf{U}\mathbf{x}) \quad (6)$$

transforms the class-conditional densities along \mathbf{w}^* to be normal densities

$$p(r_1'|\omega_i) = N(\mathbf{m}_{y|\omega_i} - \Delta\mathbf{m}_i, \sigma_{y|\omega_i}^2 \pm \Delta\sigma^2) \quad (7)$$

leaving all directions orthogonal to \mathbf{w}^* unchanged. If we set $\Delta\sigma^2 = 0$ and $\Delta\mathbf{m}_i = 0, i = 1, 2$ 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 [6, p.254] and [7, p.456]. If $\sigma_{y|\omega_i}^2 \pm \Delta\sigma^2 = 1$ and $\mathbf{m}_{y|\omega_i} - \Delta\mathbf{m}_i = 0, i = 1, 2$ we transform the class-conditional densities along \mathbf{w}^* to $N(0, 1)$ which results in full overlap of the classes along \mathbf{w}^* . This certainly eliminates the local maximum of the PF distance along \mathbf{w}^* , but it causes large changes of the distributions of the transformed data \mathbf{x}' (6) in some applications. In order to direct the local optimizer to a new maximum of $PF(\mathbf{w})$, and to keep the class-conditional densities of \mathbf{x}' (6) as close to the densities of the original data \mathbf{x} as possible we search for the smallest values of the parameters $\Delta\sigma^2$,

$\Delta \mathbf{m}_1, \Delta \mathbf{m}_2$ that result in a deflated PF distance along \mathbf{w}^* . We start our search with $\Delta \sigma^2 = 0$ and $\Delta \mathbf{m}_i = 0, i = 1, 2$ (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 $\sigma_{y|\omega_i}^2 \pm \Delta \sigma^2$ to 1. We assign the latter value to 1 if it crosses 1. For each $\Delta \sigma^2$ we compute the values of $\Delta \mathbf{m}_1$ and $\Delta \mathbf{m}_2$ by an expression proposed by us in [3,p.294].

We presented the RCS in its abstract version based on probability distributions. The application to observed data is accomplished by substituting estimates of $F(y|\omega_i)$, $\mathbf{m}_{y|\omega_i}$ and $\sigma_{y|\omega_i}^2$ over the training set X_t .

3 Batch Mode Recursive Optimization Procedure

Here we refine our recursive optimization procedure developed in [3]. The idea is to ensure optimization in a batch mode instead of the optimization in an interactive mode proposed in [3]. For this purpose we propose a procedure which performs successive modification of the training data automatically (without man-machine interactions). In order to formalize this procedure we introduce the following nomenclature:

- X_t denotes the original (sphered) training data.
- \mathbf{w}^* is the directional vector corresponding to the local maximum of the PF distance for the original data X_t .
- $\tilde{\mathbf{X}}_t$ denotes the training data used in the current iteration of the procedure.
- $\tilde{\mathbf{w}}$ is the directional vector corresponding to the local maximum of the PF distance for the current training data $\tilde{\mathbf{X}}_t$.
- $\tilde{\mathbf{X}}_t'$ denotes the modified training data which has a deflated PF distance along $\tilde{\mathbf{w}}$.
- $\tilde{\mathbf{w}}'$ is the directional vector corresponding to the local maximum of the PF distance for the modified training data $\tilde{\mathbf{X}}_t'$.

We propose the following computational procedure:

Step 1 *Starting from a directional vector we maximize the PF distance for the original training data X_t .* We save the optimal solution denoted by \mathbf{w}^* .

Step 2 *Initialization of the Reduction of the Class-Separation (RCS):* We initialize the current training data $\tilde{\mathbf{X}}_t$ with the original training data X_t ($\tilde{\mathbf{X}}_t = X_t$) and the current optimal solution $\tilde{\mathbf{w}}$ with the optimal solution \mathbf{w}^* for X_t ($\tilde{\mathbf{w}} = \mathbf{w}^*$). We set $\Delta \sigma^2 = 0$ and $\Delta \mathbf{m}_i = 0$, for $i = 1, 2$. This setting implies minimal changes of the data during the RCS.

Step 3 *Running the RCS:* We estimate the class- conditional means, variances and (cumulative) distribution functions over the projections $y_j = (\tilde{\mathbf{w}})^T \tilde{\mathbf{x}}_j$ of the current training observations $\tilde{\mathbf{x}}_j \in \tilde{\mathbf{X}}_t$ onto the current optimal vector $\tilde{\mathbf{w}}$. We substitute these estimates into (4) and (5), transform $\tilde{\mathbf{x}}_j \in \tilde{\mathbf{X}}_t$ using (6), and obtain the modified training data $\tilde{\mathbf{X}}_t' = \{\mathbf{x}_j' : \mathbf{x}_j' = \mathbf{U}^T \mathbf{q}(\mathbf{U} \tilde{\mathbf{x}}_j) \quad j = 1, 2 \dots N_t\}$ which has a deflated PF distance along the optimal solution $\tilde{\mathbf{w}}$.

Step 4 *Starting from $\tilde{\mathbf{w}}$ we maximize the PF distance for the modified training data $\tilde{\mathbf{X}}_t'$.* We save the optimal solution denoted by $\tilde{\mathbf{w}}'$.

Step 5 Starting from $\tilde{\mathbf{w}}'$ we maximize the PF distance for the original training data X_t . We save the optimal solution \mathbf{w}^* .

Step 6 Updating the control parameters $\Delta\sigma^2$, $\Delta\mathbf{m}_1$, $\Delta\mathbf{m}_2$ and the current training data $\tilde{\mathbf{X}}_t$: We compare the last two solutions \mathbf{w}^* saved in Step 5 and for the first trial in Step 1 and Step 5.

(a) If the last two solutions \mathbf{w}^* are equal, we increase $\Delta\sigma^2$ (deflate more strongly the PF distance along $\tilde{\mathbf{w}}$) and update $\Delta\mathbf{m}_1$, $\Delta\mathbf{m}_2$ by an expression proposed by us in [3,p.294]. Our experience is that an increase of $\Delta\sigma^2$ with step-size 0.1 is suitable.

(b) If the last two solutions \mathbf{w}^* are different (different local maxima of the PF distance have been identified) we update the current training data $\tilde{\mathbf{X}}_t$ with the modified training data $\tilde{\mathbf{X}}'_t$ ($\tilde{\mathbf{X}}_t = \tilde{\mathbf{X}}'_t$), update the current direction of the RCS $\tilde{\mathbf{w}}$ with the optimal solution $\tilde{\mathbf{w}}'$ for $\tilde{\mathbf{X}}'_t$ ($\tilde{\mathbf{w}} = \tilde{\mathbf{w}}'$) and restore the initial values of the control parameters $\Delta\sigma^2 = 0$, $\Delta\mathbf{m}_1 = 0$, $\Delta\mathbf{m}_2 = 0$.

Then we repeat Steps 3-6. We stop the iterations if several optimal solutions \mathbf{w}^* corresponding to different values of the $PF(\mathbf{w}^*)$ (1) are obtained.

We replicate the proposed procedure (Steps 1-6) starting from different initial vectors in Step 1. We choose them by the preliminary principal component and Fisher discriminant analysis as we did in [3] and at random which is a usual initialization in the conventional optimization. Finally we choose from the vectors \mathbf{w}^* saved in Step 5 those corresponding to large values of $PF(\mathbf{w}^*)$. We regard the selected \mathbf{w}^* , as "interesting" solutions.

4 An Interactive Run of the Recursive Optimization Procedure

Here we demonstrate the recursive optimization procedure in a run using two dimensional synthetic data. We used samples for two classes of the sample sizes $N_{t1} = N_{t2} = 50$, which were drawn from two-dimensional normal mixtures:

for class ω_1 :

$$p(x_1, x_2 | \omega_1) = \frac{1}{3}N([-1.5 \ 0]^T, \Sigma) + \frac{1}{3}N([0.5 \ -3]^T, \Sigma) + \frac{1}{3}N([-1 \ -3]^T, \Sigma),$$

for class ω_2 :

$$p(x_1, x_2 | \omega_2) = \frac{1}{3}N([-0.5 \ 3]^T, \Sigma) + \frac{1}{3}N([3 \ 0]^T, \Sigma) + \frac{1}{3}N([0.5 \ -3]^T, \Sigma).$$

Here, $N([\mu_1 \ \mu_2]^T, \Sigma)$ denotes bivariate normal density with a mean vector $[\mu_1 \ \mu_2]^T$ and a diagonal covariance matrix $\Sigma = \text{diag}(0.1, 0.2)$. Fig.2 presents the original (sphered) training data X_t .

For this data we computed the PF distances for 91 equally angled directions into the (x_1, x_2) -plane. The solid path "—" in Fig.3 presents $PF(\mathbf{w})$ (1) for the vectors \mathbf{w} directed under different angles with respect to x_1 -axis. We observe local maxima of $PF(\mathbf{w})$ at angles 19° , 64° , 105° , 128° and 162° . We ran our procedure described in Section 3. In Step 1 we chose \mathbf{w}^* directed under 105°

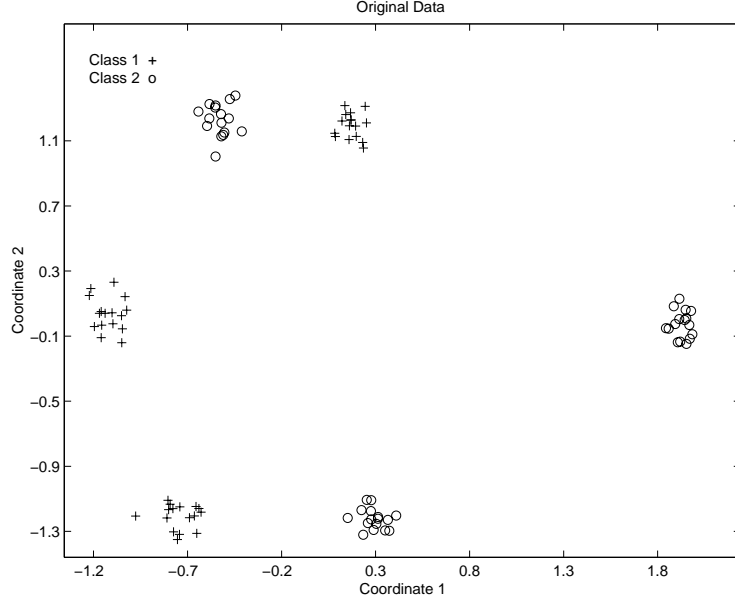


Fig. 2. Original data X_t .

with respect to x_1 -axis and observed local maximum $PF(\mathbf{w}^*) = 0.52$ (see "—" in Fig.3). Then we ran Steps 3-6 three successive times keeping $\Delta\sigma^2 = 0$, $\Delta\mathbf{m}_1 = 0$, $\Delta\mathbf{m}_2 = 0$.

Here we analyze the result obtained in the first run of the Steps 3-6. In Fig. 4 we present the transformed data $\tilde{\mathbf{X}}_t'$ obtained in Step 3. Comparing $\tilde{\mathbf{X}}_t'$ (Fig.4) with the original data X_t (Fig.2), we observe that a significant class overlap was gained along the direction under 105° for the transformed data $\tilde{\mathbf{X}}_t'$. This is a desired result because our goal was to deflate the local maximum of the PF distance at 105° in order to direct the local optimizer to another solution. We calculated $PF(\mathbf{w})$ for $\tilde{\mathbf{X}}_t'$ using different directions of \mathbf{w} and show the PF-path "... " in Fig.3. We observe that our procedure eliminated the maximum at 105° and smoothed the shape of the PF distance in the range $45^\circ - 180^\circ$ causing some restructuring of its shape. It seems reasonable to search for other data transformations which cause less restructuring of the PF distance. In [4] we proposed a neural network implementation of the RCS which by performing highly non-linear data transformation decreases the restructuring of the PF distance, but its complexity is higher than that of the procedure proposed in Section 3.

In the second and third iterations of Steps 3-6 we computed the PF distances for the successively transformed data $\tilde{\mathbf{X}}_t'$. The PF-paths are shown in Fig.3. The local maxima of these paths at 83° , 15° and 152° defined the starting points of the optimizer of PF distance for the original data X_t in Step 5. Using them our procedure converges to the solutions at 64° , 19° and 162° for the original data

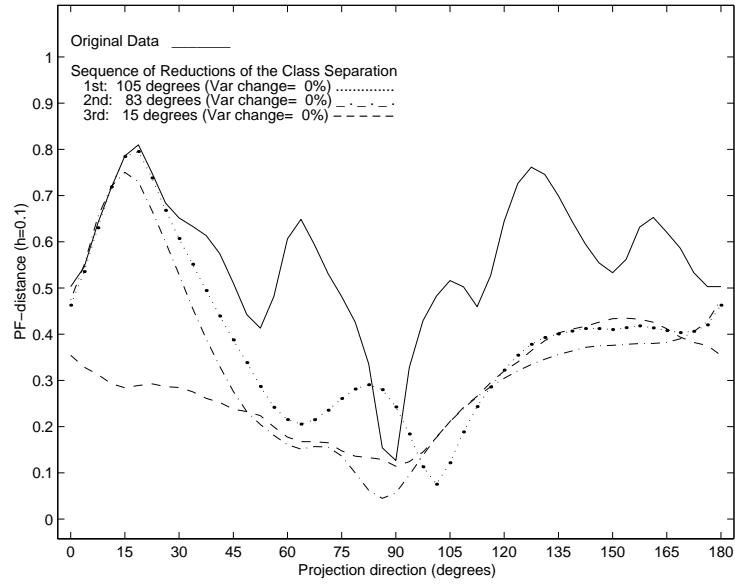


Fig. 3. PF distance for various directions into (x_1, x_2) -plane: original data —; transformed data after successive RCS's at 105° ..., 83° -.-.- and 15° - - -.

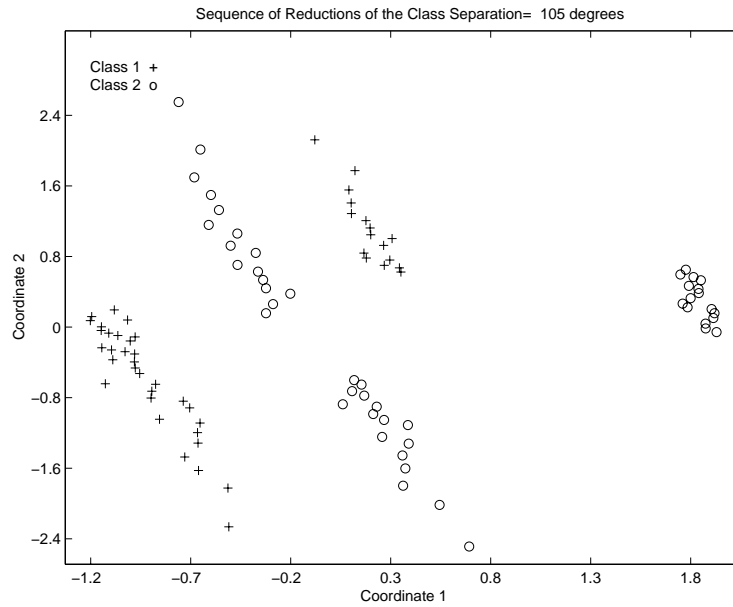


Fig. 4. Transformed data X_t' after the RCS at 105° .

”—” (see Fig.3). We found the two largest local maxima of the PF distance at 19° and 64° , which are located far away from the starting initialization 105° used in Step 1. The latter can not be obtained by conventional optimization.

5 A Comparative Study

Here we compare the discrimination qualities of the discriminant vectors \mathbf{w}^* obtained by our recursive procedure (Section 3) and those obtained by the conventional optimization with a random initialization of the starting directional vectors.

We ran experiments with observations drawn from six- dimensional distributions $p(\mathbf{x}|\omega_i) = p(x_1, x_2|\omega_i)p(x_3, x_4|\omega_i)p(x_5|\omega_i)p(x_6|\omega_i)$ for $i = 1, 2$. Here the densities were constructed with the following mixtures of the normal distributions:

for class ω_1 :

$$\begin{aligned} p(x_1, x_2|\omega_1) &= \frac{1}{3}N([0 \ 1]^T, \mathbf{I}) + \frac{1}{3}N([5 \ 3]^T, \mathbf{I}) + \frac{1}{3}N([0 \ 6]^T, \mathbf{I}) \\ p(x_3, x_4|\omega_1) &= \frac{1}{3}N([-3 \ 0]^T, 0.01\mathbf{I}) + \frac{1}{3}N([0.5 \ 3]^T, 0.01\mathbf{I}) \\ &+ \frac{1}{3}N([-0.5 \ -3]^T, 0.01\mathbf{I}) \end{aligned}$$

for class ω_2 :

$$\begin{aligned} p(x_1, x_2|\omega_2) &= \frac{1}{3}N([0 \ 3]^T, \mathbf{I}) + \frac{1}{3}N([5 \ 6]^T, \mathbf{I}) + \frac{1}{3}N([-5 \ 6]^T, \mathbf{I}) \\ p(x_3, x_4|\omega_2) &= \frac{1}{3}N([-0.5 \ 3]^T, 0.01\mathbf{I}) + \frac{1}{3}N([3 \ 0]^T, 0.01\mathbf{I}) \\ &+ \frac{1}{3}N([0.5 \ -3]^T, 0.01\mathbf{I}) \end{aligned}$$

and $p(x_5|\omega_i) = p(x_6|\omega_i) = N(0, 1)$. The classes were totally overlapped in the (x_5, x_6) -plane, partially overlapped in the (x_1, x_2) -plane and totally separated in the (x_3, x_4) -plane. We chose the data having several local maxima for $PF(\mathbf{w})$ (1). We observed two local maxima of $PF(\mathbf{w})$ into the (x_1, x_2) -plane and several local maxima into the (x_3, x_4) -plane including the global maximum of $PF(\mathbf{w})$. We set $N_{t1} = N_{t2} = 50$.

We carried out 150 runs of our procedure, starting from different initial directional vectors in Step 1. The components of the initial vectors were drawn at random from $N(0, 1)$.

We compared the discrimination quality of \mathbf{w}^* in Step 1 and Step 5. In Step 1 we carried out the conventional optimization with random initialization of the starting directional vector while in Step 5 we employed our recursive optimization.

We evaluated the discrimination qualities of \mathbf{w}^* by the resulting values of the $PF(\mathbf{w}^*)$ computed for the test (extra, validation) observations (500 per class).

In Steps 1, 4 and 5 we maximized $PF(\mathbf{w})$ (1) by a sequential quadratic programming method (routine E04UCF in the NAG Mathematical Library). We set the number of major iterations of the optimization routine E04UCF to 50. This setting was proved to be appropriate by a preliminary test.

We set $\mathbf{m}_{y|\omega_i} - \Delta\mathbf{m}_i = 0$ and $\sigma_{y|\omega_i}^2 \pm \Delta\sigma^2 = 1$ for $i=1,2$ in (4) and (5) for all runs. This setting implies that the class- conditional densities of q_1 (4) and (5) are $N(0,1)$ which results in a modified training data $\tilde{\mathbf{X}}_t'$ (Step 3) with an approximately full overlap of the classes for the previously defined discriminant directions. This certainly eliminates the local maximum of $PF(\mathbf{w})$ at the previous solutions but it causes a large restructuring of $\tilde{\mathbf{X}}_t'$ which is highly unfavorable to our procedure.

We carried out three successive runs of Steps 3-6. In order not to favor our procedure by expanding the number of iterations in the optimization, we reran Step 1 with an extended number of the major iterations of the optimization routine E04UCF. We set it to $50 \times 2 \times N_{(steps3-6)}$, with $N_{(steps3-6)}$ the number of repetitions of Steps 3-6 (in our experiments $N_{(steps3-6)} = 3$). In the comparison we used the largest value of $PF(\mathbf{w}^*)$ obtained in Step 1.

We studied the situations (initial directional vectors) in which the conventional optimization failed with the value of $PF(\mathbf{w}^*)$ smaller then 0.35 (dashed path "- - -" in Fig.5). The solid path "—" in Fig.5 presents the results obtained by our procedure. The dots in the bottom of Fig.5 indicate the sequential number of the iteration which implies the largest value of $PF(\mathbf{w}^*)$ (· - first, · · - second and · · · - third iteration). The dots which are missing indicate a case (random initializations 45, 55, 94) for which the conventional optimization was better then our procedure. Our recursive optimization (solid path) outperforms the conventional optimization in Step 1 (dashed path) for the most of the initializations.

We summarize the overall shape of the PF distance over the 100 replications by the boxplots shown in Fig.6. The boxplot in the left presents the values of $PF(\mathbf{w}^*)$ for the optimal solutions \mathbf{w}^* obtained by the conventional optimization, the central boxplot illustrates the $PF(\mathbf{w}^*)$ for \mathbf{w}^* obtained by our recursive optimization procedure, and the boxplot in the right presents the paired difference of the values of $PF(\mathbf{w}^*)$ (the difference of the solid and dashed paths of Fig.5). In Fig.6 the boxes show the values of the PF distances between quartiles; the lines represent the medians of the PF distances. Whiskers go out to the extremes of the PF distances. We observe that the values of $PF(\mathbf{w}^*)$ of our procedure tend to be larger than the values of $PF(\mathbf{w}^*)$ of the conventional optimization.

Finally we calculated the averaged difference of the values of $PF(\mathbf{w}^*)$ obtained by our recursive optimization and by the conventional optimization, which was 0.24. We evaluated the significance of this difference by the paired t-test and obtained the 99 percent confidence interval 0.24 ± 0.08 which confirms a significant increase of $PF(\mathbf{w}^*)$ for \mathbf{w}^* obtained by our procedure.

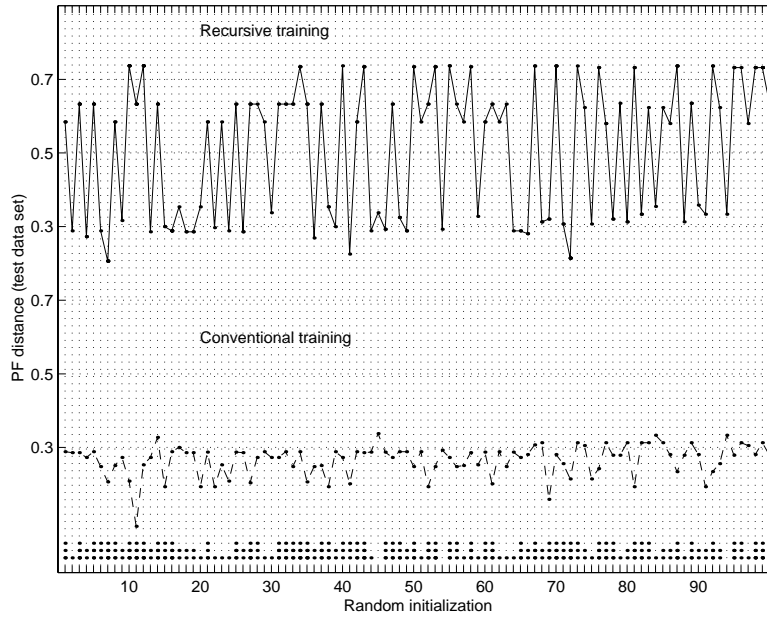


Fig. 5. Random initializations in which conventional optimization failed with the value of $PF(\mathbf{w}^*)$ smaller than 0.35.

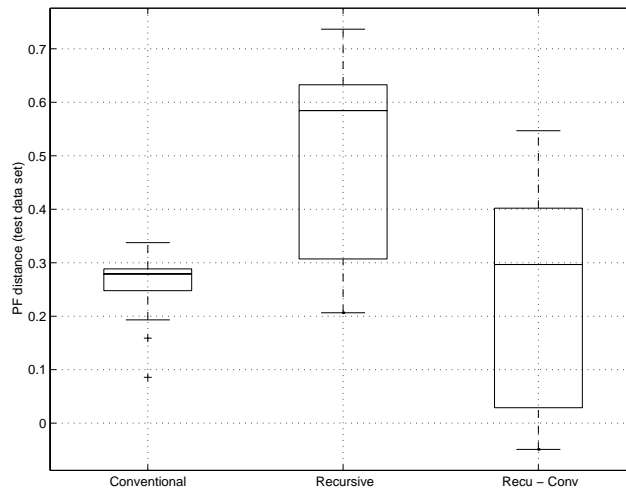


Fig. 6. Boxplots of the values of the $PF(\mathbf{w}^*)$ for \mathbf{w}^* obtained by the conventional optimization and our recursive optimization.

6 Summary and Conclusion

We have discussed a method for the nonparametric linear discriminant analysis proposed by us in [3] previously. It searches for the discriminant direction which maximizes the Patrick-Fisher (PF) distance between the projected class-conditional densities. Since the PF distance is a highly nonlinear function, a sequential search for the directions corresponding to several large local maxima of the PF distance has been used.

In this paper we refine our method [3]. We ensure optimization in a batch mode instead of optimization in an interactive mode proposed in [3]. By means of a simulation (Section 4) we have demonstrated that our procedure succeeds in finding large local maxima of the PF distance which are located far away from the starting point, and can not be found by conventional optimization. The comparative study considered in Section 5 shows that our procedure is more successful than the conventional optimization with random initialization.

References

1. Aladjem, M.E.: Multiclass discriminant mappings. *Signal Processing*. **35** (1994) 1–18
2. Aladjem, M.E.: Linear discriminant analysis for two-classes via removal of classification structure. *IEEE Trans. Pattern Anal. Mach. Intell.* **19** (1997) 187–192
3. Aladjem, M.E.: Nonparametric discriminant analysis via recursive optimization of Patrick-Fisher distance. *IEEE Trans. on Syst., Man, Cybern.* **28B** (1998) 292–299
4. Aladjem, M.E.: Linear discriminant analysis for two classes via recursive neural network reduction of the class separation. In A.Amin, D.Dori, P.Pudil and H.Freeman (eds.), *Lecture Notes in Computer Science 1451: Advances in Pattern Recognition*. (1998) 775-784
5. Devijver, P.A., Kittler, J.: *Pattern Recognition: A Statistical Approach*. Prentice-Hall International Inc., London. (1982)
6. Friedman, J.H.: Exploratory projection pursuit. *Journal of the American Statistical Association*. **82** (1987) 249–266
7. Huber, P.J.: Projected pursuit, including discussions. *The Annals of Statistics*. **13** (1985) 435-525