

# RECURSIVE OPTIMIZATION OF AN EXTENDED FISHER DISCRIMINANT CRITERION

Mayer E. Aladjem

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

## ABSTRACT

*A method for recursive optimization of an extended Fisher (ExF) discriminant criterion is proposed. The method consists of obtaining a discriminant direction which optimizes the ExF criterion, transforming the data along it into data with greater class overlap, and iteration to obtain the next discriminant direction. An application to a medical dataset indicates the potential of the proposed method for finding a sequence of oblique directions with significant class separation.*

## 1. INTRODUCTION

We discuss discriminant analysis which is carried out by the linear mapping  $\tau = \mathbf{r}\beta^T \mathbf{x}$ ,  $\mathbf{x} \in \mathbb{R}^n$ ,  $\tau \in \mathbb{R}^1$ ,  $n \geq 2$ , with  $\mathbf{x}$  an arbitrary  $n$ -dimensional observation, and  $\mathbf{r}\beta$  a direction vector (having unit length  $\mathbf{r}\beta^T \mathbf{r}\beta = 1$ ). The vector  $\mathbf{r}\beta$  optimizes an *extended Fisher (ExF) discriminant criterion* previously proposed by us [1,2]. The optimal vector  $\mathbf{r}\beta$  is called a *discriminant vector*. In this paper our goal is to obtain a sequence of discriminant vectors by successive optimization of the ExF criterion. In the past, in order to include different information in the discriminant vectors, an orthogonal constraint on the latter vectors was used. In this paper we propose a method which is free to search for discriminant directions oblique to each other. It is a modification of our method for removal of classification structures [3]. In Section 2 we describe a normalization of the data, which is required by our method. Section 3 presents the ExF criterion and the computation of the discriminant vector related to it. The new method for recursive optimization of the ExF criterion is presented in Section 4. Section 5 includes the results and analyses of an application to a medical dataset.

## 2. SPHERED DATA

Suppose we are given a set of  $N_d$  design (training) observations  $(\mathbf{z}_1, l_1), (\mathbf{z}_2, l_2), \dots, (\mathbf{z}_{N_d}, l_{N_d})$  in  $n$ -dimensional sample space  $\mathbf{z}_j \in \mathbb{R}^n$ , ( $n \geq 2$ ),  $j = 1, 2, \dots, N_d$ . 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 a decomposition of the design 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 the subset  $Z_{di}$  contains  $N_{di}$  observations properly associated with the class labeled by  $\omega_i$  ( $l_j = \omega_i \Leftrightarrow \mathbf{z}_j \in Z_i$ ) for  $i = 1, 2$ . To achieve *data sphering* [4], 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  $\mathbf{D}$  a diagonal. We then define the normalization matrix  $\mathbf{A} = \mathbf{D}^{-1/2} \mathbf{R}^T$ . In the remainder of the paper, all operations are performed on the *sphered design data*  $X_{di} = \{\mathbf{x}: \mathbf{x} = \mathbf{A}(\mathbf{z} - \mathbf{m}_Z), \mathbf{z} \in Z_{di}\}$  and *sphered new (arbitrary or test*

observations)  $\mathbf{x}=\mathbf{A}(\mathbf{z}-\mathbf{m}_z)$ ,  $\mathbf{z}\in Z_d$  with  $\mathbf{m}_z$  the mean vector of the design sample  $Z_d$ . The pooled sample covariance matrix estimated over  $X_{d_i}$  becomes the identity matrix  $\mathbf{A}\mathbf{S}_z\mathbf{A}^T=\mathbf{I}$ . This implies that for any direction vector  $\mathbf{r}$  ( $\mathbf{r}^T\mathbf{r}=1$ ) the projections  $\tau=\mathbf{r}^T\mathbf{x}$  of the sphered design observations  $\mathbf{x}\in\{X_{d_1}\cup X_{d_2}\}$  have unit pooled sample variance.

### 3. EXTENDED FISHER DISCRIMINANT CRITERION

The *extended Fisher (ExF) criterion* [1],[2] is a generalization of Malina's discriminant criterion [5], i.e.

$$G(\mathbf{r},\beta)=[(1-\beta)\mathbf{r}^T\mathbf{B}\mathbf{r}+\beta|\mathbf{r}^T\mathbf{S}^{(-)}\mathbf{r}|][\mathbf{r}^T\mathbf{S}_W\mathbf{r}]^{-1}, \quad (1)$$

with direction vector  $\mathbf{r}$ , control parameter  $\beta$  ( $0\leq\beta\leq 1$ );  $\mathbf{B}=(\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^T$  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  $i=1,2$ ;  $\mathbf{S}_W$  the pooled within-class sample covariance matrix. All of them are computed for the sphered design data sets  $X_{d_i}$ ,  $i=1,2$ . In (1) the symbol  $\mathbf{S}^{(-)}$  implies two forms of the criterion  $G(\mathbf{r},\beta)$ . The ExF discriminant vector maximizes  $G(\mathbf{r},\beta)$ . It 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  $\mathbf{S}_W^{-1}[(1-\beta)\mathbf{B}+\beta(\mathbf{S}_2-\mathbf{S}_1)]$ . 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 the values of  $\beta$  that maximize the Patrick-Fisher (PF) distance [6] between the classes along the ExF discriminant vector  $\mathbf{r}\beta$ . This gives rise to a problem of parameter optimization. Our strategy for solving it is to choose a grid of values in the interval ( $0\leq\beta\leq 1$ ), to calculate the PF distance at each value and then to choose the value with the largest PF distance as the  $\beta$ -value. From experience, a suitable size of optimization grid is 21 values (uniform grid with step 0.05). We compute the PF distance using the Parzen estimators with Gaussian kernels of the class-conditional densities of the projections  $\zeta=\mathbf{r}\beta^T\mathbf{x}$  ([6] pp.277-280). From experience, a suitable value of the smoothing parameter  $h$  (standard deviation of the Gaussian kernel) is  $h=0.1$ .

### 4. RECURSIVE OPTIMIZATION OF THE ExF CRITERION

The proposed recursive method consists of obtaining an ExF discriminant vector, transforming the data along it into data with greater class overlapping and iterating to obtain the next ExF discriminant vector. In our work, we are stimulated by an idea of Friedman [4], called "*structure removal*". We describe the method in its abstract version operating on probability distributions. The application to the data samples is obtained by substituting an estimate of the distributions over the design sets  $X_{d_1}$  and  $X_{d_2}$  ([4], p.254).

#### 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  $\zeta=\mathbf{a}^T\mathbf{x}$  of any realization  $\mathbf{x}$  we cannot gain any information about the class to which  $\mathbf{x}$  belongs. In other words, the random variable  $\mathbf{a}^T\mathbf{X}$  and the class ( $\omega_i$ ,  $i=1,2$ ) are probabilistically independent. In this case  $p_{\mathbf{a}}(\zeta|\omega_i)=p_{\mathbf{a}}(\zeta)$  (see [6], pp.198-199) with

$p_{\mathbf{a}}(\zeta|\omega_i)$  the class-conditional density of the projection  $\mathbf{a}^T\mathbf{X}$  of a random vector  $\mathbf{X}$  and  $p_{\mathbf{a}}(\zeta)$  unconditional (mixture) density of  $\mathbf{a}^T\mathbf{X}$ . It is known [4] that for most high-dimensional data, most low-dimensional projections are approximately normal. Therefore it seems reasonable to approximate  $p_{\mathbf{a}}(\zeta)$  by normal density function. Note also that in order to preserve the properties of the sphered data the random variable  $\mathbf{a}^T\mathbf{X}$  must have zero mean and unity variance. Taking into account these observations, we conclude that class-conditional densities  $p_{\mathbf{a}}(\zeta|\omega_i)$  along the zero informative direction vector  $\mathbf{a}$  can be approximated by the standard normal density  $N(0,1)$ .

#### 4.2. Reduction of the class separation along the ExF discriminant vector

The idea is to transform the class-conditional densities along the ExF vector to normal densities in order to reduce the class separation. Assume that  $\mathbf{U}$  is an orthonormal  $n \times n$  matrix with an ExF discriminant vector  $\mathbf{r}_\beta$  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 is  $\tau_1=\mathbf{r}_\beta^T\mathbf{x}$ . We denote other coordinates as  $\tau_2,\tau_3,\dots,\tau_n$  ( $\mathbf{t}=[\tau_1,\tau_2,\dots,\tau_n]^T$ ). Let  $p_{\mathbf{r}_\beta}(\tau_1|\omega_i)$ ,  $i=1,2$  be the class-conditional densities along  $\mathbf{r}_\beta$  and  $m_{\mathbf{r}_\beta|\omega_i}$ ,  $\sigma_{\mathbf{r}_\beta|\omega_i}^2$  their means and variances. We require a transformation that takes the class-conditional densities along  $\mathbf{r}_\beta$  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  $q_1(\mathbf{r}_\beta^T\mathbf{X})$  having normal class-conditional densities and  $\tau_i=q_i(\tau_i)$ ,  $i=2,3,\dots,n$  each given by identity transformation. The function  $q_1$  is obtained by the percentile transformation method:

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

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

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

$$q_1(\tau_1)=[\Phi^1(F_{\mathbf{r}_\beta}(\tau_1|\omega_2))](\sigma_{\mathbf{r}_\beta|\omega_2}^{2\pm\Delta\sigma^2})^{1/2} + (m_{\mathbf{r}_\beta|\omega_2}-\Delta m_2), \quad (3)$$

with  $\Delta\sigma^2$ ,  $\Delta m_1$ ,  $\Delta m_2$  user-supplied parameters,  $F_{\mathbf{r}_\beta}(\tau_1|\omega_i)$  the class-conditional (cumulative) distribution function along  $\mathbf{r}_\beta$  for  $i=1,2$  and  $\Phi^{-1}$  the inverse of the standard normal distribution function. Finally,

$$\mathbf{x}' = \mathbf{U}^T\mathbf{q}(\mathbf{U}\mathbf{x}) \quad (4)$$

takes  $p_{\mathbf{r}_\beta}(\tau_1'|\omega_i)=N(m_{\mathbf{r}_\beta|\omega_i}-\Delta m_i, \sigma_{\mathbf{r}_\beta|\omega_i}^{2\pm\Delta\sigma^2})$ , for  $i=1,2$  leaving all orthogonal directions of  $\mathbf{r}_\beta$  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  $\Delta\sigma^2=0$  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 ([3],p.254). If  $\sigma_{\mathbf{r}\beta|\omega_i}^2 \pm \Delta\sigma^2 = 1$  and  $m_{\mathbf{r}\beta|\omega_i} - \Delta m_i = 0$ ,  $i=1,2$  we remove totally the classification structure along  $\mathbf{r}\beta$ , making  $\mathbf{r}\beta$  a zero informative direction vector with full overlap of the classes. This causes the largest changes of the class-conditional distributions of  $\mathbf{x}'$ . We propose to make trials with progressive increasing of  $\Delta\sigma^2$ -values in the interval ( $0 \leq \Delta\sigma^2 \leq 1$ ). In order to preserve the sphering of the data we compute  $\Delta m_1$  and  $\Delta m_2$  using the sphering conditions (zero unconditional mean and unconditional variance equal to one).

#### 4.3. Recursive optimization procedure

The computation algorithm of the sequence of ExF 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 original sphered design samples.

#### Reductions of class separation:

**Step 1:** Using the sample set  $\{X_1 \cup X_2\}$ , compute the ExF vector for the  $\beta$ -value which implies the largest PF distance of the original samples  $X_{d1}$  and  $X_{d2}$ . Save the obtained ExF vector.

**Step 2:** Using the current  $\Delta\sigma^2$ -value, reduce the class separation along the ExF vector and obtain a new data set  $\{X_1' \cup X_2'\}$  (see Expr.(4)). Assign the new set to be the current sample set, i.e.  $X_1=X_1'$ ,  $X_2=X_2'$ . If only one ExF vector has been computed, repeat steps 1 and 2. Otherwise continue with step 3.

**Step 3** (Update the  $\Delta\sigma^2$ -value): Project the original samples  $X_{d1}$  and  $X_{d2}$  on to the last two ExF vectors and compare the PF distances along these vectors. If the PF distances are approximately equal, increase  $\Delta\sigma^2$ -value ( $0 < \Delta\sigma^2 \leq 1$ ). Otherwise assign  $\Delta\sigma^2=0$ .

**Repeat Steps 1-3.** We stop the iterations if several ExF vectors, with different class separations along them, are obtained.

## 5. AN APPLICATION

A real data set concerning the medical diagnosis of the neurological disease cerebrovascular accident (CVA) contains pathologo-anatomically verified CVA cases: 200 cases with haemorrhages 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 we decided to use eight largest eigenvalues of  $\mathbf{S}_Z$  in its eigenvalue-eigenvector decomposition (see Section 2). Fig.1 presents the projection of the sphered data set on to the plot spanned by the eigenvectors corresponding to the two largest eigenvalues (the principal component projection).

First we computed ExF discriminant vector for  $\beta=0$ , which implied the PF distance of maximal value 0.5046. Then we iterated by a sequence of 9 reductions of class separation following the recursive optimization procedure (Section 4.3.). We used zero  $\Delta\sigma^2$ -value in the 1st to the 6th iterations and obtained the PF distances 0.5648, 0.6302, 0.6974, 0.7501, 0.7577, 0.7592. We continued with stronger reduction of the class separation ( $\Delta\sigma^2=0.20$ , 0.65) at the 7th and the 8th iterations, and finally returned to  $\Delta\sigma^2=0.0$  at the 9th iteration observing PF distances 0.6854, 0.2461 and 0.3657 at these trials. Fig.2 presents the best result (the maximal PF distance) obtained at the 6th trial. Fig.3 presents the result of the

last (9th) trial. One can observe a destructuring of the transformed data (Fig.3a) compared with the original data (Fig.1) which decreased the class separation along 8th axis. Fig.3b shows the class-conditional densities along the ExF vector at the 9th iteration.

We analysed the discriminant information gained by the ExF vectors at the 6th and 9th trials. We projected the original data on to the plot spanned by these vectors (Fig.4) and compared the class overlap in the plot with the overlap along the best ExF vector (Fig.2). We found that the two-dimensional presentation gains larger class separation and concluded that the ExF vector at the 9th iteration adds "new" discriminant information to the best one-dimensional solution. We viewed the results of the 6th and 9th trials as "interesting" ones.

## 6. SUMMARY AND CONCLUSIONS

We have presented a method for the linear discrimination of two classes based on the extended Fisher (ExF) criterion. Just like any other projection pursuit procedure [4], our method searches for the sequence of the directions with "interesting" discriminant information. The main feature of the proposed method is its freedom to search for discriminant directions which are oblique to each other.

Our method implements Friedman's [4] procedure for recursive optimization, called structure removal. Just like Friedman's procedure we transform the densities along the found discriminant directions into normal densities. The main difference of our proposal from Friedman's [4] procedure for structure removal consists in the choice of the density which is transformed. Friedman's algorithm transforms the mixture (unconditional) density to normal density while our algorithm processes the class-conditional densities separately. The latter arises from the goal of our method. It is a tool for discriminant analysis (analysis of samples previously grouped into classes) while Friedman's procedure is oriented to cluster analysis of unclassified samples.

*Acknowledgments:* This work was supported 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 nonparametric mapping of multi-dimensional 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. of the 12th International Conference on Pattern Recognition*, vol. 2, pp.67-71, 1994.
- [4] J.H.Friedman, "Exploratory projection pursuit", *Journal of the American Statistical Association*, vol. 82, pp. 249-266, 1987.
- [5] W. Malina, "On an extended Fisher criterion for feature selection", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 3, pp. 611-614, 1981.
- [6] P.A.Devijver and J.Kittler, *Pattern Recognition: A Statistical Approach*, Prentice-Hall International, Inc., London, 1982.

Fig.1. Sphered CVA-data along the eigenvectors corresponding to the two largest eigenvalues of  $S_Z$  ("+" haemorrhages and "o" infarctions).

Fig.2. Class-conditional densities along the  $ExF$  vector at the 6th iteration.

(a)

(b)

Fig.3. Result at the 9th iteration: a) transformed data along the eigenvectors corresponding to the two largest eigenvalues of  $\mathbf{S}_Z$ , b) class-conditional densities along the ExF vector.

Fig.4. Projection of the CVA samples on to the plot spanned by the "interesting" discriminant directions.