

Two-Class Pattern Discrimination via Recursive Optimization of Patrick-Fisher Distance

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 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 PF distance and iteration to obtain the next direction. A simulation study indicates the potential of the method to find the sequence of directions with significant class separations.

1. Introduction

We discuss discriminant analysis which is carried out by the linear mapping $\tau = \mathbf{r}^T \mathbf{x}$, $\mathbf{x} \in \mathbb{R}^n$, $\tau \in \mathbb{R}^1$, $n \geq 2$, with \mathbf{x} an arbitrary sample in n -dimensional measurement space, and \mathbf{r} a direction vector (unit vector - $\mathbf{r}^T \mathbf{r} = 1$). The latter optimizes a discriminant criterion in \mathbb{R}^1 . The optimal vector \mathbf{r} is called the *discriminant vector*.

In the paper we discuss a discriminant criterion for two classes, namely the Patrick-Fisher (PF) criterion [1]. It measures the overlap of class-conditional densities of the sample projections τ . Unfortunately, it is not a unimodal function with respect to \mathbf{r} and has more than one maximum. In most applications the PF discriminant vector is searched for along the gradient of the criterion, 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 analyses, may be used for choosing a starting point for the optimization procedure. In this

This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University Beer-Sheva, Israel.

work we use a technique which combines them. It is based on an extended Fisher (ExF) criterion previously proposed by us [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-criterion 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 criterion.

Section 2 presents a normalization of the data (*sphering*) which is required by the recursive method. In Section 3 and 4 we describe the PF and ExF criteria and the computation of the discriminant vectors related to them. The new method for recursive optimization of PF criteria is presented in Section 5. Section 6 includes the results and analyses of a simulation study.

2. Sphered data

As mentioned previously we discuss the two-class problem. Suppose we are given a set of N_d *design (training) samples* $(\mathbf{z}_1, l_1), (\mathbf{z}_2, l_2), \dots, (\mathbf{z}_{N_d}, l_{N_d})$ in n -dimensional measurement space $\mathbf{z}_j \in \mathbb{R}^n, (n \geq 2), j=1, 2, \dots, N_d$. The label $l_j \in \{\omega_1, \omega_2\}$ shows that the sample \mathbf{z}_j belongs to one of the classes ω_1 or ω_2 . Using these labels 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 can be done. Let the decomposition be $Z_d = Z_{d1} \cup Z_{d2}$, where each subset Z_{di} $i=1, 2$ includes samples properly associated with the class labeled by ω_i ($l_j = \omega_i \Leftrightarrow \mathbf{z}_j \in Z_{di}$). We name N_{di} the number of design samples per class.

For the aim of data sphering (see Friedman [2], p.251), we perform eigenvalue-eigenvector decomposition of the sample covariance matrix of the class-mixture $\Sigma_Z = \mathbf{R} \mathbf{D} \mathbf{R}^T$, with \mathbf{R} an orthonormal and \mathbf{D} a diagonal ($n \times n$) matrix. We then define the normalization matrix $\mathbf{A} = \mathbf{D}^{-1/2} \mathbf{R}^T$.

In the following text all operations are performed on the *sphered design samples* with zero mean

$$X_{di} = \{\mathbf{x}: \mathbf{x} = \mathbf{A}(\mathbf{z} - \mathbf{m}_z), \mathbf{z} \in Z_{di}\}, i = 1, 2 \quad (1)$$

and *sphered new (arbitrary or test) samples* $\mathbf{x} = \mathbf{A}(\mathbf{z} - \mathbf{m}_z)$, $\mathbf{z} \in Z_d$. Here \mathbf{m}_z is the mean of the class-mixture estimated over the design samples Z_d .

The sample covariance matrix $\mathbf{\Sigma}$ of the class-mixture becomes the identity $n \times n$ matrix for the sphered design samples, i.e. $\mathbf{\Sigma} = \mathbf{A}\mathbf{\Sigma}_z\mathbf{A}^T = \mathbf{I}$. 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 samples \mathbf{x} ($\mathbf{x} \in \{X_{d1} \cup X_{d2}\}$) have unit sample variance

$$s_\tau = \mathbf{r}^T\mathbf{\Sigma}\mathbf{r} = 1. \quad (2)$$

3. Patrick-Fisher criterion

We use the discriminant criterion which is an estimator of the PF distance defined as follows (see Deijver and Kittler [7], pp.277-280):

$$D_{PF}(\mathbf{r}, h) = \left\{ \int_{-\infty}^{+\infty} [\hat{P}(\omega_1)\hat{p}_r(\zeta|\omega_1) - \hat{P}(\omega_2)\hat{p}_r(\zeta|\omega_2)]^2 d\zeta \right\}^{1/2} \quad (3)$$

with

$$\hat{p}_r(\zeta|\omega_i) = \frac{1}{h\sqrt{2\pi}N_{di}} \sum_{\mathbf{x}_{di} \in X_{di}} \exp\left\{-\frac{1}{2h^2}(\zeta - \mathbf{r}^T\mathbf{x}_{di})^2\right\}, i = 1, 2 \quad (4)$$

the Parzen estimators with Gaussian kernels of the class-conditional densities of the projections $\zeta = \mathbf{r}^T\mathbf{x}$, and $\hat{P}(\omega_i) = N_{di}/N_d$ estimators of a priori probabilities. Here \mathbf{x} is an arbitrary sample ($\mathbf{x} \in \mathbb{R}^n$), $\mathbf{x}_{di} \in X_{di}$ are ω_i -design samples, and h is a smoothing parameter.

The theoretical motivation of the PF distance is its resultant upper bound on Bayes error along direction \mathbf{r} . It is known that the PF distance induces an upper bound which is larger than those of other probabilistic class-separability measures. Nevertheless, $G_{PF}(\mathbf{r}, h)$ is more practical, because of its analytical simplification, which overcomes the numerical integration in (3).

The PF discriminant vector maximizes $G_{PF}(\mathbf{r}, h)$ for fixed h . We carry out the optimization by a sequential quadratic programming method available as a routine E04UCF in the NAG Mathematical Library. In order to search among unit direction vectors we apply maximization of $G_{PF}(\mathbf{r}, h)$ to nonlinear constraint $\mathbf{r}^T\mathbf{r} = 1$.

The primary goal is to find the global maximum of $G_{PF}(\mathbf{r}, h)$. By a naive use of the optimization algorithm, the computed value for the observed $\max\{G_{PF}(\mathbf{r}, h)\}$ 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.

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 4). In order to search for several large local maxima we propose a method for recursive optimization of $G_{PF}(\mathbf{r}, h)$ (Section 5).

4. Extended Fisher criterion

The extended Fisher (**ExF**) criterion [2] is

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

with \mathbf{r} ($\mathbf{r}^T\mathbf{r} = 1$) direction vector, β ($0 \leq \beta \leq 1$) control parameter and

$$\mathbf{B} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \quad (6)$$

$$\mathbf{\Sigma}^{(-)} = \mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \text{ or } \mathbf{\Sigma}_2 - \mathbf{\Sigma}_1 \quad (7)$$

$$\mathbf{S}_W = (N_{d1}/N_d)\mathbf{\Sigma}_1 + (N_{d2}/N_d)\mathbf{\Sigma}_2. \quad (8)$$

Here, \mathbf{B} is the sample between-class scatter matrix, \mathbf{m}_i -class-conditional sample means, $\mathbf{\Sigma}_i$ -class-conditional sample covariance matrices (scatter matrices within classes), and \mathbf{S}_W -averaged within-class sample covariance matrix. All of them are computed for the sphered design data sets X_{di} , $i = 1, 2$ (1). In (5) the symbol $\mathbf{\Sigma}^{(-)}$ implies two forms of the criterion $G_{\text{ExF}}(\mathbf{r}, \beta)$.

The ExF discriminant vector maximizes $G_{\text{ExF}}(\mathbf{r}, \beta)$. The optimization of $G_{\text{ExF}}(\mathbf{r}, \beta)$ with respect to \mathbf{r} for fixed β is carried out by solving two eigenvalue problems. The ExF discriminant vector is the eigenvector which corresponds to the largest eigenvalue of the matrices $\mathbf{S}_W^{-1}[(1-\beta)\mathbf{B} + \beta(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)]$ and $\mathbf{S}_W^{-1}[(1-\beta)\mathbf{B} + \beta(\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1)]$.

An appropriate value of the control parameter β is not known in advance. We search for it using a trial and error procedure. Our approach to model selection is oriented to a suitable starting point of the local optimizer of the PF criterion. We choose the β -value that maximizes the PF distance (3) 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 at each value and then to choose the value with the largest PF-distance as the β -value. By experience, the size of the optimization grid is taken to be 11 values (grid with equal step 0.1).

5. Recursive optimization of PF criterion

The proposed recursive method consists of obtaining a PF discriminant vector, transforming the data along it into data with greater class-overlapping (smaller PF distance), and iterating to obtain a new PF discriminant vector. We describe the method in its abstract version. That is, we operate on probability distributions. The application to the data samples is obtained by substituting an estimate of the distributions over the design sets X_{d1} and X_{d2} (1).

5.1. Zero informative direction vector

In discriminant analysis a direction vector \mathbf{a} ($\mathbf{a}^T\mathbf{a}=1$) is *zero informative* if by observing the projection $\zeta=\mathbf{a}^T\mathbf{x}$ of any sample $\mathbf{x}\in\mathbb{R}^n$ we cannot gain any information about the class to which \mathbf{x} belongs (see Devijver and Kittler [1], pp.198-199). In other words, the random variable ζ and the classes ω_i , $i=1,2$ are probabilistically independent. In this case $p_{\mathbf{a}}(\zeta|\omega_i)=p_{\mathbf{a}}(\zeta)$, with $p_{\mathbf{a}}(\zeta|\omega_i)$ the class conditional density functions of the projections $\zeta=\mathbf{a}^T\mathbf{x}$, $p_{\mathbf{a}}(\zeta)=P(\omega_1)p_{\mathbf{a}}(\zeta|\omega_1)+P(\omega_2)p_{\mathbf{a}}(\zeta|\omega_2)$ the mixture density function, and $P(\omega_i)$, $i=1,2$ the a priori probabilities of the classes ω_i , $i=1,2$.

It is known [3] that for most high-dimensional data, most low-dimensional projections are approximately normal. Therefore it seems reasonable to assume that $p_{\mathbf{a}}(\zeta)$ is approximated by a normal density function. Note also that in order to preserve the properties (1) and (2) of the sphered data, ζ must have zero mean and unity variance. Taking into account these observations, we conclude that the class conditional density functions $p_{\mathbf{a}}(\zeta|\omega_i)$, $i=1,2$ along the zero informative direction vector \mathbf{a} can be approximated by the standard normal density $N(0,1)$.

5.2. Reduction the class separation along the PF vector

Let \mathbf{r} be a vector which defines a direction with a maximum of 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 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 the direction defined by \mathbf{r} we require a transformation that takes τ_1 to normal class conditional 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 carry out this transformation: $\tau_1'=q_1(\tau_1)$ and $\tau_i=q_i(\tau_i)$, $i=2,3, \dots, n$ identity transformations. The function q_1 which results in the normal class conditional densities of τ_1' is obtained by the percentile transformation method:

- for observations \mathbf{x} from class ω_1

$$q_1(\tau_1)=[\Phi^{-1}(F_{\mathbf{r}}(\tau_1|\omega_1))](\sigma_{\mathbf{r}|\omega_1}^2\pm\Delta\sigma^2)^{1/2}+(m_{\mathbf{r}|\omega_1}-\Delta m_1) \quad (9)$$

- for observations \mathbf{x} from class ω_2

$$q_1(\tau_1)=[\Phi^{-1}(F_{\mathbf{r}}(\tau_1|\omega_2))](\sigma_{\mathbf{r}|\omega_2}^2\pm\Delta\sigma^2)^{1/2}+(m_{\mathbf{r}|\omega_2}-\Delta m_2) \quad (10)$$

with $\Delta\sigma^2$, Δm_1 , Δm_2 user-supplied parameters, $F_{\mathbf{r}}(\tau_1|\omega_i)$, $i=1,2$ the class-conditional (cumulative) distribution functions of τ_1 and Φ^{-1} the inverse of the standard normal

distribution function Φ . Finally, the transformation $\mathbf{x}'=\mathbf{U}^T\mathbf{q}(\mathbf{U}\mathbf{x})$ takes the class-conditional densities $p_{\mathbf{r}}(\tau_1|\omega_i)=N(\sigma_{\mathbf{r}|\omega_i}^2\pm\Delta\sigma^2, m_{\mathbf{r}|\omega_i}-\Delta m_i)$, $i=1,2$ to be normal densities along vector \mathbf{r} leaving all orthogonal directions unchanged.

The data sample version of (9) and (10) is implemented by substituting empirical distributions for the class-conditional distributions $F_{\mathbf{r}}(\tau_1|\omega_i)$, $i=1,2$ (see Friedman [3], p.254).

Now we are confronted with the problem of defining the values of the user-supplied parameters $\Delta\sigma^2$, Δm_1 and Δm_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 (see Friedman [3], p.254). 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 remove totally the classification structure along \mathbf{r} , making \mathbf{r} a zero informative direction vector with minimal PF distance along it (see Section 5.1.). This certainly eliminates the local maximum of the PF distance, but it causes the largest changes of the class-conditional distributions of \mathbf{x}' .

In our algorithm (Section 5.3.) we make trials with $\Delta\sigma^2$ -values in the interval ($0\leq\Delta\sigma^2\leq 1$). We chose the value of $\Delta\sigma^2$ subjectively, compromising on keeping the data structure of \mathbf{x}' as close as possible to the original one and on reducing the class separation (deflating the PF distance) along \mathbf{r} . We choose the sign (+ or -) of the change ($\pm\Delta\sigma^2$) in order to approach $\sigma_{\mathbf{r}|\omega_i}^2\pm\Delta\sigma^2$ to 1. We compute Δm_i using the sphering conditions- zero unconditional mean and unconditional variance equal to one along \mathbf{r} .

5.3. Recursive optimization procedure

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

Initialization: $X_1=X_{d1}$, $X_2=X_{d2}$ where X_{d1} , X_{d2} are the sphered design samples (1).

Step 1: Sequence of reductions of class separation

1.1. Using the sample set $\{X_1\cup X_2\}$, compute the ExF vector with the largest PF distance (see Section 4).

1.2. Starting from the ExF vector, search to a convergence point by using a local maximizer (NAG routine E04UCF) of PF criterion. The direction vector after convergence of the maximizer is a current PF vector. Save it.

1.3. Reduce the class separation along the PF vector and obtain a new set $\{X_1'\cup X_2'\}$ (Section 5.2.). Assign the new set to be the current sample set, i.e. $X_1=X_1'$, $X_2=X_2'$.

1.4. 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 into the original sphered data X_{d1} , X_{d2} . The direction vectors after convergence of the algorithm are the *adjusted PF vectors*. Save them. The vectors with largest PF distances among all the adjusted PF vectors are regarded as "*interesting*" solutions.

In step 1.3., we carry out trials with various $\Delta\sigma^2$ -values (see expressions (9),(10)). We examine the class-conditional distributions of the projected samples before and after reduction of class separation along PF vector and we choose suitable value of $\Delta\sigma^2$ subjectively.

6. Simulation study

We carried out an experiment with four-dimensional samples $\mathbf{x}=[x_1 \ x_2 \ x_3 \ x_4]^T$ drawn from class-conditional distributions $p(\mathbf{x}|\omega_i)=p(x_1,x_2|\omega_i)p(x_3,x_4|\omega_i)$, $i=1,2$ constructed from mixtures of normal distributions, i.e.:

for class ω_1 :

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

for class ω_2 :

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

We generated 150 samples per class ($N_{d1}=N_{d2}=150$). They were totally separated along a vector lying in the (x_3,x_4) -plane and directed under an angle of 11° with respect to the x_3 -axis. Fig1. presents the sphered data in the coordinate system spanned on the original x_1 -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.1b.

Following the procedure of Section 5.3, we computed the ExF discriminant vector for $\beta=0.5$, which implied a PF distance of maximal value 0.3143. Fig.2a. shows the class-conditional densities along the latter vector. Starting from it we ran the local optimizer of the PF criterion, which converged to a PF vector shown in Fig.2b. Inspecting the class-conditional densities along the latter vector (Fig.2b.), we concluded that it gains some class separation with respect to the starting ExF vector (Fig.2a).

We iterated by a sequence of three reductions of class separation with $\Delta\sigma^2=0.2$. Fig.3. shows the data after the reductions were performed. Comparing the transformed data (Fig.3) with the original (Fig.1), we observed the following destructuring of the data:

(a) A significant class-overlap into the (x_1,x_2) -plane was gained by the reduction of class separation (Fig 3a). This was a desired result because our goal was to deflate the local maxima 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.

(b) Some moving away of the encircled clusters in the (x_3,x_4) -plane was caused by the reduction of class separation (Fig.3b). 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 (Δm_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.

Nevertheless, we continued with this transformed data. We computed the ExF vector for $\beta=0.7$ which implied the maximal PF distance into the transformed data. Starting from it we ran the local optimizer of the PF criterion, which converged to a PF vector. Fig.4a shows the final solution along the adjusted (reoptimized for the original data) PF vector after the 3th reduction of class separation (see Step 2 of Section 5.3). We found a direction with large PF distances of value 0.6503, but we missed the global maximum of the value 0.7353 (see Fig.4b). 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 with $\Delta\sigma^2=0.0$. After three reductions of class separation we detected the global maximum of the PF distance. The sample distributions along the best discriminant directions are presented in Fig.4b.

7. Conclusion

We have presented a method for the linear discrimination of two classes based on the Patrick-Fisher (PF) distance. It succeeded in finding the sequence of directions with significant discriminant information for a simulation study.

Our method implements Friedman's [3] procedure for recursive optimization, called *structure removal*. The main difference of our procedure for reduction of class separation from Friedman's [2] procedure 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 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.

Our procedure, as is the case with the stochastic smoothing algorithms, may miss some largest local maxima including the global one. This was illustrated in 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, may be used.

Unfortunately they require many evaluations of the objective function which leads to long run times for high dimensional data.

References

- [1] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice-Hall International, Inc., London, 1982.
- [2] M.E. Aladjem, "Multiclass discriminant mappings", (a) *Signal Processing*, vol.35, pp.1-18, 1994 .
- [3] J.H. Friedman, "Exploratory projection pursuit", *Journal of the American Statistical Association*, vol. 82,pp. 249-266, 1987.

(b)

Fig.1. Sphered data :
(a) (x_1, x_2)-plane, (b) (x_3, x_4)-plane.

(a)

(b)

Fig.2. Class-conditional densities without reductions of class-separation:
(a) along ExF vector ($\beta=0.5$),
(b) along PF vector.

(a)

(b)

Fig.3. Transformed data after three reductions of class separation ($\Delta\sigma^2=0.2$):
(a) (x_1, x_2) -plane, (b) (x_3, x_4) -plane.

(a)

(b)

Fig.4. Class-conditional densities after three reductions of class separation:
(a) $\Delta\sigma^2=0.2$, (b) $\Delta\sigma^2=0.0$.