LINEAR DISCRIMINANT ANALYSIS ALGORITHMS

Pedro Miguel Correia Guerreiro

Instituto Superior Técnico

ABSTRACT

We propose new algorithms for computing linear discriminants to perform data dimensionality reduction from $\mathbb{R}^n$ to $\mathbb{R}^p$, with $p < n$. We propose alternatives to the classical Fisher's Distance criterion, namely, we investigate new criterions based on the: Chernoff-Distance, J-Divergence and Kullback-Leibler Divergence. The optimization problems that emerge of using these alternative criteria are non-convex and thus hard to solve. However, despite the non-convexity our algorithms guarantee global optimality for the linear discriminant when $p = 1$. This is possible due to problem reformulations and recent developments in optimization theory [8],[9]. A greedy suboptimal approach is developed for $1 < p < n$.

Index Terms— Linear Discriminants, Data Dimensionality Reduction, Fisher's Distance, Chernoff-Distance, Nonconvex strong duality results, Kullback-Leibler Divergence.

1. INTRODUCTION

Linear Discriminant Analysis (LDA) is a very important tool in a wide variety of problems. It is commonly used in machine learning problems like pattern recognition [1],[2], face recognition [4], feature extraction [3] and data dimensionality reduction.

A problem that is treated in LDA is the binary class assigning problem: given one sample in a high-dimensional space $\mathbb{R}^n$, say $x \,\epsilon\, \mathbb{R}^n$, decide to which class C_0 or C_1 it belongs to. Usually the two classes C_0 and C_1 represent two random sources. The classification process can be made in high dimension, i.e. in $\mathbb{R}^n$, using all information available. However this might be computationally heavy for certain real time applications. So, instead of using all the n entries of the sample x directly, an appropriate linear combination of them is made. With this linear combination, we try to capture some data features (hopefully those where C_0 and C_1 differ most), and then perform the data classification. Making these linear combinations, leads generically to information loss, and consequently increases the probability of erroneous classifications. This problem can be attenuated, by making more than one linear combination, and collect them in a vector y, to perform the classification. The number of linear combinations is denoted by p, where $p < n$. That is

$$y = Qx \tag{1}$$

where $Q \,\epsilon\, \mathbb{R}^{p\times n}$ is called the linear discriminant, $y \,\epsilon\, \mathbb{R}^p$ is the vector that collects the p linear combinations, and $x \,\epsilon\, \mathbb{R}^n$ is the sample to be classified. The classification process is made trough the low-dimensional vector $y \,\epsilon\, \mathbb{R}^p$, which works like a signature of the sample x.

The key issue here is the design of the linear discriminant Q. This design process is generically formulated as an optimization problem, where the objective function measures class separability in the projected space $\mathbb{R}^p$, i.e.

$$\max_{Q \,\epsilon\, \mathbb{R}^{p\times n}} \quad f(Q). \tag{2}$$

The choice of the cost function in (2) plays a critical role. An obvious proposal for such cost function, would be $f(Q) = -P(e)(Q)$, where $P(e)(Q)$ stands for the probability of error of the optimum detector in $\mathbb{R}^p$, for the given setup, the minus sign has to do with the fact, that the optimization problem in (2), has been written as a maximization problem. However, in general there is no closed form expression for $P(e)(Q)$. This motivates the introduction of alternative suboptimum choices, which are nonetheless tractable.

2. PREVIOUS WORK

We now give a precise formulation of the problem to be solved and review previous works in this area.

In what follows, the two classes C_0 and C_1 introduced in section 1 are identified with two random sources, that are here denoted by S_0 for source 0, and by S_1 for source 1. We focus on the Gaussian case.

Given the two independent n-dimensional Gaussian distributed sources

$$\begin{array}{l} S_0 : \, x \sim F_0 = N(\mu_0, \Sigma_0) \\ S_1 : \, x \sim F_1 = N(\mu_1, \Sigma_1) \end{array} \tag{3}$$

we wish to find the linear discriminant Q, for data dimensionality reduction, minimizing erroneous classification of the samples generated by these sources in low dimension.

Being $x \,\epsilon\, \mathbb{R}^n$ a sample generated by one of the n-dimensional sources S_0 or S_1 that are considered to be equally probable, a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^p$ is made with the linear discriminant $Q \,\epsilon\, \mathbb{R}^{p\times n}$, i.e.

$$y = Qx.$$

Due to this linear mapping, we have

$$\begin{array}{l} s_0 : y \sim f_0 = N(Q\mu_0, Q\Sigma_0 Q^T) \\ s_1 : y \sim f_1 = N(Q\mu_1, Q\Sigma_1 Q^T) \end{array} \tag{4}$$

where s_0 and s_1 denote the p-dimensional sources that result from the dimensionality reduction induced by the linear discriminant $Q \,\epsilon\, \mathbb{R}^{p\times n}$.

Whenever a sample x is available, it has to be classified. The classification is made with the maximum likelihood criteria, that is more well known in this context as the Neyman-Pearson detector. The linear map Q is applied to the sample x, forming $y = Qx$, and then the maximum likelihood criterion is applied to the random variable y. If $N(Q\mu_0, Q\Sigma_0 Q^T)(y) > N(Q\mu_1, Q\Sigma_1 Q^T)(y)$, y is considered to have been generated by the p-dimensional source s_0 and x is therefore considered to have been generated by the n-dimensional source S_0 and vice-versa.

In the following, we discuss, several proposals for the cost function $f(Q)$ in (2), and we analyze the strengths and weaknesses of previous works, that utilize such cost functions.

A popular choice for $f(Q)$ is the Fisher's Distance which is now reviewed.

We wish to optimally separate in Fisher's sense, the signatures y from S_0, from the signatures from S_1. Intuitively this is equivalent, to separate as much as possible the probability density functions f_0 and f_1 of the signatures as in (4).

The general optimization problem in (2) under Fisher's Distance Maximization criterion [6] is

$$\begin{array}{ll} \max & \mathsf{tr}\{(Q(\Sigma_0+\Sigma_1)Q^T)^{-1} \\ & (Q(\mu_0-\mu_1)(\mu_0-\mu_1)^T Q^T)\} \\ Q \,\epsilon\, \mathbb{R}^{p\times n} & \end{array} \tag{5}$$

where the objective function is the Fisher's Distance between the p-dimensional distributions f_0 and f_1.

In order to better understand what Fisher's Distance measures, the case where $Q \,\epsilon\, \mathbb{R}^{1\times n}$ is presented. Putting $Q = [q^T]$, where $q \,\epsilon\, \mathbb{R}^n$, (5) boils down to:

$$\begin{array}{ll} \max & \dfrac{q^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T q}{q^T(\Sigma_0+\Sigma_1)q} \\ q \,\epsilon\, \mathbb{R}^n & \end{array} \tag{6}$$

Now, it's easy to understand that, the outer class variance $q^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T q = [q^T\mu_0 - q^T\mu_1]^2$ is being maximized while the total inner class variance $q^T(\Sigma_0+\Sigma_1)q$, is being minimized.

The solution Q for (5), can be obtained by doing the eigenvalue decomposition of $(\Sigma_0+\Sigma_1)^{-\frac{1}{2}}(\mu_0-\mu_1)(\mu_0-\mu_1)^T(\Sigma_0+\Sigma_1)^{-\frac{1}{2}}$, and taking for its p rows, the p eigenvectors associated to the p largest eingenvalues, (see [6]). However, since $(\Sigma_0+\Sigma_1)^{-\frac{1}{2}}(\mu_0-\mu_1)(\mu_0-\mu_1)^T(\Sigma_0+\Sigma_1)^{-\frac{1}{2}}$ has rank 1, it is easy to see that the optimum discriminant for $p > 1$ achieves the same performance, as measured by (5), as the optimum discriminant for $p = 1$. That is, there is no gain in projecting to spaces whose dimension $p > 1$. For $p = 1$, the optimum descriminant is $Q = [q^T]$ where q is a solution of (6), that is

$$q = (\Sigma_0+\Sigma_1)^{-1}(\mu_0-\mu_1). \tag{7}$$

In sum, Fisher's Distance Maximization criterion enjoys a closed form solution and a very intuitive interpretation. However, it only allows dimensionality reduction to $p = 1$.

It was said previously that in general there is no closed form expression for the classification error rate. This leads to the utilization of suboptimal measures for it. The theoretical basis for the cost functions or measures used in [6] and [7] is now presented.

Stein's Lemma. [10] Suppose we have k statically independent samples from the same source, and the classification is made trough the maximum-likelihood detector, then we have

$$\lim_{k\to+\infty} \frac{\log P_F(k)}{k} = -D_{KL}(f_0||f_1) \text{ for fixed } P_M \tag{8}$$

$$\lim_{k\to+\infty} \frac{\log P_e(k)}{k} = -C(f_0,f_1) \tag{9}$$

$$\lim_{k\to+\infty} \frac{\log P_e(k)}{k} \geq -JD(f_0,f_1) \tag{10}$$

where $P_F(k)$ is the probability of false alarm, $P_e(k)$ is classification error probability and P_M is the missing probability, when k samples from the same source are used to make the classification. Note that f_0 and f_1 are the p-dimensional probability density functions of the signatures y.

The exponents $D_{KL}(f_0||f_1)$, $JD(f_0,f_1)$, $C(f_0,f_1)$ in (8), are the Kullback-Leibler Divergence, the J-Divergence and the Chernoff Distance, whose definitions for generic p-dimensional probability density functions f_0, f_1 are

$$D_{KL}(f_0||f_1) = \int_{\mathbb{R}^p} f_0(y)\log\frac{f_0(y)}{f_1(y)}dy \tag{11}$$

$$JD(f_0,f_1) = \frac{D_{KL}(f_0||f_1) + D_{KL}(f_1||f_0)}{2} \tag{12}$$

$$C(f_0,f_1) = \max_{0\leq t\leq 1} -\log\left(\int_{\mathbb{R}^p} f_0(y)^t f_1(y)^{1-t} dy\right) \tag{13}$$

respectively.

Stein's Lemma gives asymptotic expressions for $P_F(k)$ and $P_e(k)$. Here, the motivation to use the error exponents as cost functions is that, hopefully, for small k these asymptotic expressions already represent good approximations of P_F and

P_e. Stein's Lemma fills heuristically the lack of closed form expression for these probabilities.

Looking at Stein's Lemma statement in (8), it can be seen that in order to minimize $P_F(k)$ and $P_e(k)$, $D_{KL}(f_0||f_1)$, $JD(f_0, f_1), C(f_0, f_1)$ must be maximized. The $D_{KL}(f_0||f_1)$, $JD(f_0, f_1)$, $C(f_0, f_1)$ are measures of the dissimilarity of the two probability density functions f_0 and f_1. This is the approach taken in [7] with the utilization of the J-Divergence, and in [6] with the Chernoff Distance. In [7] the J-Divergence is maximized for the case where the means of f_0 and f_1 are equal, and in [6] the Chernoff Distance is maximized, but it is not guaranteed that the linear discriminant Q found by their iterative algorithm is globally optimal.

3. KULLBACK-LEIBLER DIVERGENCE MAXIMIZATION

Inserting the probability density functions $f_0 = N(Q\mu_0, Q\Sigma_0 Q^T)$ and $f_1 = N(Q\mu_1, Q\Sigma_1 Q^T)$ of the signatures y defined in (4),in the Kullback-Leibler definition in (11), yields

$$\begin{aligned} D_{KL}(f_0||f_1)(q) = \; & \tfrac{1}{2}\Big(\tfrac{q^T\Sigma_0 q}{q^T\Sigma_1 q} - \log\tfrac{q^T\Sigma_0 q}{q^T\Sigma_1 q} \\ & + \tfrac{[q^T(\mu_0-\mu_1)]^2}{q^T\Sigma_1 q} - 1\Big) \end{aligned} \tag{14}$$

The goal is to find the global maximizer q of (14), i.e.

$$\begin{aligned} q = \; \underset{q \neq 0}{\arg\max} \; & \tfrac{1}{2}\Big(\tfrac{q^T\Sigma_0 q}{q^T\Sigma_1 q} - \log\tfrac{q^T\Sigma_0 q}{q^T\Sigma_1 q}. \\ & + \tfrac{[q^T(\mu_0-\mu_1)]^2}{q^T\Sigma_1 q} - 1\Big) \end{aligned} \tag{15}$$

It's easy to verify, that (14) doesn't depend on the norm of q. So a restriction that doesn't eliminate any direction for q, is admissible.

In order to simplify the objective function of the optimization problem in (15), and without eliminating any direction for q, the restriction $q^T\Sigma_1 q = 1$ is chosen. Applying the restriction, the optimization problem in (15) becomes

$$\begin{aligned} q = \underset{q^T\Sigma_1 q = 1}{\arg\max} \; & q^T[\Sigma_0 + (\mu_0-\mu_1)(\mu_0-\mu_1)^T]q. \\ & - \log q^T\Sigma_0 q \end{aligned} \tag{16}$$

In what follows, $\Sigma_0 + (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T$ is substituted by R, resulting

$$q = \underset{q^T\Sigma_1 q = 1}{\arg\max} \quad q^T Rq - \log q^T\Sigma_0 q. \tag{17}$$

The optimization problem in (17) is non-convex. In order to deal with the non-convexity, a reformulation of the problem is made by introducing the variables x and y

$$x = q^T Rq \tag{18}$$

$$y = q^T\Sigma_0 q \tag{19}$$

resulting for (17) in

$$\max_{(x,y)\,\epsilon\, C = \{(q^T Rq,\, q^T\Sigma_0 q)\,:\, q^T\Sigma_1 q = 1\}} \quad x - \log y. \tag{20}$$

Reformulating the optimization problem this way, the optimization is just made in two variables $(x, y)\,\epsilon\, C$. However, the complexity of the original problem is hidden in the definition of the set C. The strategy to solve (17) consists in finding the solution (x^*, y^*) for (20), and then computing a corresponding q, i.e, a q that solves the following system of quadratic equations:

$$q : \quad \begin{aligned} q^T Rq &= x^* \\ q^T\Sigma_0 q &= y^* \\ q^T\Sigma_1 q &= 1 \end{aligned} \; . \tag{21}$$

The set C is compact and connected, because it's the image of an ellipsoid by a continuous map. This implies that if $(x, y)\,\epsilon\, C$, then x and y belong to closed and bounded intervals on $\mathbb{R}$.

The variable x lies in the interval $[x_{\min}, x_{\max}]$ where

$$x_{\min} = \min_{q^T\Sigma_1 q = 1} \quad q^T Rq \tag{22}$$

$$x_{\max} = \max_{q^T\Sigma_1 q = 1} \quad q^T Rq \tag{23}$$

The solutions to (22) and (23) are $x_{\min} = \lambda_{\min}(\Sigma_1^{-\frac{1}{2}} R \Sigma_1^{-\frac{1}{2}})$ and $x_{\max} = \lambda_{\max}(\Sigma_1^{-\frac{1}{2}} R \Sigma_1^{-\frac{1}{2}})$ respectively.

Knowing this, the strategy to solve (20) consists in discretizing the above interval fixing a value for x, and optimizing over the y variable. Given the objective function in (20), this corresponds to minimize y.

This procedure has to be done for all points x of the discretization of $[x_{\min}, x_{\max}]$. Once this procedure is finished, the best pair (x^*, y^*) is chosen and the corresponding q defined in (21), is the one that solves (17).

Fixing a value for $x\,\epsilon\,[x_{\min}, x_{\max}]$ and attending to (18), the problem related with the y variable optimization is

$$\min_{\substack{q^T Rq = x \\ q^T\Sigma_1 q = 1}} \quad q^T\Sigma_0 q \tag{24}$$

This problem is non-convex and is solved trough duality theory.

In the process of finding the pair (x^*, y^*) that solves the optimization problem in (20), for a fixed value of x, it is just needed to know the value of the best attainable value of y (calculated in (24)). It can be shown that strong duality exists for (24) when $x\,\epsilon\,]x_{\min}, x_{\max}[$. The values of y variable are calculated trough the dual problem, that is:

$$\max_{\substack{\Sigma_0 + \lambda_1 R + \lambda_2\Sigma_1 \geq 0 \\ \mathsf{var}\,:\,(\lambda_1, \lambda_2)\,\epsilon\,\mathbb{R}^2}} \quad -\lambda_1 x - \lambda_2 1 \tag{25}$$

In order to obtain the solution q for (17) and knowing that it verifies $q^T Rq = x^*$, the bi-dual problem of (24) is used where the restriction $\mathsf{tr}(RQ) = x^*$ is the bi-dual equivalent of $q^T Rq = x^*$

$$\begin{array}{ll} \min & \mathsf{tr}(\Sigma_0 Q) \\ \mathsf{tr}(RQ) = x^* & \\ \mathsf{tr}(\Sigma_1 Q) = 1 & \\ Q \succeq 0 & \end{array} \tag{26}$$

Provided Slater conditions are verified for (26) that can be shown to be equivalent to $x^* \,\epsilon\,]x_{\min}, x_{\max}[$, its solution Q is a rank-1 semidefinite positive matrix, and its only eigenvector is the solution q for the problem i.e., it is the linear discriminant that optimizes the Kullback-Leibler Divergence criterion.

Similar approaches can be followed to solve for the J-Divergence in (12) and the Chernoff Distance in (13).

4. GREEDY ALGORITHMS

The algorithms to be developed here perform a dimensionality reduction from n dimensions to p dimensions, where $p > 1$, through the linear discriminant matrix $Q \,\epsilon\, R^{p\times n}$, i.e.

$$y = Qx \tag{27}$$

where x is the n-dimensional sample, Q is the linear discriminant matrix, and $y \,\epsilon\, \mathbb{R}^p$ is the signature of the sample x, used in the classification procedure. The algorithms operate in a greedy manner. We focus on theKullback-Leibler Divergence criterion. For the other criteria the algorithms follow a similar pattern, which will not be repeated here.

The optimal linear discriminant $Q \,\epsilon\, \mathbb{R}^{p\times n}$, that maximizes the Kullback-Leibler Divergence between the p-dimensional probability density functions $f_0(Q) = N(Q\mu_0, Q\Sigma_0 Q^T)$ and $f_1(Q) = N(Q\mu_1, Q\Sigma_1 Q^T)$, is found by solving the optimization problem

$$\begin{array}{ll} \max\limits_{Q \,\epsilon\, \mathbb{R}^{p\times n}} & D_{KL}(f_0||f_1)(Q) \end{array} \tag{28}$$

where

$$\begin{array}{l} D_{KL}(f_0||f_1)(Q) = \\ \frac{1}{2}(\log \frac{|Q\Sigma_1 Q^T|}{|Q\Sigma_0 Q^T|} - tr((Q\Sigma_1 Q^T)^{-1}(Q\Sigma_0 Q^T)) \\ +(\mu_0 - \mu_1)^T Q^T (Q\Sigma_1 Q^T)^{-1} Q(\mu_0 - \mu_1) - p). \end{array} \tag{29}$$

The main problem with this approach is the non-convexity of the objective function. Although, the case $p = 1$ could be treated trough a series of reformulations and simplifications which made possible finding the solution efficiently, we were not able to extend this procedure for $p > 1$. So a sub-optimal approach to solve (28) is taken. This approach consists in compute the p rows of $Q \,\epsilon\, \mathbb{R}^{p\times n}$ one by one, by solving p 1-dimensional optimization problems, like the one in (15) for the case of the Kullback-Leibler Divergence.

It can be shown that without loss of optimality, the matrix Q in (28) can be taken to be Stiefel, i.e., with orthonormal rows. The fact that Q can be a Stiefel matrix motivates the following procedure to compute its p rows.

$$Q = \left[\begin{array}{c} -\, q_1^T \,- \\ \vdots \\ -\, q_p^T \,- \end{array} \right]$$

The first row q_1^T coincides with the linear discriminant q transposed, for the 1-dimension problem (see 15), i.e.

$$\begin{array}{lll} q_1 = & q = & \arg\max\limits_{q \neq 0} \; \frac{1}{2}(\frac{q^T\Sigma_0 q}{q^T\Sigma_1 q} - \log \frac{q^T\Sigma_0 q}{q^T\Sigma_1 q} \\ & & \qquad + \frac{[q^T(\eta_0 - \eta_1)]^2}{q^T\Sigma_1 q} - 1) \end{array} \tag{30}$$

The second row is computed by running again the algorithm, but now imposing that such row is orthogonal to the first, i.e:

$$q_2 = O_1 g \tag{31}$$

where $O_1 \,\epsilon\, \mathbb{R}^{n\times(n-1)}$ is a matrix, whose columns generate the orthogonal complement of the subspace generated by q_1, and $g \,\epsilon\, \mathbb{R}^{n-1}$ is the vector that collects the coefficients of the linear combination of the columns of O_1.

In order to compute q_2, a modified version of (30) is solved, i.e.

$$\begin{array}{ll} g = & \arg\max\limits_{g \neq 0} \; \frac{1}{2}(\frac{g^T O_1^T \Sigma_0 O_1 g}{g^T O_1^T \Sigma_1 O_1 g} - \log \frac{g^T O_1^T \Sigma_0 O_1 g}{g^T O_1^T \Sigma_1 O_1 g} \\ & \qquad + \frac{[g^T O_1^T(\eta_0 - \eta_1)]^2}{g^T O_1^T \Sigma_1 O_1 g} - 1) \end{array} . \tag{32}$$

The modification introduced, was the substitution of the q in (30), by $O_1 g$. This imposes orthogonality condition. Note that this optimization problem has exactly the same form of the 1-dimensional problem in (30), being therefore solved in exactly the same way. Note that $q_2 = O_1 g$.

To solve for row i, it's just a matter of substituting O_1 by O_{i-1}, being O_{i-1}, the matrix that generates the orthogonal complement to the subspace generated by the $i-1$ rows, previously calculated.

It's important to note, that the complexity of the subproblems solved to compute the p rows is decreasing. This is due to the fact that the optimization is being made in subspaces whose dimensions are decreasing.

5. COMPUTER SIMULATIONS: DIMENSIONALITY REDUCTION TO $\mathbb{R}$

As an illustration of the superiority of these new methods compared to the classical Fisher's Distance criterion, we show the results for the case of projecting the n-dimensional samples to $\mathbb{R}$. The results of the simulations are presented for three distinct cases concerning the parameters of the sources: distinct means and distinct covariance matrices (table 1), equal means and distinct covariance matrices (table 2), distinct means and equal covariance matrices (table 3), with increasing data dimensionality $n = 10, 20, 30, 40, 50$.

Table 1. Distinct Means and Distinct Covariance Matrices

	JD	KLD	CHF	FLDA
n=10	0.8508	0.8309	0.8607	0.5969
n=20	0.9870	0.9870	0.9415	0.6412
n=30	0.9010	0.9013	0.9376	0.7056
n=40	0.9935	0.9936	0.9426	0.6931
n=50	0.9891	0.9893	0.9430	0.7088

Table 2. Equal Means and Distinct Covariance Matrices

	JD	KLD	CHF	FLDA
n=10	0.9852	0.9853	0.9404	0.6006
n=20	0.9827	0.9830	0.9411	0.5159
n=30	0.9820	0.9821	0.9400	0.5255
n=40	0.9867	0.9868	0.9403	0.5243
n=50	0.9583	0.9586	0.9376	0.5053

Table 3. Distinct Means and Equal Covariance Matrices

	JD	KLD	CHF	FLDA
n=10	0.6740	0.6743	0.6743	0.6737
n=20	0.9255	0.9254	0.9253	0.9253
n=30	0.9203	0.9202	0.9204	0.9206
n=40	0.9665	0.9667	0.9666	0.9668
n=50	1.0000	1.0000	1.0000	1.0000

As we can see from tables 1 and 2, the FLDA always correspond to the worst performance.

6. CONCLUSIONS AND FUTURE WORK

In this thesis we proposed new criteria for designing linear discriminants for data dimensionality reduction prior to the application of a binary detector. We also developed algorithms to solve the non-convex optimization problems corresponding to the design of these new linear discriminants. These algorithms compute the linear discriminants that maximize the Chernoff Distance, the J-Divergence and Kullback-Leibler Divergence between the probability density functions that characterize the low-dimensional signatures of the original data.

The optimization problems that result from maximizing these measures of dissimilarity of the two sources are nonconvex. However it was possible to solve them efficiently (global optimality), through reformulations and the use of duality theory, for the case where the n-dimensional samples are mapped to $\mathbb{R}$. A suboptimal strategy was proposed for the case of mapping the samples to $\mathbb{R}^p$, with p greater than one.

The results obtained proved uniquivocally that the new techniques outperform the Fisher's Distance Criteria. This is due to the fact that the new criteria can discriminate the probability density functions through their variance. It's important to note that a Gaussian probability density function is characterized by its two first moments, the covariance matrix and the mean. Thus a good discriminator should use both to distinguish them. This is secured in the Chernoff Distance, J-Divergence and Kullback-Leibler Divergence criteria.

7. REFERENCES

[1] R.O. Duda, P.E. Hart, D.H. Stork. *Pattern Classification (2nd ed.)*. Wiley Interscience, (2000).

[2] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley-Interscience; New Ed edition (August 4, 2004).

[3] Jian Yang, Hui Ye, Zhang David. *A new LDA-KL combined method for feature extraction and its generalisation* Pattern Analysis and Applications, archive Volume 7 , Issue 2 (July 2004) table of contents Pages: 225 - 225 Year of Publication: 2004 ISSN:1433-7541

[4] Yongping Li , Josef Kittler , Jiri Matas. "Effective Implementation of Linear Discriminant Analysis for Face Recognition and Verification." Proceedings of the 8th International Conference on Computer Analysis of Images and Patterns, p.234-242, September 01-03, 1999.

[5] J.A.Thomas, T.M.Cover. *Elements of Information Theory*. John Wiley and Sons, Inc.,1991.

[6] Luis Rueda, Myriam Herrera. " A New Linear Dimensionality Reduction Technique Based on Chernoff Distance" IBERAMIA-SBIA 2006: 299-308

[7] Louis L. Scharf. *Statistical Signal Processing: Detection, Estimation and Time series Analysis*. Addision and Wesley, 1991.

[8] B. T. Polyak. "Convexity of Quadratic Transformations and Its Use in Control and Optimization". Journal of Optimization Theory and Applications. Vol. 99, No. 3, pp.553-583, December 1998.

[9] Lloyd L. Dines. "On the Mapping of Quadratic Forms." Bull. Amer. Math. Soc. Volume 47, Number 6 (1941), 494-498.

[10] Don H. Johnson, Sinan Sinanovic. *Symetrizing the Kullback-Leibler Distance*. Computer and Information Technology Institute. Department of Electrical Engineering. Rice University. Houston