A maximum uncertainty LDA-based approach for limited sample size problems – with application to face recognition

Carlos Eduardo Thomaz, Edson Caoru Kitani
Department of Electrical Engineering
Centro Universitário da FEI, São Paulo, Brazil
cet@fei.edu.br

Duncan Fyfe Gillies
Department of Computing
Imperial College, London, UK

Abstract *A critical issue of applying Linear Discriminant Analysis (LDA) is both the singularity and instability of the within-class scatter matrix. In practice, particularly in image recognition applications such as face recognition, there are often a large number of pixels or pre-processed features available, but the total number of training patterns is limited and commonly less than the dimension of the feature space. In this study, a new LDA-based method is proposed. It is based on a straightforward stabilisation approach for the within-class scatter matrix. In order to evaluate its effectiveness, experiments on face recognition using the well-known ORL and FERET face databases were carried out and compared with other LDA-based methods. The classification results indicate that our method improves the LDA classification performance when the within-class scatter matrix is not only singular but also poorly estimated, with or without a Principal Component Analysis intermediate step and using less linear discriminant features. Since statistical discrimination methods are suitable not only for classification but also for characterisation of differences between groups of patterns, further experiments were carried out in order to extend the new LDA-based method to visually analyse the most discriminating hyper-plane separating two populations. The additional results based on frontal face images indicate that the new LDA-based mapping provides an intuitive interpretation of the two-group classification tasks performed, highlighting the group differences captured by the multivariate statistical approach proposed.*

Keywords: *Linear Discriminant Analysis (LDA); small sample size; face recognition*

1 Introduction

The Fisher Discriminant Analysis, also called the Linear Discriminant Analysis (LDA), has been used successfully as a statistical feature extraction technique in several classification problems.

A critical issue in using LDA is, however, the singularity and instability of the within-class scatter matrix. In practice, particularly in image recognition applications such as face recognition, there are often a large number of pixels or pre-processed features available, but the total number of training patterns is limited and commonly less than the dimension of the feature space. This implies that the within-class scatter matrix either will be singular if its rank is less than the number of features or might be unstable if the total number of training patterns is not significantly larger than the dimension of the feature space.

A considerable amount of research has been devoted to the design of other Fisher-based methods, for targeting small sample and high dimensional problems [1, 3, 21, 25, 26, 27, 28]. However, less attention has been paid to problems where the dimensionality of the feature space is comparable to the total number of training examples. In this situation, the within-class scatter matrix is full rank but poorly estimated.

In this study, a new Fisher-based method is proposed. It is based on the straightforward maximum entropy covariance selection approach [23] that overcomes both the singularity and instability of the within-class scatter matrix when LDA is applied in limited sample and high dimensional problems. In order to evaluate its effectiveness, experiments on face recognition using the well-known ORL and FERET face databases were carried out and compared with other LDA-based methods. The classification results indicate that our method improves the LDA classification performance when the within-class scatter matrix is singular as well as poorly estimated, with or without a Principal Component Analysis (PCA) intermediate step and using less linear discriminant features.

Since statistical discrimination methods are suitable not only for classification but also for characterisation of differences between groups of patterns, further experiments were carried out in order to extend the new LDA-based method to visually analyse the most discriminating hyper-plane separating two populations. The additional results based on frontal face images indicate that the new LDA-based mapping provides an intuitive interpretation of the two-group classification tasks performed, highlighting the group differences captured by the multivariate statistical approach proposed.

2 Linear Discriminant Analysis (LDA)

The primary purpose of the Linear Discriminant Analysis is to separate samples of distinct groups by maximising their between-class separability while minimising their within-class variability. Although LDA does not assume that the populations of the distinct groups are normally distributed, it assumes implicitly that the true covariance matrices of each class are equal because the same within-class scatter matrix is used for all the classes considered [11].

Let the between-class scatter matrix S_b be defined as

$$S_b = \sum_{i=1}^{g} N_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \qquad \textbf{(1)}$$

and the within-class scatter matrix S_w be defined as

$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T \qquad \textbf{(2)}$$

where $x_{i,j}$ is the n-dimensional pattern j from class π_i, N_i is the number of training patterns from class π_i, and g is the total number of classes or groups. The vector $\bar{x}_i$ and matrix S_i are respectively the unbiased sample mean and sample covariance matrix of class π_i [7]. The grand mean vector $\bar{x}$ is given by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{g} N_i \bar{x}_i = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{N_i} x_{i,j}\ , \qquad \textbf{(3)}$$

where N is the total number of samples, that is, $N = N_1 + N_2 + \cdots + N_g$. It is important to note that the within-class scatter matrix S_w defined in equation (2) is essentially the standard pooled covariance matrix multiplied by the scalar $(N - g)$, that is

$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = (N - g) S_p\ . \qquad \textbf{(4)}$$

The main objective of LDA is to find a projection matrix P_{lda} that maximizes the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix (Fisher's criterion), that is

$$P_{lda} = \arg\max_{P} \frac{\left| P^T S_b P \right|}{\left| P^T S_w P \right|}\ . \qquad \textbf{(5)}$$

Devijver and Kittler [5] have shown that P_{lda} is in fact the solution of the following eigensystem problem:

$$S_b P - S_w P \Lambda = 0\ . \qquad \textbf{(6)}$$

Multiplying both sides by S_w^{-1}, equation (6) can be rewritten as

$$
\begin{aligned}
&S_w^{-1}S_bP - S_w^{-1}S_wP\Lambda = 0 \\
&S_w^{-1}S_bP - P\Lambda = 0 \\
&(S_w^{-1}S_b)P = P\Lambda
\end{aligned}
\qquad \textbf{(7)}
$$

where P and Λ are respectively the eigenvectors and eigenvalues of $S_w^{-1}S_b$. In other words, equation (7) states that if S_w is a non-singular matrix then the Fisher's criterion described in equation (5) is maximised when the projection matrix P_{lda} is composed of the eigenvectors of $S_w^{-1}S_b$ with at most $(g-1)$ nonzero corresponding eigenvalues. This is the standard LDA procedure.

The performance of the standard LDA can be seriously degraded if there are only a limited number of total training observations N compared to the dimension of the feature space n. Since the within-class scatter matrix S_w is a function of $(N-g)$ or less linearly independent vectors, its rank is $(N-g)$ or less. Therefore, S_w is a singular matrix if N is less than $(n+g)$, or, analogously, might be unstable if N is not at least five to ten times $(n+g)$ [9].

In the next section, recent LDA-based methods proposed for targeting limited sample and high dimensional problems are described. A novel method of combining singular and non-singular covariance matrices for solving the singularity and instability of the within-class scatter matrix is proposed in section 4.

3 LDA Limited Sample Size Methods

A critical issue for the standard LDA feature extraction technique is the singularity and instability of the within-class scatter matrix. Thus, a considerable amount of research has been devoted to the design of other LDA-based methods, for overcoming the limited number of samples compared to the number of features. In the following sub-sections, recent LDA-based methods with application to face recognition are described. Since the face recognition problem involves small training sets, a large number of features, and a large number of groups, it has become the most used application to evaluate such limited sample size approaches [1, 3, 21, 25, 26, 27, 28].

3.1 Fisherfaces Method

The Fisherfaces [1, 28] method is one of the most successful feature extraction approaches for solving limited sample size problems in face recognition. It is also called the Most Discriminant Features (MDF) method [21].

The Fisherfaces or MDF method is essentially a two-stage dimensionality reduction technique. First the face images from the original vector space are projected to a lower dimensional space using Principal Component Analysis (PCA) [24] and then LDA is applied next to find the best linear discriminant features on that PCA subspace.

More specifically, the MDF projection matrix P_{mdf} can be calculated as

$$
P_{mdf} = P_{lda} * P_{pca}, \qquad \textbf{(8)}
$$

where P_{pca} is the projection matrix from the original image space to the PCA subspace, and P_{lda} is the projection matrix from the PCA subspace to the LDA subspace obtained by maximising the ratio

$$
P_{lda} = \arg\max_P \frac{\left|P^T P_{pca}^T S_b P_{pca} P\right|}{\left|P^T P_{pca}^T S_w P_{pca} P\right|}. \qquad \textbf{(9)}
$$

As described in the previous section, equation (9) analogously states that if $P_{pca}^T S_w P_{pca}$ is a non-singular matrix then the Fisher's criterion is maximised when the projection matrix P_{lda} is composed of the eigenvectors of $(P_{pca}^T S_w P_{pca})^{-1}(P_{pca}^T S_b P_{pca})$ with at most $(g-1)$ nonzero corresponding eigenvalues.

The singularity problem of the within-class scatter matrix S_w is then overcome if the number of retained principal components varies from at least g to at most $N-g$ PCA features [1, 21, 28].

3.2 Chen et al.'s Method (CLDA)

Chen et al. [3] have proposed another LDA-based method, here called CLDA, that overcomes the singularity problems related to the direct use of LDA in small sample size applications, particularly in face recognition.

The main idea of their approach is to use either the discriminative information of the null space of the within-class scatter matrix to maximise the between-class scatter matrix whenever S_w is singular, or the eigenvectors corresponding to the set of the largest eigenvalues of matrix $(S_b+S_w)^{-1}S_b$ whenever S_w is non-singular. Fukunaga [7] has proved that the eigenvectors of $(S_b+S_w)^{-1}S_b$ are the same as $S_w^{-1}S_b$.

The CLDA algorithm for calculating the projection matrix P_{clda} can be summarised as follows [3]:

i. Calculate the rank r of the within-class scatter matrix S_w;

ii. If S_w is non-singular, that is $r = n$, then P_{clda} is composed of the eigenvectors corresponding to the largest eigenvalues of $(S_b + S_w)^{-1} S_b$;

iii. Otherwise, calculate the eigenvectors matrix $V = [v_1, ..., v_r, v_{r+1}, ..., v_n]$ of the singular within-class scatter matrix S_w. Let Q be the matrix that spans the S_w null space, where $Q = [v_{r+1}, v_{r+2}, ..., v_n]$ is an $n \text{ x } (n-r)$ sub-matrix of V;

iv. The projection matrix P_{clda} is then composed of the eigenvectors corresponding to the largest eigenvalues of $QQ^T S_b (QQ^T)^T$. Chen et al. have proved that those eigenvectors obtained through the transformation QQ^T are the most discriminant vectors in the original sample space [3].

Although their experimental results have shown that CLDA improves the performance of a face recognition system compared with Liu et al.'s approach [12] and the standard template matching procedure [10], Chen et al.'s approach will select the same linear discriminant features as the standard LDA when S_w is non-singular but poorly estimated.

3.3 Yu and Yang's Method (DLDA)

Yu and Yang [27] have developed a direct LDA algorithm (DLDA) for high dimensional data with application to face recognition that diagonalises simultaneously the two symmetric matrices S_w and S_b [7].

The key idea of their method is to discard the null space of S_b by diagonalising S_b first and then diagonalising S_w. As pointed out by Yu and Yang [27] the traditional LDA procedure takes the reverse order and consequently discards the null space of S_w which contains discriminative information [3]. This diagonalisation process also avoids the singularity problems related to the use of the pure LDA in high dimensional data where the within-class scatter matrix S_w is likely to be singular [27].

The DLDA algorithm for calculating the projection matrix P_{dlda} can be described as follows [27]:

i. Diagonalise S_b, that is calculate the eigenvector matrix V such that $V^T S_b V = \Lambda$;

ii. Let Y be the first m columns of V corresponding to the S_b largest eigenvalues, where $m \leq rank(S_b)$. Calculate $D_b = Y^T S_b Y$, where D_b is the diagonal $m \text{ x } m$ sub-matrix of the eigenvalues matrix Λ;

iii. Let $Z = Y D_b^{-1/2}$ be a whitening transformation of S_b that also reduces its dimensionality from n to m, i.e $Z^T S_b Z = (Y D_b^{-1/2})^T S_b (Y D_b^{-1/2}) = I$;

iv. Diagonalise $Z^T S_w Z$, that is compute U and D_w such that $U^T (Z^T S_w Z) U = D_w$;

v. Calculate the projection matrix P_{dlda} given by $P_{dlda} = D_w^{-1/2} U^T Z^T$.

Using computational techniques to handle large scatter matrices, Yu and Yang's [27] experimental results have shown that DLDA can be applied on the original vector space of face images without any explicit intermediate dimensionality reduction step. However, they pointed out [27] that by replacing the between-class scatter matrix S_b with the total scatter matrix S_T, given by $S_T = S_b + S_w$, the first two steps of their algorithm becomes exactly the PCA dimensionality reduction technique.

3.4 Yang and Yang's Method (YLDA)

More recently, Yang and Yang [26] have proposed a linear feature extraction method, here called YLDA, which is capable of deriving discriminatory information of the LDA criterion in singular cases.

Analogous to the Fisherfaces method described previously in the subsection 3.1, the YLDA is explicitly a two-stage dimensionality reduction technique. That is, PCA [24] is used firstly to reduce the dimensionality of the original space and then LDA, using a particular Fisher-based linear algorithm called Optimal Fisher Linear Discriminant (OFLD) [25], is applied next to find the best linear discriminant features on that PCA subspace.

The OFLD algorithm [25] can be described as follows:

i. In the *m*-dimensional PCA transformed space, calculate the within-class and between-class scatter matrices S_w and S_b;

ii. Calculate the eigenvectors matrix $V = [v_1, v_2, ..., v_m]$ of S_w. Suppose the first q eigenvectors of S_w correspond to its non-zero eigenvalues;

iii. Let a projection matrix be $P_1 = [v_{q+1}, v_{q+2}, ..., v_m]$, which spans the null space of S_w. Form the transformation matrix Z_1 composed of the eigenvectors of $P_1^T S_b P_1$. The first k_1 YLDA discriminant vectors are given by $P_{ylda}^1 = P_1 Z_1$, where generally $k_1 = g - 1$;

iv. Let a second projection matrix be $P_2 = [v_1, v_2, ..., v_q]$. Form the transformation matrix Z_2 composed of the eigenvectors corresponding to the k_2 largest eigenvalues of $(P_2^T S_w P_2)^{-1}(P_2^T S_b P_2)$. The remaining k_2 YLDA discriminant vectors are given by $P_{ylda}^2 = P_2 Z_2$, where k_2 is an input parameter that can extend the final number of LDA features beyond the $(g-1)$ nonzero S_b eigenvalues;

v. Form the projection matrix P_{ylda} given by the concatenation of P_{ylda}^1 and P_{ylda}^2.

Yang and Yang [26] have proved that the number m of principal components to retain for a best LDA performance should be equal to the rank of the total scatter matrix S_T, given, as reminder, by $S_T = S_b + S_w$ and calculated on the original space [26]. However, no procedure has been shown to determine the optimal value for the parameter k_2. This parameter is context dependent and consequently can vary according to the application studied. Moreover, although YLDA addresses the PCA+LDA problems when the total scatter matrix S_T is singular, such PCA strategy does not avoid the within-class scatter instability when S_T is non-singular but poorly estimated.

4 Maximum Uncertainty LDA

In order to avoid both the singularity and instability critical issues of the within-class scatter matrix S_w when LDA is used in limited sample and high dimensional problems, we propose a new LDA-based approach based on a straightforward covariance selection method for the S_w matrix [23].

4.1 Related Methods

In the past, a number of researchers [2, 4, 17, 19] have proposed a modification in LDA that makes the problem mathematically feasible and increases the LDA stability when the within-class scatter matrix S_w has small or zero eigenvalues.

The idea is to replace the pooled covariance matrix S_p of the scatter matrix S_w (equation (4)) with a ridge-like covariance estimate of the form

$$\hat{S}_p(k) = S_p + kI, \qquad \textbf{(10)}$$

where I is the *n* by *n* identity matrix and $k \ge 0$. DiPillo [4] attempted to determine analytically the optimal choice for the value k. However, such solution has been shown intractable in practice and several researchers have performed simulation studies to choose the best value for k [4, 17, 19].

According to Rayens [19], a reasonable grid of potential simulation values for the optimal k could be

$$\lambda_{\min} \le k \le \lambda_{\max}, \qquad \textbf{(11)}$$

where the values $\lambda_{\min}$ and $\lambda_{\max}$ are respectively the nonzero smallest and largest eigenvalues of the pooled covariance matrix S_p. Rayens [19] has suggested that a more productive searching process should be based on values near $\lambda_{\min}$ rather than $\lambda_{\max}$. However, this reasoning is context-dependent and a time-consuming leave-one-out optimisation process is necessary to determine the best multiplier for the identity matrix.

Other researchers have imposed regularisation methods to overcome the singularity and instability in sample based covariance estimation, especially to improve the Bayes Plug-in or QDF classification performance [6, 8, 22]. Most of these works have used shrinkage parameters that combine linearly a singular or unstable covariance matrix, such as S_p, to a multiple of the identity matrix.

According to these regularisation methods, the ill posed or poorly estimated S_p could be replaced with a convex combination matrix $\hat{S}_p(\gamma)$ of the form

$$\hat{S}_p(\gamma) = (1-\gamma)S_p + (\gamma)\bar{\lambda}I, \qquad \textbf{(12)}$$

where the shrinkage parameter γ takes on values $0 \le \gamma \le 1$ and could be selected to maximise the leave-one-out classification accuracy. The identity matrix multiplier would be given by the average eigenvalue $\bar{\lambda}$ of S_p calculated as

$$\bar{\lambda} = \frac{1}{n}\sum_{j=1}^{n}\lambda_j = \frac{tr(S_p)}{n}, \qquad \textbf{(13)}$$

where the notation “tr” denotes the trace of a matrix.

The regularisation idea described in equation (12) would have the effect of decreasing the larger eigenvalues and increasing the smaller ones, thereby counteracting the biasing inherent in sample-based estimation of eigenvalues [6].

4.2 The Proposed Method

The proposed method considers the issue of stabilising the S_p estimate with a multiple of the identity matrix by selecting the largest dispersions regarding the S_p average eigenvalue. It is based on our maximum entropy covariance selection idea developed to improve quadratic classification performance on limited sample size problems [23].

Following equation (10), the eigen-decomposition of a combination of the covariance matrix S_p and the *n* by *n* identity matrix I can be written as [14]

$$\begin{aligned}\hat{S}_p(k) &= S_p + kI \\ &= \sum_{j=1}^{r} \lambda_j \phi_j (\phi_j)^T + k\sum_{j=1}^{n} \phi_j (\phi_j)^T \\ &= \sum_{j=1}^{r} (\lambda_j + k)\phi_j (\phi_j)^T + \sum_{j=r+1}^{n} k\phi_j (\phi_j)^T\end{aligned} \qquad \textbf{(14)}$$

where r is the rank of S_p ($r \le n$), λ_j is the *j*th non-zero eigenvalue of S_p, ϕ_j is the corresponding eigenvector, and k is an identity matrix multiplier. In equation (14), the following alternative representation of the identity matrix in terms of any set of orthonormal eigenvectors is used [14]

$$I = \sum_{j=1}^{n} \phi_j (\phi_j)^T \ . \qquad \textbf{(15)}$$

As can be seen from equation (14), a combination of S_p and a multiple of the identity matrix I as described in equation (10) expands all the S_p eigenvalues, independently whether these eigenvalues are either null, small, or even large.

A possible regularisation method for LDA could be the one that decreases the larger eigenvalues and increases the smaller ones, as briefly described by equation (12) of the previous sub-section. According to this idea, the eigen-decomposition of a convex combination of S_p and the *n* by *n* identity matrix I can be written as

$$\begin{aligned}\hat{S}_p(k) &= (1-\gamma)S_p + \gamma\bar{\lambda}I \\ &= (1-\gamma)\sum_{j=1}^{r} \lambda_j \phi_j (\phi_j)^T + \gamma\sum_{j=1}^{n} \bar{\lambda}\phi_j (\phi_j)^T\end{aligned} \qquad \textbf{(16)}$$

where the mixing parameter γ takes on values $0 \le \gamma \le 1$ and $\bar{\lambda}$ is the average eigenvalue of S_p.

Despite the substantial amount of computation saved by taking advantage of matrix updating formulas [6, 19, 22], the regularisation method described in equation (16) would require the computation of the eigenvalues and eigenvectors of an *n* by *n* matrix for each training observation of all the classes in order to find the best mixing parameter γ. In recognition applications where several classes and a large total number of training observations are considered, such as face recognition, this regularisation method might be unfeasible.

Yet, equation (16) describes essentially a convex combination between a singular or poorly estimated covariance matrix, the pooled covariance matrix S_p, and a non-singular or well-estimated covariance matrix: the identity matrix I. Therefore, the same idea described in [23] of selecting the most reliable linear features when blending such covariance matrices can be used.

Since the estimation errors of the non-dominant or small eigenvalues are much greater than those of the dominant or large eigenvalues [7], we propose the following selection algorithm in order to expand only the smaller and consequently less reliable eigenvalues of S_p, and keep most of its larger eigenvalues unchanged:

i. Find the Φ eigenvectors and Λ eigenvalues of S_p, where $S_p = S_w/[N-g]$;

ii. Calculate the S_p average eigenvalue $\bar{\lambda}$ using equation (13);

iii. Form a new matrix of eigenvalues based on the following largest dispersion values

$$\Lambda^* = diag[\max(\lambda_1, \bar{\lambda}), \ldots, \max(\lambda_n, \bar{\lambda})]; \qquad \textbf{(17a)}$$

iv. Form the modified within-class scatter matrix

$$S_w^* = S_p^*(N-g) = (\Phi\Lambda^*\Phi^T)(N-g) \ . \qquad \textbf{(17b)}$$

The maximum uncertainty LDA (MLDA) is constructed by replacing S_w with S_w^* in the Fisher's criterion formula described in equation (5). It is a straightforward method that overcomes both the singularity and instability of the within-class scatter matrix S_w when LDA is applied directly in limited sample and high dimensional problems. MLDA also avoids the computational costs inherent to the aforementioned shrinkage processes.

The main idea of the proposed LDA-based method can be summarised as follows. In limited sample size and high dimensional problems where the within-class scatter matrix is singular or poorly estimated, it is reasonable to expect that the Fisher's linear basis found by minimizing a more difficult "inflated" within-class S_p^* estimate would also minimize a less reliable "shrivelled" within-class S_p estimate.

5 Experiments

In order to evaluate the effectiveness of the maximum uncertainty LDA-based method (MLDA) on face recognition, comparisons with the standard LDA (when possible), Fisherfaces, CLDA, DLDA, and YLDA, were performed using the well-known Olivetti-Oracle Research Lab$^{\zeta}$ (ORL) and FERET [18] face databases.

A simple Euclidean distance classifier was used to perform classification in the projective feature space, analogously to the other approaches we investigated. Although other classifiers, such as the k nearest neighbour classifier and Mahalanobis distance classifier [7, 11], could be used to perform such classification, we would expect that the relative performance of the different methods investigated would not be affected by the choice of that classifier. Each experiment was repeated 25 times using several features. Distinct training and test sets were randomly drawn, and the mean and standard deviation of the recognition rate were calculated. The classification of the ORL 40 subjects was computed using for each individual 5 images to train and 5 images to test. In the FERET database with 200 subjects, the training and test sets were respectively composed of 3 and 1 frontal images.

For implementation convenience, the ORL face images were resized to 32x32 pixels, representing a recognition problem where the within-class scatter matrix is singular, that is the total number of training observations was $N = 200$ and the dimensionality of the original images was $n = 1024$. The FERET images were resized to 16x16 pixels in order to pose an alternative pattern recognition problem where the within-class scatter matrix is non-singular but poorly estimated, i.e. $N = 600$ and $n = 256$.

To determine the number of principal components to be retained in the intermediate step of Fisherfaces, experimental analyses were carried out based on the best classification accuracy of several PCA features in between the corresponding interval $(g, N-g)$. The best results were obtained when the ORL and FERET original images were first reduced respectively to 60 and 200 PCA features.

For the purpose of establishing the number of the YLDA best discriminant vectors derived from the within-scatter matrix eigenvectors space, we used for the ORL database the eigenvectors corresponding to the remaining 10 largest eigenvalues, as suggested by Yang and Yang's work [26]. For the FERET database, the eigenvectors corresponding to the remaining 20 largest eigenvectors were sufficient to determine the respective YLDA best discriminant vectors. We assumed that an eigenvalue λ is positive if $round(\lambda) > 0$.

6 Results

Tables 1 and 2 present the maximum test average recognition rates (with standard deviations) of the ORL and FERET databases over the corresponding number of PCA (when applicable) and LDA features.

Since the ORL face database contains only 40 subjects to be discriminated, the LDA features of the Fisherfaces, CLDA, DLDA, and MLDA were limited to 39 components. Using the remaining 10 largest eigenvalues, the number of YLDA discriminant vectors could be extended from 39 to 49 LDA features. Also, the notation "-" in the standard LDA (LDA) row of the Table 1 indicates that the within-class scatter matrix was singular and consequently the standard LDA could not be calculated.

	Features		
Method	PCA	LDA	Recognition Rate
Fisherfaces	60	39	94.9% (1.9%)
YLDA	199	45	96.1% (1.4%)
LDA	-	-	-
CLDA		39	95.4% (1.5%)
DLDA		39	94.9% (1.6%)
MLDA		39	95.8% (1.6%)

Table 1. ORL (32x32 pixels) LDA classification results.

Table 1 shows that the maximum uncertainty LDA (MLDA) led to higher classification accuracies than the other one-stage approaches (CLDA and DLDA). The overall best classification result was reached by Yang and Yang's approach (YLDA) – 96.1% (1.4%) – which was not significantly greater than the MLDA one – 95.8% (1.6%). However, the YLDA used a much larger two-stage linear transformation matrix compared to the one-stage methods. In terms of how sensitive the MLDA results were to the choice of the training and test sets, it is fair to say that the new LDA standard deviations were similar to the other methods.

[ζ] Available on the following web site: http://www.cl.cam.ac.uk/Research/DTG/attarchive/facedatabase.html

Table 2 presents the results of the FERET database. In this application, the within-class scatter was non-singular but poorly estimated and the standard LDA (LDA) could be applied directly on the face images. As can be seen from Table 2, the overall best classification result was achieved by MLDA – 95.4% (1.4%) – using remarkably only 10 features. Again, regarding the standard deviations, MLDA showed to be as sensitive to the choice of the training and test sets as the other approaches investigated.

	Features		
Method	PCA	LDA	Recognition Rate
Fisherfaces	200	20	91.5% (1.9%)
YLDA	256	92	94.7% (1.4%)
LDA		20	86.2% (1.9%)
CLDA		20	86.2% (1.9%)
DLDA		20	94.5% (1.3%)
MLDA		10	95.4% (1.4%)

Table 2. FERET (16x16 pixels) LDA classification results.

7 Memory Issues

According to Samal and Iyengar [20], images with 32x32 pixels and at least 4 bits per pixel are sufficient for face identification problems. However, it is possible that memory computation problems would arise when scatter matrices larger than 1024x1024 elements are used directly in the optimisation of the Fisher's criterion described in equation (5).

In fact, the PCA intermediate step that has been applied to project images from the original space into the face subspace has made not only some of the aforementioned LDA-based approaches mathematically feasible in limited sample size and high-dimensional classification problems, but also has allowed the within-class S_w and between-class S_b scatter matrices to be calculable in computers with a normal memory size [13].

In the experiments described previously, our attention was focused on evaluating the new LDA-based performance in situations where the within-class scatter matrix was either singular or poorly estimated, without a PCA intermediate step of dimensionality reduction. However, it would be important to assess the proposed method in higher resolution images where the PCA intermediate step is made necessary to avoid such memory computation difficulties.

Thus, we discuss here experimental results that evaluate the previous top 2 MLDA and YLDA approaches when the standard resolutions of 64x64 pixels and 96x64 pixels were used respectively for the ORL and FERET face images. Analogous to the previous experiments, the classification of the ORL 40 subjects was computed using in total 200 examples for training (5 images per subject) and the remaining 200 examples (5 images per subject) for testing. In the FERET database with 200 subjects, the total number of training and test sets were respectively composed of 600 (3 images per subject) and 200 (1 image per subject) images. Following the Yang and Yang's work [26], we used again the eigenvectors corresponding to the remaining 10 largest eigenvalues to extend the number of YLDA discriminant vectors. For the FERET database, the eigenvectors corresponding to the remaining 25 largest eigenvalues were sufficient to determine the respective YLDA best discriminant vectors.

As described previously, the total number of principal components to retain for a best LDA performance should be equal to the rank of the total scatter matrix $S_T = S_w + S_b$ [26]. When the total number of training examples N is less than the dimension of the original feature space n, the rank of S_T can be calculated as [15]

$$\begin{aligned} rank(S_T) &\le rank(S_w) + rank(S_b) \\ &\le (N-g) + (g-1) \\ &\le N-1. \end{aligned} \tag{18}$$

In order to avoid the high memory rank computation of such large scatter matrices and because both MLDA and YLDA deal with the singularity of the within-class scatter matrix, we used equation (18) to assume that the rank of S_T in both applications was $N-1$. Therefore, we first projected the original ORL and FERET images into the corresponding 199 and 599 largest principal components and secondly we applied the MLDA and YLDA feature classification methods.

Table 3 shows the maximum test average recognition rates (with standard deviations) of the ORL and FERET datasets over the corresponding number of PCA and LDA features. To determine the number of linear discriminant features to be retained, experimental analyses were carried out based on the best classification accuracy of several LDA features. As can be seen, likewise the previous experiments, the best classification results for the ORL dataset was achieved by the Yang and Yang's approach (YLDA), which was slightly better than the MLDA one. However, the YLDA used a larger two-stage linear transformation matrix. In the FERET application, where the higher resolution images improved the classification results of both YLDA and MLDA approaches, the MLDA achieved clearly the best classification performance, using impressively only 10 LDA features after the PCA dimensionality reduction.

Dataset	Features		
Method	PCA	LDA	Recognition Rate
ORL			
YLDA	199	46	96.1% (1.5%)
MLDA	199	39	95.7% (1.5%)
FERET			
YLDA	599	220	95.5% (1.2%)
MLDA	599	10	97.6% (1.1%)

Table 3. ORL (64x64 pixels) and FERET (96x64 pixels) LDA classification results.

8 Visual Analysis of the Discriminative Information

In the generic discrimination problem, where the training sample consists of the class membership and observations for N patterns, the outcome of interest fall into g classes and we wish to build a rule for predicting the class membership of an observation based on n variables or features. However, as mentioned earlier, statistical discrimination methods are suitable not only for classification but also for characterisation of differences between groups of patterns. For example, in face recognition we might want to understand the differences between male and female face images by exploring the discriminating hyper-plane found by the linear multivariate statistical classifier.

In this section, we present some initial experiments to visually analyse the linear discriminant feature found by the MLDA approach. We have used frontal images of a face database maintained by the Department of Electrical Engineering at FEI. This database contains a set of face images taken between June and December 2005 at the Artificial Intelligence Laboratory in São Bernardo do Campo, with 14 images for each of 118 individuals, a total of 1652 images[*].

In order to estimate the MLDA separating hyperplane, we have used training examples and their corresponding labels to construct the classifier. First a training set is selected and the average image vector of all the training images is calculated and subtracted from each n-dimensional vector. Then the training matrix composed of zero mean image vectors is used as input to compute the PCA transformation matrix. The columns of this n x m transformation matrix are eigenvectors, not necessarily in eigenvalues descending order. Analogously to the previous section, we have retained all the PCA eigenvectors with non-zero eigenvalues, that is, $m = N - 1$. The zero mean image vectors are projected on the principal components and reduced to m-dimensional vectors representing the most expressive features of each one of the n-dimensional image vector. Afterwards, this N x m data matrix is used as input to calculate the MLDA discriminant eigenvector, as described in the section 4.2. Since in these experiments we have limited ourselves to two-group classification problems, there is only one MLDA discriminant eigenvector. The most discriminant feature of each one of the m-dimensional vectors is obtained by multiplying the N x m most expressive features matrix by the m x 1 MLDA linear discriminant eigenvector. Thus, the initial training set of face images consisting of N measurements on n variables, is reduced to a data set consisting of N measurements on only 1 most discriminant feature.

Once the two-stage PCA+MLDA classifier has been constructed, we can move along its corresponding projection vector and extract the discriminant differences captured by the classifier. Any point on the discriminant feature space can be converted to its corresponding n-dimensional image vector by simply: (1) multiplying that particular point by the transpose of the corresponding linear discriminant vector previously computed; (2) multiplying its m most expressive features by the transpose of the principal components matrix; and (3) adding the average image calculated in the training stage to the n-dimensional image vector. Therefore, assuming that the spreads of the classes follow a Gaussian distribution and applying limits to the variance of each group, such as $\pm 2s_i$, where s_i is the standard deviation of each group, we can move along the MLDA most discriminant features and map the results back into the image domain.

Figure 1 presents the PCA+MLDA most discriminant feature of the FEI face database for a facial expression experiment. In this two-group classification task, we have used 33 examples of smiling and 33 examples of non-smiling frontal face images (female only). Figure 1 displays the image regions captured by the MLDA classifier that change when we move from one side (smiling) of the dividing hyper-plane to the other (non-smiling), following limits to the standard deviation ($\pm 2s_i$, whenever possible) of each sample group. As can be seen, the PCA+MLDA hyper-plane effectively extracts the group differences, showing clearly the changes in the facial expression of the images. It is important to note that these differences

[*] All these images are available upon request (cet@fei.edu.br).

could be very subtle on samples that are very close to the dividing boundary, such as the feature points $+2s_1$ and $-1s_2$, and consequently difficult to characterise as belonging to one of the groups when mapped back to the original image space.

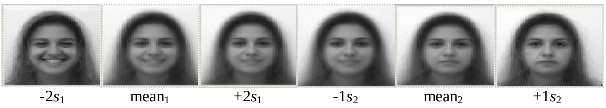

Figure 1. Visual analysis of the regions captured by the classifier that change when we move from one side (smiling) of the dividing PCA+MLDA hyper-plane to the other (non-smiling), following limits of $\pm 2s_i$ standard deviations (whenever possible) for each sample group. We can see clearly the changes in the facial expression of the images.

Analogously to the previous experiments, Figure 2 illustrates the PCA+MLDA most discriminant feature using 33 examples of male and 33 examples of female face images. In this two-group classification task, we intend to visually analyse the differences captured by the MLDA classifier between male and female frontal samples. As can be seen, the mapping procedure provides again an intuitive interpretation of the classification experiments and highlights the changes when we move from one side (female) of the dividing hyper-plane to the other (male).

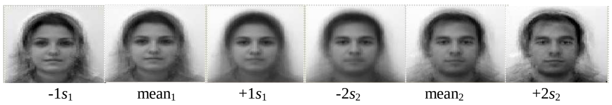

Figure 2. Visual analysis of the regions captured by the classifier that change when we move from one side (female) of the dividing PCA+MLDA hyper-plane to the other (male), following limits of $\pm 2s_i$ standard deviations (whenever possible) for each sample group. We can see clearly the changes between the female and male frontal samples.

9 Conclusions

In this paper, we extended the idea of the maximum entropy selection method used in Bayesian classifiers to overcome not only the singularity but also the instability of the LDA within-class scatter matrix in limited sample, high dimensional problems.

The new LDA-based method (MLDA) is a straightforward approach that considers the issue of stabilising the ill posed or poorly estimated within-class scatter matrix with a multiple of the identity matrix. Although such modification has been used before, our method is based on selecting the largest and consequently most informative dispersions. Therefore, it avoids the computational costs inherent to the commonly used optimisation processes, resulting in a simple and efficient implementation for the maximisation of the Fisher's criterion.

Classification experiments were carried out to evaluate this approach on face recognition, using the standard ORL and FERET databases. Comparisons with similar methods, such as Fisherfaces [1, 28], Chen et al.'s [3], Yu and Yang's [27], and Yang and Yang's [25, 26] LDA-based methods, were made. In both databases, our method improved the LDA classification performance with or without a PCA intermediate step and using less linear discriminant features. Regarding the sensitivity to the choice of the training and test sets, the maximum uncertainty LDA gave a similar performance to the compared approaches.

In addition, further experiments of the new LDA-based method (MLDA) were presented to analyse the most discriminating hyper-plane separating two populations. The PCA+MLDA multivariate statistical approach has shown to be an efficient way of mapping multivariate classification results of the whole images back into the original image domain for visual interpretation. We limited our analyses to two-group problems, such as smiling/non-smiling and male/female frontal face classification. In both experiments, the two-stage mapping procedure provided an intuitive interpretation of the classification experiments and highlighted the group differences captured by the MLDA classifier.

We have shown that in limited sample size and high dimensional problems where the within-class scatter matrix is singular or poorly estimated, the Fisher's linear basis found by minimising a more difficult but appropriate "inflated" within-class scatter matrix would also minimise a less reliable "shrivelled" within-class estimate. We believe that such LDA modification might be suitable for solving not only the singularity and instability issues of the linear Fisher methods, but also the Fisher discriminant analysis with kernels [16] where the non-linear mapping of the original space to a higher dimensional feature space would commonly lead to a ill-posed within class scatter matrix.

Acknowledgments

The authors would like to thank Leo Leonel de Oliveira Junior for acquiring and normalizing the FEI face database under the grant FEI-PBIC 32/05. Also, we would like to thank the referees for their constructive comments which helped to improve this paper.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[2] N.A. Campbell, “Shrunken estimator in discriminant and canonical variate analysis”, *Applied Statistics*, vol. 29, pp. 5-14, 1980.

[3] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu, “A new LDA-based face recognition system which can solve the small sample size problem”, *Pattern Recognition*, 33 (10), pp. 1713-1726, 2000.

[4] P.J. Di Pillo, “Biased Discriminant Analysis: Evaluation of the optimum probability of misclassification”, *Communications in Statistics-Theory and Methods*, vol. A8, no. 14, pp. 1447-1457, 1979.

[5] P.A. Devijver and J. Kittler, *Pattern Classification: A Statistical Approach.* Prentice-Hall, Englewood Cliffs, N. J., 1982.

[6] J.H. Friedman, “Reguralized Discriminant Analysis”, *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165-175, March 1989.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second edition. Boston: Academic Press, 1990.

[8] T. Greene and W.S. Rayens, “Covariance pooling and stabilization for classification”, *Computational Statistics & Data Analysis*, vol. 11, pp. 17-42, 1991.

[9] A. K. Jain and B. Chandrasekaran, “Dimensionality and Sample Size Considerations in Pattern Recognition Practice”, *Handbook of Statistics*, P.R. Krishnaiah and L.N. Kanal Eds, vol. 2, pp. 835-855, North Holland, 1982.

[10] A. K. Jain, R. P. W. Duin and J. Mao, “Statistical Pattern Recognition: A Review”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, January 2000.

[11] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, fourth edition. New Jersey: Prentice Hall, 1998.

[12] K. Liu, Y. Cheng, and J. Yang, “Algebraic feature extraction for image recognition based on an optimal discriminant criterion”, *Pattern Recognition*, 26 (6), pp. 903-911, 1993.

[13] Y. Li, J. Kittler, and J. Matas, “Effective Implementation of Linear Discriminant Analysis for Face Recognition and Verification”, *Computer Analysis of Images and Patterns: 8th International Conference CAIP'99*, Springer-Verlag LNCS 1689, pp. 232-242, Ljubljana, Slovenia, September 1999.

[14] S.L. Marple, *Digital Spectral Analysis with Applications*. Englewood Cliffs, N.J: Prentice-Hall, 1987.

[15] J.R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, revised edition. Chichester: John Wiley & Sons Ltd., 1999.

[16] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. –R. Muller, “Fisher discriminant analysis with kernels”, *IEEE Neural Networks for Signal Processing IX*, pp. 41-48, 1999.

[17] R. Peck and J. Van Ness, “The use of shrinkage estimators in linear discriminant analysis”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, no. 5, pp. 531-537, September 1982.

[18] P. J. Phillips, H. Wechsler, J. Huang and P. Rauss, “The FERET database and evaluation procedure for face recognition algorithms”, *Image and Vision Computing Journal*, vol. 16, no. 5, pp. 295-306, 1998.

[19] W.S. Rayens, “A Role for Covariance Stabilization in the Construction of the Classical Mixture Surface”, *Journal of Chemometrics*, vol. 4, pp. 159-169, 1990.

[20] A. Samal and P. Iyengar, “Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey”, *Pattern Recognition*, 25 (1), pp. 65-77, 1992.

[21] D. L. Swets and J. J. Weng, “Using Discriminant Eigenfeatures for Image Retrieval”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831-836, 1996.

[22] S. Tadjudin, “Classification of High Dimensional Data With Limited Training Samples”, PhD thesis, Purdue University, West Lafayette, Indiana, 1998.

[23] C. E. Thomaz, D. F. Gillies and R. Q. Feitosa. "A New Covariance Estimate for Bayesian Clas-

sifiers in Biometric Recognition", *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, vol. 14, no. 2, pp. 214-223, February 2004.

[24] M. Turk and A. Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, vol. 3, pp. 71-86, 1991.

[25] J. Yang and J. Yang, "Optimal FLD algorithm for facial feature extraction", *SPIE Proceedings of the Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision*, vol. 4572, pp. 438-444, 2001.

[26] J. Yang and J. Yang, "Why can LDA be performed in PCA transformed space? ", *Pattern Recognition*, vol. 36, pp. 563-566, 2003.

[27] H. Yu and J. Yang, "A direct LDA algorithm for high dimensional data – with application to face recognition", *Pattern Recognition*, vol. 34, pp. 2067-2070, 2001.

[28] W. Zhao, R. Chellappa and A. Krishnaswamy, "Discriminant Analysis of Principal Components for Face Recognition", in *Proc. 2nd International Conference on Automatic Face and Gesture Recognition*, 336-341, 1998.