A Novel FLDA Formulation for Numerical Stability Analysis

Fernando VILARIÑO, Debora GIL, Petia RADEVA
Centre de Visió per Computador – Dept. Informàtica, UAB
Edifici O – Campus UAB, 08193 Bellaterra, Barcelona, Catalonia, Spain
{fernando;petia}@uab.es

Abstract. Fisher Linear Discriminant Analysis (FLDA) is one of the most popular techniques used in classification applying dimensional reduction. The numerical scheme involves the inversion of the within-class scatter matrix, which makes FLDA potentially ill-conditioned when it becomes singular. In this paper we present a novel explicit formulation of FLDA in terms of the eccentricity ratio and eigenvector orientations of the within-class scatter matrix. An analysis of this function will characterize those situations where FLDA response is not reliable because of numerical instability. This can solve common situations of poor classification performance in computer vision.

Keywords. Supervised Learning, Linear Discriminant Analysis, Numerical Stability, Computer Vision

1 Introduction

FISHER Linear Discriminant Analysis (FLDA) is a well known technique used in classification problems applying dimensional reduction of high-dimensional feature spaces. The main idea of this dimensional reduction is to seek for a feature subspace in which the classification problem is more manageable. In the case of linear discriminant analysis (LDA), this is achieved through a linear transformation. FLDA approach looks for a compromise in which the class clusters are mapped onto a reduced space maximizing its distance, while trying to keep the class clusters as compact as possible. In this subspace, the projected samples are labeled to their corresponding class by standard assignment criteria (defined metrics, nearest neighborhood, etc.). Because of its simplicity, FLDA has been widely used in the fields of face recognition [19], [18], [1] and automatic speech recognition [13] (just to mention a few). Further, FLDA is used in more sophisticated discriminant schema, such as Discriminant Component Analysis (DCA) [21], and non-linear classification using kernels [15], [16].

The FLDA scheme encodes distance between clusters and its compactness into the between- and within-class scatter matrices. In the solution to the FLDA problem, this set of matrices can be found as a part of a generalized eigenvalue problem [4], or as part of a two step-wise orthogonalization procedure, as pointed out by Fukunaga [6]. In any case, FLDA numerical scheme involves the inversion, either explicitly [4] or implicitly [6], of the within-class scatter matrix. Therefore, special attention should be paid to those situations where the latter matrix becomes non-invertible, as they turn the scheme ill-posed and lead

to numerical instability. This phenomenon is produced by two main reasons: a low number of samples in comparison to the feature space dimensionality, and a low scatter for certain directions.

On one hand, working with a large set of samples is not always possible, and it is a requirement rarely met in very high dimensional spaces [17], as it occurs in the framework of face recognition [3]. On the other hand, sample distributions with few scatter in some dimensions are frequent when directional filter banks are used for feature extraction [5], [20] and [8]. Further, low scatter directions are generally related to noise. Hence, they constitute a main hindrance in a whitening scheme [6], because they are amplified and lead to undesired overfitting [1]. Among the different solutions to the former non-invertibility problems ([1],[3], [14], [18], [11]) the most popular are those that use a previous PCA in order to discard low scatter directions [1],[18] and [11]. Special care must be taken in this step, since FLDA projection space might include directions of small scatter rejected by PCA. This dichotomy is explained in [18], where it is shown that PCA is suitable for encoding the Most Expressive Features, while FLDA is related to the Most Discriminant Features. In those situations where the former spaces are complementary, a previous PCA step will eliminate the features that actually should efficiently discriminate. Most approaches look for a compromise between the number of dimensions rejected and the efficiency of the remaining set to discriminate [11], [12], [10]. Besides, they assume that it is achieved in terms of spectral energy criteria, regardless of whether there has been any real improvement in the numerical stability.

Our effort has been focused on characterizing those situations such that FLDA presents numerical instability by means of error propagation principles [2]. Following the latter, we will establish a stability criterion in terms of output precision, considering that a generic procedure is non reliable if little changes in the input data yield a significant variation in the resulting response. This ratio of variation is measured by the first derivatives of the function encoding such procedure. Therefore, in this paper we will develop an explicit formulation for FLDA response in terms of the eccentricity ratio -related to the eigenvalues magnitude- as well as eigenvectors orientations of the within-class matrix. On one hand, bounds on the first derivatives serve to delimit the non-stability regions (NSR) in the former configuration space, where the FLDA approach yields non reliable results. This criterion assures that we can apply the FLDA robustly from a numerical point of view. On the other hand, a study of the formula in the two-dimensional case yields that not only eigenvalues induce numerical errors, but also orientation of eigenvectors. In fact, we will show that the angular orientation of the within-class eigenvectors plays a decisive role in numerical stability. Further, we will link our numerical stability scheme to PCA dimensionality reduction used as a previous step to FLDA. Our novel point of view will provide with an efficient criterion for choosing the number of dimensions rejected by PCA.

The organization of the paper is as follows: Section II presents the general FLDA formulation and a geometrical interpretation of the mathematical objects present on it. In section III, we describe the propose an explicit formula for FLDA based on geometric parameters: the orientation angle of the within-class scatter matrix and the eccentricity ratio In the conclusions of section IV, we address the explicit characterization of the stability configuration space, based on the angular and eccentricity parameters derived exclusively from the within-class scatter matrix for the analysis of numerical stability, for being exhaustively presented in future papers.

2. Linear Discriminant Analysis Formulation

Given a set of samples, X, in an n-dimensional feature space that can be labeled into c classes, a discriminant analysis seeks for a transformation to another feature space so that the classes are best separated. In classical Linear Discriminant Analysis transformations are assumed to be linear. Under this assumption, the problem reduces to finding an m-dimensional subspace, W, and a corresponding linear manifold achieving the optimal separability between the projected classes. That is, we look for a matrix Wmxn such that we can perform the optimal separation between classes in the subspace where we have the projected samples Y = WX. In the case of FLDA, this optimal separation is based in two expectations: 1) to maximize distance between class means and 2) to get the concentration of all of the samples that belong to the same class around its mean. Applying this separability criterion, FLDA looks for the subspace, W, that maximizes the Fisher ratio quotient:

$$J(\mathbf{W}) = \frac{|\mathbf{W}^t \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^t \mathbf{S}_W \mathbf{W}|} \tag{1}$$

where SB is the between-class scatter matrix and SW is the within-class scatter matrix. Under the assumption of normal distributions, SB and SW are built as follows:

1. $\mathbf{S_B} = \sum_{\mathbf{i=1}}^{\mathbf{c}} (\mu_\mathbf{i} - \mu)(\mu_\mathbf{i} - \mu)^\mathbf{t}$ with sub index i for the mean vector for the i-class, is symmetric and positive semi definite. Since it is the sum of c outer products of vectors, its rank is at most c-1 and it is quite singular [17]
2. $\mathbf{S}_W = \sum_{\mathbf{i=1}}^{\mathbf{c}} \sum_{\mathbf{j=1}}^{\mathbf{N_i}} (\mathbf{Y_j} - \mu_\mathbf{i})(\mathbf{Y_j} - \mu_\mathbf{i})^\mathbf{t}$ with Ni the number of samples for class i, is the sum of the scatter matrices for the c-classes. It is symmetric and positive semi definite, but when dealing with a large set of samples (in comparison to the dimension, n, of the feature space) it is in general positive definite, and so, non-singular.

Notice that once we have the projected data, Y = WX, our classification problem becomes a problem of finding minimum distance to the projected mean, nearest neighbor assignment, etc. In order to give an interpretation to the ratio J(W), we analyze each of the matrices in the next section.

2.1 Geometrical interpretation

The matrix S_B, as a linear application, encodes the projection from the n-dimensional space into an affine variety M generated by the c-classes mean vectors. Its at most c-1 non-null eigenvectors define this variety and correspond to the main scatter directions of the mean points set. Since the product WtS_BW cancels for vectors perpendicular to M, the subspace maximizing this quantity is a basis of the vector subspace that defines M. Hence, maximizing the numerator of (1) implies giving the projection onto the linear variety that maximizes distances between classes mean points.

Statistically, S_W is the scatter matrix of a normal distribution equivalent to the one build up as the sum of the c-classes scatter matrices. The resulting distribution takes into account the scatter of each class, but not their spatial location -this information is incorporated in S_B-.

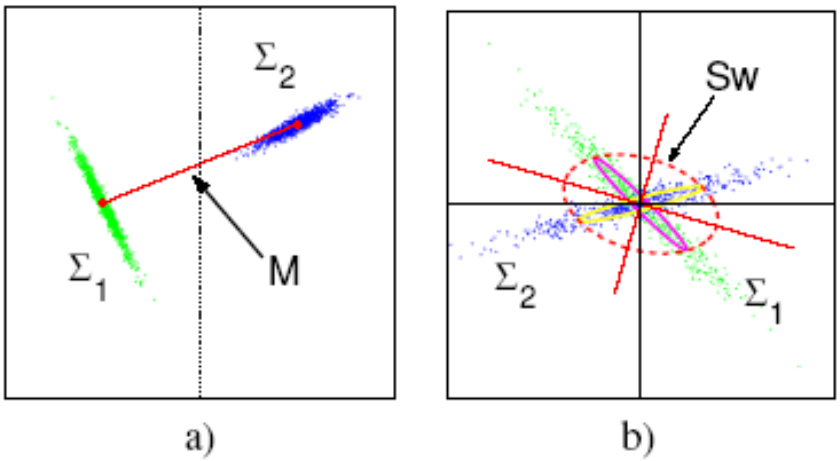

Fig. 1. a) Two normal distributions with different orientations and eccentricity ratio. b) S_w geometrical interpretation. The axes of the ellipses has been drawn with standard deviation length.

Thus, S_w cannot be viewed as the total scatter matrix for all samples, but as a weighted mean of the classes scatter, as the graphical representation of Fig. 1b shows. The quadratic form associated to S_W can be represented by an ellipsoid. As a linear application, S_W transforms vectors from the unitary sphere onto this associated ellipsoid. Its n non-null orthogonal eigenvectors describe the directions of maximum scatter of the distribution associated to S_W, which coincide with the principal axes of the ellipsoid representing the quadratic form. The corresponding eigenvalues, larger for sparser data, are the length of the ellipsoid axes. The determinant $|WtS_WW|$ is minimum for the scatter direction of lowest scatter and maximum for the highest one. Hence, $|WtS_BW|$ is minimum for the direction in which the data projection minimizes the projected points scatter, and maximum for the direction of maximum scatter.

Notice that the quotient J(.) is maximum for the subspace that maximizes the numerator and minimizes the denominator. Consequently, by the former analysis, maximizing (1) is equivalent to finding the projection subspace, W, that satisfies a compromise between maximizing the distance between class means and comprising the projected classes [17].

An example of FLDA solution for the case of two normal distributions is shown in Fig. 2. The green and blue normal distributions of Fig.2a have different orientations and eccentricities. The yellow ellipse represents the distribution associated to SW. The dashed line shows the projecting line of the FLDA solution. Plot 2b shows the value of J(.) as a function of the solution vector orientation. Notice that, by its own definition as scalar products of unitary vectors, both numerator and denominator of J(.) are periodic functions, with a phase delay depending on the relative orientations between the principal eigenvectors S_W and S_B. On one hand, the magnitude of $|WS_BW|$ is a function of the distance between lass means, and its phase delay depends on the orientation of M. On the other hand, the magnitude of $|WtS_WW|$ is related to the eccentricity ratio of S_W, tending to be constant for an isotropic –spherical distribution. Its phase delay depends on the orientation of the major axis. When the major axis of S_W is perpendicular to M, we reach the optimal classification outcome for FLDA: maximum between-class distance and minimum within-class scatter.

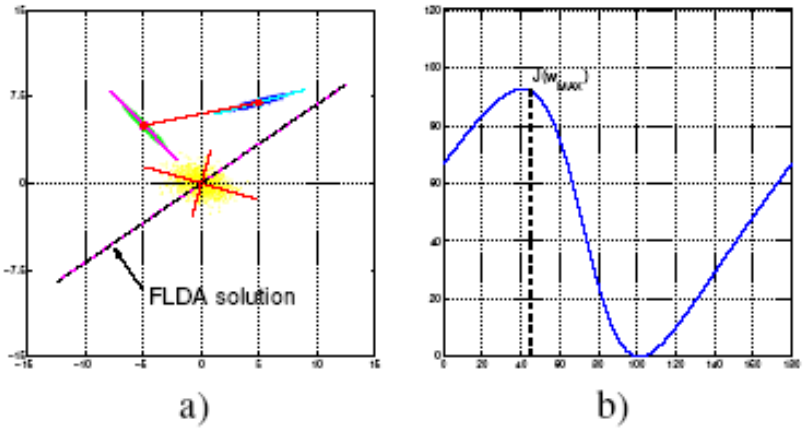

Fig. 2. a) S_W orientation and FLDA response. b) Plot of J(.) vs. angle β.

2.2 Fisher eigenvector solution

It can be proved [4], [9] that J(.) is maximized when the column vectors of S_W are the eigenvectors of $S_W^{-1}S_B$ associated with the largest eigenvalues. In fact, the eigenvalue solution is exactly the value of $J(W) = \lambda$, for the W that maximizes J(.).

We may find numerical problems with this eigenvalue scheme when S_W becomes singular due to -i,e., with low sample scatter in some direction-. One possible solution is to perform an intermediate dimensional reduction with a principal component analysis [19]. PCA seeks for the most representative features in the statistical scatter sense, taking apart those features with the lowest statistical scatter. In general, these taken-apart features are assumed to be useless for class representation, or simply noisy features. Unfortunately, when our class samples are defined in the rejected feature space, the criterion of maximizing J(.) will not give optimal solutions in terms of class separability. An alternative formulation to cope with S_W invertibility problems is to state a generalized eigenvalue problem $S_BW = \lambda S_WW$.

However, numerical instability is still present, since it comes from both the eigenvalue ratio and the orientation of its corresponding eigenvectors.

In [6] it is shown that solving the generalized eigenvalue problem is equivalent to finding the eigenvectors of the transformation $Y = A'X$, where $A = \Phi\Theta^{-1/2}\Psi$. This transformation diagonalizes both S_W and S_B, and can be performed in two steps. The first step -namely the whitening transform- transforms S_W into the identity matrix applying $Y = A1'X$, where $A1 = \Phi\Theta^{-1/2}$, for Φ and Θ the eigenvector and eigenvalue matrices of S_W. The second step applies $Z = A_2'Y$ where $A_2 = \Psi$, with Ψ the eigenvectors of $A_1'S_BA$, transforming S_B into a diagonal matrix –S_w keeps the identity so it is invariant to orthogonal transforms-. We may notice the reader that, in this framework, numerical instability problems hide in the whitening transformation itself rather than in the transformed feature domain. Our numerical stability approach applies to FLDA solution based on the primitive feature space.

From now on, we will focus in the qualitative study of FLDA for the 2-class classification problem in a two dimensional feature space. The goal is to present the solution vector for FLDA as a function of the eccentricity ratio of S_W and the orientation of

the main axes of the normal distribution associated. The following section will use this result to analyze the situations where FLDA presents non-optimal solutions.

3 Geometrical FLDA Formulation

In the problem of 2 classes, we have rank(S_B) = 1, and the solution for the FLDA is defined by the unique eigenvector w of non-null eigenvalue λ, given by the generalized eigenvector formulation:

$$(\mathbf{S}_W{}^{-1}\mathbf{S}_B)\mathbf{w} = \lambda\mathbf{w} \tag{2}$$

Further, since we can assume that ||w|| = 1 and we work in a 2D space, we have that w=(cos α, sin α) and, so, solving (2) reduces to finding the angular orientation of w with respect to the line connecting the class mean points, M. The main steps in the deduction of a formula for α, are the following.

First, since the eigenvalue solution is invariant under orthogonal transforms -see [6] for an analysis of its derived consequences-, we can assume, without lost of generality, that the line connecting the class means is parallel to the x-axis. This step, that only requires a rotation of the main coordinate frame, implies that

$$\mathbf{S_b} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

and that angles (including the solution α) are defined with respect to the x-axis.
Second, since Sw is symmetric, it can be expressed in diagonal form [7] as:

$$\mathbf{Sw} = \begin{pmatrix} C_\beta & -S_\beta \\ S_\beta & C_\beta \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} C_\beta & S_\beta \\ -S_\beta & C_\beta \end{pmatrix} = \begin{pmatrix} \lambda_1 C_\beta^2 + \lambda_2 S_\beta^2 & (\lambda_2 - \lambda_1) S_\beta^2 C_\beta^2 \\ (\lambda_2 - \lambda_1) S^2 \beta C_\beta^2 & \lambda_1 C_\beta^2 + \lambda_2 S_\beta^2 \end{pmatrix} \tag{3}$$

where λ_i are the eigenvalues of S_W in increasing order ($\lambda_1 <= \lambda_2$), β is the angle of the ellipse minor axis and C_β; S_β are abbreviations for cos β and sin β -see Fig. 3 for a graphical representation-.

Using the above notation, we can express the tangent of the angle solution to (2) in terms of the eccentricity ratio of S_W, defined by $e = \lambda_1/\lambda_2$, and β by the function:

$$f(e, \beta) = \tan(\alpha) = \frac{(1-e) S_\beta C_\beta}{e S_\beta^2 + C_\beta^2} \tag{4}$$

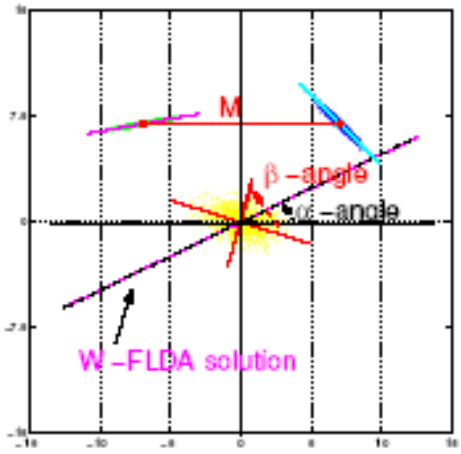

Fig. 3. Distributions and angles: The system is rotated to make the line that passes through the mean points parallel to the x-axis, the yellow ellipse represents S_W, β is the angle of the minor axis of S_W and is the angle of w.

Therefore, studding numerical properties of FLDA reduces to analyzing the properties of (4) as a function of the two variables (e, β). Defining an stability criterion based on this geometrical parameters, as also empirical results based in real world data, is related to our next publications.

4 Conclusions

In this paper we have presented a novel formulation for FLDA based on a geometric interpretation of the parameters. This formulation takes into account only the angle of orientation of the within-class scatter matrix and the eccentricity ratio for the 2-class problem in 2D. One of its main advantages is that it relates very closely geometrical distribution of data with the mathematical arguments passed to the formula, making more clear the interpretation of results. We base our outcome in a deep explanation of geometrical and mathematical issues around the Fisher discriminant quotient. The main objective of this result is to apply this formulation to the calculus of numerical stability of FLDA –work to be done in next papers-. The mathematical development of the formula will be deeply presented as an extended publication. This results are useful for poor classification performance in computer vision problems –face recognition, speech recognition, medical imaging, etc- in order to detect the numerical instability due to the high eccentric geometrical distribution of samples sets.

References

[1] Peter N. Belhumeur, J. Hespanha, and David J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. In ECCV (1), pages 45-58, 1996.

[2] G. Dahlquist and A. Bjorck. Numerical Methods. Prentice-Hall, 1974.

[3] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. J. Opt. Soc. Am. A, 14:1724-1733, 1997.

[4] R.A. Fisher. The statistical utilization of multiple mesurements. In Annals of Eugenics, volume 8, pages 376-386, 1938.

[5] W. T. Freeman and E. H. Adelson. The design and use of steerable lters. IEEE TPAMI, 13(9):891-906, 1991.

[6] Keinosuke Fukunaga. Introduction To Statistical Pattern Recognition. Academic Press, 2 edition, 1991.

[7] David A. Harville. Matrix Algebra Form a Statician's Perspective. Springer, 1999.

[8] Petia Radeva J.M Pardo. Discriminant snakes for 3d reconstruction in medical images. In ICPR2000 Proc. volume 4, pages 336-339.

[9] J.T. Kent K.V. Mardia and J.M. Bibby. Multivariate Analysis. Academic Press, 1995.

[10] Gwi-Tae Lee, Sung-Og. Park. Pose invariant view-based enhanced sher linear discriminant models for face recognition. In ICCAS Proceedings, 2001.

[11] C. Liu and H. Wechsler. Enhanced Fisher linear discriminant models for face recognition. In In Proc. the 14th ICPR, 1998.

[12] Chengjun Liu and Harry Wechsler. A shape and texture based enhanced sher classi er for face recognition. IEEE Transactions IP, 10(4):598-608, 2001.

[13] H. Dolfing M. Katz, H.G. Meier and D. Klakow. Robustness of linear discriminant analysis in automatic speech recognition. In PatternRecognition, 2002. Proceedings. 16th International Conference on, volume 3, pages 374, 2002.

[14] Aleix M. Martinez and Avinash C. Kak. PCA versus LDA. IEEE TPAMI, 23(2):228-233, 2001.

[15] S. Mika, G. R¨atsch, J. Weston, B. Scholkopf, and K.-R. M¨uller. Fisher discriminant analysis with kernels. In Neural Networks for Signal Processing IX, p 41-48. 1999.

[16] S. Mika, A. Smola, and B. Scholkopf. An improved training algorithm for kernel Fisher discriminants. In Proc. AISTATS, 2001.

[17] David G. Stork Richard O. Duda, Peter E. Hart. Pattern Classification. Wiley-Interscience, 2 edition, 2001.

[18] Daniel L. Swets and Juyang Weng. Using discriminant eigenfeatures for image retrieval. IEEE TPAMI, 18(8):831-836, 1996.

[19] M. Turk and A. Pentland. Eigenfaces for recognition. Journal of Congitive Neuroscience, 3(1):71-86, 1991.

[20] Fernando Vilariño and Petia Radeva. Patch-optimized discriminant active contour for medical image segmentation. VIII Conferencia Ibreroamericana de Inteligencia Artificial. Sevilla. Spain, 2002.

[21] Wenyi Zhao. Discriminant component analysis for face recognition.