International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 22, No. 6 (2008) 1–31
© World Scientific Publishing Company

World Scientific
www.worldscientific.com

NONLINEAR FEATURE EXTRACTION USING FISHER CRITERION

MATÍAS A. BUSTOS, MANUEL A. DUARTE-MERMOUD* and NICOLÁS H. BELTRÁN†

Electrical Engineering Department
University of Chile, Av. Tupper 2007
Casilla 412-3, Santiago, Chile
**mduartem@ing.uchile.cl*
†nicolas.beltran@die.uchile.cl

In this paper the problem of nonlinear feature extraction based on the optimization of the Fisher criterion is analyzed. A new nonlinear feature extraction method is proposed. The method does not make use of numerical algorithms and it has an analytical (closed-form) solution. Moreover, no assumptions on the class probability distribution functions are imposed. The proposed method is applied to some standard pattern recognition problems and compared with other classical methodologies already proposed in the literature. The performance of the proposed method turned out to be superior when compared with the other methods studied.

Keywords: Feature extraction; nonlinear feature extraction; Fisher transformation; nonlinear Fisher transformation; kernel Fisher.

1. Introduction

Advances in computation have made possible to handle large amount of data to solve numerous problems in several engineering areas that not so long ago seemed untractable. Pattern recognition and data mining are some of the fields benefited by these developments. Currently it is common to find pattern recognition applications involving feature vectors of large dimension.[8,22] It is also known that feature vectors of large dimension cause training troubles in classification algorithms such as the "dimensionality coarse"[3,21] and overtraining, affecting directly the classifier generalization capacity.

One alternative to face this dimension problem is to use feature extraction techniques. These techniques generate new variables of lower dimension maintaining the discrimination power of the original data. Feature extraction has been the object of several studies resulting in numerous algorithms and methods of feature extraction.[19,29,34] The most popular, due to their simplicity and robustness, are

*Author for correspondence

Principal Component Analysis (PCA) and Fisher Transformation (FT) or Fisher Discriminant Analysis (FDA). Both methods extract the features through a linear projection of the original data but optimizing different criteria. FDA generates new feature vectors preserving the discrimination power of the original data and drastically diminishing their dimension, but losing other kind of information contained in the data (e.g. physical meaning, redundancy, brightness, etc.). PCA instead looks for the best representation of high dimension data, in the mean square sense, in a subspace of lower dimension. Another difference is that FDA is a supervised method, i.e. uses information about the class where each training pattern belongs, whereas PCA is a nonsupervised method.[2,14]

Since the proposed method will be based on the Fisher criterion, in what follows we will present a brief historical background of Fisher developments. In 1937 Fisher introduced FDA for two class problems.[10,11] This procedure allows to find a data projection where the quotient between the distance of the projected class means (inter-class distance) and the sum of the projected class scatter around the projected class means, (intra-class distance) is minimized. Rao[30] generalized this procedure to the case of C classes based on the data projection onto a C-1 space, through a matrix. Among the advantages of Rao's procedure is the obtainment of an analytical solution solving an eigenvalue–eigenvector problem.

Forty years later, Campell[6] proved that FDA solution is equivalent to the solution obtained by using the maximum *a posteriori* (MAP) rule for the case when classes have normal distributions with equal covariance matrices. In 1990 Fukunaga[14] presented an extensive study about the properties of Fisher criterion (base of FDA for C classes) and proposed several alternative criteria. Later, in 1996, the generalization of FDA was attempted through numerical methods[15,17] and also using the maximum expectation algorithm.[16,18] Three years later, in 1999, Mika[25] proposed a nonlinear extension of FDA for two-class problems, using the same ideas used by Scholkopf to generalize PCA.[31] This method is known as kernel Fisher (KFDA). Several improvements have been proposed to the training algorithms of this method.[24,26] Later, Baudat[1] published the first attempt to solve KFDA for C classes, but it was only in 2002 when Navarrete *et al.*[27] solved the multiclass KFDA.

Feature extraction methods based on transformations of the input samples (measurements) produce a new set of features in the transformed space that can exhibit high "information packing" properties compared with the original input samples. The basic reasoning behind transform-based features is that appropriately chosen transformation can exploit discrimination and remove redundancies, which usually exist in a set of samples obtained by measuring devices.[34] A certain degree of class separation can be achieved in the domain of transformed features when linear transformations are used, which might be enough for certain type of classification problems. However, there are some pattern recognition problems where classes are quite inbred and using linear transformations is not enough to get good classification results. In these cases nonlinear transformations of original samples are called for in order to improve class separation properties. Another important fact supporting nonlinear FDA is that FDA does not work properly when class means are different or when the information for classification purposes lies on data variance rather than in the mean. It is important to point out that there exist others approaches for nonlinear feature extraction not based on Fisher criterion; e.g. the one proposed by Zhang on polygonal principal curves[39] or those based on neural networks.[34]

In this paper the optimization of the Fisher criterion in a space nonlinearly related to the original data is studied. In Sec. 2 the problem is solved using calculus of variations finding an analytical solution that needs the knowledge of *a posteriori* probability density that a vector (pattern) X belongs to each class. Since probability densities are in general unknown, the solution is then restricted to transformation that can be written as a linear combination of basis functions, finding a closed-form solution. A procedure is presented in Sec. 3 associated with the proposed solution to substantially diminish the computational load. In Sec. 4, the classification behavior of the combination of three feature extraction methods (no feature extraction, FDA and quadratic FDA) together with a classifier based on Linear Discriminant Analysis (LDA), will be compared. Six experiments will be performed using six standard data sets encountered in pattern recognition literature.[4] In order to compare the behavior of the feature extraction methods, the classification error using LDA[34] as classifier is computed. LDA was chosen because it is the simplest statistical classifier able to generate only linear decision boundaries, and therefore the effect of the feature extraction methods will be highlighted in the classification rate of the system. In all cases, the use of the proposed feature extraction method significantly improved classification rates.

2. Non Linear Fisher Transformation

In this section a new approach to Fisher criterion optimization is presented. This approach, based on calculus of variations[9], does not restrict the feature extraction function to a linear transformation, case solved by Fisher[10,11] and Rao[30] in the thirties and forties respectively.

2.1. *Optimization of Fisher criterion in function spaces*

Let $\pounds^2$ be the space of functions defined on $\Re^n \to \Re^m$, having continuous partial derivatives of order 1 and 2. Our objective is to find a function $Z(X) \in \pounds^2$ (with $Z \in \Re^m$ and $X \in \Re^n$), such that the Fisher index evaluated in the space generated by $Z(X)$ is maximum. For notation purposes the symbol $\sim$ over the original variables will be used to denote variables in the transformed space. The Fisher index in the transformed space will be given by

$$J = \mathrm{tr}\{\tilde{S}_w^{-1}\tilde{S}_b\} \tag{2.1}$$

where

$$\tilde{S}_w = \sum_{i=1}^{C} P(w_i)\tilde{\Sigma}_i \tag{2.2}$$

$$\tilde{S}_b = \sum_{i=1}^{C} P(w_i)(\tilde{\mu}_i - \tilde{\mu}_0)(\tilde{\mu}_i - \tilde{\mu}_0)^T \tag{2.3}$$

$$\tilde{\mu}_i = \int_{-\infty}^{\infty} Z(X)p(X/w_i)dX, \quad \tilde{\mu}_0 = \int_{-\infty}^{\infty} Z(X)p(X)dX \tag{2.4}$$

$$\tilde{\Sigma}_i = \int_{-\infty}^{\infty} Z(X)Z(X)^T p(X/w_i)dX \tag{2.5}$$

$\tilde{S}_w \in \Re^{m\times m}$ is the within-class scatter matrix, $\tilde{\Sigma}_i \in \Re^{mm}$ is the covariance matrix for class w_i, $P(w_i)$ is *a priori* probability of class w_i, $\tilde{S}_b \in \Re^{m\times m}$ is the between-class scatter matrix, $\tilde{\mu}_i \in \Re^m$ is the mean of class w_i, $\tilde{\mu}_0 \in \Re^m$ is the global mean vector and $p(X/w_i)$ is the conditional density. C is the number of classes.

It is possible to prove (see Appendix A) that the function $Z(X)$ that optimizes criterion (2.1) satisfies the following relationship

$$2\frac{\partial J}{\partial \tilde{S}_w} Z(X) = -\sum_{i=1}^{C}\left[\hat{p}(X/w_i)\frac{\partial J'}{\partial \tilde{\mu}_i}\right] \tag{2.6}$$

From (2.6) it is observed that $Z(X)$ explicitly depends on *a posteriori* probability density functions $\hat{p}(X/w_i)$ defining the probability of X belonging to each class in the Bayes sense.

However, this result is not applicable to real pattern recognition problems since the form and parameters of $\hat{p}(X/w_i)$ are unknown. Even if $\hat{p}(X/w_i)$ were explicitly known, we would need numerical methods to solve (2.6) due to the dependence of $Z(X)$ on matrices $\partial J/\partial\tilde{S}_w$ and $\partial J/\partial\tilde{\mu}_i$. Moreover, in the case of known there would be no necessity of feature extraction since classification would be done directly using the Bayes rule, assuring minimum error probability in the classification process.

2.2. *Constrained solution to NLFT problem*

In this section we present a methodology for a nonlinear extension of the Fisher transformation. To this extent we restrict the class of functions $Z(X)$ to be considered in the solution of (2.1) to those that can be written as a linear combination of K functions $\{\varphi_i(X)\}_{i=1}^{K}$ with $\varphi_i(X) : \Re^n \to \Re$, i.e.

$$Z(X) = \begin{bmatrix} Z_1(X) \\ Z_2(X) \\ \cdot \\ \cdot \\ Z_m(X) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{K} \alpha_1^i \varphi_i(X) \\ \sum_{i=1}^{K} \alpha_2^i \varphi_2(X) \\ \cdot \\ \cdot \\ \sum_{i=1}^{K} \alpha_m^i \varphi_i(X) \end{bmatrix} \in \mathbb{R}^m \tag{2.7}$$

or equivalently

$$Z(X) = \Omega^T \Phi(X) \tag{2.8}$$

where

$$\Omega^T = \begin{bmatrix} \alpha_1^1 & \alpha_1^2 & \cdots & \alpha_1^K \\ \alpha_2^1 & \alpha_2^2 & \cdots & \alpha_2^K \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_m^1 & \alpha_m^2 & \cdots & \alpha_m^K \end{bmatrix} \qquad \Phi(X) = \begin{bmatrix} \varphi_1(X) \\ \varphi_2(X) \\ \cdot \\ \cdot \\ \varphi_K(X) \end{bmatrix} \in \Re^K \tag{2.9}$$

Vector $\Phi(X) \in \Re^K$ is a vector function whose components $\varphi_i(X)$ are nonlinear scalar functions of the elements of the original feature space X. Matrix $\Omega \in \Re^{K\times m}$ contains the parameters of the transformation.

The idea of using functions of the form (2.8) to find nonlinear solutions to Fisher criterion is not new. This idea has been successfully used in the kernel Fisher feature extraction.[27,36] However, in those works function $\Phi(X)$ is introduced with the argument that data separation in the space generated by the nonlinear transformation will be improved. The latter is based on the fact that we are implicitly using high order correlations in the original space.[7]

In the present work a second interpretation will be given to function $\Phi(X)$, where each component corresponds to a basis function. Thus, each component of the optimal solution of the Fisher criterion in $\pounds^2$ is approximated by these basis functions.

It can be proved (see Appendix B) that matrix $\Omega \in \Re^{K\times m}$ optimizing the Fisher criterion satisfies the following relationship

$$(S_w^{-1}S_b)(\Omega B) = (\Omega B)\Delta \tag{2.10}$$

where B is defined in (B.19). Equation (2.10) shows that the elements of matrix Δ and the columns of matrix ΩB correspond to the first m eigenvalues and eigenvectors respectively of matrix $S_w^{-1}S_b$.

Since matrix S_b is the sum of C independent matrices of rank 1, $C-1$ of which are independent, then S_b is at most of rank $C-1$.[14] Thus, matrix $S_w^{-1}S_b$ has at most rank $C-1$, where C is the number of classes. Therefore matrix $S_w^{-1}S_b$ has $C-1$ nonzero eigenvalues. Based on this and considering (2.10) we can determine the dimension of the transformed space, we rewrite J using the fact that the trace of a matrix is equal to the sum of the eigenvalues, i.e.

$$J = \mathrm{tr}\{S_w^{-1}S_b\} = \underbrace{\sum_{i=1}^{C-1}\lambda_i}_{\neq 0} + \underbrace{\sum_{j=C}^{K}\lambda_j}_{=0} \tag{2.11}$$

From (B.20) and (B.21) we have

$$J = \mathrm{tr}\{BS_w'^{-1}S_b'B^{-1}\} = \mathrm{tr}\{S_w'^{-1}S_b'\} = \underbrace{\sum_{i=1}^{C-1}\lambda_i}_{\neq 0} + \underbrace{\sum_{j=C}^{K}\lambda_j}_{=0} \tag{2.12}$$

Then if we choose $m = C - 1$, the value of J in the space $\Phi(X)$ of dimension K is the same to that obtained using the minimum number of dimensions (i.e. without zero eigenvalues).

2.3. *Relationship between the general and constrained solution of NLFT problem*

In Sec. 2.1 we solved the problem

$$Z^*(X) = \max_{Z\in\mathbb{C}^2}\{J(Z(X))\} \tag{2.13}$$

over all $Z(X) \in \mathbb{C}^2$. However, the solution depends explicitly on *a posteriori* probability density of each class. To avoid this difficulty in Sec. 2.2 we restrict the solutions to functions $Z(X)$ whose components $Z_i(X)$ can be expressed as

$$Z_i(X) = \sum_{j=1}^{K}\alpha_i^j\varphi_j(X) \tag{2.14}$$

where $\{\varphi_j(X)\}_{j=1}^K$ is a set of basis functions each one defined on $\Re^n \rightarrow \Re$. Thus, (2.13) can be written as

$$\begin{aligned} &Z^*(X) = \max_{Z\in\mathbb{C}^2}\{J(Z(X))\} \\ &\text{subject to} \\ &Z(X) = \Omega^T\Phi(X) \end{aligned} \tag{2.15}$$

Clearly for $\Phi(X)$ fixed, (2.15) is equivalent to

$$Z^*(X) = \max_{\Omega}\{J(\Omega^T\Phi(X))\} \tag{2.16}$$

which is a parametric optimization problem.

The solution obtained applying this constraint to calculus of variations problems is known as Ritz approximation[13] and was proposed in 1908 by Ritz based on the previous work by Lord Rayleigh. The main property of the solution using the Ritz approximation is that the ith component of the solution corresponds to the projection of the ith component of the general solution into the space generated by the basis functions $\{\varphi_i(X)\}_{i=1}^K$.[23] Based on this property, it is clear that what we are doing by considering nonlinear extensions of the Fisher transformation, employing non linear transformations of data of the form (2.14), is approximating the general solution (2.6), through a linear combination of the functions defining the transformation. That is to say, the coefficients maximizing the Fisher index satisfy

$$\Omega = R_{\varphi\varphi}^{-1} R_{\varphi Z} \tag{2.17}$$

where $R_{\varphi\varphi}$ corresponds to the auto-correlation matrix of functions $\varphi_i(X)$ and $R_{\varphi Z}$ is the cross-correlation matrix between $\varphi_i(X)$ and $Z_i^*(X)$ where $Z_i^*(X)$ corresponds to the ith component of the general solution problem.

3. Nonlinear Fisher Transformation in High Dimensional Spaces

In general the scatter matrices are of high dimension. In image processing, there exist some techniques to face the matrix inversion problem of S_w and to handle the number of computations associated to these matrices.[7] Almost all of them are based on a procedure that combines PCA and LDA.[2] Basically these procedures consider projection matrices of the form

$$A = A_{\text{LDA}} \cdot A_{\text{PCA}} \tag{3.1}$$

PCA is used to project the original data onto a subspace with the aim of decreasing the pattern dimension and where matrix S_w is nonsingular, so that the computation of eigenvectors of $S_w^{-1}\ S_b$ can be easily done. Although these techniques allow obtaining a solution, in the first projection performed by PCA some directions of the originals space, containing relevant information for classification purposes, can be disregarded. In fact, Chen in 2000 proved that the null subspace of S_w contains the information with the most discriminatory power.[7] Then by using PCA and getting a nonsingular S_w in the projected space, we are eliminating the null subspace of S_w in the original space and therefore this type of algorithm is not optimal.

Hua Yu and Jie Yang[38] proposed a different solution known as Direct LDA that does not eliminate the null space of S_w and that will be used in the NLFT context as a tool to diminish the amount of computations. This method is called Inverse Simultaneous Diagonalization (ISD) and it is explained in Sec. 3.1.

3.1. *Inverse simultaneous diagonalization*

The main idea is to use the property that the matrix whose columns are the eigenvectors of $S_w^{-1}\ S_b$ is the same that allows the simultaneous diagonalization of S_w and S_b, i.e. if we find a matrix A such that

$$A^T S_w A = D \tag{3.2}$$

and

$$A^T S_b A = I \tag{3.3}$$

with D a diagonal matrix, then A is the matrix formed by the eigenvectors of $S_w^{-1}\ S_b$. Next we present the procedure to find the transformation matrix A without inverting matrix S_w.

(1) Diagonalization of S_b:
To find a matrix $V \in \Re^{n\times n}$ such that

$$V^T S_b V = \Delta \tag{3.4}$$

where $V^T V = I$ and Δ is a diagonal matrix whose elements are ordered in decreasing order. Matrix V can be found using eigenvalue–eigenvector computation, i.e. V is formed by the eigenvectors of S_b and Δ contains the eigenvalues of S_b on its diagonal. Since S_b can be singular some of the eigenvalues can be zero. It is necessary to eliminate these zero eigenvalues and the corresponding eigenvectors since the scatter between classes along these directions are zero, and contain no discrimination power. Since the range of S_b is $C-1$, where C is the number of classes, then there exist $C-1$ nonzero eigenvalues.

Let Y be the $C-1$ first columns of V, ($Y \in \Re^{n\times(C-1)}$) then we can write

$$Y^T S_b Y = D_b > 0 \tag{3.5}$$

where D_b corresponds to the main submatrix of $(C-1)\times(C-1)$ of matrix Δ and it is a positive definite diagonal matrix, without zero elements on its diagonal.

(2) Let Z be defined as $Z = YD_b^{1/2}$ with $Z \in \Re^{n\times(C-1)}$. Clearly

$$(YD_b^{1/2})^T S_b (YD_b^{1/2}) = I_{(C-1)} \Rightarrow Z^T\ S_b\ Z = I_{(C-1)} \tag{3.6}$$

Matrix Z diagonalizes S_b and reduces its dimension from $n\times n$ to $(C-1)\times(C-1)$.

(3) Diagonalization of $Z^T\ S_w\ Z$.
To find a matrix $U \in \Re^{(C-1)\times(C-1)}$such that

$$U^T Z^T S_w Z U = D_w \tag{3.7}$$

with $U^T U = 1$. Again it is possible to find $D_w \in \Re^{(C-1)\times(C-1)}$ and U through an eigenvalue eigenvector analysis of matrix $Z^T S_w Z$. Notice that D_w can contain zero elements on its diagonal.

Matrix Z diagonalizes S_b and reduces its dimension from $n\times n$ to $(C-1)\times(C-1)$.

(4) Let A be defined as $A = U^T Z^T$, with $A \in R^{(C-1)\times n}$. Then matrix A simultaneously diagonalizes S_b and S_w and reduces its dimension to $(C-1)$. A corresponds to the matrix formed by the eigenvectors associated with the $(C-1)$ nonzero eigenvalues of $S_w^{-1}S_b$, i.e. the solution for the linear optimization of Fisher criterion.

3.2. *Analysis of eigenvalues–eigenvectors in high dimensional spaces*

As shown in Sec. 3.1, using ISD it is possible to compute the matrix transformation even in the case when S_b is not invertible. Although simultaneous inverse diagonalization reduce computations as compared with the traditional approach, in the first stage it is necessary to perform an eigenvalue–eigenvector analysis of the between-class scatter matrix S_b, which can be of a very high dimension. This analysis has to be explicitly done in the space generated by $\Phi(X)$, to explicitly compute the nonlinear transformation $\Phi(X)$. Thus we can use the structure of the scatter matrices to reduce the amount of computations of the eigenvalue–eigenvector analysis.

To this extent we will use the method stated by Fukunaga in Ref. 14, by Kirby and Sirovich in Ref. 20 and by Turk and Penland in Ref. 35, to efficiently compute the eigenvalues and eigenvectors of matrices S_b and $Z^T S_b\, Z$. The method exploits the fact that scatter matrices can be expressed as the product of a matrix and its transpose, i.e.

$$S_b = \sum_{i=1}^{C} P(w_i)(\mu_i - \mu_0)(\mu_i - \mu_0)^T = \Psi_b \Psi_b^T \tag{3.8}$$

with

$$\Psi_b = [\sqrt{P(w_1)}(\mu_1 - \mu_0), \sqrt{P(w_2)}(\mu_2 - \mu_0), \ldots, \sqrt{P(w_C)}(\mu_C - \mu_0)] \tag{3.9}$$

Notice that Ψ_b is an nx C matrix, where n is the pattern size and C is the number of classes. We now state the following Lemma due to Turke and Penland.[35]

Lemma 3.1. *Let L be any $(n \times m)$ matrix. Then the function $V = Lv$ is a one-to-one mapping from the eigenvalues of $L^T L \in \Re^{m\times m}$ to the eigenvectors of $LL^T \in \Re^{n\times n}$.*

Proof. See Ref. 35 for the proof. □

We can directly use Lemma 3.1 in the diagonalization of S_b in (3.4) by considering $L = \Psi_b$. Since $\Psi_b \in \Re^{n\times(C-1)}, \Psi_b^T \Psi_b$ has $(C-1)$ eigenvectors, the same we need to compute for matrix Y in (3.5). The main advantage of computing the eigenvectors through Lemma 3.1 is that it allows direct computation of the C eigenvectors associated with the C nonzero eigenvalues of matrix S_b and not the n eigenvalues of it. Usually the number of classes is around dozens whereas the pattern size can be of the order of hundred. Thus, the saving in computations can be significant.

To compute the eigenvectors of matrix $Z^T S_w Z$ in (3.7) through Lemma 3.1 we write $Z^T S_w Z$ as the product of a matrix and its transpose. To this extent we first rewrite S_w in the form

$$S_w = \Psi_w \Psi_w^T \tag{3.10}$$

where $\Psi_w \in \Re^{n\times(C-1)}$ is defined as

$$\Psi_w = [\psi_1, \psi_2, \ldots, \psi_C] \tag{3.11}$$

and $\psi_i \in \Re^n$ is given by

$$\psi_i = \sqrt{P(w_i)} \sum_{X\in w_i} (X - \mu_i) \tag{3.12}$$

Then we can write

$$Z^T S_w Z = Z^T \Psi_w \Psi_w^T Z = \left(Z^T \Psi_w\right)\left(Z^T \Psi_w\right)^T \tag{3.13}$$

Under these conditions we can use Lemma 3.1 with $L = Z^T \Psi_w$. However, there still exists the problem of computing $\Phi(X)$. If a large number of basis functions are considered, a large computational effort is needed in computing $\Phi(X)$. In the next section we analyze the particular case of linear and quadratic terms only.

3.3. *Quadratic fisher transformation*

In what follows, a new feature extraction method is proposed, based on the previous developments, which use linear and quadratic basis functions. The method will be called Quadratic Fisher Discriminant Analysis (QFDA).

If we considerer the optimization problem (2.1) and choose a transformation of the form

$$Z(X) = \begin{bmatrix} A_1^T X + X^T B_1 X \\ A_2^T X + X^T B_2 X \\ \cdot \\ A_m^T X + X^T B_m X \end{bmatrix}, \quad Z \in \Re^m,\ \ X \in \Re^n,\ \ A_i \in \Re^n,\ \ B_i \in \Re^{n\times n} \tag{3.14}$$

which can be written as

$$Z(X) = \Omega^T \Phi(X) \tag{3.15}$$

where

$$\Phi(X) = \begin{bmatrix} x_1^2 \\ x_1 x_2 \\ \vdots \\ x_1 x_n \\ x_2^2 \\ x_2 x_1 \\ \vdots \\ x_2 x_n \\ \vdots \\ x_n x_1 \\ x_n x_2 \\ \vdots \\ x_n^2 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \Re^{n^2+n} \tag{3.16}$$

$\Phi(X)$ is a vector of dimension $n^2 + n$ and $\Omega \in \Re^{m\times(n^2+n)}$. The dimension of $\Phi(X)$can be reduced considering only one cross term and not all of them, imposing extra conditions on matrices B_i, i.e. $B_i = B_i^T$. Since $\Phi\,(X)$dimension is high we can use the method stated in Sec. 3.1 to solve the Fisher optimization problem. Moreover, we can use the method in Sec. 3.2 to perform the eigenvalue–eigenvector analysis.

Notice that in QFDA the dimension of $\Phi(X)$ increases as $n^2, (n = \dim(X))$, so computation problem can become complex if the dimension of X is high.

3.4. *Summary of the proposed method*

Our objective is to find matrix Ω that maximizes the Fisher criterion in the transformed space using the transformation given by (3.8). We have proved that this matrix satisfies Eq. (3.10). If we consider, for example a practical problem with feature vectors of dimension $n = 64$, matrix Ω will be of dimension $K \times m$, with $K = n^2 + n = 4160$ and $m = C - 1$, where C is the number of classes.

In order to solve Eq. (3.10) for Ω, the following algorithm can be used to reduce the amount of computations.

(a) Use the ISD procedure presented in Sec. 3.1 to find the eigenvalues and eigenvectors of matrix $S_w^{-1}\ S_b$.

(b) Use the information on eigenvalues and eigenvectors of $S_w^{-1}\ S_b$ to find matrix ΩB from Eq. (3.10) and then find Ω, since B is nonsingular and of low dimension.

(c) In the ISD procedure of step (a) it will be necessary to compute the eigenvalues and eigenvectors of high dimension matrices (S_b and $Z^T\ S_w\ Z$). To reduce the amount of work use Lemma 3.1 given in Sec. 3.2.

3.5. *Proposed method and kernel Fisher multiclasses*

In what follows we will apply the proposed method to solve kernel Fisher multiclasses and we will see that it converges to the standard kernel Fisher formulation for the two-class problems.

In the two-class problems the transformation between the $\Phi\,(X)$ space and the original feature space is of dimension $(m \times 1)$ where m corresponds to the dimension of vectors in $\Phi(X)$. From kernel theory we deduce that if ω_j is the jth column of transformation Ω, then we can write it in the form

$$\omega_j = \sum_{i=1}^{N} \alpha_i \Phi(X_i) \tag{3.17}$$

i.e. vector ω_j belongs to the subspace spanned by the nonlinear transformation of the training examples X_i.[25] Then matrix Ω corresponds to a vector defined as

$$\Omega = \sum_{i=1}^{M} \alpha \Phi(X_i) \tag{3.18}$$

It is easy to see that matrix Ω can be written as

$$\Omega = \left[\sum_{i=1}^{N} \alpha_i^1 \Phi(X_1) \cdots \sum_{i=1}^{N} \alpha_i^N \Phi(X_N) \right] \tag{3.19}$$

Using (3.19) and the definition of μ_i we have:

$$\Omega^T \mu_i = \begin{bmatrix} \frac{1}{N_i} \sum_{j=1}^{N} \sum_{m=1}^{N_i} \alpha_j^1 K(X_j, X_m^i) \\ \vdots \\ \frac{1}{N_i} \sum_{j=1}^{N} \sum_{m=1}^{N_i} \alpha_j^N K(X_j, X_m^i) \end{bmatrix} = \begin{bmatrix} \alpha_1^T M_i \\ \alpha_2^T M_i \\ \vdots \\ \alpha_n^T M_i \end{bmatrix} = \alpha^T M_i \tag{3.20}$$

where

$$(M)_i = \frac{1}{N} \sum_{j=1}^{N} K(X_j, X_i) \tag{3.21}$$

$$(M_i)_j = \frac{1}{N} \sum_{k=1}^{N} K(X_j, X_k^i) \tag{3.22}$$

Then we can write

$$\Omega^T S_b \Omega = \alpha^T Q \alpha \tag{3.23}$$

with $\Omega \in \Re^{m \times 1}$ and

$$Q = \sum_{i=1}^{2} (M_i - M)(M_i - M)^T \tag{3.24}$$

Similarly for S_w we have

$$\Omega^T S_w \Omega = \alpha^T R \alpha \tag{3.25}$$

where

$$R = K_1(I - 1_{N_1})K_1^T + K_2(I - 1_{N_2})K_2^T \tag{3.26}$$

$$(K_i)_{n,m} = K(X_n, X_m^i) \tag{3.27}$$

$(K_i)_{n,m}$ is known as the Kernel Matrix of class i, I denotes the identity matrix and 1_{N_i} denotes a matrix whose elements are $1/N_i$. Then the Fisher index in the transformed space can be written as

$$\mathrm{tr}\{(\alpha^T R \alpha)^{-1}(\alpha^T Q \alpha)\} \tag{3.28}$$

Matrices R and Q obtained by this procedure correspond to matrices R and Q stated by Mika in his first work on kernel Fisher for two-class problems.[25]

In Sec. 4 QFDA is used and compared with other classical methods when solving different pattern recognition problems.

4. Experiments and Comparisons of NLFT with Classical Methods

Throughout this paper it has been emphasized the necessity of extending the Fisher Transform (FT) or Fisher Discriminant Analysis (FDA) to the nonlinear case, motivated by the fact that FDA does not work properly when the mean of the classes do not coincide or when the essential information for classification lies on data variance rather than in the mean. Furthermore there exists some previous work[25] where the classification has significantly improved by using kernel Fisher. For this reason in this section a series of experiments are presented where the objective is to evaluate the usefulness of QFDA in real pattern recognition problems.

4.1. *Brief description of data sets used*

In what fallows a brief description of data sets used in this study for comparison purposes is presented. All of them correspond to real data and they are considered standard data sets in the literature.[4] The first five of them are part of the database repository of University of California at Irvine (UCI) and the last one is part of the benchmark repository of Carnegie Mellon University (CMU).

(i) Wisconsin Breast Cancer Database (WBCD): This is one of the three breast cancer data set repository of UC Irving. The information was collected at the Wisconsin University by W. U. Wolberg.[37] The problem is to predict from a patient tumor tissue if this is malign benign. The data set has two classes, nine attributes and 742 observations. Since 16 observations presented no attributes they were discarded, using then only 716 observations. From these, 485 examples (65.5%) correspond to Class 1 (malign) and the remaining 241 patterns (34.5%) belong to Class 2 (benign).

(ii) PIMA Indian Diabetes Database (PIMAIDD): This data set contains information from women older than 21 years descending from Pima tribe, living in the Phoenix Arizona area.[33] The problem is to predict if the patient presents diabetes based on medical and psychological exams. In this case there are 14 properties and 768 observations. Class 1 (positive) contains 500 examples (65.1%), whereas Class 2 (negative) has 268 patterns (34.9%).

(iii) Thyroid Disease Database (TDD): Here the problem consists in determining if a patient presents thyroid disorders based on medical exams. In the data set there are three classes (normal, hyperthyroidism and hypothyroidism). This problem presents 21 features and it is organized in 3772 observations for training and 3428 for validation. In the simulations performed in this paper both data sets were put together since the evaluation of the methods was done using 10 fold cross-validation.

(iv) Ionosphere Data base (ID): This data set corresponds to measures of the ionosphere radar echo.[32] The problem consists in determining if a radar signal was able to capture the ionosphere structure or if this signal does not contain information about ionosphere. Data was acquired in Goose Bay, Labrador USA, using a 16-array of high frequency radars transmitting a total power of 6.4 KW. The data set belongs to John Hopkins University. The problem has two classes, 34 characteristics and 341 measurements.

(v) StatLog Vehicle Silhouette Database (SVSD): This data set belongs to UCI database repository and was developed by the Turing Institute in Glasgow, Scotland. The problem consists in predicting the type of a vehicle based on geometric attributes of the vehicle silhouette obtained from image processing. The vehicles included are autobus, Chevrolet Van, Saab 9000 and Opel Manta 400. The problem has four classes, 18 attributes and 846 patterns.

(vi) Sonar Database (SD): This data set belongs to CMU benchmark repository. The problem consists in determining if the object is a stone or a mine from the information contained in the power spectrum of sonar signals. The data set has 208 examples, 60 properties and two classes.

4.2. *Methodology used in simulations*

Since the objective of this section is to study the advantages of the proposed feature extraction method, all of the data sets already described will be classified using the Quadratic Fisher Transformation (QFDA) in the feature extraction stage, previous to the classification stage by LDA. Each experiment will be performed three times; first without using any feature extraction method, then using linear Fisher extraction (FDA) and finally employing QFDA. For the first case (no feature extraction) the dimension of the input vectors corresponds to the number of original attributes mentioned in Sec. 4.1 for each database, whereas in the other two cases (FDA and QFDA) the dimension is reduced to $C - 1$ where C is the number of classes of the dataset, also indicated in Sec. 4.1. The classification rate will be estimated through cross-validation with ten subsets (ten fold cross-validation) and to measure significant differences in the methods the McNemar Test of Hypothesis[12] will be used.

Since the purpose of the study is to compare the behavior of the classification varying the feature extraction method, the same classification method will be used in all simulations. To this extent the simplest statistical classifier, linear discriminant analysis (LDA),[34] will be used to realize the effects of the feature extraction method being used. LDA consists of applying the maximum *a posteriori* (MAP) rule assuming normal data distribution, as well as that each class has the same covariance matrix. Using these two assumptions the MAP rule is simplified as:

To assign the pattern X to class w_i if and only if

$$C_i \geq C_j \quad \forall i \neq j \tag{4.1}$$

where

$$C_k = 2X^T \sum^{-1} \mu_k + \mu_k \sum^{-1} \mu_k - 2\log(P(w_k)) \tag{4.2}$$

Σ corresponds to the data covariance matrix, μ_k is the mean of class w_k and $P(w_k)$ is *a priori* probability of class w_k. Assuming that all classes have the same covariance matrix, LDA is a simplification of the MAP rule for normal distribution that in pattern recognition area is known as quadratic discriminant analysis (QDA).[34] The structure of the linear classification implies that decision boundaries are hyperplanes. Thus, this algorithm can only classify correctly problems where data is linearly separable.

4.3. *Simulations and results*

Simulation results obtained by applying the three methods on the six databases described in Sec. 4.1 are presented in what follows.

(i) Simulation results using the Wisconsin Breast Cancer Database

In what follows the results using the Wisconsin Breast Cancer Database (WBCD) are presented, using cross-validation with ten sets. Table 1 shows the correct classification rates using three feature extraction methods. First no feature extraction is used i.e. LDA is directly applied to data. Next, FDA is used as feature extraction and finally QFDA is used a feature extraction method.

From Table 1 we can deduce that feature extraction positively affects the classifier behavior and QFDA introduces an improvement of 4% in the classification rate if compared with the cases with no feature extraction or when using FDA.

In Table 2 presents the p-value of McNemar Test of Hypothesis[12] of the statistical significance of the three methods studied. We recall that in this Test, the classification rate of two classifiers is statically different with a 95% of certainty, if the p-value of the Test is greater than 3.84. From Table 2 it is observed that no significant differences between the first two schemes (without feature extraction and FDA) whereas there is a significant difference in the classification rates when QFDA is considered a feature extraction method.

Table 1. Classification rates for WBCD.

Method	Classification Rate	Standard Deviation
LDA	0.93	0.010
FDA + LDA	0.93	0.008
QFDA + LDA	0.97	0.003

Table 2. McNemar Test of Hypothesis for WBCD.

Method	LDA	FDA + LDA	QFDA + LDA
LDA		0.8	13.1
FDA + LDA	0.8		12.5
QFDA + LDA	13.1	12.5	

Table 3. Confusion matrices for WBCD.

	LDA		FDA + LDA		QFDA + LDA	
	Classified as		Classified as		Classified as	
	Cancer	No Cancer	Cancer	No Cancer	Cancer	No Cancer
Cancer Sample	0.88	0.12	0.88	0.12	0.95	0.05
No Cancer Sample	0.02	0.98	0.02	0.98	0.01	0.99

Table 4. Classification rate for PIMAIDD.

Method	Classification Rate	Standard Deviation
LDA	0.77	0.003
FDA + LDA	0.78	0.001
QFDA + LDA	0.81	0.001

Table 3, shows the confusion matrices for each method studied (without feature extraction, FDA and QFDA). As seen from Table 3, QFDA significantly decreases the number of confusions.

(ii) Simulation results using PIMA Indian Diabetes Database
Table 4 presents the results of the three classification methods studied, when applied to the PIMA Indian Diabetes Database (PIMAIDD). From Table 4 we can conclude that the use of a feature extraction method does positively affect the behavior of the classifier, diminishing at least the variance of the classification rate which can be attributed to a better determination of classifier parameters when operating in lower dimension spaces. On the other hand, we can see that QFDA gives an improvement of 3% in the classification rate if compared with FDA

Table 5 summarizes the p-value of McNemar Test of Hypothesis[12] over the significance in the differences of the classification rates. From Table 5, we conclude that there exist significant differences between the three classification schemes.

Table 6 slows the confusion matrices for the three cases studied (without extraction, FDA and QFDA). From this table we can conclude that QFDA improved the classification rate and significantly diminished the number of confusions in at least one class.

(iii) Simulation results using Thyroid Disease Database
Table 7 the classification results of the three methods studied when using the Thyroid Disease Database (TDD) are presented. A noticeable improvement is observed in the classification rate when using QFDA of about 30%.

Table 5. McNemar Test of Hypothesis for PIMAIDD.

Meted	LDA	FDA + LDA	QFDA + LDA
LDA		5.14	12.01
FDA + LDA	5.14		6.96
QFDA + LDA	12.01	6.96	

Table 6. Confusion matrices for PIMAIDD.

	LDA		FDA + LDA		QFDA + LDA	
	Classified as		Classified as		Classified as	
	Diabetes	No Diabetes	Diabetes	No Diabetes	Diabetes	No Diabetes
Sample with Diabetes	0.88	0.12	0.89	0.11	0.91	0.09
Sample without Diabetes	0.43	0.57	0.42	0.58	0.39	0.61

Table 7. Classification rate for TDD.

Method	Classification Rate	Standard Deviation
LDA	0.50	0.06
FDA + LDA	0.51	0.07
QFDA + LDA	0.81	0.03

Table 8. McNemar Test de Hypothesis for TDD.

Method	LDA	FDA+LDA	QFDA+LDA
LDA		0.5	8.3
FDA + LDA	0.5		7.9
QFDA + LDA	8.3	7.9	

Table 8 shows the p-values of McNeman Test of Hypothesis[12] from which we can state that there are significant differences between QFDA and the two other methods, and there is no significant difference between FDA and the case when no feature extraction is used

The confusion matrices for all three cases studied are presented in Table 9. From this table we can see that QFDA significantly diminished the number of confusions. In this experiment QFDA allowed an improvement in the classification rate of 30% (see Table 7) diminishing mainly the confusion between classes 2 and 3 (hyperthyroidism and hypothyroidism) as seen in Table 9.

In Fig. 1 is plotted the projection of the original data through the linear Fisher transformation.

Figure 2 shows the projection of the original data through the quadratic Fisher transformation. Since this problem has three classes ($C = 3$), the transformed data belongs to $\Re^2$, i.e. $(C - 1)$. From Figs. 1 and 2, an improvement can be seen in the class separation using QFDA, making the classification problem simpler.

(iv) Simulation results using Ionosphere Database

In this section we present the classification results of the three methods analyzed when applied to the Ionosphere Database (ID), using cross-validation with ten sets. Table 10 summarizes the classification rates for the three cases studied, where it can be seen that QFDA gives 100% of correct classification, improving in almost 14% the classification obtained with the other two methods.

Table 9. Confusion matrices for TDD.

	LDA			LDA + FDA			LDA + QFDA		
	Classified as			Classified as			Classified as		
	Normal	Hyperthyroidism	Hypothyroidism	Normal	Hyperthyroidism	Hypothyroidism	Normal	Hyperthyroidism	Hypothyroidism
Normal	0.51	0.08	0.41	0.53	0.10	0.37	0.88	0.04	0.078
Hyperthyroidism	0.02	0.01	0.97	0.02	0.01	0.97	0.0	0.56	0.44
Hypothyroidism	0.005	0.005	0.99	0.0	0	1.0	0.0	0.0	1.0

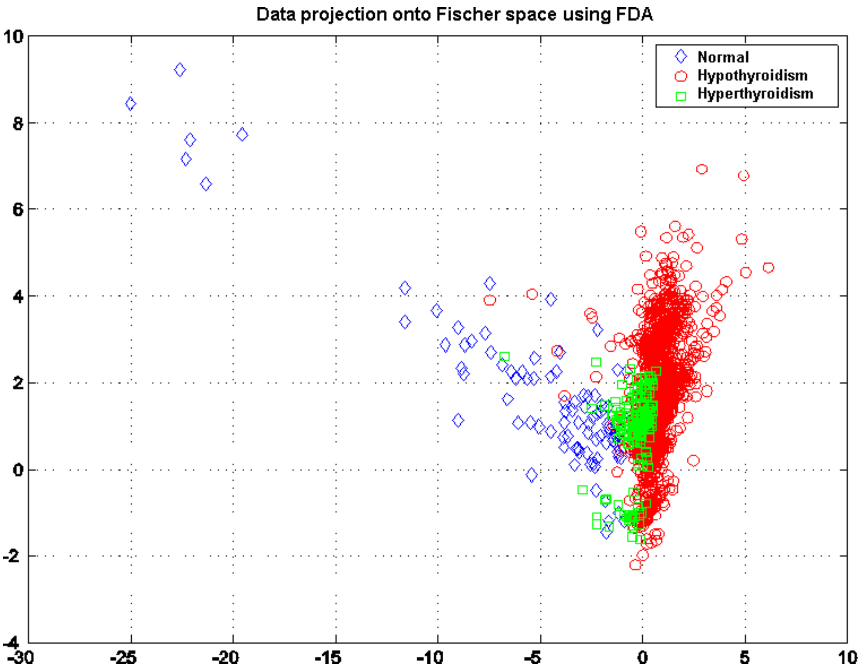

Fig. 1. Data projection onto Fisher space using FDA for TDD.

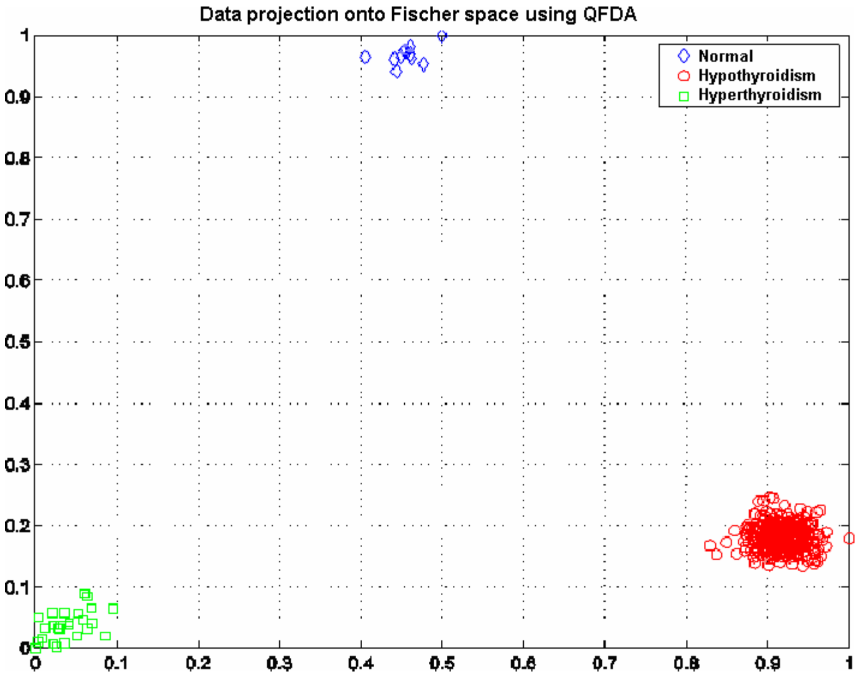

Fig. 2. Data projection onto Fisher space using QFDA for TDD.

Table 10. Classification rate for ID.

Method	Classification Rate	Standard Deviation
LDA	0.81	0.002
FDA + LDA	0.86	0.002
QFDA + LDA	1	0.000

Table 11. McNemar Test of Hypothesis for ID.

Method	LDA	FDA + LDA	QFDA + LDA
LDA		5.2	24.1
FDA + LDA	5.2		18.6
QFDA + LDA	24.1	18.6	

Table 12. Confusion matrices for ID.

	LDA		LDA + FDA		LDA + QFDA	
	Classified as		Classified as		Classified as	
	Information	Noise	Information	Noise	Information	Noise
Information signal	0.66	0.34	0.76	0.24	1	0
Noise signal	0.03	0.97	0.03	0.97	0	1

Table 13. Classification rates for SVSD.

Method	Classification Rate	Standard Deviation
LDA	0.77	0.06
FDA +LDA	0.78	0.003
QFDA + LDA	0.96	0.005

Table 14. McNemar Test of Hypothesis for SVSD.

Method	LDA	FDA + LDA	QFDA + LDA
LDA		5.7	12.8
FDA + LDA	5.7		11.5
QFDA + LDA	12.8	11.5	

Table 11 the p-value of McNemar Test of Hypothesis[12] is presented in the three cases studied. They indicate that the three methods are statically different.

Table 12 presents the confusion matrices for the three methods compared. It can be seen that QFDA improved the classification rate by significantly diminishing the number of confusions

(v) Simulation results using the Statlog Vehicle Silhouette database

Table 13 presents the classification rate of the three methods studied when applied to Statlog Vehicle Silhouette Database (SVSD) using cross-validation with ten sets.

Table 14 shows the p-values of McNemar Test of Hypothesis.[12] From this we can conclude that all three methods have significant differences from the statistical point of view.

Table 15 presents the confusion matrices for the three cases analyzed.

Table 15. Confusion matrices for SVSD.

	LDA				LDA + FDA				LDA + QFDA			
	Classified as				Classified as				Classified as			
	Opel	Saab	Van	Bus	Opel	Saab	Van	Bus	Opel	Saab	Van	Bus
Opel	0.6000	0.0368	0.3211	0.0421	0.6368	0.0316	0.2947	0.0368	0.9474	0.0053	0.0474	0
Saab	0.0169	0.9605	0.0113	0.0113	0.0169	0.9661	0.0056	0.0113	0.0113	0.9661	0.0226	0
Van	0.3122	0.0582	0.5820	0.0476	0.3016	0.0529	0.5979	0.0476	0.0212	0.0106	0.9683	0
Bus	0.0204	0.0102	0.0102	0.9592	0.0204	0.0102	0.0102	0.9592	0.0051	0	0.0051	0.9898

From Tables 13–15, it is concluded that QFDA improved the classification rates in about 18% if compared with FDA and 19% if compared with the case when no extraction method is used. Also the number of confusions in the system is noticeably diminished when using QFDA.

In Fig. 3 is plotted the original data projection when using FDA and Fig. 4 the original data projection when using QFDA. It can be seen that QFDA produces a subspace with better class separation than FDA.

(vi) Simulation results using Sonar Database

Table 16 shows the classification rates of the three methods studied when applied to Sonar Data Set (SD) and using cross-validation with ten sets.

A 100% of correct classification is achieved with QFDA, improving the results obtained with FDA and without extraction methods in 10% and 25% respectively. In Table 17, the p-values of McNemar Test of hypothesis[12] is summarized for the three methods and we can conclude that they are statistically different.

Table 18 presents the confusion matrices for the three methods studied where the reduction can be seen in the number of confusion when using QFDA as compared with the other two methods.

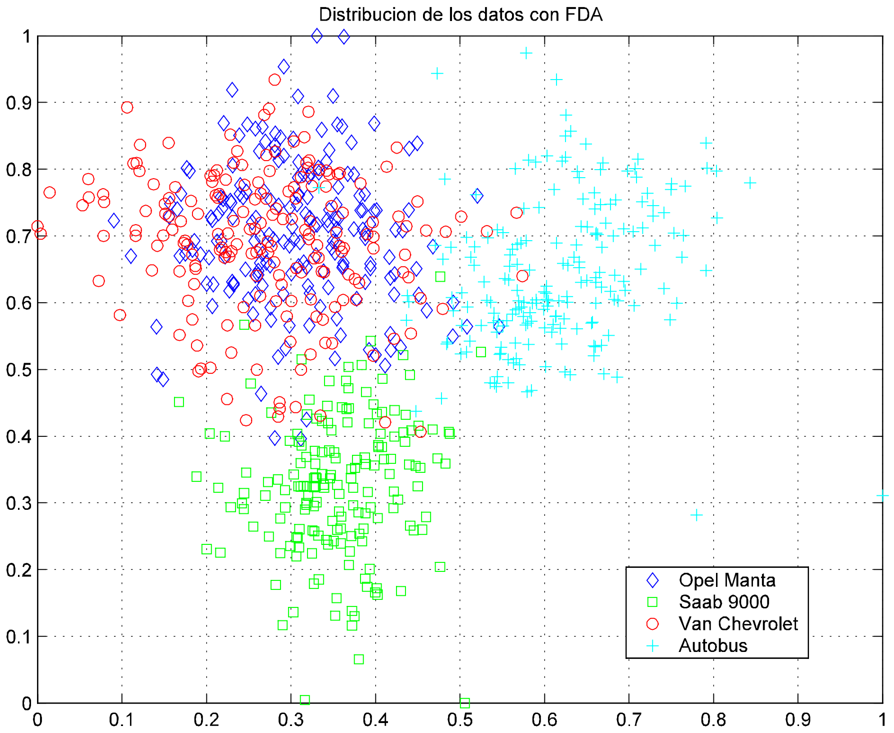

Fig. 3. Data projection onto Fisher space when using FDA for SVSD.

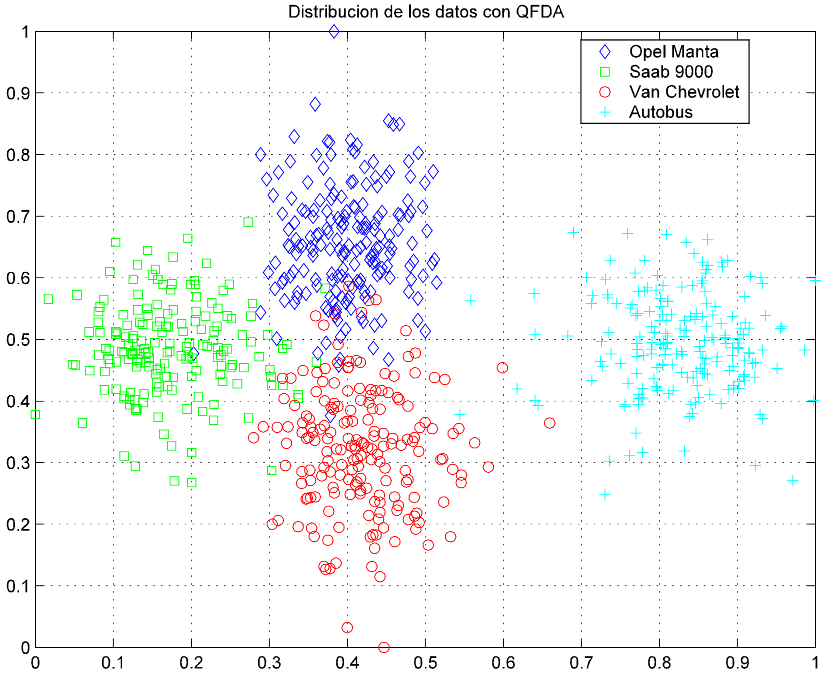

Fig. 4. Data projection onto Fisher space when using QFDA for SVSD.

Table 16. Classification rates for SD.

Method	Classification Rate	Standard Deviation
LDA	0.75	0.002
FDA + LDA	0.90	0.005
QFDA + LDA	1.00	0.001

Table 17. McNemar Test of Hypothesis for SD.

Method	LDA	FDA + LDA	QFDA + LDA
LDA		9.03	19.01
FDA + LDA	9.03		8.05
QFDA + LDA	19.01	8.05	

Table 18. Confusion matrices for SD.

	LDA		LDA + FDA		LDA + QFDA	
	Classified as		Classified as		Classified as	
	Stone	Mine	Stone	Mine	Stone	Mine
Stone	0.7216	0.2784	0.8763	0.1237	1.0	0.0
Mine	0.2162	0.7838	0.0721	0.9279	0.0	1.0

5. Conclusions

In this paper the optimization of the Fisher criterion in a space nonlinearly related to the original data was studied. First, the problem was solved using calculus of variations in the function space, concluding that although it is possible to solve the problem, its solution need to know *a posteriori* probability density that a vector (pattern) X belongs to each class. This implies that the results cannot be used in real pattern recognition problems since probability densities are, in general, unknown.

As a way of avoiding this explicit dependence of the solution on probability densities, the solution (transformation) was restricted to functions that can be written as a linear combination of basis functions. By restricting the solutions to this type of functions, it is possible to solve the problem without using the knowledge of the probability densities, obtaining a closed-form analytical solution. Thus, the obtained solution corresponds to the projection of each component of the general solution in the space function of continuous second derivatives, onto the space generated by the functions $\varphi_i(X)$, components of $\Phi(X)$.

Although the solution found to the nonlinear optimization of the Fisher criterion does not depend on the class probability densities (form and parameters), the computational procedure associated to this solution can be quite demanding. A procedure and Lemma were presented in order to substantially diminish the computational load of the proposed solution and to make its implementation in real problems simpler.

Simulation results presented using six standard data sets in pattern recognition literature showed that QFDA significantly improved the classification rates in all six cases and diminished the number of confusions in the system.

As future work it is proposed to use wavelet decomposition instead of a Taylor Series to approximate the components of the general optimization solution of the Fisher criterion. Due to the wavelets property of approximating functions with lower number of coefficients, the computational load associated with the method will diminish. Besides, the potential advantages in dimension reduction, due to the particular form of the Haar wavelet will probably reduce the amount of computations importantly.

Acknowledgments

The results presented in this work were supported by CONYCIT-Chile, under the grant FONDEF D01-1016, “Chilean Red Wine Classification by means of Intelligent Instrumentation”

Appendix A.

The first variation of J, given by (2.1), due to a change δZ in Z is given by[13]

$$\delta J = J[Z(X) + \delta Z(X)] - J[Z(X)] \tag{A.1}$$

Using a Taylor series expansion of (2.1) together with (A.1) we get:

$$\delta J = \sum_{i=1}^{C} \left[\frac{\partial J}{\partial \tilde{\mu}_i} \delta \tilde{\mu}_i + \mathrm{tr} \left\{ \frac{\partial J}{\partial \tilde{\Sigma}_i} \delta \tilde{\Sigma}_i \right\} \right] + O(\delta^3) \tag{A.2}$$

From definitions (2.4) and (2.5) we can compute the variations in $\tilde{\mu}_i$ and $\tilde{\Sigma}_i$ when Z is changed in δZ, obtaining,

$$\delta \tilde{\mu}_i = \int_{-\infty}^{\infty} \delta Z(X) p(X/w_i) dX \tag{A.3}$$

$$\delta \tilde{\Sigma}_i = \int_{-\infty}^{\infty} \left(\delta Z(X) Z(X)^T + Z(X) \delta Z(X)^T \right) p(X/w_i) dX \tag{A.4}$$

Neglecting the high order terms in (A.2) and using relations (A.3) and (A.4) we have

$$\begin{aligned} \delta J = \sum_{i=1}^{C} \Bigg[& \frac{\partial J}{\partial \tilde{\mu}_i} \int_{-\infty}^{\infty} \delta Z(X) \cdot p(X/w_i) dX \\ & + 2 \cdot \mathrm{tr} \left\{ \frac{\partial J}{\partial \tilde{\Sigma}_i} \int_{-\infty}^{\infty} \delta Z(X) Z(X)^T p(X/w_i) dX \right\} \Bigg] \end{aligned} \tag{A.5}$$

Permuting the trace and integral functions and using the property that $\mathrm{tr}(AVU^T) = V^T A^T U$ where $V, U \in \Re^n$ and $A \in \Re^{n \times n}$[5] the second term of (A.5) can be written as

$$2 \int_{-\infty}^{\infty} \delta Z(X)^T \left[\frac{\partial J}{\partial \tilde{\Sigma}_i} \right]^T Z(X) p(X/w_i) dX \tag{A.6}$$

Thus, (A.5) can be written as

$$\delta J = \int_{-\infty}^{\infty} \delta Z(X)^T \sum_{i=1}^{C} \left[\frac{\partial J}{\partial \tilde{\mu}_i} + 2 \frac{\partial J}{\partial \tilde{\Sigma}_i} Z(X) \right] \cdot p(X/w_i) dX \tag{A.7}$$

To find the extreme of (A.7) the following must be satisfied[9]

$$\delta J = 0, \quad \forall \delta Z \tag{A.8}$$

Thus, any $Z(X)$ maximizing (2.1) should satisfy

$$2 \sum_{i=1}^{C} \left[p(X/w_i) \frac{\partial J}{\partial \tilde{\Sigma}_i} \right] Z(X) = - \sum_{i=1}^{C} \left[p(X/w_i) \frac{\partial J}{\partial \tilde{\mu}_i} \right] \tag{A.9}$$

Using definitions (2.1)–(2.3), we see that J depends on $\tilde{\Sigma}_i$ only through $\tilde{S}_w$ and does not explicitly depend on $\tilde{\Sigma}_i$. This allows computing $\partial J / \partial \tilde{\Sigma}_i$, as follows

$$\frac{\partial J}{\partial \tilde{\Sigma}_i} = \frac{\partial J}{\partial \tilde{S}_w} \cdot \frac{\partial \tilde{S}_w}{\partial \tilde{\Sigma}_i} \tag{A.10}$$

From (2.2) the partial derivative $\partial \tilde{S}_w / \partial \tilde{\Sigma}_i$ can be computed as

$$\frac{\partial \tilde{S}_w}{\partial \tilde{\Sigma}_i} = \frac{\partial}{\partial \tilde{\Sigma}_i} \left[\sum_{j=1}^{C} P(w_j) \cdot \tilde{\Sigma}_j \right] = P(w_i) \tag{A.11}$$

Then (A.10) becomes

$$\frac{\partial J}{\partial \tilde{\Sigma}_i} = P(w_i) \cdot \frac{\partial J}{\partial \tilde{S}_w} \tag{A.12}$$

Thus, condition (A.9) can be expressed as

$$2\sum_{i=1}^{C} [p(X/w_i)P(w_i)] \frac{\partial J}{\partial \tilde{S}_w} Z(X) = -\sum_{i=1}^{C} \left[p(X/w_i) \frac{\partial J}{\partial \tilde{\mu}_i} \right] \tag{A.13}$$

Applying the total probability theorem[28] we recognize that the term

$$\sum_{i=1}^{C} P(w_i)p(X/w_i) = p(X) \tag{A.14}$$

corresponds to the total probability density of X. Thus, we can write (A.13) as

$$2p(X) \frac{\partial J}{\partial \tilde{S}_w} Z(X) = -\sum_{i=1}^{C} \left[p(X/w_i) \frac{\partial J}{\partial \tilde{\mu}_i} \right] \tag{A.15}$$

If we define

$$\hat{p}(X/w_i) = \frac{P(X)p(X/w_i)}{p(X)} \tag{A.16}$$

and

$$\frac{\partial J'}{\partial \tilde{\mu}_i} = \frac{1}{P(X)} \frac{\partial J}{\partial \tilde{\mu}_i} \tag{A.17}$$

we can express (A.15) in the following form

$$2\frac{\partial J}{\partial \tilde{S}_w} Z(X) = -\sum_{i=1}^{C} \left[\hat{p}(X/w_i) \frac{\partial J'}{\partial \tilde{\mu}_i} \right] \tag{A.18}$$

Appendix B.

Let us consider (2.8) under the conditions given in Sec. 2.2. The first variation of $Z(X)$ for fixed $\Phi(X)$ is defined as

$$\delta Z = \delta \Omega^T \Phi(X) \tag{B.1}$$

Replacing (2.8) and (B.1) in (A.7) we have

$$\delta J = \int_{-\infty}^{\infty} \Phi^T(X) \delta\, \Omega \sum_{i=1}^{C} \left[\frac{\partial J}{\partial \tilde{\mu}_i} + 2\frac{\partial J}{\partial \tilde{\Sigma}_i} \Omega^T \Phi(X) \right] p(X|w_i)dX \tag{B.2}$$

Integrating and factorizing (B.2) we get

$$\delta J = \mathrm{tr}\left\{ \delta\Omega \sum_{i=1}^{C} \left[\frac{\partial J}{\partial \tilde{\mu}_i} \mu_i^T + 2\frac{\partial J}{\partial \tilde{\Sigma}_i} \Omega^T \Sigma_i \right] \right\} \tag{B.3}$$

Imposing the extreme condition on J

$$\delta J = 0, \quad \forall \delta Z \tag{B.4}$$

we get

$$2\sum_{i=1}^{C} \frac{\partial J}{\partial \tilde{\Sigma}_i} \Omega^T \Sigma_i = -\sum_{i=1}^{C} \frac{\partial J}{\partial \tilde{\mu}_i} \mu_i^T \tag{B.5}$$

Using the same argument as in (A.10), the partial derivative can $\partial J/\partial \tilde{\Sigma}_i$ be computed as (see (A.12))

$$\frac{\partial J}{\partial \tilde{\Sigma}_i} = P(w_i) \frac{\partial J}{\partial \tilde{S}_w} \tag{B.6}$$

On the other hand $\partial J/\partial \tilde{S}_w$ can be expressed as

$$\frac{\partial J}{\partial \tilde{S}_w} = \frac{\partial \left(\operatorname{tr} \left\{ \tilde{S}_w^{-1} \tilde{S}_b \right\} \right)}{\partial \tilde{S}_w} = -\tilde{S}_w^{-1} \tilde{S}_b \tilde{S}_w^{-1} \tag{B.7}$$

Replacing (B.7) in (B.6) we get

$$\frac{\partial J}{\partial \tilde{\Sigma}_i} = -P(w_i) \tilde{S}_w^{-1} \tilde{S}_b \tilde{S}_w^{-1} \tag{B.8}$$

Similarly, we can compute $\partial J/\partial \tilde{\mu}_i$ as

$$\frac{\partial J}{\partial \tilde{\mu}_i} = \frac{\partial J}{\partial \tilde{S}_b} \frac{\partial \tilde{S}_b}{\partial \tilde{\mu}_i} \tag{B.9}$$

where

$$\frac{\partial J}{\partial \tilde{S}_b} = \frac{\partial \left(\operatorname{tr} \left\{ \tilde{S}_w^{-1} \tilde{S}_b \right\} \right)}{\partial \tilde{S}_b} = \tilde{S}_w^{-1} \tag{B.10}$$

and from (2.3)

$$\begin{aligned} \frac{\partial \tilde{S}_b}{\partial \tilde{\mu}_i} &= \frac{\partial \left(\sum_{i=1}^{C} P(w_i)(\tilde{\mu}_i - \tilde{\mu}_0)(\tilde{\mu}_i - \tilde{\mu}_0)^T \right)}{\partial \tilde{\mu}_i} && \text{(B.11)} \\ &= \frac{\partial \left(\sum_{i=1}^{C} P(w_i)(\tilde{\mu}_i \tilde{\mu}_i^T - \tilde{\mu}_0 \tilde{\mu}_i^T - \tilde{\mu}_i \tilde{\mu}_0^T + \tilde{\mu}_0 \tilde{\mu}_0^T) \right)}{\partial \tilde{\mu}_i} && \text{(B.12)} \\ &= 2P(w_i)(\tilde{\mu}_i - \tilde{\mu}_0) && \text{(B.13)} \end{aligned}$$

Then, replacing (B.10) and (B.13) in (B.9) we have

$$\frac{\partial J}{\partial \tilde{\mu}_i} = 2P(w_i) \tilde{S}_w^{-1} (\tilde{\mu}_i - \tilde{\mu}_0) \tag{B.14}$$

Replacing (B.8) and (B.14) in (B.5), the extreme condition can be written as

$$\sum_{i=1}^{C} P(w_i)\tilde{S}_w^{-1}\tilde{S}_b\tilde{S}_w^{-1}\Omega^T\Sigma_i = \sum_{i=1}^{C}\tilde{S}_w^{-1}P(w_i)(\tilde{\mu}_i - \tilde{\mu}_0)\mu_i^T \tag{B.15}$$

$$\tilde{S}_b\tilde{S}_w^{-1}\sum_{i=1}^{C}\Omega^T P(w_i)\Sigma_i = \Omega^T\sum_{i=1}^{C}P(w_i)(\mu_i - \mu_0)(\mu_i - \mu_0)^T \tag{B.16}$$

$$\tilde{S}_b\tilde{S}_w^{-1}\Omega^T S_w = \Omega^T S_b \tag{B.17}$$

Since S_b and S_w are symmetric matrices (B.17) can be expressed as

$$(S_w^{-1}S_b)\Omega = \Omega(\tilde{S}_w^{-1}\tilde{S}_b) \tag{B.18}$$

Since the Fisher index is invariant under nonsingular transformation, we can use the Simultaneous Matrix Diagonalization Lemma[14] in the transformed space to transform equation (B.18) into an eigenvalue eigenvector equation without altering the solution.

Let us consider the change of variable

$$Y' = B^T Y(X) \tag{B.19}$$

where $Y' \in \Re^m$, $B \in \Re^{m\times m}$ is a nonsingular matrix and $Y \in \Re^m$. The scatter matrices in the space Y' are defined as

$$\tilde{S}'_w = B^T\tilde{S}_w B = I_m \tag{B.20}$$

$$\tilde{S}'_b = B^T\tilde{S}_b B = \Delta \tag{B.21}$$

where I_m denotes the $(m\times m)$ identity matrix and Δ is an $(m\times m)$ diagonal matrix containing the eigenvalues of $\tilde{S}_b$. Since $\tilde{S}'_w$ is the identity, it is easy to observe that the elements of the diagonal matrix Δ correspond to the m eigenvalues of $\tilde{S'}_w^{-1}\tilde{S'}_b$ ordered in descending order. After the change of variable, Eq. (B.18) can be expressed as

$$(S_w^{-1}S_b)(\Omega B) = (\Omega B)(\tilde{S'}_w^{-1}\tilde{S}'_b) \tag{B.22}$$

Replacing (B.20) and (B.21) in (B.22) we get

$$(S_w^{-1}S_b)(\Omega B) = (\Omega B)\Delta \tag{B.23}$$

References

1. G. Baudat and F. Anouar, Generalized discriminant analysis using a kernel approach, *Neural Comput.* **10** (2000) 2385–2404.
2. P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, Eigenfaces v/s Fisherface: Recognition using class specific linear projection, *IEEE Trans. PAMI* **7** (1997) 711–720.
3. R. E. Bellman, *Dynamic Programming* (Princeton University Press, 1957).
4. C. Blake and C. Merz, UCI repository of machine learning databases, Technical Report University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

5. H. Brunzell and J. Eriksson, Feature reduction for classification of multidimensional data, *Patt. Recogn.* **33** (2000) 1741–1748.
6. N. A. Campbell, Canonical variate analysis- a general model formulation, *Aust. J. Stat.* **26** (1984) 86–96.
7. L. Chen, H. Liao, M. Ko, J. Lin and G. Yu, A new LDA based face recognition system which can solve the small size problem, *Patt. Recogn.* **33** (2002) 1713–1726.
8. W. S. Chen, P. C. Yuen, J. Huang and B. Fang, Two-step single parameter regularization Fisher discriminant method for face recognition, *Int. J. Patt. Recogn. Artif. Intell.* **20** (2006) 189–207.
9. G. M. Ewing, *Calculus of Variations with Applications* (Dover Publications, New York, 1985).
10. R. A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. Eugen.* **7** (1936) 111–132.
11. R. A. Fisher, The statistical utilization of multiple measurements, *Ann. Eugen.* **8** (1938) 376–386.
12. J. L. Fleiss, *Statistical Methods for Rates and Proportions.* Second Edition (Wiley, 1981).
13. M. J. Forray, *Variational Calculus in Science and Engineering* (McGraw-Hill, 1968).
14. K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Second Edition (Academic Press, New York, 1990).
15. T. J. Hastie, A. Buja and R. Tibshirani, Penalized discriminant analysis by optimal score, AT&T Bell labs Technical Report, 1993.
16. T. J. Hastie and R. Tibshirani, Discriminant analysis by Gaussian mixtures, AT&T Bell Labs Technical Report, 1994.
17. T. J. Hastie and R. Tibshirani, Flexible discriminant analysis by optimal scoring, AT&T Bell Labs Technical Report, 1996.
18. T. J. Hastie, R. Tibshirani and A. Buja, Flexible discriminant and mixture models, in *Proc. Neural Networks and Statistics*, eds. J. Kay and D. Titterington, (Oxford University Press, 1995).
19. L. B. Holder, I. Russell, Z. Markov, A. G. Pipe and B. Carse, Current and future trends in feature selection and extraction for classification problems, *Int. J. Patt. Recogn. Artif. Intell.* **19** (2005) 133–142.
20. M. Kirby and M. Sirovich, Application of the Karhunen–Loeve procedure for the characterization of human faces, *IEEE Trans. PAMI* **12** (1990) 103–108.
21. N. Kumar and A. G. Andreou, On generalization of linear discriminant analysis, Technical Report, JHU/ECE-9607, Johns Hopkins University, 1996.
22. R. S. Lin and L. H. Chen, A new approach for audio classification and segmentation using Gabor wavelets and Fisher linear discriminator, *Int. J. Patt. Recog. Artif. Intell.* **19** (2005) 807–822.
23. C. Magers, Least squares approach to the Rayliegh–Ritz method, *Proc. Louisiana-Mississippi Section of the Mathematical Association of America*, 2001.
24. S. Mika, G. Ratsch and K. R. Muller, A mathematical programming approach to the Kernel Fisher algorithm, in *Adv. Neural Infor. Process. Syst.* **13** (2001) 591–598.
25. S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K. R. Muller, Fisher discriminant analysis with kernels, *Proc. IEEE Neural Networks for Signal Processing Workshop*, Madison, USA, 1999, pp. 41–48.
26. S. Mika, A. J. Smola and B. Scholkopf, An improved training algorithm for kernel Fisher discriminants, *Proc. AISTATS 2001*, 2001, pp. 98–104.
27. P. Navarrete and J. Ruiz-del-Solar, On the generalization of Kernel machines, *Int. Workshop on Pattern Recognition with Support Vector Machines — SVM2002*, Niagara Falls, Canada, 2002.

28. A. Papoulis, *Probability and Statistics* (Pearson Education, 1990).
29. W. L. Poston and D. J Marchete, Recursive dimensionality reduction using Fisher discriminant analysis, *Patt. Recogn.* **31** (1998) 881–888.
30. C. R. Rao, The utilization of multiple measurements in problems of biological classification, *J. Roy. Stat. Soc.* **10** (1948) 159–203.
31. B. Scholkopf, A. J. Smola and K. R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Comput.* **10** (1998) 1299–1319.
32. V. G. Sigillito, S. P. Wing, L. V. Hutton and K. B. Baker, Classification of radar returns from the ionosphere using neural networks, *Johns Hopkins APL Techn. Dig.* **10** (1989) 262–266.
33. J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler and R. S. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, *IEEE Proc. Symp. Computer Applications and Medical Care*, 1988, pp. 261–265.
34. S. Theodoridis and K. Koutroumbas, *Pattern Recognition* (Academic Press, New York, 1999).
35. M. Turk and A. Penland, Eigenfaces for recognition, *J. Cogn. Neurosci.* **3** (1991) 71–86.
36. T. Van Gestel, J. A. K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor and J. Vandewalle, Bayesian framework for least squares support vector machine classifiers, Gaussian processes and kernel Fisher discriminant analysis, *Neural Comput.* **14** (2002) 1115–1147.
37. W. H. Wolberg and O. L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, *Proc. Nat. Acad. Sci.* **87** (1990) 9193–9196.
38. H. Yu and J. Yang, A direct LDA algorithm for high-dimensional data — with application to face recognition, *Patt. Recog.* **34** (2001) 2067–2070.
39. F. Zhang, Nonlinear feature extraction and dimension reduction by polygonal principal curves, *Int. J. Patt. Recogn. Artif. Int.* **20** (2006) 63–78.

Matías A. Bustos was born in Santiago, Chile in 1977. He received his B.Sc. and M.Sc in Electrical Engineering from the Universidad de Chile in 2004. Currently he is working as senior research engineer in the Services and Technology division of the Mining and Metallurgical Research Center of Chile.

He is involved in the development of computer vision and image processing algorithms used in soft sensor applications for mining processes.

Manuel A. Duarte-Mermoud received the degree of Civil Electrical Engineer from the University of Chile in 1977 and the M.Sc., M.Phil. and the Ph.D. degrees, all in electrical engineering, from Yale University in 1985, 1986 and 1988 respectively.

From 1977 to 1979, he worked as Field Engineer at Santiago Subway. In 1979 he joined the Electrical Engineering Department of University of Chile, where he is currently Professor.

His main research interests are in robust adaptive control (linear and nonlinear systems), system identification, signal processing and pattern recognition. He is focused on applications to mining and wine industry, sensory systems and electrical machines and drives.

Dr. Duarte is senior member of the IEEE and IFAC. He is past Treasurer and past President of ACCA, the Chilean National Member Organization of IFAC, and past Vice-President of the IEEE-Chile.

Nicolás H. Beltrán got his Electrical Engineering degree at the University of Chile where he graduated in 1974. He earned both his Master of Electrical Engineering in 1981 and his Doctoral degree in Applied Sciences in 1985, from the Katholieke Universiteit Leuven (KUL), Belgium. He was with the Venezuelan Institute for Scientific Research (IVIC) in Caracas, Venezuela from 1975 to 1979.

Since 1985 he is with the Electrical Engineering Department of the University of Chile where he has conducted research in his areas of interest, ceramic sensors and intelligent instrumentation. Dr. Beltran is senior member of IEEE.