Dimensionality Reduction, Classification and Reconstruction Problems in Statistical Learning Approaches

Gilson A. Giraldi [1]
Paulo S. Rodrigues [2]
Edson C. Kitani [3]
Carlos E. Thomaz [4]

Abstract:
Statistical learning theory explores ways of estimating functional dependency from a given collection of data. The specific sub-area of supervised statistical learning covers important models like Perceptron, Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA). In this paper we review the theory of such models and compare their separating hypersurfaces for extracting group-differences between samples. Classification and reconstruction are the main goals of this comparison. We show recent advances in this topic of research illustrating their application on face and medical image databases.

1 Introduction

The main goal of this paper is to present advanced aspects in supervised statistical learning for image analysis. The basic pipeline to follow in this subject is: (a) Dimensionality reduction; (b) Choose a learning method to compute a separating hypersurface, that is, to solve the classification problem; (c) Reconstruction problem, that means, to consider how good a low dimensional representation might look like. We show recent advances in each one of these steps illustrating their application on face and medical image databases.

Statistical learning plays a key role in many areas of science and technology. This field explores ways of estimating functional dependency from a given collection of data [14]. It covers important topics in classical statistics such as discriminant analysis, regression methods, and the density estimation problem [15, 11, 22]. Moreover, pattern recognition and classification can be seen from the viewpoint of the general statistical problem of function estimation

[1]Department of Computer Science, LNCC, Petrópolis, Rio de Janeiro, Brazil
{gilson@lncc.br}
[2]Department of Computer Science, FEI, São Bernardo do Campo, São Paulo, Brazil
{psergio@fei.edu.br}
[3]Department of Electrical Engineering, USP, São Paulo, São Paulo, Brazil
{ekitani@lsi.usp.br}
[4]Department of Electrical Engineering, FEI, São Bernardo do Campo, São Paulo, Brazil
{cet@fei.edu.br}

from empirical data. Therefore, such research topics fit very well in the statistical learning framework [9, 10].

In image databases it is straightforward to consider each data point (image) as a point in a n-dimensional space, where n is the number of pixels of each image. Therefore, dimensionality reduction may be necessary in order to discard redundancy and simplify further computational operations. The most known technique in this subject is the Principal Components Analysis (PCA) [10]. However, other criteria have been proposed contrary to the idea of selecting the principal components with the largest eigenvalues for dimensionality reduction [4, 17]. We review these works theoretically and present some investigations that we have done in this area. Specifically, if we navigate on each principal component we observe that the most expressive components for reconstruction (i.e., the components with the largest eigenvalues) are not necessarily the most discriminant ones for classification. This can be seen as a special kind of the reconstruction problem and will be discussed using face images [19, 31, 18].

In the classification problem, we try to distinguish separated subsets (classes) and find an approach to automatically label data points according to their corresponding classes [9]. Such goal can be achieved through the separating hyperplanes generated by supervised statistical learning methods like Perceptron, SVM and LDA [32, 14, 13]. In the recent years, these methods have played an important role for characterizing differences between a reference group of patterns using image samples of patients [24, 26, 27, 28, 12] as well as face images [5, 19, 25, 21]. Besides, their extensions for the non-linear case, as well as the Maximum uncertainty LDA (MLDA) approach to address the limited sample size problem, have been reported in a number of works in the literature [32, 14, 13, 12, 24, 19, 31, 26, 28, 18]. An important point here is which learning method to use. We review the theory behind the cited methods, their common points, and discuss why SVM is in general the best technique for classification but not necessarily the best for extracting discriminant information. A case study of face recognition [31, 19, 25] will be used in order to help this discussion [20]. Moreover, such analysis points towards the idea of using the discriminant weights given by separating hyperplanes to select the most important features of a data set for classification. Another case study of breast lesion classification in ultrasound images will be used in order to help this discussion [20].

In what follows, we review in section 2 principal components analysis. Next, in section 3, we describe the Perceptron, SVM and LDA statistical learning approaches. Then, in section 4, we consider the classification and reconstruction problems from the viewpoint of SVM and LDA methods. Next, the experimental results used to help the discussion of the paper are presented, with two case studies: face image database analysis (section 5) and breast cancer classification (section 6). Finally, in section 7, we conclude the paper, summarizing its main contributions and describing possible future works.

2 Principal Components Analysis

In this section we review results in principal components analysis (PCA). When applying PCA for dimensionality reduction before classification, it has been common practice to use components with the largest eigenvalues. Such idea can be justified if we clarify the main principles of the PCA analysis. For completeness, we will start this section reviewing firstly some points of this theory.

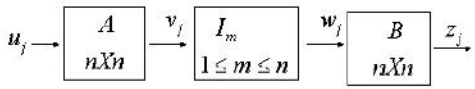


Figure 1. KL Transform formulation. Reprinted from [16].

Principal Component Analysis, also called Karhunen-Loeve or KL method, can be seen as a method for data compression or dimensionality reduction [1] (see also [16], section 5.11). Thus, let us suppose that the data to be compressed consist of N measurements or samples, from a n-dimensional space. Then, PCA searches for k n-dimensional orthonormal vectors that can best represent the data, where $k \leq n$. Thus, let $U = \{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N\}$ be the data set. By now, let us suppose that the centroid of the data set is the center of the coordinate system,that is,

$$C_M = \frac{1}{N}\sum_{i=1}^{N} \mathbf{u}_i = \mathbf{0}. \tag{1}$$

To address the issue of compression, we need a projection basis that satisfies a proper optimization criterium. Following [16], consider the operations in Figure 1. The vector $\mathbf{u}_j$ is first transformed to a vector $\mathbf{v}_j$ by the transformation matrix A. Thus, we truncate $\mathbf{v}_j$ by choosing the first m elements of $\mathbf{v}_j$. The obtained vector $\mathbf{w}_j$ is just the transformation of $\mathbf{v}_j$ by I_m, that is a matrix with '1s' along the first m diagonal elements and zeros elsewhere. Finally, $\mathbf{w}_j$ is transformed to $\mathbf{z}_j$ by the matrix B. Let the square error be defined as follows:

$$J_m = \frac{1}{N}\sum_{j=0}^{N} \|\mathbf{u}_j - \mathbf{z}_j\|^2 = \frac{1}{n} Tr\left[\sum_{j=0}^{N} (\mathbf{u}_j - \mathbf{z}_j)(\mathbf{u}_j - \mathbf{z}_j)^{*T}\right], \tag{2}$$

where Tr means the trace of the matrix between the square brackets and the notation $(*T)$ means the transpose of the complex conjugate of a matrix. Following Figure 1, we observe that $\mathbf{z}_j = BI_mA\mathbf{u}_j$. Thus we can rewrite (2) as:

$$J_m = \frac{1}{N} Tr \left[\sum_{i=0}^{N} \left(\mathbf{u}_j - BI_m A\mathbf{u}_j \right) \left(\mathbf{u}_j - BI_m A\mathbf{u}_j \right)^{*T} \right], \tag{3}$$

which yields

$$J_m = \frac{1}{N} Tr \left[\left(I - BI_m A \right) R \left(I - BI_m A \right)^{*T} \right], \tag{4}$$

where

$$S = \sum_{i=0}^{N} \mathbf{u}_j \mathbf{u}_j^{*T}. \tag{5}$$

Following the literature, we call S the covariance matrix. We can now state the optimization problem by saying that we want to find the matrices A, B that minimizes J_m. The next theorem gives the solution for this problem.

Theorem 1: The error J_m in expression (4) is minimum when

$$A = \Phi^{*T}, \quad B = \Phi, \quad AB = BA = I, \tag{6}$$

where Φ is the matrix obtained by the orthonormalized eigenvectors of S arranged according to the decreasing order of its eigenvalues. *Proof.* See [16].

Therefore, from this result, if our aim is data compression than we must choose the components with the largest eigenvalues. However, we can change the *information measure* in order to get another criterium. This is performed in [4] by using the following idea. Let $\mathbf{u}$ be a n-dimensional random vector with a mixture of two normal distributions with means μ_1 and μ_2, mixing proportions of p and $(1-p)$, respectively, and a common covariance matrix Σ. Let Δ denote the Mahalanobis distance between the two sub-populations, and S be the covariance matrix of $\mathbf{u}$. Then, it can be shown that [4]:

$$S = p(1-p)\mathbf{d}\mathbf{d}^T + \Sigma, \tag{7}$$

where $\mathbf{d} = \mu_1 - \mu_2$.

In [4], instead of the optimization criterium of choosing the components that minimizes J_m, it is assumed that the effectiveness of using a set of variables can be measured by the Mahalanobis distance computed on the basis of these variables. This distance will be called the information contained in the variables.

So, let $a_i, i = 0, 1, ..., n-1$ be the eigenvectors of S with eigenvalues $\lambda_0, \lambda_1, , ..., \lambda_{n-1}$, not necessarily sorted in decreasing order. For a given $B_m = (a_1, a_2, ..., a_m)$, $m \leq n$ we

denote Δ_m the distance between the sub-populations using $B_m\mathbf{u}$. For instance, for $m = 1$, we can write:

$$a_1^T S a_1 = p(1-p) a_1^T \mathbf{d}\mathbf{d}^T a_1 + a_1^T \Sigma a_1. \tag{8}$$

Therefore, using the fact that $Sa_1 = \lambda_1 a_1$, the expression (8) becomes:

$$\lambda_1 = p(1-p)\left(a_1^T \mathbf{d}\right)^2 + a_1^T \Sigma a_1. \tag{9}$$

So, we find that:

$$a_1^T \Sigma a_1 = \lambda_1 - p(1-p)\left(a_1^T \mathbf{d}\right)^2. \tag{10}$$

But, the Mahalanobis distance computed in the variable defined by the component a_1 is:

$$\Delta_1^2 = \mathbf{d}^T a_1 \left(a_1^T \Sigma a_1\right) a_1^T \mathbf{d} = \frac{\left(a_1^T \mathbf{d}\right)^2}{a_1^T \Sigma a_1}, \tag{11}$$

which can be rewritten as:

$$\Delta_1^2 = \frac{\left(a_1^T \mathbf{d}\right)^2}{\lambda_1 - p(1-p)\left(a_1^T \mathbf{d}\right)^2}, \tag{12}$$

by using the equation (10). Finally, if we divide the numerator and denominator of equation (12) by λ_1 we obtain:

$$\Delta_1^2 = \frac{\frac{\left(a_1^T \mathbf{d}\right)^2}{\lambda_1}}{1 - p(1-p)\frac{\left(a_1^T \mathbf{d}\right)^2}{\lambda_1}}. \tag{13}$$

This result can be generalized for $m > 1$ by the following result [4]:

$$\Delta_m = \frac{\sum_{i=0}^{n-1} \frac{\left(a_i^T \mathbf{d}\right)^2}{\lambda_i}}{1 - p(1-p)\sum_{i=0}^{n-1} \frac{\left(a_i^T \mathbf{d}\right)^2}{\lambda_i}}. \tag{14}$$

By now, the following consequence may be drawn: The component with the largest amount of separating information between two sub-populations is not necessarily the one with the largest eigenvalue. This is because Δ_i is a monotonic function of $\left(a_i^T \mathbf{d}\right)^2 / \lambda_i$ instead of λ_i (see Δ_1 in equation (13)). The best subset of m principal components is the one with the largest Δ_m.

The relationship between this criterium and the last one that selects principal components in decreasing order of eigenvalues is obtained by demonstrating that the information is distributed in m (or less than m) principal components if Σ has m distinct eigenvalues. In other words, the use of traditional PCA is justified when the information is concentrated in a few principal components with a large sample size, but such most expressive components do not necessarily represent the most discriminant information between sample groups as we will illustrate in section 5.

3 Statistical Learning Models

In this section we discuss some aspects os statistical learning theory related to the Support Vector Machine (**SVM**), Perceptron and Linear Discriminate Analysis (**LDA**) methods [32, 2]. The goal is to set a framework for further comparisons. The material to be presented follows the references [32, 14, 2].

Statistical learning theory explores ways of estimating functional dependency from a given collection of data. It covers important topics in classical statistics such as discriminant analysis, regression methods, and the density estimation problem [15, 11, 22]. Statistical learning is a kind of statistical inference (also called inductive statistics). It encompasses a rigorous qualitative/quantitative theory to set the necessary conditions for consistency and convergence of the learning process as well as principles and methods based on this theory for estimating functions, from a small collection of data [14, 32].

Pattern recognition and classification problems belongs to the general statistical problem of function estimation from empirical data. Therefore, they fit very well in the statistical learning framework [9, 10].

Statistical inference has more than 200 years, including names like Gauss and Laplace. However, the systematic analysis of this field started only in the late 1920s. By that time, an important question to be investigated was finding a reliable method of inference, that means, to solve the problem: *Given a collection of empirical data originating from some functional dependency, infer this dependency* [32].

Therefore, the analysis of methods of statistical inference began with the remarkable works of Fisher (parametric statistics) and the theoretical results of Glivenko and Cantelli (convergence of the empirical distribution to the actual one) and Kolmogorov (the asynp-

totically rate of that convergence). These events determined two approaches to statistical inference: The particular (*parametric*) inference and the *general* inference [32].

The parametric inference aims to create statistical methods for solving particular problems. Regression analysis is a know technique in this class. On the other hand, the general inference aims to find one induction method that can be applied for any statistical inference problem. Learning machines, like Perceptron [2] and SVM [32, 3], and the LDA approach are nice examples in this area. They are discussed next.

3.1 Perceptron Model

The Perceptron is the first logical neuron. Its development starts with the work of W. S. McCulloch and W.A. Pitts in 1943 [2]. It describes the fundamentals functions and structures of a neural cell reporting that a neuron will fire an impulse only if a threshold value is exceeded.

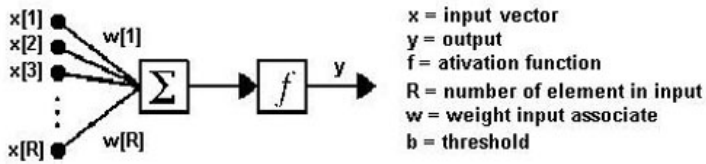


Figure 2. McCulloch-Pitts neuron model.

Figure 2 shows the basic elements of McCulloch-Pitts model: x is the input vector, w are weights associated, y is output, R is number of elements in input and f is the activation function, named *decision function* in statistical learning theory, that determines the value in output. A simple choice for f is the signal function $sgn(.)$. In this case, if the sum, across all the inputs with its respective weights exceeds the threshold b the output is 1 else the value of y is -1, that is:

$$y = sgn(\sum_{i=1}^{R} w_i x_i - b). \tag{15}$$

But the McCulloch-Pitts neuron did not have a mechanisms for *learning*. Based on biological evidences, D.O. Hebb suggested a rule to adapt the weights input, which is interpreted as learning rule for the system [2]. This biological inspired procedure can be expressed in the following manner:

$$w_i^{new} = w_i^{old} + \Delta w_i; \quad \Delta w_i = \eta(y^{desired} - y)x_i, \tag{16}$$

where w^{new} and w^{old} are adapted weights and initials weights respectively, η is a real parameter to control the rate of learning and $y^{desired}$ is the desired (know) output. This *learning rule* plus the elements of Figure 2 is called the perceptron model for a neuron. It was proposed by F. Rosenblatt, at the end of of 1950s.

Then, the learning typically occurs through training, or exposure to a know set of input/output data. The training algorithm iteratively adjusts the connection weights $\{w_i\}$ analogous to synapses in biological nervous. These connection weights store the knowledge necessary to solve specific problems.

Geometrically, the connection weights $\{w_i\}$ and the the threshold b define a plane in a high dimensional space that separates the samples of distinct groups. Such *separating hyperplane* can be further used for classification on a new input vector x. Therefore, the learning process means to be able to adjust initial weights towards the separating hyperplane. Besides, it can be demonstrated that, if we can choose a small margin δ, such that:

$$\sum_{i=1}^{R} w_i x_i - b > \delta, \quad if \quad y = 1, \tag{17}$$

$$\sum_{i=1}^{R} w_i x_i - b < -\delta, \quad if \quad y = -1, \tag{18}$$

then, the number of times, T, that the rule defined by expression (16) is applied (number of iteractions) is bounded by:

$$T \leq \frac{1}{\delta^2}. \tag{19}$$

More precise bounds can be found in [32].

3.2 Support Vector Machines

In section 3.1 it becomes clear the importance of separating hyperplanes methods for learning algorithms. In this section we will present a special type of separating hyperplanes with optimality properties. So, given a training set:

$$S = \{(y_1, x_1), ..., (y_m, x_m)\}, \quad x \in \Re^n, \quad y \in \{-1, 1\}, \tag{20}$$

we say that the subset I for which $y = 1$ and the subset II for which $y = -1$ are separable by the hyperplane:

$$x * \phi = c, \tag{21}$$

if there exists both a unit vector ϕ ($|\phi| = 1$) and a constant c such that the inequalities:

$$x_i * \phi > c, \quad x_i \in I, \tag{22}$$

$$x_j * \phi < c, \quad x_j \in II, \tag{23}$$

hold true ("$*$" denotes the usual inner product in $\Re^m$). Besides, let us define for any unit vector ϕ the two values:

$$c_1(\phi) = \min_{x_i \in I} (x_i * \phi), \tag{24}$$

$$c_2(\phi) = \max_{x_j \in II} (x_j * \phi). \tag{25}$$

The Figure 3 represents the dataset and the hyperplanes defined by ϕ and the values c_1, c_2 defined in expressions (24)-(25):

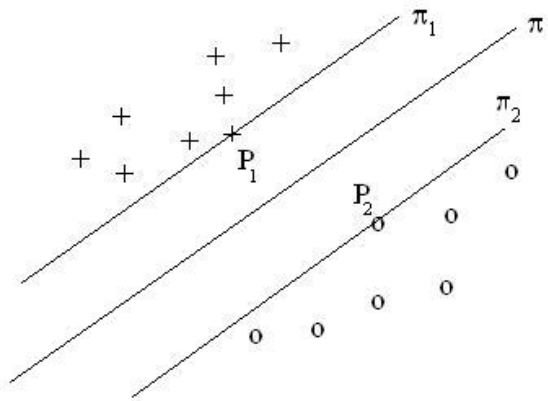


Figure 3. Separating hyperplane π and its offsets π_1, π_2.

In this figure the points P_1 and P_2 gives the solutions of problems (24)-(25), respectively, and the planes π_1 and π_2 are defined by:

$$x * \phi = c_1, \tag{26}$$

$$x * \phi = c_2. \tag{27}$$

Now, le us consider the plane π, parallel to π_1,π_2, with the property:

$$d_\pi\left(P_1\right) = d_\pi\left(P_2\right), \tag{28}$$

where $d_\pi\left(P\right)$ means the Euclidean distance from a point P to a plane π. This plane is the hyperplane that separates the subsets with maximal marging. Expression (28) can be written as:

$$\left|\frac{P_1 * \phi - c}{|\phi|}\right| = \left|\frac{P_2 * \phi - c}{|\phi|}\right|. \tag{29}$$

If we suppose $P_1 * \phi - c \geq 0$ then we have $P_2 * \phi - c \leq 0$. So, by remembering that $|\phi| = 1$, the expression (29) becomes:

$$\left(P_1 * \phi - c\right) + \left(P_2 * \phi - c\right) = 0, \tag{30}$$

then, by using expressions (26)-(27) we finally obtain:

$$c = \frac{c_1\left(\phi\right) + c_2\left(\phi\right)}{2}. \tag{31}$$

Besides, let us call the $d_{\pi_1}\left(\pi_2\right)$ the distance between the planes π_1 and π_2, which can be computed through the distance between the point P_1 and the plane π_2, given by:

$$d_{\pi_1}\left(\pi_2\right) \equiv d_{\pi_2}\left(P_1\right) = \frac{\left(P_1 * \phi - c_2\right)}{|\phi|}, \tag{32}$$

By using expression (26), this equation becomes:

$$d_{\pi_1}\left(\pi_2\right) = c_1 - c_2. \tag{33}$$

We call the *maximum margin hyperplane* or the *optimal hyperplane* the one defined by the unit vector ϕ_0 that maximizes the function:

$$\rho(\phi) = \frac{c_1(\phi) - c_2(\phi)}{2}, \tag{34}$$

$$|\phi| = 1. \tag{35}$$

The corresponding separating plane π has a constant c given by equation (31).

Now let us consider another version of the optimization problem above. Let us consider a vector ψ such that $\psi/|\psi| = \phi$. So, equations (26)-(27) become:

$$x_i * \psi > |\psi| c_1, \quad x_i \in I, \tag{36}$$

$$x_j * \psi < |\psi| c_2, \quad x_j \in II \tag{37}$$

Let us suppose that there is a constant b_0 such that $|\psi| c_1 \geq 1 - b_0$ and $|\psi| c_2 \leq -1 - b_0$. Then, we can rewrite expressions (36)-(37) as:

$$x_i * \psi + b_0 \geq 1, \quad y_i = 1, \tag{38}$$

$$x_j * \psi + b_0 \leq -1, \quad y_j = -1. \tag{39}$$

To understand the meaning of b_0 it is just a matter of using the fact that the equality in (38) holds true for P_1 and the equality in (39) is true for P_2. Therefore, it is straightforward to show that:

$$b_0 = -|\psi| \left(\frac{c_1(\phi) + c_2(\phi)}{2} \right) = -|\psi| c. \tag{40}$$

So, by substituting this equation in expressions (38)-(39) one obtains:

$$x_i * \phi \geq c + \frac{1}{|\psi|}, \quad y_i = 1, \tag{41}$$

$$x_i * \phi \leq c - \frac{1}{|\psi|}, \quad y_j = -1. \tag{42}$$

These expressions mean that we suppose that we can relax the constant c through the value $(1/|\psi|)$ without loosing the separating property. But, the vector ψ is not an unit one. Therefore the distance (32) can be obtained by:

$$d_{\pi_1}(\pi_2) = \frac{(1 - b_0) - (-b_0 - 1)}{|\psi|} = \frac{2}{|\psi|}. \tag{43}$$

In order to maximize this distance (and also maximize the function $\rho(\phi)$ in equation (34)) we must minimize the denominator in expression (43). So, we get an equivalent statement to define the optimal hyperplane: Find a vector ψ_0 and a constant (threshold) b_0 such that they satisfy the constraints:

$$x_i * \psi_0 + b_0 \geq 1, \quad y_i = 1, \tag{44}$$

$$x_j * \psi_0 + b_0 \leq -1, \quad y_j = -1. \tag{45}$$

and the vector ψ_0 has the smallest norm:

$$|\psi| = \psi * \psi. \tag{46}$$

We shall simplify the notation by rewriting the constraints (44)-(45) in the equivalent form:

$$y_i (x_i * \psi_0 + b_0) \geq 1, \quad i = 1, 2, ..., m. \tag{47}$$

In order to solve the quadratic optimization problem stated above, it is used in [32] the Kühn-Tucker Theorem, which generalizes the Lagrange multipliers for convex optimization. The corresponding Lagrange function is:

$$L(\psi, b, \alpha) = \frac{1}{2}\psi * \psi - \sum_{i=1}^{m} \alpha_i (y_i ((x_i * \psi_0) + b_0) - 1), \tag{48}$$

where α_i are the Lagrange multipliers. Following the usual theory, the minimum points of this functional must satisfy the conditions:

$$\frac{\partial L}{\partial \psi} = \psi - \sum_{i=1}^{m} y_i \alpha_i x_i = 0, \tag{49}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{m} y_i \alpha_i = 0. \tag{50}$$

If we substitute (49) into the functional (48) and take into account the result (50) we finally render the following objective function:

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j (x_i * x_j). \tag{51}$$

We must maximize this expression in the nonnegative quadrant $\alpha_i \geq 0, i = 1, 2, ..., m$, under the constraint (50) . In [32] it is demonstrate that the desired solution is given by:

$$\psi_0 = \sum_{i=1}^{m} y_i \alpha_i x_i, \tag{52}$$

$$b_0 = \max_{|\phi|=1} \rho(\phi), \tag{53}$$

subject to:

$$\sum_{i=1}^{m} y_i \alpha_i = 0, \tag{54}$$

$$\alpha_i (y_i ((x_i * \psi_0) + b_0) - 1) = 0, \quad i = 1, 2, ..., m, \tag{55}$$

$$\alpha_i \geq 0. \tag{56}$$

The expression (55) states the Kuhn-Tucker conditions. By observing these conditions one concludes that the nonzero values of α_i, $i = 1, 2, ..., m$, correspond only to the vectors x_i that satisfy the equality:

$$y_i ((x_i * \psi_0) + b_0) = 1. \tag{57}$$

These vectors are the closest to the optimal hyperplane. They are called *support vectors*. The separating hyperplane can be written as:

$$f\left(x,\alpha_0\right)=\sum_{i=1}^{m}y_i\alpha_i^0\left(x_i*x\right)+b_0, \tag{58}$$

where α_i^0, $i=1,2,...,m$,satisfy the constraints (54)-(56). So, we can construct a decision function that is nonlinear in the input space:

$$f\left(x,\alpha\right)=sign\left(\sum_{i=1}^{m}y_i\alpha_i^0\left(x_i*x\right)+b\right), \tag{59}$$

Now we describe two generalizations for the above approach..

3.2.1 Generalizing the SVMs: According to Vapnik [32], a *support vector machine* implement the following idea: "It maps the input vectors x into a high-dimensional *feature space* Z through some nonlinear mapping, chosen a priori. In the space Z an optimal separating hyperplane is constructed."

The key idea behind this proposal comes from the inner product "$*$" in equation (58). Firstly, if we map a vector $x\in\Re^n$ into a Hilbert space Z with coordinates $(z_1,z_2,...)$ we get another representation for the feature space given by:

$$z_1\left(x\right),z_2\left(x\right),...,z_n\left(x\right),..., \tag{60}$$

Then, taking the usual inner product in the Hilbert space we get an equivalent representation for the inner product in the $\Re^n$:

$$z^1*z^2=\sum_{i=1}^{\infty}a_iz_i^1\left(x^1\right)z_i^2\left(x^2\right)\Longleftrightarrow K\left(x^1,x^2\right),\quad a_i\geq 0 \tag{61}$$

where $K\left(x^1,x^2\right)$ is a symmetric function satisfying the condition:

$$\int_C\int_C K\left(u,v\right)g\left(u\right)g\left(v\right)dudv\geq 0, \tag{62}$$

for all $g\in L^2\left(C\right)$, C being a compact subset of $\Re^n$. In this case we say that $K\left(u,v\right)$ is the kernel that generates the inner product for the feature space.

Therefore, we can generalize expression (59) by using the inner product defined by the kernel K :

$$f(x,\alpha) = sign\left(\sum_{supportvectors} y_i \alpha_i^0 K(x_i * x) + b\right), \tag{63}$$

or, equivalently, we can use the linear decision function in the feature space Z :

$$f(x,\alpha) = sign\left[\sum_{supportvectors} y_i \alpha_i^0 \left(\sum_{r=1}^{\infty} a_r z_r\left(x^i\right) z_r(x)\right) + b\right], \tag{64}$$

These expressions define the SVM method [32, 3]. In summary, SVM seeks to find the hyperplane defined by equation (64) which separates positive and negative observations with the maximum margin.

3.2.2 General Nonseparable Case: Sometimes the subsets may be nonseparables, that means, we can not find a small constant δ such that conditions (17)-(18) hold true. In this case, one solution is to work with a more convenient optimization problem. Following [32], we will generalize the optimal hyperplane definition by using a linear optimization procedure. We must observe that the heart of the quadratic optimization problem of section 3.2 is that $\alpha_i \geq 0$, $i = 1, 2, .., m$, and conditions (47), in the sense that they define a separator with good properties. So, we can relax the later constraints by using the following idea. Minimize the functional:

$$L = \sum_{i=1}^{m} \alpha_i + C \sum_{i=1}^{m} \xi_i, \tag{65}$$

subject to:

$$\alpha_i \geq 0, \quad \xi_i \geq 0, \quad i = 1, 2, ..., m, \tag{66}$$

$$y_i\left((x_i * \psi_0) + b_0\right) \geq 1 - \xi_i, \quad i = 1, 2, ..., m. \tag{67}$$

The constant C is a given value. Besides we can also apply an equivalent technique when using a generalized inner product defined by the kernel K in equation (63). In this case, the decision rule has the form given by expression (63). We can solve an optimization problem defined by the minimization of the functional (65) subject to the constraints:

$$\alpha_i \geq 0, \quad \xi_i \geq 0, \quad i = 1, 2, ..., m, \tag{68}$$

$$y_i \left(\sum_{j=1}^{m} y_j \alpha_j K \left(x_i * x_j \right) + b \right) \geq 1 - \xi_i. \tag{69}$$

However, we can not guarantee that these SV machines possess all the nice properties of the machines defined on the basis of the optimal hyperplane.

3.3 Linear Discriminant Analysis (LDA)

The primary purpose of the Linear Discriminant Analysis, or simply LDA, is to separate samples of distinct groups by maximizing their between-class separability while minimizing their within-class variability. LDA 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.

Let the between-class scatter matrix S_b be defined as

$$S_b = \sum_{i=1}^{g} N_i (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})^T \tag{70}$$

and the within-class scatter matrix S_w be defined as

$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \overline{x}_i)(x_{i,j} - \overline{x}_i)^T \tag{71}$$

where $x_{i,j}$ is the n-dimensional pattern (or sample) 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 $\overline{x}_i$ and matrix S_i are respectively the unbiased sample and sample covariance matrix of class π_i [10]. The grand mean vector $\overline{x}$ is given by

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{g} N_i \overline{x}_i = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{N_i} x_{i,j}, \tag{72}$$

where N is, as described earlier, the total number of samples, that is, $N = N_1 + N_2 + \ldots + N_g$. It is important to note that the within-class scatter matrix S_w defined in equation (71) is

essentially the standard pooled covariance matrix S_p multiplied by the scalar $(N-g)$, where S_p can be written as

$$S_p = \frac{1}{N-g}\sum_{i=1}^{g}(N_i-1)S_i = \frac{(N_1-1)S_1+(N_2-1)S_2+\ldots+(N_g-1)S_g}{N-g}. \quad (73)$$

The main objective of LDA is to find a projection matrix W_{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 criterium), that is,

$$W_{lda} = \arg\max_{W}\frac{\left|W^T S_b W\right|}{\left|W^T S_w W\right|}. \quad (74)$$

The Fisher's criterium described in equation 74 is maximized when the projection matrix W_{lda} is composed of the eigenvectors of $S_w^{-1}S_b$ with at most $(g-1)$ nonzero corresponding eigenvalues [10, 9]. In the case of a two-class problem, the LDA projection matrix is in fact the leading eigenvector w_{lda} of $S_w^{-1}S_b$, assuming that S_w is invertible.

However, in limited sample and high dimensional problems, such as in face images analysis, S_w is either singular or mathematically unstable and the standard LDA cannot be used to perform the separating task. To avoid both critical issues, we have calculated w_{lda} by using a maximum uncertainty LDA-based approach (MLDA) that considers the issue of stabilizing the S_w estimate with a multiple of the identity matrix [30, 29, 31]. In a previous study [29, 31] with application to the face recognition problem, Thomaz et. al showed that the MLDA approach improved the LDA classification performance with or without a PCA intermediate step and using less linear discriminant features.

The MLDA algorithm can be described as follows:

1. Find the Φ eigenvectors and Λ eigenvalues of S_p, where $S_p = \frac{S_w}{N-g}$;
2. Calculate the S_p average eigenvalue $\overline{\lambda}$, that is,

$$\overline{\lambda} = \frac{1}{n}\sum_{j=1}^{n}\lambda_j = \frac{Tr(S_p)}{n}; \quad (75)$$

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

$$\Lambda^* = diag[max(\lambda_1,\overline{\lambda}), max(\lambda_2,\overline{\lambda}), \ldots, max(\lambda_n,\overline{\lambda})]; \quad (76)$$

4. Form the modified within-class scatter matrix

$$S_w^* = S_p^*(N-g) = (\Phi\Lambda^*\Phi^T)(N-g). \tag{77}$$

The MLDA method is constructed by replacing S_w with S_w^* in the Fisher's criterium formula described in equation 74. It is based on the idea 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 minimizing a more difficult but appropriate *inflated* within-class scatter matrix would also minimize a less reliable *shrivelled* within-class estimate.

4 Classification versus Reconstruction

In this section, we consider classification and reconstruction problems from the viewpoint of LDA and SVM methods. As described previously in section 3, both linear discriminant methods seek to find a decision boundary that separates data into different classes as well as possible. Figures 4 and 5 picture a hypothetical data set composed of two classes. Theses figures represent the data set, the PCA components (PCAx and PCAy) and the separating plane obtained respectively by the LDA and SVM methods.

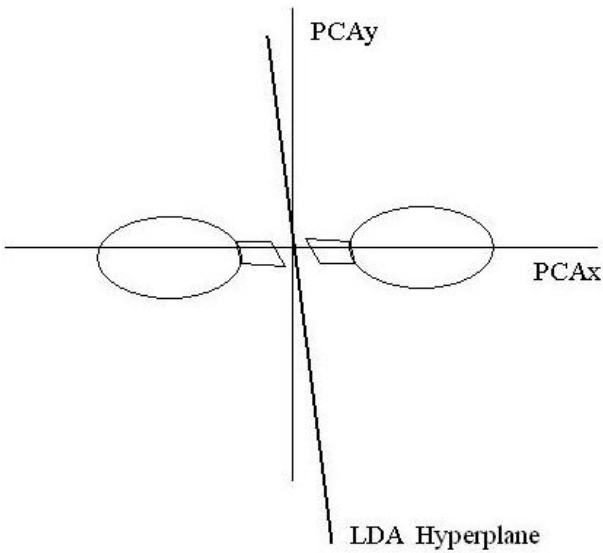


Figure 4. LDA separating hyperplane.

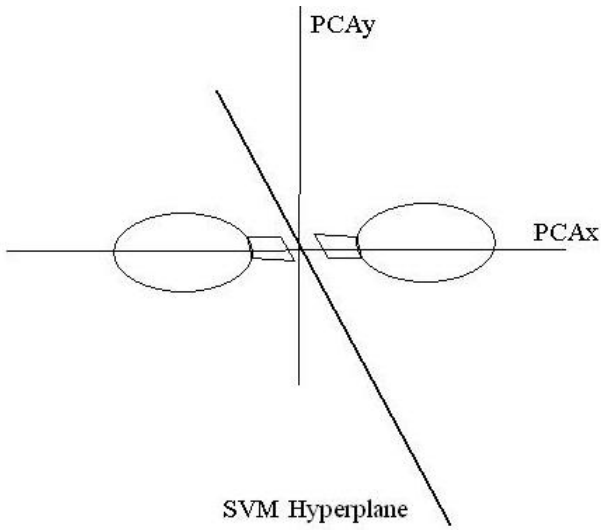


Figure 5. SVM separating hyperplane.

The LDA solution is a spectral matrix analysis of the data and is based on the assumption that each class can be represented by its distribution of data, that is, the corresponding mean vector (or class prototype) and covariance matrix (or spread of the sample group). In other words, LDA depends on all of the data, even points far away from the separating hyperplane and consequently is less robust to gross outliers [14]. This is the reason why the LDA method may give misclassified data points nearby the boundary of the classes.

The description of the SVM solution, on the other hand, does not make any assumption on the distribution of the data, focusing on the observations that lie close to the opposite class, that is, on the observations that most count for classification. In other words, SVM discriminant direction focuses on the data at the boundary of the classes, extracting group-differences that are less perceivable on the original image space. This is emphasized in Figure 5 which indicates that SVM is more robust to outliers, giving a zoom into the subtleties of group differences.

However, according to the above discussion, if we take a one-dimensional point over the SVM and LDA discriminant directions (normal vector of the separating planes) and project it back into the image domain, we expect to observe a better reconstruction result for the LDA case. To clarify this fact, we shall remember that LDA works by maximizing the between-class separability while minimizing their within-class variability, that means, the method tries to collapse the classes into single points as separated as possible. Therefore,

the LDA discriminant direction takes into account all the data performing a more informative reconstruction process in terms of extracting group differences, as we will illustrate in the next section.

5 Case Study 1: Face Image DataSet

We present in this section experimental results on face images analysis. These experiments illustrate the reconstruction problem based on the most expressive principal components and compare the discriminative information extracted by the MLDA and SVM linear approaches. Since the face recognition problem involves small training sets and a large number of features, common characteristics in several pattern recognition applications, and does not require a specific knowledge to interpret the differences between groups, it seems an attractive application to investigate and discuss the statistical learning methods reviewed in this study.

We have used frontal images of a face database maintained by the Department of Electrical Engineering at FEI to carry out the experiments. This database contains a set of face images taken between June 2005 and March 2006 at the Artificial Intelligence Laboratory in São Bernardo do Campo, with 14 images for each of 200 individuals - a total of 2800 images. All images are colour and taken against a white homogenous background in an upright frontal position with profile rotation of up to about 180 degrees. Scale might vary about 10% and the original size of each image is 640×480 pixels. To minimize image variations that are not necessarily related to differences between the faces, we aligned first all the frontal face images of the database to a common template so that the pixel-wise features extracted from the images correspond roughly to the same location across all subjects. For implementation convenience, all the frontal images were cropped to 260×360 pixels, resized to 64×64 pixels, and then converted to 8-bit grey scale.

We have carried out the following two-group statistical analyzes: female versus male (gender) experiments, and non-smiling versus smiling (expression) experiments. The idea of the first discriminant experiment is to evaluate the statistical approaches on a discriminant task where the differences between the groups are evident. The second experiment poses an alternative analysis where there are subtle differences between the groups. Since the number of female images was limited and equal to 49 when we carried out these experiments, we have composed the gender training set of 49 frontal female images and 49 frontal male images. For the expression experiments, we have used the 49 frontal male images previously selected and their corresponding frontal smiling images.

5.1 PCA Reconstruction Results

As the average face image is an n-dimensional point ($n = 4096$) that retains all common features from the training sets, we could use this point to understand what happens statistically when we move along the principal components and reconstruct the respective coordinates on the image space. Analogously to the works by Cootes et al. [8, 7, 6], we have reconstructed the new average face images by changing each principal component separately using the limits of $\pm\sqrt{\lambda_i}$, where λ_i are the corresponding largest eigenvalues.

Figure 6 illustrates these transformations on the first three most expressive principal components using the gender training set. As can be seen, the first principal component (on the top) captures essentially the variations in the illumination and gender of the training samples. The second principal component (middle), in turn, models variations related to the grey-level of the faces and hair, but it is not clear which specific variation this component is actually capturing. The last principal component considered, the third component (bottom), models mainly the size of the head of the training samples. It is important to note that as the gender training set has a very clear separation between the groups, the principal components have kept this separation and when we move along each principal component axis we can see this major difference between the samples, even though subtly, such as in the third principal component illustrated.

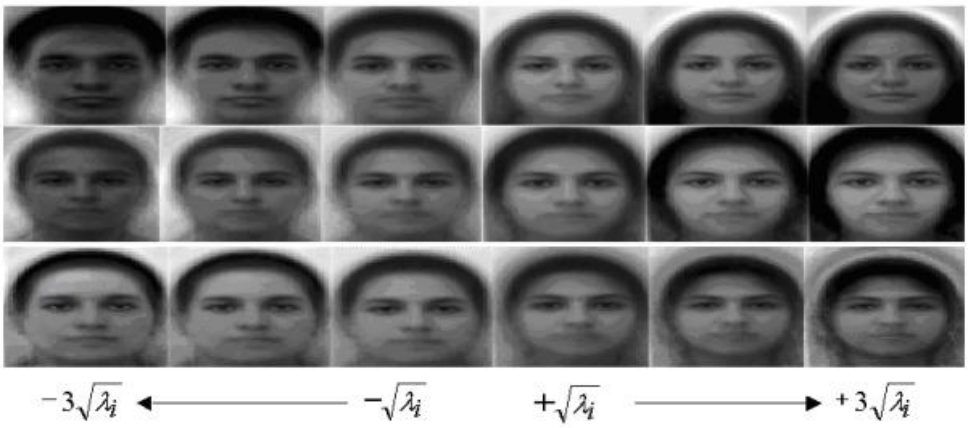


Figure 6. PCA reconstruction results using the gender training set.

Figure 7 presents the three most expressive variations captured by PCA using the expression training set, which is composed of male images only. Analogously to the gender experiments, the first principal component (on the top) captures essentially the changes in

illumination, the second principal component (middle) models variations particularly in the head shape, and the third component (bottom) captures variations in the facial expression among others.

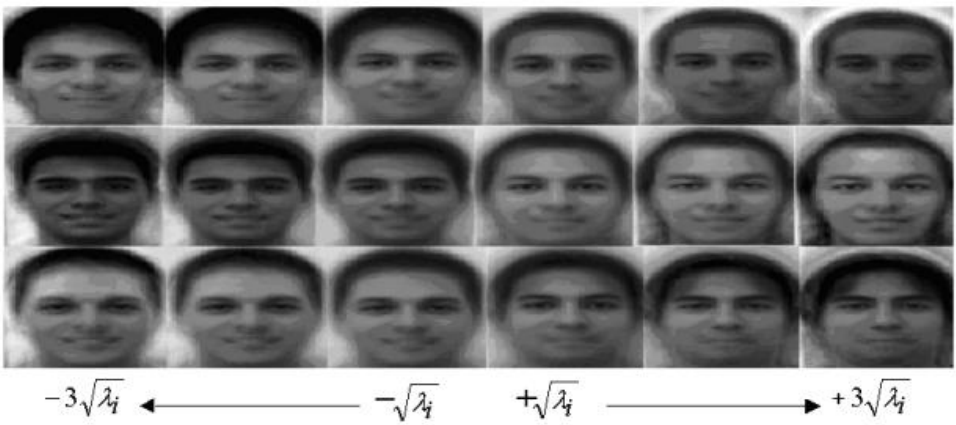


Figure 7. PCA reconstruction results using the expression training set (males images only).

As we should expect, these experimental results show that PCA captures features that have a considerable variation between all training samples, like changes in illumination, gender, and head shape. However, if we need to identify specific changes such as the variation in facial expression solely, PCA has not proved to be a useful solution for this problem. As can be seen in Figure 7, although the third principal component (bottom) models some facial expression variation, this specific variation has been captured by other principal components as well including other image artifacts. Likewise, as Figure 6 illustrates, although the first principal component (top) models gender variation, other changes have been modeled concurrently, such as the variation in illumination. In fact, when we consider a whole grey-level model without landmarks to perform the PCA analysis, there is no guarantee that a single principal component will capture a specific variation alone, no matter how discriminant that variation might be.

5.2 Classification and Information Extraction Results

To compare the classification and information extraction results achieved by the MLDA and SVM linear approaches, we present in this subsection results on the same gender and expression training sets used in the previous subsection.

Table 1 shows the leave-one-out recognition rates of the MLDA and SVM classifiers

on the gender and facial expression experiments using all the non-zero principal components of the intensity features. As can be seen, both methods discriminate gender examples of males (sensitivity) from those of females (specificity) with similar recognition rates. However, for the facial expression experiments, where the differences between the groups are less evident, the SVM achieved clearly the best recognition rates, showing higher sensitivity (non-smiling), specificity (smiling), and accuracy rates than the MLDA approach.

	Gender Experiments			Facial Expression Experiments		
Classifier	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
MLDA	89.80%	85.71%	87.76%	71.43%	69.39%	70.41%
SVM	87.76%	91.84%	89.80%	81.63%	85.71%	83.67%

Table 1. MLDA and SVM classification results.

Since in the experiments of this study we have limited ourselves to two-group classification problems, there is only one MLDA and SVM discriminant direction. Therefore, assuming that the spreads of the classes follow a Gaussian distribution and applying limits to the variance of each group, such as $\pm 3sd_i$, where sd_i is the standard deviation of each group, we can move along the MLDA and SVM most discriminant features and map the results back into the image domain for visual analysis.

The visual analyzes of the linear discriminant feature found by the MLDA and SVM classifiers are summarized in Figure 8. Figures 8.a and 8.b show the group-differences captured by the multivariate statistical classifiers using respectively all the gender and facial expression training samples. These images correspond to one-dimensional points on the MLDA and SVM feature spaces projected back into the image domain and located at 3 standard deviations of each sample group. As can be seen, both MLDA and SVM hyperplanes similarly extracts the gender group differences, showing clearly the features that mainly distinct the female samples from the male ones, such as the size of the eyebrows, nose and mouth, without enhancing other image artifacts. Looking at the facial expression spatial mapping, however, we can visualize that the discriminative direction found by the MLDA has been more effective with respect to extracting group-differences information than the SVM one. For instance, the MLDA most discriminant direction has predicted facial expressions not necessarily present in our corresponding expression training set, such as the "definitely non-smiling" or may be "anger" status and the "definitely smiling" or may be "happiness" status represented respectively by the images -$3sd_1$ and +$3sd_2$ in Figure 8.b.

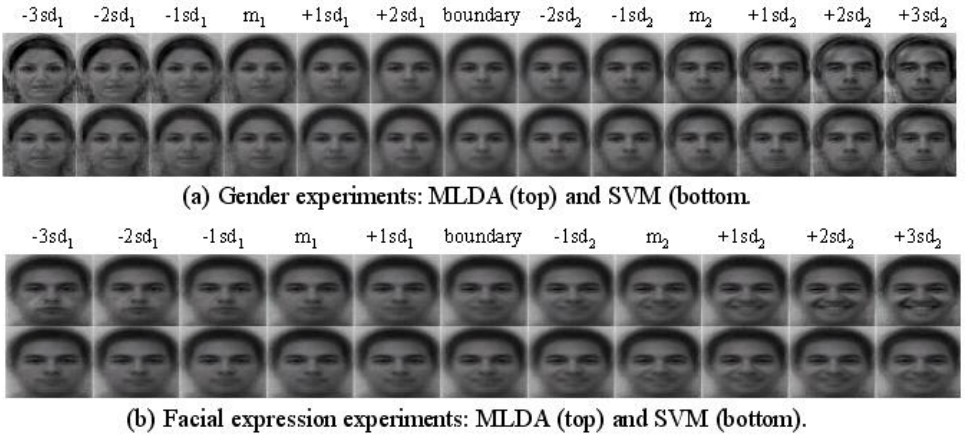


Figure 8. MLDA and SVM information extraction results.

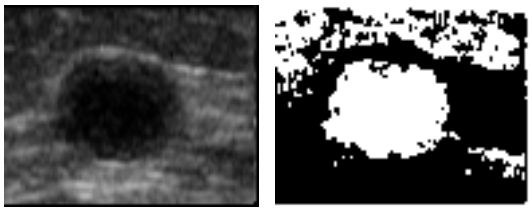
6 Case Study 2: Breast Cancer Classification

When the dimensionality is not a critical point in the application, a classification approach over the target vectorial space can be directly applied. However, some problems regarding low dimensionality in applications based on extracted features may hold. One of such problems is to find out which features have been more influence at the classification step and, as such, should be avoided due to poor influence. On the other hand, what features have weighted the final classification?

In this section, we present an application to the case of classification under low dimensionality using a non-linear kernel embedded in a SVM framework. The application presented here considers the classification in a Computer Aided Diagnosis System (CAD). Ultra Sound (US) exams play an import role in the help of radiologists in the diagnostic of breast cancer, and CAD systems are a power tool in the task of early diagnostic. As per American Cancer Society, breast cancer ranks second in the list of women's cancer. Even though the rate of breast cancer has risen since 1980, the mortality rates have declined by 2.3 since 1990. The reduction in mortality rate has been due to early detection and improvement in technology for treatment, which justifies the development of CAD approaches.

Such systems are normally composed of several different steps, coming from low level to high level vision systems. The application presented here was proposed by Rodrigues et al in [20] and is an automatic methodology for breast lesion classification in ultrasound images based on the following five-step framework:(a) Non-extensive entropy segmentation algo-

rithm; (b) Morphological cleaning to improve segmentation result; (c) Accurate boundary extraction through level set framework; (d) Feature extraction based on information provided by radiologists; (e) Non-linear classification using the breast lesion features as inputs.



(a) (b)

Figure 9. (a) Original ultrasound benign image; (b) NESRA segmentation.

More especifically, the first step of the framework performs an initial segmentation of the ultrasound image using a generalization of the well known Boltzman-Gibbs-Shannon entropy. Rodrigues et al. [20] have presented an algorithm, called NESRA (Non-Extensive Segmentation Recursive Algorithm) to detect the main regions of the ultrasound images (say, the tumor and the background ones) as well as the narrow region around the tumor. These regions are fundamental to further extracting the tumor features. Figure 9 shows an image example of an original benign lesion used in this work (on the left) and the corresponding result after the non-extensive segmentation with the NESRA algorithm (on the right). The justification and implication of why using a non-extensive segmentation algorithm for ultrasound images can be found in [20] and references therein.

As described in the framework, in the second step we have used a morphological chain approach in order to extract the ROI from the background. This was accomplished through the following rule. Considering the binary image generated by NESRA (e.g Figure 9-b), let α and β be the total ROI's area and the total image area, respectively. If $\alpha \geq \xi\beta$ an erosion is carried out and if $\alpha \leq \delta\beta$ a dilation is peformed. Assuming that the ROI has a geometric point near to the image center, we apply a region growing algorithm which defines the final ROI's boundary. In this paper, we follow the same parameters proposed by Rodrigues et al. [20] and have chosen $\xi = 0.75$ and $\delta = 0.25$ to extract most the ROIs. The result of this morphological rule applied in the image of Figure 9-b is illustrated in Figure 10-a. As can be seen, the region generated by the morphological chain rule is a coarse representation of the lesion region. Then, we have applied a level set framework [20] using as initialization this region's boundary [23]. The result can be seen in Figure 10-b, which was accomplished with only 10 iterations of the level set approach.

The result of the lesion extraction illustrated in Figure 10 has served as input to calculate the tumor features commonly used by radiologists in diagnosis. Then, the next step

(a) (b)

Figure 10. (a) ROI after morphological step; (b) final ROI after the level set approach.

is the feature extraction of the ROI. In the work presented in [20], three radiologists stated five features which have high probability to work well as a discriminator between malignant and benign lesions. Then, we have used these features and tested them in order to achieve the best combination in terms of performance. The feature space has been composed by the following attributes:

- Area (AR): The first feature considered is the lesion area. As indicated by the radiologists, since malignant lesions generally have large areas in relation to benign ones, this characteristic might be an important discriminant feature. We have normalized it by the total image area.

- Circularity (CT): The second characteristic is related to the region circularity. Since benign lesions generally have more circular areas compared with the malignant ones, also this can be a good discriminant feature. Then, we have taken the ROI's geometric center point and compute the distance from each boundary point (x_i, y_i) to it. We should expect that malignant lesions tend to have high standard deviations of the average distances in relation to the benign ones. Also, this feature is normalized by total image area.

- Protuberance (PT): The third feature is the size distribution of the lobes in a lesion. A boundary's lobe is a protuberant region on the boundary. We have computed the convex hull of the ROI and the lobe as a protuberance between two valleys. The lobe areas are computed and only those greater than 10% of the lesion area are considered. This feature is taken as the average area of the lobes. We might expect, according to the radiologists, that malignant lesions have high average area in relation to benign ones.

- Homogeneity (HO): The next feature is related to the homogeneity of the lesion. Malignant lesions tend to be less homogeneous than benign ones. Then, we take the Boltzman-Gibbs-Shannon entropy – taken over the gray scale histogram – relative to the maximum entropy as the fourth discriminant feature. In this case, we should expect

that as higher the relative entropy less homogeneous is the lesion region and, consequently, higher is the chance to be a malign lesion.

- Acoustic Shadow (AS): The last feature is related with a characteristic called acoustic shadow. In benign lesions there are many water particles and, as a consequence, dark areas below such lesions are likely to be detected. On the other hand, when the lesion is more solid (a malignant characteristic), there is a tendency in forming white areas below it. We have computed the relative darkness between both areas (lesion's area and area below the lesion) and have taken it as the fifth lesion feature.

These features are the input to a SVM classifier that separates the breast lesions between malignant and benign types. The applied SVM utilizes B-spline as a kernel in its framework. In [20], Rodrigues et al. have justified the use of a B-Spline as a kernel for the SVM by comparing its performance with polynomial and exponential kernels. Additionally, in [20], ROC analyzes of several combinations of the five-feature set have been performed to determine the best recognition performance of the framework. Although the results have shown that not all of these five features should be used to improve the classification accuracy, no theoretical justification has been presented in order to select a specific subset of the original feature space for optimum performance.

In this case study, we will analyze these results further using the discriminant information provided by the SVM separating hyperplane used to classify previously the sample groups. Thus, we repeat the same experiments carried out in [20], which have used a 50 pathology-proven cases database – 20 benign and 30 malignant. Each case is a sequence of 5 images of the same lesion. Then, we tested 100 images of benign lesion and 150 of malignant ones. Since the detection of a malignant lesion between five images of the same case indicates a malignant case, it is reasonable to consider 250 different cases.

1	Area (AR)
2	Acoustic Shadow (AS)
3	Homogeneity (HO)
4	Circularity (CT)
5	Protuberance (PT)

Table 2. SVM most discriminant features in decreasing order.

Since the SVM separating hyperplane has been calculated on the original feature space, it is possible to determine the discriminant contribution of each feature by investigating the weights of the most discriminant direction found by the SVM approach. Table 2 lists the features in decreasing order of discriminant power selected by the SVM separating hyperplane using all the samples available. As can be seen, SVM has selected the AR feature

as the most discriminant feature, followed by AS, HO, CT and PT. In other words, we should expect a better performance of the classifier when using, for instance, two features only if we select the pair of features (AR, AS) rather than (CT, PT).

Following the discriminant order suggested in Table 2, we can guide our classification experiments by combining the features according to those which improve the groups separation. Then, we carried out experiments with the following features combination:

1. AR + AS
2. AR + AS + HO
3. AR + AS + HO + CT
4. AR + AS + HO + CT + PT

We have used the cross-validation method to evaluate these experiments. Therefore, the ultrasonic images are firstly divided randomly into five groups. We first set the first group as a testing set and use the remaining four groups to train the SVM. After training, the SVM is then tested on the first group. Then, we set the second group as a testing group and the remaining four groups are trained and then the SVM is tested on the second. This process is repeated until all the five groups have been set in turn as testing sets.

The ROC curves of these experiments are shown in Figure 11. As can be seen in Figure 11, the best combination which yields the largest Az value (area under the ROC curve) is clearly the one composed of the features AR, AS and HO, with $Az = 92\%$. The other combinations show lower Az values and worst sensitivity and specificity ratios compared to the AR+AS+HO features set.

As mentioned in the previous case study, the other main task that can be carried out by the statistical learning method is to reconstruct the most discriminant feature described by the SVM separating hyperplane. Assuming that the clouds of the malignant and benign points follow a multidimensional Gaussian distribution and applying limits to the variation of each cloud, we can move along this most discriminant hyperplane and map the result back into the original domain to understand the discriminant information captured by the classifier.

Figure 12 presents the SVM most discriminant feature of the five-feature dataset using all the examples as a training samples. It displays the differences on the original features captured by the classifier that change when we move from one side (malignant or group 1) of the dividing hyperplane to the other (benign or group 2), following limits to ± 3 standard deviations of each sample group. We can see clearly differences in the AR as well as AS and HO features. That is, the changes on the AR, AS, and HO features are relatively more significant to discriminate the sample groups than the CT and PT features, as reported in

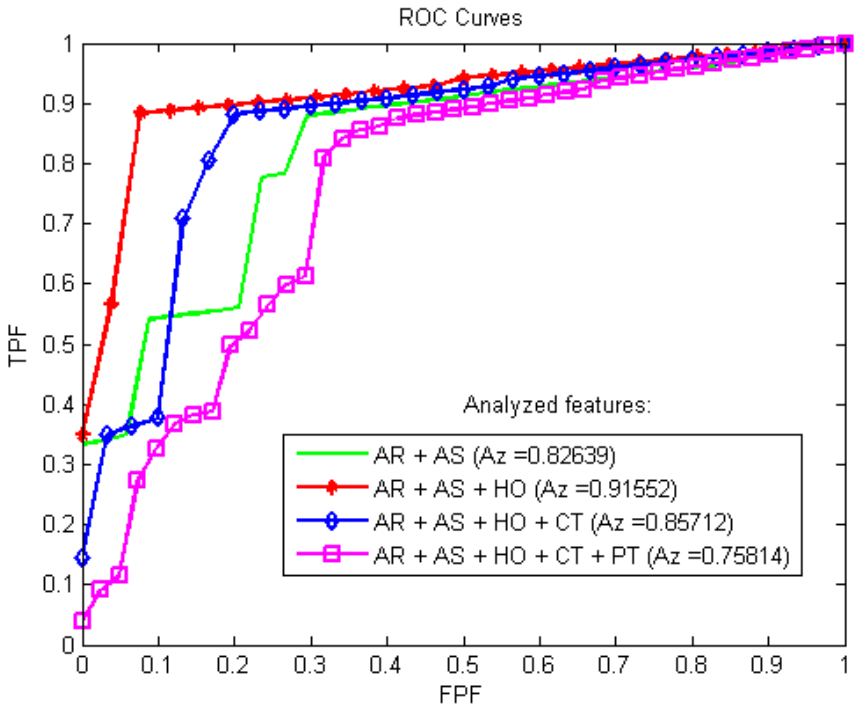


Figure 11. ROC curves for the discriminant combinations of features in classification of tumor in ultrasound images.

the ROC analysis illustrated in the previous Figure 11. Additionally, Figure 12 illustrates that when we move from the definitely benign samples (on the right) to the definitely malign samples (on the left), we should expect an relative increase on the lesion area (AR), and a relative decrease on the acoustic shadow (AS) and homogeneity (HO) of the lesion. All these results are plausible and provide a quantitative measure to interpreting the discriminant importance and variation of each feature in the classification experiments.

7 Conclusion and Perspectives

In this paper we describe the Peceptron, SVM and LDA methods. Both perceptron and SVM methods seek for a separating hyperplane. The former is an iterative method and the obtained hyperplane may be not the optimal one. SVM is based on an explicit optimization

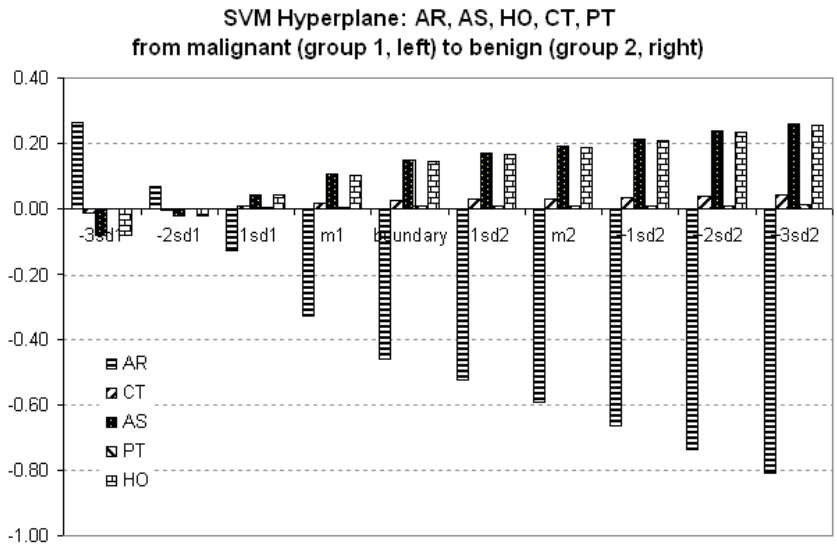


Figure 12. Statistical differences between the malign (on the left) and benign (on the right) samples captured by the SVM hyperplane.

method in order to achieve optimality properties.

Besides, we have compared the LDA and SVM separating hyperplanes for extracting group-differences between face images. Our experimental results indicate that SVM is a more robust technique for classification than LDA. This is not a surprising result since the LDA-based method is not directly related to classification accuracy and its optimality criterion is based on prototypes, covariance matrices, and normally distributed classes. However, the hyperplanes found by both approaches are different from the point of view of detecting discriminative information from images. The findings of this study support the hypothesis that variations within each group are very useful for extracting discriminative information, especially the points that are far way from the boundary. Statistically, we can understand such points as coming definitely from either class and consequently more informative for characterizing group-differences. In contrast, the SVM discriminative direction that focuses on the data at the boundary of the classes defines differences that might be less perceivable on the original image space. We believe that fundamentally LDA (or the MLDA) seeks direction that maximizes group differences, whereas SVM seeks direction that maximizes boundary differences. This is an important distinction between the two approaches, particularly in

extracting discriminative information from groups of patterns where ground truth might be unknown, such as medical images of a particular brain disorder.

The generalizations proposed in sections 3.2.1 and 3.2.2 puts the SVM technique in a linear optimization context. For instance, we can use a simplex method to minimize expression (65), subject to the constraints (66)-(67). Then, we could compare SVM and Perceptron both from the viewpoint of convergence rate and quality of the obtained result. Also, the sensitivity to the size (m) of the input set is another point to be considered. Moreover, we can perform the experiment illustrated in Figure 8 for a general separating hypersurface (non-linear case). In this case, the normal direction changes when we travel along the boundary which may bring new aspects for the reconstruction process.

The KL transform is the approach behind the PCA analysis, as we saw in section 2. Dimensionality reduction can be also achieved by others unitary transforms, like the cosine one. Besides, when we travel in a normal direction of the separating hyperplane and perform the reconstruction it is important to track some structures in the image space. This track can be performed by a deformable model which position must be corrected frame by frame to improve efficiency.

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. In addition, the authors would like to thank the support provided by PCI-LNCC, FAPESP (grant 2005/02899-4), and CNPq (grants 473219/04-2 e 472386/2007-7).

References

[1] V. Algazi and D. Sakrison. On the optimality of karhunen-loeve expansion. *IEEE Trans. Information Theory*, pages 319–321, 1969.

[2] R. Beale and T. Jackson. *Neural Computing*. MIT Press, 1994.

[3] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[4] W.-C. Chang. On using principal components before separating a mixture of two multivariate normal distributions. *Appl. Statist.*, 32(3):267–275, 1983.

[5] 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. *Paterns Recognition*, 33:1713–1726, 2000.

[6] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. In *4th International Conference on Automatic Face and Gesture Recognition*, pages 227–232, 2000.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV'98*, pages 484–498, 1998.

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models- their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

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

[10] K. Fukunaga. Introduction to statistical pattern recognition. *Boston: Academic Press*, second edition, 1990.

[11] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[12] P. Golland, W. Grimson, M. Shenton, and R. Kikinis. Detection and analysis of statistical differences in anatomical shape. *Medical Image Analysis*, 9:69–86, 2005.

[13] P. Golland, W. E. L. Grimson, M. E. Shenton, and R. Kikinis. Deformation analysis for shape based classification. *Lecture Notes in Computer Science*, 2082, 2001.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[15] C. Huberty. *Applied Discriminant Analysis*. John Wiley & Sons, INC., 1994.

[16] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, Inc., 1989.

[17] I. T. Jolliffe, B. J. T. Morgan, and P. J. Young. A simulation study of the use of principal component in linear discriminant analysis. *Journal of Statistical Computing*, 55:353–366, 1996.

[18] E. C. Kitani and C. E. Thomaz. Analise de discriminantes lineares para modelagem e reconstrucao de imagens de face (in portuguese). In *6th Encontro Nacional de Inteligencia Artificial ENIA'07*, pages 962–971, 2007.

[19] E. C. Kitani, C. E. Thomaz, and D. F. Gillies. A statistical discriminant model for face interpretation and reconstruction. In *SIBGRAPI'06, IEEE CS Press*, pages 247–254, 2006.

[20] P. Rodrigues, G. Giraldi, R.-F. Chang, and J. Suri. Non-extensive entropy for cad systems of breast cancer images. In *In Proc. of International Symposium on Computer Graphics, Image Processing and Vision - SIBGRAPI'06*, Manaus, Amazonas, Brazil, 2006.

[21] P. Shih and C. Liu. Face detection using discriminating feature analysis and support vector machine. *Pattern Recogn.*, 39(2):260–276, 2006.

[22] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

[23] J. S. Suri and R. M. Ragayyan. *Recent Advances in Breast Imaging, Mammography and Computer Aided Diagnosis of Breast Cancer*. SPIE Press, April 2006.

[24] C. Thomaz, J. Boardman, D. Hill, J. Hajnal, D. Edwards, M. Rutherford, D. Gillies, and D. Rueckert. Using a maximum uncertainty lda-based approach to classify and analyse mr brain images. In *International Conference on Medical Image Computing and Computer Assisted Intervention MICCAI04*, pages 291–300, 2004.

[25] C. Thomaz, P. Rodrigues, and G. Giraldi. Using face images to investigate the differences between lda and svm separating hyper-planes. In *II Workshop de Visao Computacional*, 2006.

[26] C. E. Thomaz, N. A. O. Aguiar, S. H. A. Oliveira, F. L. S. Duran, G. F. Busatto, D. F. Gillies, and D. Rueckert. Extracting discriminative information from medical images: A multivariate linear approach. In *SIBGRAPI'06, IEEE CS Press*, pages 113–120, 2006.

[27] C. E. Thomaz, J. P. Boardman, S. Counsell, D. Hill, J. V. Hajnal, A. D. Edwards, M. A. Rutherford, D. F. Gillies, and D. Rueckert. A whole brain morphometric analysis of changes associated with preterm birth. In *SPIE International Symposium on Medical Imaging: Image Processing*, volume 6144, pages 1903–1910, 2006.

[28] C. E. Thomaz, J. P. Boardman, S. Counsell, D. Hill, J. V. Hajnal, A. D. Edwards, M. A. Rutherford, D. F. Gillies, and D. Rueckert. A multivariate statistical analysis of the developing human brain in preterm infants. *Image and Vision Computing*, 25(6):981–994, 2007.

[29] C. E. Thomaz and D. F. Gillies. A maximum uncertainty lda-based approach for limited sample size problems - with application to face recognition. In *SIBGRAPI'05, IEEE CS Press*, pages 89–96, 2005.

[30] C. E. Thomaz, D. F. Gillies, and R. Q. Feitosa. A new covariance estimate for bayesian classifiers in biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, 14(2):214–223, 2004.

[31] C. E. Thomaz, E. C. Kitani, and D. F. Gillies. A maximum uncertainty lda-based approach for limited sample size problems - with application to face recognition. *Journal of the Brazilian Computer Society (JBCS)*, 12(2):7–18, 2006.

[32] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, INC., 1998.