working papers

UNIVERSIDAD CARLOS III DE MADRID

DIMENSIONALITY REDUCTION WITH IMAGE DATA

Mónica Benito and Daniel Peña*

Abstract

A common objective in image analysis is dimensionality reduction. The most common often used data-exploratory technique with this objective is principal component analysis. We propose a new method based on the projection of the images as matrices after a Procrustes rotation and show that it leads to a better reconstruction of images.

Keywords: Eigenfaces; multivariate linear regression; singular value decomposition; principal component analysis; generalized procrustes analysis.

*Benito, Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: mbenito@est-econ.uc3m.es, Peña, Departamento de Estadística, Universidad Carlos III de Madrid, e-mail: dpena@est-econ.uc3m.es. Mónica Benito was supported by FPU grandt of the Ministery of Education of Spain. She is very grateful to Steve Marron for his hospitality and insightful comments during the time she was working on this research at the University of North Carolina in Chapel Hill.

Dimensionality Reduction with Image Data

Mónica Benito and Daniel Peña

Universidad Carlos III de Madrid, Spain

Summary. A common objective in image analysis is dimensionality reduction. The most often used data-exploratory technique with this objective is principal component analysis. We propose a new method based on the projection of the images as matrices after a Procrustes rotation and show that it leads to a better reconstruction of images.

Keywords: Eigenfaces; Multivariate linear regression; Singular value decomposition; Principal component analysis; Generalized proscrustes analysis.

1. Introduction

Exploratory image studies are generally aimed at data inspection and dimensionality reduction. One of the most popular approaches to reduce dimensionality and derive useful compact representations for image data is Principal Component Analysis (PCA). Kirby & Sirovich (1990) proposed using PCA to reduce the dimensionality when representing human faces. The performance of this method on aligned and scaled human faces is very good, but it does not work well for non-aligned faces. Alternative approaches using Independent Component Analysis (ICA) for face representation have been proposed by Barlett and Sejnowski (1997). Wu & Zhou (2002) have also demostrated that a pre-processing step of the image sample improves the PCA performance. In the last two decades, PCA has been especially popular in the object recognition community, where it has succesfully been employed by Turk & Pentland (1991), Valentin et. al (1996) and Swets & Weng (1996).

Address for correspondence: Mónica Benito, Departamento de Estadística, Universidad Carlos III de Madrid, c/ Madrid 126, 28903 Getafe (Madrid), Spain. E-mail: mbenito@est-econ.uc3m.es

The problem we are interested in is as follows. We have a set of images which represent similar objects, for instance, human faces, temporal images of the same scene, objects in a process of quality control, and so on. Any particular image (say the $n-th$ image) is represented by a matrix X_n of I rows and J columns. We assume that the sample contains the set of N images, $X_1, X_2, ..., X_N$. Each matrix consists of elements x_{ij}, with $i = 1, ..., I$ and $j = 1, ..., J$, that represent the pixel intensities extracted from digitized images. All the elements x_{ij} are in the range between 0 and 255, where the value 0 represents black color, and the value 255 white. Suppose that each matrix is transformed into a vector $\mathbf{x}_n$ by row (or column) concatenation. Therefore, we have a set of N vectors in a high dimensional space, specifically, $\mathbf{x}_n \in \Re^d$ where $d = I \times J$, $n = 1, ..., N$. For convenience, the vectors are assumed to be normalized, so that $\mathbf{x}_n^T \mathbf{x}_n = 1$. Note that this set of vectors can be represented by an $N \times d$ matrix X in which the $n-th$ row is equal to $\mathbf{x}_n$. When dealing with high-dimensional observations, linear mappings are often used to reduce dimensionality of the data by extracting a small (compared to the original dimensionality of the data) number of linear features. Among all linear, orthonormal transformations, principal component analysis is optimal in the sense that it minimizes, in mean square sense, the errors in the reconstruction of the original signal $\mathbf{x}_n$ from its low-dimensional representation, $\widehat{\mathbf{x}}_n$. As is well known, PCA is based on finding directions of maximal variability. In this paper we propose an alternative way of projecting the original data on a subspace of lower dimension. Instead of concatenating rows or columns, we keep the structure of the matrix in the projection. The rest of the paper is organized as follows. In the next section, we briefly introduce the basic theory of the standard method used for dimensionality reduction based on principal component in the context of image data. In Section 3 we propose a new approach which keeps the internal structure of the image and we show that this procedure has important advantages compared to classical PCA. In section 4 we discuss the problems of aligning and scaling images before the dimension reduction is carried out, and introduce a generalized proscrustes rotation to solve this problem. Finally, in Section 5 we present the experimental results of the procedure when applied to a human face data base.

2. Principal Component Analysis (PCA)

Assume that $\mathbf{x}_1^T$, $\mathbf{x}_2^T$, ..., $\mathbf{x}_N^T$, is a set of $d-$dimensional input data vectors which are assumed to be zero mean and unit variance and let X be the matrix $N \times d$ whose $n - th$ row is the vector $\mathbf{x}_n^T$. The purpose of PCA is to find p $(p < d)$ standardized linear combinations of the original variables $X\mathbf{u}_1$, $X\mathbf{u}_2$, ...,$X\mathbf{u}_p$ which are uncorrelated and have maximal variance. These vectors are obtained by the orthogonal transformation (see Mardia et. al., 1992, for more details),

$$Z = X U_p \tag{1}$$

where U_p is an orthogonal matrix with columns vectors $\mathbf{u}_i$ which are eigenvectors linked to the p-th largest eigenvalues of the matrix $X'X$. The optimal prediction $\widehat{X}_p$ with a matrix of rank p of the X matrix with the least squares criterion is based on the singular value decomposition of this matrix X. Suppose that the rank of X is r, $r = \min\{N, d\}$. The best reconstruction of this matrix using $p \leq r$ dimensions is given by

$$\widehat{X}_p = X U_p U_p^T \tag{2}$$

and this implies that each vector $\mathbf{x}_n$ is predicted as,

$$\widehat{\mathbf{x}}_n = \sum_{i=1}^{p} (\mathbf{x}_n^T \mathbf{u}_i) \mathbf{u}_i. \tag{3}$$

Note that in order to predict the N images, which implies a total of IJN scalar pixel predictions, we only need the p vectors $\mathbf{u}_i$ of dimension IJ and the p scalar values $z_i = (\mathbf{x}_n^T \mathbf{u}_i)$ for each image. Thus, the number of scalar values required for the prediction is:

$$pIJ + pN = IJN(\frac{p}{N} + \frac{p}{IJ}) = IJNc_1 \tag{4}$$

where c_1 represents the factor of reduction. If p/IJ is small, so that c_1 is small, the reduction of dimension required for the reconstraction of the images can be very important. This decomposition was used for image dimension reduction by Turk & Pentland (1991), and it is often refered to as the eigenface method. It is the most common approach for dimension reduction with images.

3. An Alternative approach based on matrix projections

We are interested in a projection method which keep the matrix structure of the image. Yang & Yang (2002) proposed the projection of the rows of the matrix in the context of

feature extraction. Here we follow a similar approach. Assume without loss of generality that $I > J$. Then, given $\mathbf{a}$ a unit norm $J \times 1$ vector, we can project the rows of X_n on the $\mathbf{a}$ direction by,

$$\mathbf{w}_n = X_n \mathbf{a} \tag{5}$$

We will call this $I-$dimensional projected vector $\mathbf{w}_n$ the projected feature vector of X_n. Suppose that we project all the images in this way and obtain a set of vectors, $\mathbf{w}_n$, $n = 1, ..., N$. In order to find a good projection direction, let us call S_r the $I \times I$ covariance matrix for these vectors representing the rows, (the subindex r is due to the projection of the rows. We will discuss later the projection of the columns). This matrix is given by

$$S_r = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{w}_n - \overline{\mathbf{w}} \right) \left(\mathbf{w}_n - \overline{\mathbf{w}} \right)^T , \tag{6}$$

where $\overline{\mathbf{w}}$ is the mean of the projected vectors. The two most often used measures to describe scatter about the mean in multivariate data are the total variation, given by the trace of the covariance matrix, and the generalized variance, given by the determinant of this matrix. For simplicity let us find the direction $\mathbf{a}$ which maximizes the total variation given by the trace of S_r. Then

$$max \ \ tr(S_r) = \ max \ \ tr \left(\frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{w}_n - \overline{\mathbf{w}} \right) \left(\mathbf{w}_n - \overline{\mathbf{w}} \right)^T \right) \tag{7}$$

and using the definition (5),

$$tr(S_r) = tr \left(\frac{1}{N} \sum_{n=1}^{N} \left(X_n \mathbf{a} - \overline{X} \mathbf{a} \right) \left(X_n \mathbf{a} - \overline{X} \mathbf{a} \right)^T \right) \tag{8}$$

where $\overline{X} = \dfrac{1}{N} \sum_{n=1}^{N} X_n$ denotes the mean image. As

$$max \ \ tr(S_r) = \ max \ \frac{1}{N} \times tr \left(\mathbf{a}^T \left[\sum_{n=1}^{N} \left(X_n - \overline{X} \right)^T \left(X_n - \overline{X} \right) \right] \mathbf{a} \right) \tag{9}$$

it follows that vector $\mathbf{a}$ is the eigenvector linked to the largest eigenvalue of the matrix

$$\Sigma_c = \frac{1}{N} \sum_{n=1}^{N} \left(X_n - \overline{X} \right)^T \left(X_n - \overline{X} \right) \quad ; \quad \Sigma_c \in \Re^{J \times J} \tag{10}$$

As we need more than one direction of projection to characterize the sample, we compute the set of eigenvectors $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p$, which constitute a basis for $\Re^p$ from which the data

can be estimated using a subspace of lower dimension, $p \leq \min \{I, J\}$. It is easy to see that the same criterion is obtained if we start projecting the columns instead of the rows. Given $\mathbf{b}$ a unit norm $I \times 1$ vector, the projection of the columns of X_n on the $\mathbf{b}$ direction is given by,

$$\mathbf{z}_n = X_n^T \mathbf{b} \tag{11}$$

and this $J-$dimensional projected vector $\mathbf{z}_n$ is the projected feature vector of X_n. The covariance matrix between the projected vectors $\mathbf{z}_n$ is defined by

$$S_c = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{z}_n - \overline{\mathbf{z}}) (\mathbf{z}_n - \overline{\mathbf{z}})^T \tag{12}$$

where $\overline{\mathbf{z}}$ is the average projected columns, and maximizing the trace of S_c leads to finding the eigenvectors linked to the largest eigenvalues of the matrix

$$\Sigma_r = \frac{1}{N} \sum_{n=1}^{N} \left(X_n - \overline{X} \right) \left(X_n - \overline{X} \right)^T \quad ; \quad \Sigma_r \in \Re^{I \times I} \tag{13}$$

which have the same non null eigenvalues as Σ_c.

3.1. Prediction by Multivariate Regression

Let W_n be the feature vectors obtained as the solution of (7),

$$W_n = [X_n \mathbf{a}_1, ..., X_n \mathbf{a}_p] = X_n A_p \quad ; \quad W_n \in \Re^{I \times p} \tag{14}$$

we can use these data to predict the matrix X_n by the multivariate regression model

$$X_n = W_n \beta_n + \varepsilon_n \tag{15}$$

where the matrix X_n is predicted from its feature vectors W_n using some parameters $\beta_n = \left[\beta_n^1, ..., \beta_n^J \right] \in \Re^{p \times J}$, which depend on the image. The least squares estimate is given by $\widehat{\beta}_n = \left(W_n^T W_n \right)^{-1} W_n^T X_n$ and the prediction of the matrix X_n with this model is

$$\widehat{X}_n = H_n X_n \tag{16}$$

where $H_n = W_n \left(W_n^T W_n \right)^{-1} W_n^T$ is the perpendicular projection operator onto the column space of W_n . Observe that although the set of design matrices $W_1, W_2, ..., W_N$ are specific for each image, this set has been obtained using a common orthonormal base of projection

based on the singular value decomposition of the covariance matrix by columns between the original images, Σ_c, as we discussed above.

The reconstruction of each image with IJ numbers requires the matrix W_n, with dimension Ip plus the vector β_n of dimension pJ leading to

$$N(Ip + pJ) = IJN(\frac{p}{J} + \frac{p}{I}) = IJNc_2 \tag{17}$$

and we see that if both I and J are large with relation to p the reduction in the dimension of the problem can be important. Suppose that the projection has been done using the columns of X instead of the rows, then, the feature matrix Z_n has dimension $J \times p$ and the perpendicular projection operator $G_n = Z_n \left(Z_n^T Z_n \right)^{-1} Z_n^T$ is used to reconstruct each image. That reconstruction requires the same number of parameters as that defined in (17). Let be $r = \max\{I, J\}$, we suggest using the projection by rows when $r = I$ and on the contrary, project the columns when $r = J$. This criterion is based on the idea that we want to use as much information as possible to reconstruct the images, so we are interested in a feature matrix which has the highest dimension.

4. Image Registration

When dealing with a set of homogeneous objects, as in the case of the human face database, the different ilumination and facial expressions greatly increase the difficulty of the reconstruction task. To avoid this problem, some procedures have been proposed to pre-process the images focussing on smoothing techniques, Wu and Zhou (2002). Nevertheless, when the problem focuses on shape normalization, these techniques are not useful under such variations. Alternatively, the sample can be seen as a set of shapes with respect to a local 2D coordinates system. We can combine these different local coordinate systems into a common system in order to have a normalized sample of objects before they are analyzed by subspace techniques. This geometric transformation process is known as registration. Depending of the complexity of the object, it may require two or more viewpoints, also called landmarks, to register it appropiately. From this point of view, we would like all the images to satisfy some constraints. For instance, these constraints may include the line between the two eyes being parallel to the horizontal axis, the inter-ocular distance being set to a fixed value, or the size of the face being fixed. In this sense, when some important points or landmarks (for instance, the eye location) are fixed, we would expect exactly all

the sample centered in the middle of the image. This can be solved easily by an affine transformation,

$$\mathbf{b}_i = D\mathbf{a}_i + \mathbf{s} \quad , \quad i = 1, ..., d \tag{18}$$

where d is the number of pixels, $d = I \times J$ and the vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ belong to $\Re^2$, due the pixel's spatial coordinates. Thus, for any pixel in the image, say the $i - th$, this transformation maps the vector $\mathbf{a}_i$ to $\mathbf{b}_i$. The affine transformation is able to rescale the image by the matrix D as well as a possible rotation of the face. The vector $\mathbf{s}$ describes the translation. Next, we will discuss how the parameters $(D, \mathbf{s})$ for the desidered transformation can be determined.

4.1. Affine Transformation

What we call affine transformation is a procedure in which the landmark points are selected so that these three landmarks have the same coordinates in all the images. To describe the face location and orientation in an image, the best selection seems to be the points which describe the position of the left and right eye, and the point that marks the lower end of the chin (see Nadenau, 1997). Let $\mathbf{a}_{1n}, \mathbf{a}_{2n}$ and $\mathbf{a}_{3n}$ be the points which describe these positions in the $n - th$ image of the sample, $n = 1, ..., N$. We want to transform them to fixed points given by $B = [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3]$. Let $A_n = [\mathbf{a}_{1n}, \mathbf{a}_{2n}, \mathbf{a}_{3n}]$, then

$$B = DA_n + \mathbf{s} \quad , \quad n = 1, ..., N \tag{19}$$

and the matrix A and the vector $\mathbf{s}$ are obtained by solving this system of 6 equations and 6 unknowns. This approach has two main limitations. The first one is that we can select only three points to fix the object normalization. The second is that we are not keeping the relative distances among the landmarks in the transformation. As an alternative, we propose a new procedure to estimate the similarity transformation that avoids these two liminations.

4.2. Procrustes Analysis

Procrustes analysis theory is a set of mathematical tools to directly estimate and perform simultaneous similarity transformations among the objects landmarks up to their maximal agreement. Based on this idea, we can focus on a goodness of fit measure used to compare

N configurations of points. The basic procedure is as follows. Let A_n be the $r \times 2$ matrix of coordinates of r landmarks in the $n - th$ image, $n = 1, ..., N$. We wish to find simultaneous translations, rotations, and scale factors of these N sets of points into positions of best fit with respect to each other. The functional model of the transformation is stated as follows,

$$\widehat{A}_n = c_n A_n T_n + \mathbf{1} \mathbf{t}_n^T \quad , \quad n = 1, ..., N \tag{20}$$

where c_n is the scale factor, T_n is 2×2 orthogonal rotation matrix, $\mathbf{t}_n$ is a 2×1 traslation vector, and $\mathbf{1}$ is a 2×1 unit vector. The N matched configurations are measured by means of the residual sum of squares between each point of each configuration and the corresponding point of the average configuration or common coordinate system. For this task, Generalized Orthogonal Procrustes Analysis (Gower, 1975) provides least-squares correspondence of more than two point matrices. According to Goodall (1991), there is a matrix B, also called consensus matrix, which contains the true coordinates of the r points defined in a mean and common coordinate system. The solution of the problem can be thought of as the search for the unknown optimal matrix B. Defining C as the geometrical centroid of the transformed matrices $\widehat{A}_1, ..., \widehat{A}_N$,

$$C = \frac{1}{N} \sum_{n=1}^{N} \widehat{A}_n \tag{21}$$

the solution of the registration problem is achieved by using the following minimum condition

$$\sum_{n=1}^{N} tr \left\{ \left[\widehat{A}_n - C \right]^T \left[\widehat{A}_n - C \right] \right\} \tag{22}$$

in an iterative computation scheme of centroid C. To start the algorithm we need to define an initial centroid, which will be used to fit all the models A_n, one at time, to this temporary centroid C (this will be explained in section 4.2.1). The idea behind the procrustes solution in (19) is an iterative updating of C and $\widehat{A}_n$ according to equations (21) and (20) until global convergence, i.e., until the centroid configurations variations between two subsequent iterations are smaller than a pre-defined threshold ϵ. Hence, the final solution of the centroid corresponds to the least squares estimation $\widehat{B}$ and shows the final coordinates of r points in the maximal agreement with respect to least squares objective function. Finally, the unknown similarity transformation parameters $(T_n, \mathbf{t}_n, c_n)$, $n = 1, ..., N$, are then determined using the procrustes algorithm procedure for fitting two given sets of points, A_n and C, as we

did early in each iteration of the algorithm but using the temporary centroid. Next, we will explain the process of fitting two configurations of points and how the initial centroid is defined.

4.2.1. *Fitting two configurations of points*

To solve the problem of transforming a given matrix A into a given matrix B by an orthogonal matrix T, a traslation vector $\mathbf{t}$ and a scale factor c, Schoenemann and Carroll (1970) proposed a least squares method called Extended Orthogonal Procrustes (EOP) which consists of minimizing the sum of squares of the residual matrix

$$E = cAT + \mathbf{1}\mathbf{t}^T - B \tag{23}$$

Matrices A and B are $r \times p$ dimensional, containing r corresponding points in the $p-$dimensional space. In the image data context, $p = 2$. Similarly as in (20), $\mathbf{1}$ is the $r \times 2$ unit vector and $\mathbf{t}$ is 2×1 traslation vector. In order to obtain the least squares estimation of the unknown parameters $(T, \mathbf{t}, c)$ the solution must satisfy the following condition,

$$min\ tr\left\{E^T E\right\} =\ min\ tr\left\{\left(cAT + \mathbf{1}\mathbf{t}^T - B\right)^T \left(cAT + \mathbf{1}\mathbf{t}^T - B\right)\right\} \tag{24}$$

Let $E = \widehat{A} - B$. Then the criterion condition (22) with the orthogonality condition $T^T T = TT^T = I$, imply the Lagrangean function

$$L = tr\left\{E^T E\right\} + tr\{\Lambda(T^T T - I)\} \tag{25}$$

where Λ is a matrix of Lagrangean multipliers. As the derivations of the Lagrangean function with respect to unknowns must be zero in order to obtain a least squares estimation, we have

$$\frac{\partial L}{\partial T} = 2c^2 A^T AT + 2cA^T \mathbf{1}\mathbf{t}^T - 2cA^T B + T(\Lambda + \Lambda^T) = 0 \tag{26}$$

$$\frac{\partial L}{\partial \mathbf{t}} = 2cT^T A^T \mathbf{1} - 2B^T \mathbf{1} + 2\mathbf{1}^T \mathbf{1}\mathbf{t} = 0 \tag{27}$$

$$\frac{\partial L}{\partial c} = 2c\ tr\{T^T A^T AT\} + 2tr\{T^T A^T \mathbf{1}\mathbf{t}^T\} - 2tr\{A^T BT\} = 0 \tag{28}$$

Let us say $S = A^T \left(I - \frac{\mathbf{1}\mathbf{1}^T}{\mathbf{1}^T \mathbf{1}}\right) B$. Since the matrices SS^T and $S^T S$ have the same singular values, defining

$$svd\left\{SS^T\right\} = VDV^T \tag{29}$$

and

$$svd\left\{S^T S\right\} = WDW^T \tag{30}$$

where $svd\{\}$ stands for Singular Value Decomposition, we can solve the unknown orthogonal transformation matrix T as

$$T = VW^T \tag{31}$$

Finally, the solution of the similarity transformation parameters $(c, \mathbf{t})$ can be computed by

$$c = \frac{tr\left\{T^T A^T \left(I - \frac{\mathbf{1}\mathbf{1}^T}{p}\right) B\right\}}{tr\left\{A^T \left(I - \frac{\mathbf{1}\mathbf{1}^T}{p}\right) B\right\}} \tag{32}$$

and

$$\mathbf{t} = \frac{(B - cAT)^T \mathbf{1}}{p} \tag{33}$$

For simplicity, in the calculation phase one can use the following result according to Schoenemann and Carroll (1970) between the matrices $S^T S$, SS^T and S,

$$svd\left\{S\right\} = VD_s W^T \quad ; \quad D_s \neq D \tag{34}$$

4.2.2.   Computation of centroid C

To obtain the initial centroid C, we should define one of the coordinates matrices A_n as fixed, and sequently link the others by means of the Extended Orthogonal Procrustes algorithm (EOP) instead of registering pairs of single models (Beinat and Crosilla, 2001). This process is as follows. First, A_1 is fixed as the initial model and A_2 is rotated, translated and scaled to fit this model using (EOP). Similarly, A_3 is transformed into positions of best fit with respect to $\widehat{A}_2$, the coordinates matrix A_2 after the transformation. Next, A_4 is transformed with respect to $\widehat{A}_3$, and so on. All the models $\widehat{A}_n$, $n = 1, ..., N$, are used to estimate the aproximated shape of the whole object (landmarks), which provides an initial value for the centroid C, C^0

$$C^0 = \frac{1}{N} \sum_{n=1}^{N} \widehat{A}_n \tag{35}$$

where the superscript 0 is due the initial iteration of the algorithm. Once the initial centroid is estimated, the unknown similarity transformation parameters in (20), $(T_n^0, \mathbf{t}_n^0, c_n^0)$, can be determined by means of the Extended Orthogonal Procrustes (EOP) calculation of each model point matrix A_n to the centroid C^0. Then, the centroid is iterative updated after

the calculation of each matrix $\widehat{A}_n$ until the stabilization of the centroid C, i.e. until the L2-norm between the centroid in the current iteration and the previous one is less than a fixed threshold ϵ,

$$||C^{iter} - C^{iter-1}||^2 < \epsilon \qquad (36)$$

5. Experiments

In the first example the method proposed in (16) for dimension reduction is compared to the standard eigenface technique on a gray-level database. We compare the dimensionality reduction performance when a frontal view face database is used, showing that the new technique leads to a better result for the data analyzed. In the second example we show that the proposed Procrustes analysis works well for the image registration problem.

5.1. Example 1

We use a gray-level frontal view face database that comprises 114 full-face pictures, 56 males and 58 females ($N = 114$). Each image is digitized in a gray-scale, with a resolution of 248×186, i.e. 248 rows and 186 columns ($I = 248, J = 186$). We compare the reconstruction performance of the traditional method with the new one when the number of singular values used (i.e. dimension of the subspace) increase gradually. The quality of the reconstruction, as the efficiency of representing the data by the subspace, is measured by the mean square error (MSE). The maximum dimension of the subspace using the traditional method, is $p = N$, i.e., the sample size, while with the new method is $p = min\{I, J\} = J$. Recall that the number of singular values chosen determines the dimensionality of the subspace used to reconstruct the face images. In Figure 1 we plot the average reconstruction error ($AMSE$) for the training sample when the number of estimated parameters k increase as a function of the number of singular values used, p. Let $k^{St} = IJN(\frac{p}{N} + \frac{p}{IJ})$ and $k^{New} = IJN(\frac{p}{J} + \frac{p}{I})$ be the number of parameters used in the reconstruction by the standard method and the new one. For simplicity, we only consider $p = 1, ..., 40$. Figure 1 is a 3D graph, in which each point has three coordinates, $(x, y, z) = (k, AMSE, p)$. Thus, when the number of singular values are fixed, the x-axis represents the amount of parameters needed to reconstruct the image, and the average mean square error (AMSE) in the reconstruction is computed (y-axis). The upper plotted points correspond to the singular values used by the standard

method, and the lower points are the ones used by the proposed method. This graph demostrates that the quality of the reconstruction by the new procedure is better than the traditional one. If we compare the number of dimensions needed to reconstruct the image with a fixed amount of average mean square error, for instance, $AMSE \approx 200$, we need to use $p = 6$ singular values with the new procedure, against the 28 needed with the older one, which increase the number of estimated parameters substantially. To visualize in more detail the performance of the reconstruction by both methods, Figure 2 gradually shows the reconstruction of one individual of the sample when the number of singular values is $p = 5, 10, 20$ and 50. Its reconstruction accuracy is measured by the MSE.

These figures clearly demostrate that when the dimensionality of the subspace is the same, the new method always perform better than the standard eigenface technique. In order to further analyze the results depicted in Figure 2, we compare the distances between pairs of reconstructed images in $\Re^p$ (low-dimensional subspace) with the corresponding distances in the original high-dimensional space. In Table 1 we show the average L1-norm between original and projected images, when the dimensionality of the subspace increases from 5 to 50. The L1-norm is defined as,

$$||\delta - \widehat{\delta}||^1 = \sum_{n=1}^{N} \left| \delta_i - \widehat{\delta}_i \right| \tag{37}$$

where $m = \frac{n(n-1)}{2}$ indicates all the combinations of the N elements taken by pairs, and δ_i is the euclidean distance between the original $i - th$ pair in the sample, $i = 1, ..., m$. The distances between the reconstructed $i - th$ pair by the standard and the new method are denoted by $\widehat{\delta}_i^{St}$ and $\widehat{\delta}_i^{New}$, respectively. Figure 3 shows the L1-norm between distances when the reconstruction has been done by the standard method and by the new one, when the subspace of projection increases from 1 to 50 (horizontal axis). It can be observed that the proposed method provides an important improvement in comparison with the standard method in replicating the original distances between observations.

5.2. Example 2

In this example, we will show that the proposed image registration procedure is more effective than the affine transformation. For this purpose, we will register the face database used in example 1 in order to work with normalized objects. We choose as control points (landmarks) the coordinates associated to the left and right eyes and the end point of the
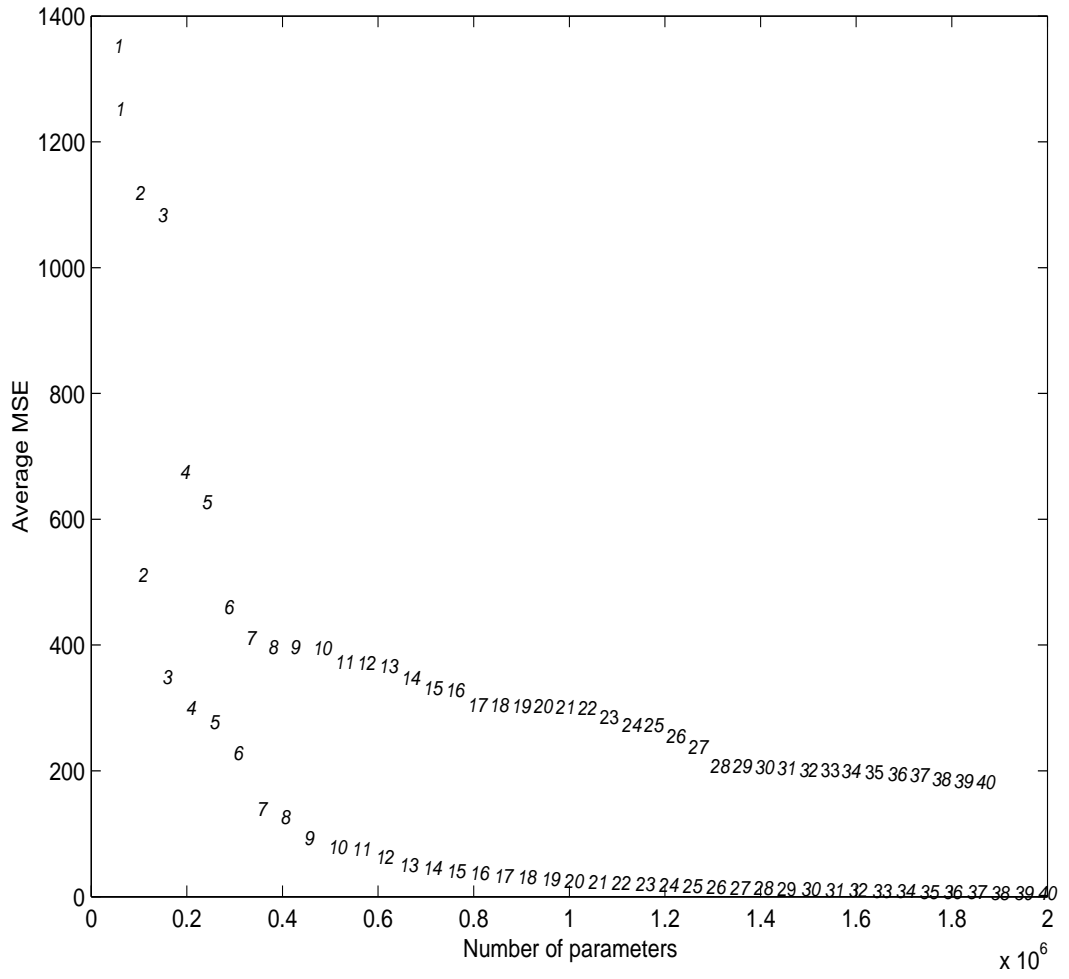
Fig. 1. Comparision of the average mean square error between eigenface method (upper points) and the proposed method (lower points) when the number of singular values used increases from 1 to 40.

p=5, MSE=628   p=5, MSE=277

p=10, MSE=396   p=10, MSE=80

p=20, MSE=303   p=20, MSE=26

p=50, MSE=134   p=50, MSE=3

Fig. 2. Image Reconstruction by means of the standard method (left panels) and by the new method (right panels) using $p = 5, 10, 20$ and 50 singular values

Table 1. Average L1-norm be-
tween pairs of original and re-
constructed images when the sub-
space increase from 5 to 50

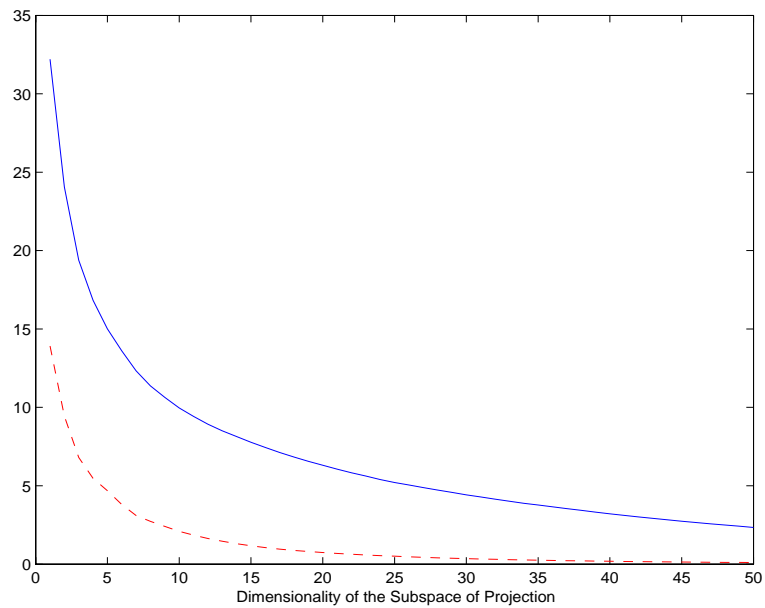| p	$\|\|\delta - \widehat{\delta}_i^{St}\|\|^1$	$\|\|\delta - \widehat{\delta}_i^{New}\|\|^1$
5	14.99	4.92
10	9.86	2.24
20	5.74	0.75
30	3.68	0.36
50	1.52	0.10

Fig. 3. L1-norm between original and reconstructed images by the standard method (continuous line) and by the new one (dash line)

Table 2. Model Coordinates Matrices in the Sample

$Landmarks$	A_1	A_2	A_3	A_4	A_5
left eye	$(124, 74)$	$(121, 70)$	$(123, 65)$	$(127, 71)$	$(123, 74)$
right eye	$(124, 130)$	$(120, 129)$	$(121, 125)$	$(125, 126)$	$(123, 126)$
chin	$(232, 107)$	$(229, 105)$	$(234, 100)$	$(221, 102)$	$(229, 98)$
$Landmarks$	A_6	A_7	A_8	A_9	A_{10}
left eye	$(121, 63)$	$(118, 66)$	$(115, 63)$	$(126, 63)$	$(125, 68)$
right eye	$(122, 123)$	$(116, 125)$	$(115, 123)$	$(123, 117)$	$(124, 124)$
chin	$(231, 93)$	$(234, 100)$	$(228, 91)$	$(228, 91)$	$(229, 101)$

chin. Thus, each image X_n has associated to a coordinate matrix A_n, $n = 1, ..., N$, where $A_n \in \Re^{3 \times 2}$. The accuracy locating these points is critical because they are used to estimate the transformation parameters. For simplicity, we only consider in this example $N = 10$ and in Table 2 we show the landmarks selected. In order to solve the normal equation system (19), we need to fix the output coordinates, B, to fit the sample matrices A_n, $n = 1, ..., N$. We will use the average coordinates in the sample,

$$B = \frac{1}{N} \sum_{n=1}^{N} A_n = \left[\begin{array}{c} (121, 68) \\ (120, 125) \\ (228, 99) \end{array} \right] \tag{38}$$

For the procrustes analysis, we fix the threshold ϵ to stop the algorithm at 0.001. The least squares estimation of the true coordinates defined in a common coordinate system are,

$$C = \left[\begin{array}{c} (125, 74) \\ (124, 131) \\ (231, 106) \end{array} \right] \tag{39}$$

As an illustration, Figure 4 shows the solution of the registration problem for the $10 - th$ image in the sample. Comparing the coordinate matrix A_{10} with the target B, the image has to be moved down. The left panel in Figure 4 shows the original image. The middle panel shows the image registration by means of the affine transformation and the right panel by means of the procrustes analysis. Notice that while in the middle panel the classical affine transformation procedure deforms the original image, in the left image the procrustes algorithm perfectly reproduces the image.

Fig. 4. Image Registration of one individual in the sample

References

Barlett, M.S. and Sejnowski, T.J. (1997). Independent components of face images: a representation for face recognition. *Proceedings of the Fourth Annual Joint Symposium on Neural Computation, CA.*

Beinat, A. and Crosilla, F. (2001). Generalized Procrustes analysis for size and shape 3-D object reconstruction. *Optical 3-D Measurements Techniques V, Viena*, pp. 345-353.

Bierens, H.J. and Hartog, J. (1988). Non-linear regression with discrete explanatory variables, with an application to the earnings function. *Journal of Econometrics,* **38**, pp.269-299

Christensen, R. (1991). *Linear Models for Multivariate, Time Series and Spatial Data.* New York: Springer-Verlag.

Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal Royal Statistical Society Series B*, **53**, pp.285-339

Gower, J.C. (1975). Generalized Procrustes analysis. *Psychometrika*, **40**, pp.33-51

Kirby, M. and Sirovich, L. (1990). Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Machine Intell.*, **12**, pp.103-108

Krzanowski, W.J. and Marriot, F.H.C. (1994). *Multivariate Analysis: Distributions, ordination and inference.* UK: Edward Arnold.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1992). *Multivariate Analysis.* CA: Academic Press.

Nadenau, M. (1997). Localization and Shape Normalization for Face Recognition. Available from: *http://dewww.epfl.ch/˜nadenau/Research/Paper/Diploma.pdf).*

Schoenemann, P.H. and Carroll, R. (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika*, **35**, pp.245-255

Swets, D. and Weng, J. (1996). Using Discriminant Eigenfeatures for Image Retrieval. *Tech. Rep.*

Theobald, C.M., Glasbey, C.A., Horgan, G.W. and Robinson, C.D. (2004). Principal component analysis of landmarks from reversible images. *Journal Royal Statistical Society Series C*, **53**, pp.163-175.

Turk,M. and Pentland, A. (1999). Face recognition using Eigenfaces. *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pp.586-591.

Valentin, D., Abdi, H. and O'Toole, A. (1996). Principal Component and Neural Network Analysis of face images. *Progress in Mathematical Psycology.*

Wu, J. and Zhou, Z. (2002). Face Recognition with one training image per person. *Pattern Recognition Letters*, **23**, pp.1711-1719

Yang, J. and Yang, J. (2002). From Image vector to matrix: a straightforward image projection technique. *Pattern Recognition*, **35**, pp.1997-1999