Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 38 (2005) 2400–2408
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

A fast approach for dimensionality reduction with image data

Mónica Benito[a,*,1], Daniel Peña[b]

[a]*Department of Statistics, Universidad Carlos III de Madrid, Spain*
[b]*Department of Statistics, Universidad Carlos III de Madrid, Spain*

Received 9 August 2004; received in revised form 14 March 2005; accepted 14 March 2005

Abstract

An important objective in image analysis is dimensionality reduction. The most often used data-exploratory technique with this objective is principal component analysis, which performs a singular value decomposition on a data matrix of vectorized images. When considering an array data or tensor instead of a matrix, the high-order generalization of PCA for computing principal components offers multiple ways to decompose tensors orthogonally. As an alternative, we propose a new method based on the projection of the images as matrices and show that it leads to a better reconstruction of images than previous approaches.

Keywords: Eigenfaces; Multivariate linear regression; N-mode PCA; Principal component analysis; Singular value decomposition; Tensors

1. Introduction

Exploratory image studies generally aim at data inspection and dimensionality reduction. The algebra of matrices has been employed successfully in the context of image analysis. In particular, one of the most popular approaches to reduce dimensionality and derive useful compact representations for image data is principal component analysis (PCA). In 1990 Ref. [1] proposed to use PCA to reduce the dimensionality when representing human faces. In the last two decades, PCA has been very popular in the object recognition community, [2–4]. The performance of this method on aligned and scaled human faces is very good, but it does not work well for non-aligned faces. In particular, when external factors, such as lighting, viewpoint and expression, are permitted to modify facial images, more sophisticated mathematical approaches are needed. A refinement of PCA for face representation is independent component analysis (ICA), see Ref. [5]. It is shown in Ref. [6] that a pre-processing step of the image sample improves the PCA performance. PCA-based algorithms take advantage of the functionality and simplicity of matrix algebra. As was pointed out in Ref. [7], the natural representation of a collection of images is a three-dimensional array, or third-order tensor, rather than a matrix of vectorized images. In addressing the problem of dimensionality reduction with array data, the multilinear algebra, the algebra of higher-order tensors, offers a powerful mathematical framework for analyzing the multifactor structure of images that can account explicitly for each of the multiple factors inherent to image information. In this sense Ref. [8] proposed a multilinear modeling technique employing a tensor extension of the conventional matrix singular value decomposition (SVD), known as the N-mode SVD. However, for high-dimension tensors one can find computational problems because this approach

* Corresponding author. Tel.: +34 91 624 9859; fax: +34 91 624 9849.

[1] Mónica Benito was supported by FPU grant of the Ministry of Education of Spain. She is very grateful to Steve Marron for his hospitality and insightful comments during the time she was working in this research at the University of North Carolina at Chapel Hill.

0031-3203/$30.00
doi:10.1016/j.patcog.2005.03.022

involves solving large minimization problems and in such circumstances the resulting matrices are often unfeasible to compute because of its high dimensionality. Hence, we are interested in finding a fast and feasible method to solve the problem of dimensionality reduction with high reconstruction quality rates when dealing with such high-dimensional data. In this paper, we propose an alternative method in which the original data is projected on a subspace of lower dimension keeping the internal structure of the matrices in the projection. 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 nth 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, $X_1, X_2, \ldots, X_N$.Each matrix consists of elements x_{ij}, with $i=1,\ldots,I$ and $j=1,\ldots,J$, 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 the black color, and the value 255 the white color. For convenience, we normalize the pixel intensities, so $x_{ij} \in [0, 1]$. 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 components. In Section 3, we explain the high-order generalization of PCA (N-mode PCA) and the singular value decomposition (N-mode SVD) of matrices for computing principal components. In Section 4, we propose a new approach which keeps the internal structure of the images. Finally, in Section 5 we report on the experimental results using a human face database.

2. Principal component analysis

Suppose that each matrix X_n, $n=1,\ldots,N$, is transformed into a vector $\mathbf{x}_n^{\mathrm{T}}$ by row (or column) concatenation. Therefore, we have a set of N vectors in a high-dimensional space, specifically, $\mathbf{x}_n^{\mathrm{T}} \in \Re^d$, where $d=I\times J$, $n=1,\ldots,N$. For convenience, the vectors are assumed to be normalized, so that $\sum_{n=1}^{N} \bar{\mathbf{x}}_n = 0$ and $\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n = 1$. Note that this set of vectors can be represented by an $N \times d$ matrix X in which the nth row is equal to $\mathbf{x}_n^{\mathrm{T}}$. 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, $\widehat{\mathbf{x}}_n$. The purpose of PCA is to find p $(p<d)$ standardized linear combinations of the original variables $X\mathbf{u}_1, X\mathbf{u}_2, \ldots, X\mathbf{u}_p$ which are uncorrelated and have maximum variance. This procedure is illustrated in Fig. 1. We have observations $\mathbf{x}_n$ in $\Re^2$ and the first principal component $\mathbf{u}_1$ defines the direction where the variance of the projected points, $z_n = \mathbf{x}_n^{\mathrm{T}}\mathbf{u}_1$, is maximum. The second principal component must be orthogonal to the first one, and is given by $\mathbf{u}_2$. In general, the principal components are given by (see Ref. [9] for more details)

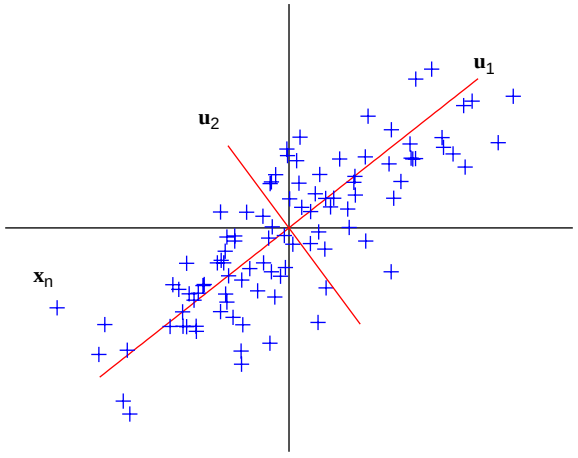

Fig. 1. Scatter plot in a bivariate gaussian sample, where u1 and u2 are the two first directions of maximum variability.

$$Z = XU_p, \tag{1}$$

where U_p is an $N \times p$ matrix with columns vectors $\mathbf{u}_i$, $i=1,\ldots,p$, which are the eigenvectors linked to the pth largest eigenvalues of the matrix $X^{\mathrm{T}}X$. The optimal prediction $\widehat{X}$ 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 $p \leqslant r$ dimensions is given by

$$\widehat{X} = XU_pU_p^{\mathrm{T}}, \tag{2}$$

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

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

Let

$$X = V_rG_rU_r^{\mathrm{T}} \tag{4}$$

be the singular value decomposition of the matrix X, where V_r contains the eigenvectors of XX^{T} linked to non-null eigenvalues, U_r contains those of $X^{\mathrm{T}}X$ and G_r is a diagonal matrix with the squared root of the non-null eigenvalues of $X^{\mathrm{T}}X$. Then Eq. (2) can be also be written as

$$\widehat{X} = V_pG_pU_p^{\mathrm{T}}, \tag{5}$$

where V_p and U_p include $p<r$ eigenvectors and G_p is diagonal and contains the p largest eigenvalues of $X^{\mathrm{T}}X$. Note that in order to predict the N images, which implies a

total of IJN scalar predictions, we only need the p vectors $\mathbf{u}_i$ of dimension IJ and the p scalar values $z_i=(\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i)=(\mathbf{v}_i g_i^{1/2})$ for each image, which implies

$$pIJ + pN = IJN\left(\frac{p}{N} + \frac{p}{IJ}\right) = IJNc_1, \tag{6}$$

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 reconstruction of the images can be very important. This decomposition was used for image dimension reduction in Ref. [2], and it often referred to as the eigenface method. It is the most common approach for dimension reduction with images.

3. High-order generalization of PCA

3.1. Multi-way models

We define multi-way data as data with elements arranged as $x_{ijk\ldots}$, $i=1,\ldots,I$; $j=1,\ldots,J$; $k=1,\ldots,K$; $\ldots$ where the number of indices may vary. Each index is called way or mode, and the number of levels in the mode is called the dimension of that mode. The number of ways is the geometrical dimension of the array. In particular, when dealing with a sample of N images with size $I\times J$, we have three indices and the data can be geometrically arranged in a box of dimension $I\times J\times N$, where index i refers to a row-mode, index j refers to a column-mode and index k refers to a tube-mode. For expressing models using algebra notation we can rearrange a multi-way array to a matrix. This transformation is known as unfolding. For instance, the $I\times J\times N$ array $\underline{X}$ can be unfolded to an $I\times JN$ matrix X.

When considering the case of three-way analysis, a decomposition of the array is made into trilinear components. In 1966 Tucker [10] developed a series of three-way models called three-mode PCA. In the context of image data, when analyzing a collection of N images we are interested in reducing the two first modes (i.e., the index i and the index j) leaving the third mode k uncompressed. Generalizing the singular value decomposition of a matrix we can write the following model for a given image k:

$$\widehat{X}_k = AH_kB^{\mathrm{T}}; \quad k=1,\ldots,N, \tag{7}$$

where A and B are $I\times J$ matrices with orthonormal columns, so that $AA^{\mathrm{T}}=I$ and $BB^{\mathrm{T}}=I$ (the subindex k is due the observation in the sample), and H_k is a squared matrix. Note that the matrices A and B are the same for all the images, whereas the matrix H_k depends on the image. This model can be seen as a weighted sum of outer products where the weight of the outer product between the ith component from A and the jth component from B for the kth image is determined by the element h_{ijk}. A particular case of the three-way model assuming a model of rank D $(D<I)$ in the first mode and E $(E<J)$ in the second mode, is the Tucker2 model [11], which generalizes Eq. (7). Calling X to the array $\underline{X}$ arranged as an $N\times IJ$ matrix, we have

$$\widehat{X} = H(A\otimes B)^{\mathrm{T}}, \tag{8}$$

where H is the array $\underline{H}=H_1\times H_2\times\cdots\times H_N$ arranged as an $N\times DE$ matrix. The $I\times D$ matrix A and the $J\times E$ matrix B are the loading or component matrices for each mode, respectively. This expression is equivalent to the above defined in Eq. (7). The estimation of this model is achieved via an alternating least-squares (ALS) algorithm which minimizes the loss function (see Ref. [12])

$$\min_{\mathrm{A,B,H}_k} \sum_{k=1}^{N} \|X_k - AH_kB^{\mathrm{T}}\|_F^2 \tag{9}$$

for the unknown matrices A, B and H_k, where $\|.\|_F^2$ stands for the Frobenious norm. Thus,

$$\min_{\mathrm{A,B,H}_k} \sum_{k=1}^{N} \|X_k - \widehat{X}_k\|_F^2 = \min_{\mathrm{A,B,H}_k} \sum_{k=1}^{N} trace[(X_k-\widehat{X}_k)^{\mathrm{T}}(X_k-\widehat{X}_k)]. \tag{10}$$

Using the expression (8) this minimization problem is equivalent to

$$\min_{\mathrm{A,B,H}} \|X - H(A\otimes B)^{\mathrm{T}}\|_F^2. \tag{11}$$

ALS consists of dividing the parameters into several sets in which each set is estimated in a least-squares sense conditionally on the remaining parameters. The estimation is repeated iteratively until no change is observed in the parameter values or in the fit of the model to the data. Next, we will explain an algorithm, the ALS algorithm, to solve this problem.

3.2. ALS algorithm

Each subproblem in Eq. (9) can be solved easily given the matrices A and B. From Eq. (7) the core matrix H_k can be found by

$$\widehat{H}_k = A^{\mathrm{T}}X_kB; \quad k=1,\ldots,N. \tag{12}$$

From the definition of H_k and using Eq. (12) it follows that $\widehat{X}_k$ can be expressed as follows:

$$\widehat{X}_k = AA^{\mathrm{T}}X_kBB^{\mathrm{T}}. \tag{13}$$

For a given B it follows that finding the optimal A is equal to minimizing the norm of $(X_k-AA^{\mathrm{T}}M)$, where $M=X_kBB^{\mathrm{T}}$. The problem we have to solve is

$$\min_A \|X_k - AA^{\mathrm{T}}M\|_F^2 = \min_A [trace\{(X_k-AA^{\mathrm{T}}M)(X_k-AA^{\mathrm{T}}M)^{\mathrm{T}}\}]; \quad k=1,\ldots,N. \tag{14}$$

Using the properties of the trace, the above expression is equivalent to

$$\min_A [trace\{X_k X_k^T\} + trace\{AA^T MM^T AA^T\} - 2trace\{AA^T MX_k^T\}], \quad (15)$$

and this is equivalent to minimizing

$$F = trace\{AA^T MM^T AA^T\} - 2trace\{AA^T MX_k^T\}. \quad (16)$$

Using that $MM^T = MX_k^T$ and using the properties of the trace, we have

$$F = trace\{A^T MX_k^T A\} - 2trace\{A^T MX_k^T A\} = -trace\{A^T MX_k^T A\}, \quad (17)$$

leading to the maximization problem

$$\max_A [trace\{A^T MX_k^T A\}] = \max_A [trace\{A^T MM^T A\}]. \quad (18)$$

Hence the optimal A is the eigenvectors linked to the D largest eigenvalues of MM^T. The algorithm now uses this matrix A in Ref. [3] to obtain a new matrix B. This is iterated until convergency. Each iteration solves an eigenvalue–eigenvector problem of a dimension equal to the number of components for the mode in question. Finally, the core matrix H_k is estimated by Eq. (12) and the unfolded X_k as defined in Eq. (13) obtaining the reconstructed array $\underline{\widehat{X}}$. If the model has been fitted successfully, the core array $\underline{\widehat{H}}$ contains exact the same information as $\underline{X}$ merely expressed using different coordinates. The ALS algorithm for fitting the model is among the fastest of the multi-way algorithms. However, for large problems the fitting procedure requires increasing computational efforts.

To start the algorithm several kinds of initializations has been proposed; Ref. [13] advocates for choosing B at random and Ref. [14] suggests to choose B from an SVD on the $J \times I$ matrix X_k^T.

Assuming equal number of factors for the first and second modes, i.e., $D = E = p$, the reconstruction of the array requires the matrix A, B and H, which leads a total of $Ip + Jp + Np^2$ parameters.

4. An alternative approach based on matrix projections

We are interested in a projection method which keeps the matrix structure of the image. Here, we follow a similar approach to Ref. [15], which proposed the projection of the rows of the matrix in the context of feature extraction. Assume without loss of generally 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

$$\mathbf{w}_n = X_n \mathbf{a}. \quad (19)$$

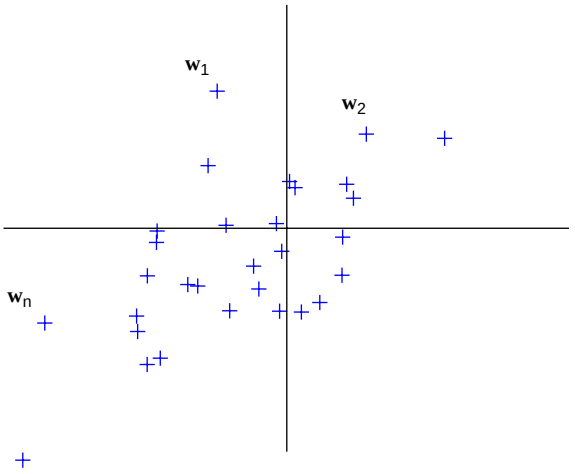

Fig. 2. Scatter plot of the images projected in some direction $\mathbf{a}$.

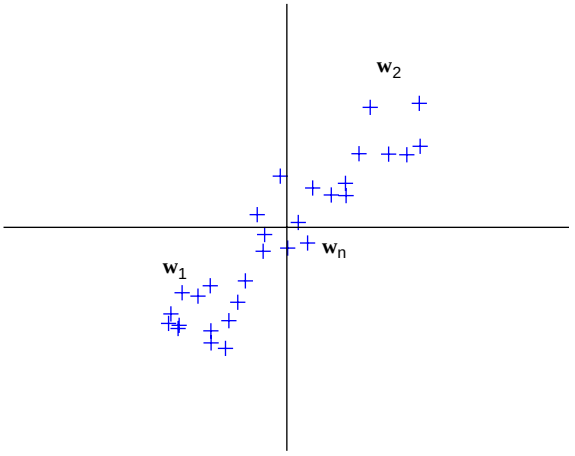

Fig. 3. Scatter plot of the images projected in another direction $\mathbf{a}$.

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, $\mathbf{w}_n$, $n = 1, \ldots, N$. The idea is to find a direction of projection in such a way that these set of vectors $\mathbf{w}_n$ are as separated as possible from the rest. In essence, we try to find a direction $\mathbf{a}$ such that when projecting the rows of the nth image, $\mathbf{f}_{n1}^T, \mathbf{f}_{n2}^T, \ldots, \mathbf{f}_{nI}^T$, where $\mathbf{f}_{ni}^T \in \Re^J$, $i = 1, \ldots, I$, the vector obtained, $\mathbf{w}_n = [\mathbf{f}_{n1}\mathbf{a}\ \mathbf{f}_{n2}\mathbf{a} \cdots \mathbf{f}_{nI}\mathbf{a}]$, is as different as possible from the other vectors $\mathbf{w}_k$, $k \neq n$. This is illustrated in Figs. 2 and 3. Suppose, for instance, that we have images with only two rows ($I = 2$), and $\mathbf{w}_n \in \Re^2$, $n = 1, \ldots, N$. The projection direction which leads to Fig. 2 provides larger variability among the vectors $\mathbf{w}_n$ than the one which leads to Fig. 3. We want to find a direction $\mathbf{a}$ which maximizes the sum of the variances of the components of these vectors, $\mathbf{w}_n$.

In order to find a good projection direction, let us call S_r to the $I \times I$ covariance matrix for these vectors representing the rows (the subindex r is due the projection of the rows; we will discuss later the projection of the columns). This matrix is given by

$$S_r = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{w}_n - \overline{\mathbf{w}})(\mathbf{w}_n - \overline{\mathbf{w}})^{\mathrm{T}}, \tag{20}$$

where $\overline{\mathbf{w}}$ is the mean of the projected images. 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

$$\max \mathit{trace}\{S_r\} = \max \mathit{trace} \left\{ \frac{1}{N} \sum_{n=1}^{N} (\mathbf{w}_n - \overline{\mathbf{w}})(\mathbf{w}_n - \overline{\mathbf{w}})^{\mathrm{T}} \right\}, \tag{21}$$

and using the definition (19),

$$\mathit{trace}\{S_r\} = \mathit{trace} \left\{ \frac{1}{N} \sum_{n=1}^{N} (X_n\mathbf{a} - \overline{X}\mathbf{a})(X_n\mathbf{a} - \overline{X}\mathbf{a})^{\mathrm{T}} \right\}, \tag{22}$$

where $\overline{X} = \frac{1}{N} \sum_{n=1}^{N} X_n$ denotes the mean image. Using the properties of the trace operator,

$$\max \mathit{trace}\{S_r\} = \max \frac{1}{N} \times \mathit{trace} \left\{ \mathbf{a}^{\mathrm{T}} \left[\sum_{n=1}^{N} (X_n - \overline{X})(X_n - \overline{X})^{\mathrm{T}} \right] \mathbf{a} \right\}, \tag{23}$$

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

$$\Sigma_c = \frac{1}{N} \sum_{n=1}^{N} (X_n - \overline{X})^{\mathrm{T}}(X_n - \overline{X}); \quad \Sigma_c \in \Re^{J \times J}. \tag{24}$$

As we need more than one direction of projection to characterize the sample, we compute the set of eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$, which constitute a basis for $\Re^p$ from which the data can be estimated using a subspace of lower dimension, $p \leqslant \min\{I, J\}$.

Similarly, the same result 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

$$\mathbf{z}_n = X_n^{\mathrm{T}}\mathbf{b}, \tag{25}$$

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

$$S_c = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{z}_n - \overline{\mathbf{z}})(\mathbf{z}_n - \overline{\mathbf{z}})^{\mathrm{T}}, \tag{26}$$

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

$$\Sigma_r = \frac{1}{N} \sum_{n=1}^{N} (X_n - \overline{X})(X_n - \overline{X})^{\mathrm{T}}; \quad \Sigma_r \in \Re^{I \times I}. \tag{27}$$

Note that the trace of S_r and the trace of S_c are the same.

4.1. Prediction by multivariate regression

Let W_n be the feature vectors obtained as the solution of Eq. (21),

$$W_n = [X_n\mathbf{a}_1, \ldots, X_n\mathbf{a}_p] = X_n A_p; \quad W_n \in \Re^{I \times p}. \tag{28}$$

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

$$X_n = W_n\beta_n + \varepsilon_n, \tag{29}$$

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

$$\widehat{X}_n = W_n \widehat{\beta}_n. \tag{30}$$

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

$$N(Ip + pJ) = IJN\left(\frac{p}{J} + \frac{p}{I}\right) = IJNc_2, \tag{31}$$

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.

Alternatively, we can take into account the array structure of the data to reconstruct all the samples using a common matrix β in the regression model. Let $\underline{X}$ be the $I \times J \times N$ array of images and $\underline{W}$ be the $I \times p \times N$ array of features defined in Eq. (28). Thus, the following regression model is stated, which reduce the number of parameters to estimate

$$\begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_N \end{bmatrix} = \begin{bmatrix} W_1 \\ W_2 \\ \ldots \\ W_N \end{bmatrix} \beta + \varepsilon, \tag{32}$$

or similarly,

$$X = W\beta + \varepsilon, \tag{33}$$

where X is the unfolded $IN \times J$ array $\underline{X}$ and W is the unfolded $IN \times p$ array $\underline{W}$. The least-squares estimation of this model is given by

$$\widehat{\beta} = (W^{\mathrm{T}}W)^{-1}W^{\mathrm{T}}X. \tag{34}$$

Finally, each matrix X_n is reconstructed using its feature vectors W_n and the matrix of parameters $\beta \in \Re^{p\times J}$,

$$\widehat{X}_n = W_n\widehat{\beta}; \quad n = 1, \ldots, N, \tag{35}$$

leading to $NIp + pJ$ parameters. Although using a common matrix β will decrease the goodness, the number of parameters is reduced considerably.

Similarly, the same result is obtained if we start projecting the columns instead of the rows in Eq. (19). 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 $G_n = Z_n(Z_n^{\mathrm{T}}Z_n)^{-1}Z_n^{\mathrm{T}}$ is used to reconstruct each image. Let $r = \max\{I, J\}$. We suggest to use the projection by rows when $r = I$ and alternatively, 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.

5. Experiments

To illustrate the method proposed in Eq. (35) we compare the proposed method with the standard eigenface technique and the N-way PCA. The data will be a frontal view face database.

As we stated earlier, when dealing with a set of homogeneous objects, image formation depends on scene geometry, viewpoint and illumination conditions, which greatly increase the difficulty of the reconstruction task. Since the image sample can be seen as a set of shapes with respect to a local 2D coordinate 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. In the next paragraph, we propose a registration method based on the Procrustes analysis theory as a prior step to the dimensionality reduction task.

5.1. Image registration

Depending on the complexity of the object it may require two or more viewpoints (or landmarks) to register it appropriately. Procrustes analysis theory is a set of mathematical tools to directly estimate and perform simultaneous similarity transformations among the object 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 nth image, $n = 1, \ldots, N$. 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:

$$\widehat{A}_n = c_n A_n T_n + \mathbf{1}\mathbf{t}_n^{\mathrm{T}}; \quad n = 1, \ldots, N, \tag{36}$$

where c_n is the scale factor, T_n is 2×2 orthogonal rotation matrix, $\mathbf{t}_n$ is a 2×1 translation 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 (see Ref. [16]) provides least-squares correspondence of more than two point matrices. According to Ref. [17] 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 as the search of the unknown optimal matrix B. Defining C as the geometrical centroid of the transformed matrices $\widehat{A}_1, \ldots, \widehat{A}_N$, the solution of the registration problem is achieved by using the following minimum condition:

$$\sum_{n=1}^{N} trace\{(\widehat{A}_n - C)^{\mathrm{T}}(\widehat{A}_n - C)\} \tag{37}$$

in an iterative computation scheme of centroid C. 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 $(T_n, \mathbf{t}_n, c_n)$, $n = 1, \ldots, N$, are determined using the extended orthogonal procrustes (EOP) procedure for fitting two given set of points, A_n and $\widehat{B}$ (see Ref. [18]).

5.2. Example

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 choose as control points (landmarks) to register the images the coordinates associated with the left and right eyes and the end point of the chin. Thus, each image X_n has associated a coordinate matrix A_n, $n = 1, \ldots, 114$, where $A_n \in \Re^{3\times 2}$. We compare the reconstruction performance of the traditional method called eigenface (PCA) and the three-way PCA with the new one when the dimension of the subspace increases gradually. The quality of the reconstruction is measured by the mean squared error (MSE).

In Fig. 4, we plot the average reconstruction error (AMSE) for the training sample when the dimensionality of the subspace in the projection increases as a function of the number of singular values used, p. For simplicity, we only consider

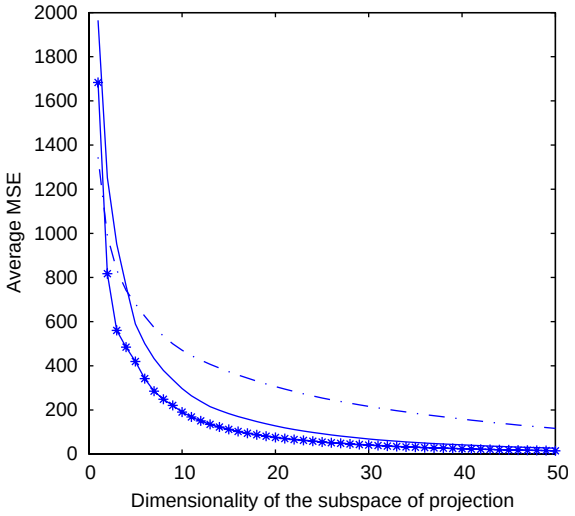

Fig. 4. Comparison of the average mean square error between the eigenface method (dash line), three-way PCA (solid line) and the proposed method (start points) when the number of singular values increases from 1 to 50.

$p = 1, \ldots, 50$. The dash line correspond to the singular values used by the standard method, the solid line is used by the three-way PCA and the line with start points corresponds to the new method. This graph shows that the quality of the reconstruction by the new procedure is better than by the others (Fig. 4).

To see in more detail the performance of the reconstruction by the three methods, Figs. 5–8 show gradually the reconstruction of one individual in the sample when the number of singular values increases, $p = 5, 10, 20$ and 50. The reconstruction accuracy is measured by the MSE.

These figures clearly show that when the dimensionality of the subspace is the same, the new method performs better than the others. In order to further analyze these results, 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. Table 1 shows the average L1-norm between original and projected images, given by

$$\|\delta - \widehat{\delta}\|^1 = \sum_{i=1}^{m} |\delta_i - \widehat{\delta}_i|, \tag{38}$$

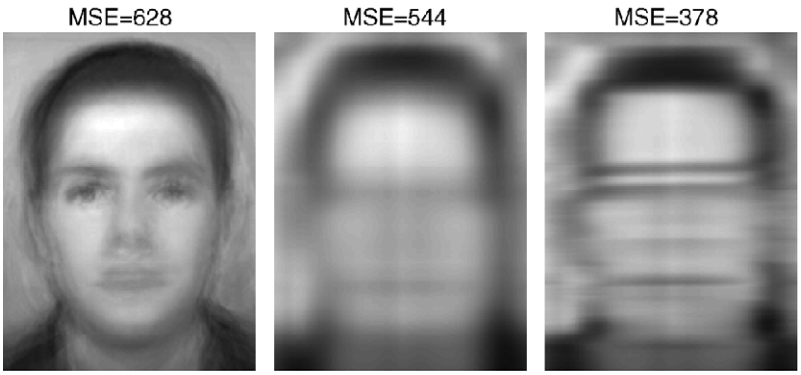

Fig. 5. Image reconstruction by means of the standard method (left panel), the three-way PCA (middle panel) and by the new method (right panel) using $p = 5$ singular values.

Fig. 6. Image reconstruction by means of the standard method (left panel), the three-way PCA (middle panel) and by the new method (right panel) using $p = 10$ singular values.

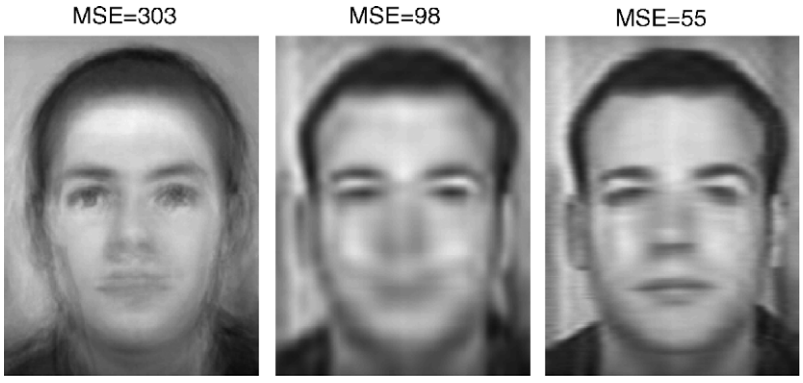

Fig. 7. Image reconstruction by means of the standard method (left panel), the three-way PCA (middle panel) and by the new method (right panel) using $p = 20$ singular values.

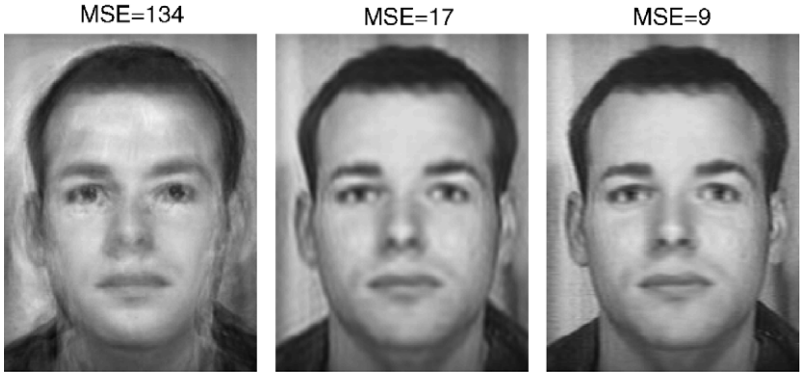

Fig. 8. Image reconstruction by means of the standard method (left panel), the three-way PCA (middle panel) and by the new method (right panel) using $p = 50$ singular values.

Table 1
Average L1-norm between pairs of original and reconstructed images when the subspace increases from 5 to 50

p	$\|\delta - \widehat{\delta}^{Pca}\|^1$	$\|\delta - \widehat{\delta}^{3Pca}\|^1$	$\|\delta - \widehat{\delta}^{New}\|^1$
5	14.99	10.94	4.92
10	9.86	6.19	2.24
20	5.74	2.19	0.75
30	3.68	1.01	0.36
50	1.52	0.35	0.10

when the dimensionality of the subspace increases from 5 to 50. In this equation $m = N(N-1)/2$ is the total number of pairs from the N elements, δ_i is the Euclidean distance between the elements in the ith pair in the sample, $i = 1, \ldots, m$, and $\widehat{\delta}_i$ is the estimated distance by some reconstruction method. The distances between the reconstructed ith pair by the standard PCA, the three-way PCA and the new method are denoted by $\widehat{\delta}_i^{Pca}$, $\widehat{\delta}_i^{3Pca}$ and $\widehat{\delta}_i^{New}$, respectively.

It can be observed that the proposed method gives an important improvement in comparison to the others in replicating the original distances between images. This property is important in applications where the analysis is based on distances between objects, as in multidimensional scaling or discriminant analysis applications.

References

[1] M. Kirby, L. Sirovich, Application of the Karhunen–Loeve procedure for the characterization of human faces, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1990) 103–108.
[2] M. Turk, A. Pentland, Face recognition using eigenfaces, in: Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition, 1999, pp. 586–591.
[3] D. Valentin, H. Abdi, A. O'Toole, Principal component and neural network analysis of face images, Progress in Mathematical Psycology, 1996.
[4] D. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, Technical Report, 1996.

[5] M.S. Barlett, T.J. Sejnowski, Independent components of face images: a representation for face recognition, in: Proceedings of the Fourth Annual Joint Symposium on Neural Computation, CA, 1997.
[6] J. Wu, Z. Zhou, Face recognition with one training image per person, Pattern Recog. Lett. 23 (2002) 1711–1719.
[7] A. Shashua, A. Levin, Linear image coding for regression and classification using the tensor-rank principle, in: Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition, 2001.
[8] M. Vasilescu, Terzopoulus, Multilinear analysis of image ensembles: TensorFaces, ECCV (2002) 447–460.
[9] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, CA, 1992.
[10] L.R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika 31 (1966) 279–311.
[11] P. Kroonenberg, J. Leeuw, Principal component analysis of three-mode data by means of alternating least squares algorithms, Psychometrika 45 (1) (1980) 69–97.
[12] P. Kroonenberg, Three-Mode Principal Component Analysis. Theory and Applications, DSWO Press, Leiden, 1983.
[13] R.A. Harshman, M.E. Lundy, The PARAFAC model for three-way factor analysis and multidimensional scaling, in: H.G. Law, C.W. Snyder, J.A. Hattie, R.P. McDonald (Eds.), Research Methods for Multimode Data Analysis, New York, 1984.
[14] P.M. Kroonenberg, J.M.F. Ten Berge, P. Brouwer, H.A.L. Kiers, Gram-Schmidt versus Bauer–Rutishauser in alternating least squares algorithms for three-mode principal component analysis, Comp. Stat. Quart. 2 (81) (1989).
[15] J. Yang, J. Yang, From Image vector to matrix: a straightforward image projection technique, Pattern Recog. 35 (2002) 1997–1999.
[16] J.C. Gower, Generalized procrustes analysis, Psychometrika 40 (1975) 33–51.
[17] C. Goodall, Procrustes methods in the statistical analysis of shape, J. Roy. Statist. Soc. Ser. B 53 (1991) 285–339.
[18] P.H. Schoenemann, R. Carroll, Fitting one matrix to another under choice of a central dilation and a rigid motion, Psychometrika 35 (1970) 245–255.

Further Reading

[19] R. Sands, F. Young, Component models for three-way data: an alternating least squares algorithm with optimal scaling features, Psychometrika 45 (1) (1980) 39–67.
[20] A. Kapteyn, H. Neudecker, T. Wansbeek, An approach to n-mode component analysis, Psychometrika 51 (1) (1986) 269–275.
[21] H.J. Bierens, J. Hartog, Non-linear regression with discrete explanatory variables, with an application to the earnings function, J. Econometrics 38 (1988) 269–299.
[22] R. Christensen, Linear Models for Multivariate, Time Series and Spatial Data, Springer, Berlin, 1991.
[23] L. Shousong, P. Gemperline, Eliminating complex eigenvectors and eigenvalues in multiway analyses using the direct trilinear decomposition method, J. Chemometr. 7 (1993) 77–88.
[24] D.S. Burdick, An introduction to tensor products with applications to multiway data analysis, Chemometr. Intell. Lab. Syst. 28 (1995).
[25] M. Nadenau, Localization and shape normalization for face recognition. Available from: http://dewww.epfl.ch/~nadenau/Research/Paper/Diploma.pdf, 1997.
[26] C.A. Anderson, R. Bro, Improving the speed of multi-way algorithms. Part i: Tucker3, Chemometr. Intell. Lab. Syst. 42 (1998) 93–103.
[27] A. Beinat, F. Crosilla, Generalized Procrustes analysis for size and shape 3-D object reconstruction. Optical 3-D Measurements Techniques V, Viena, 2001, pp. 345–353.
[28] T.G. Kolda, Orthogonal tensor decompositions, SIAM J. Matrix Anal. Appl. 23 (1) (2001) 243–255.

About the Author—MÓNICA BENITO received her M.Sc. in Statistical Sciences at the Universidad Carlos III of Madrid in 1999. At the moment she is working on her Ph.D. thesis in the field of Multivariate Analysis for image data. Her interests include Dimensionality Reduction, Statistical Pattern Recognition and Quality Control.

About the Author—DANIEL PEÑA is professor of Statistics at Universidad Carlos III of Madrid. He has published more than 150 articles on Multivariate Methods, Time Series and Robustness and wrote several books on statistical methods.